

# Dendograms of Mixing Measures for Softmax-Gated Gaussian Mixture of Experts: Consistency without Model Sweeps

Do Tien Hai<sup>\*,1,2</sup>

Nhat Ho<sup>6</sup>

Trung Nguyen Mai<sup>\*,2,3</sup>

Binh T. Nguyen<sup>1,2</sup>

Trung Tin Nguyen<sup>\*,†,4,5</sup>

Christopher Drovandi<sup>4,5</sup>

<sup>1</sup>Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam.

<sup>2</sup>Vietnam National University Ho Chi Minh City, Vietnam.

<sup>3</sup>Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam.

<sup>4</sup>ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems.

<sup>5</sup>School of Mathematical Sciences, Queensland University of Technology, Brisbane City, Australia.

<sup>6</sup>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 dendograms 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.,

$\epsilon$ -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.

## Contents

<b>1 INTRODUCTION</b>	<b>2</b>
<b>2 PRELIMINARIES</b>	<b>4</b>
<b>3 FAST-RATE-AWARE EXPERT AGGREGATION IN SGMOE</b>	<b>5</b>
3.1 Why Merge Experts? The Rate Gap . . . . .	5
3.2 A Fast-Rate-Aware Voronoi Distance . . . . .	5
3.3 A Merge Operator Tailored to SGMOE . . . . .	5
3.4 The Aggregation Path: A Hierarchical View . . . . .	6

\*Co-first author, †Corresponding author.

3.5	Heights and Likelihood Along the Path . . . . .	6
3.6	Choosing the Number of Experts via a Height-Likelihood Rule . . . . .	7
3.7	Proof Sketches . . . . .	7
<b>4</b>	<b>SIMULATION STUDIES</b>	<b>8</b>
<b>5</b>	<b>CONCLUSION</b>	<b>8</b>
<b>A</b>	<b>REAL DATA ILLUSTRATION</b>	<b>14</b>
<b>B</b>	<b>ILLUSTRATION OF VORONOI CELLS AND MERGE STEPS FOR SGMoE</b>	<b>15</b>
<b>C</b>	<b>THEORETICAL CHALLENGES: MORE DETAILS</b>	<b>15</b>
<b>D</b>	<b>PROOF OF MAIN RESULTS</b>	<b>16</b>
D.1	Proof Sketches . . . . .	17
D.1.1	Proof Sketch of Lemma 1 . . . . .	17
D.1.2	Proof Sketch of Theorem 1 . . . . .	17
D.1.3	Proof Sketch of Theorem 2 . . . . .	18
D.1.4	Proof Sketch of Theorem 3 . . . . .	18
D.2	Proof of Lemma 1 . . . . .	18
D.3	Proof of Theorem 1 . . . . .	21
D.4	Proof of Theorem 2 . . . . .	29
D.5	Proof of Theorem 3 . . . . .	30
D.6	Proof of Theorem 4 . . . . .	36

## 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; 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., 2025, 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 EM 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; 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, 2024b).

**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; 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 (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., 2024c; Zens, 2019). A recent alternative leverages

dendograms of mixing measures for selection without exhaustive sweeps in (Do et al., 2024; Thai et al., 2025).

**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  $\epsilon$ -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 generated by a SGMoE model of order  $K_0$ , whose conditional density is

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

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

necessarily normalized) mixing measure

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

where  $\boldsymbol{\eta}_k^0 := (\omega_{0k}^0, \boldsymbol{\omega}_{1k}^0, \mathbf{a}_k^0, b_k^0, \sigma_k^0) \in \Theta \subset \mathbb{R} \times \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}_{>0}$ . Assume  $\Theta$  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).

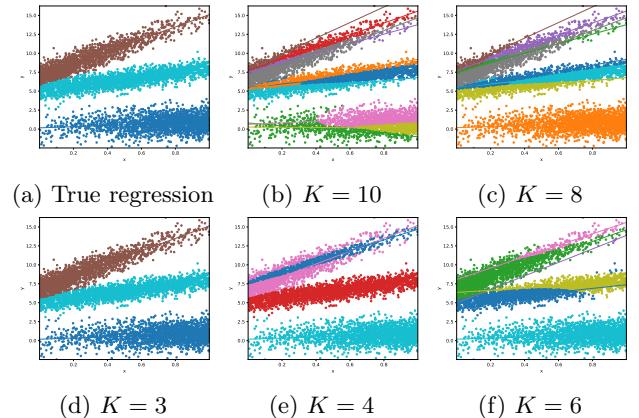


Figure 1: Merging procedure from  $K = 10$  to  $K = 3$  of true mixing measure  $G_0(3)$  with  $K_0 = 3$  components, defined in eq. (11).

**Maximum Likelihood Over At Most  $K$  Experts.** When the true order  $K_0$  is unknown, we estimate within

$$\mathcal{O}_K(\Theta) := \left\{ G = \sum_{k=1}^{K'} \exp(\omega_{0k}) \delta_{(\boldsymbol{\omega}_{1k}, \mathbf{a}_k, b_k, \sigma_k)} : 1 \leq K' \leq K, (\omega_{0k}, \boldsymbol{\omega}_{1k}, \mathbf{a}_k, b_k, \sigma_k) \in \Theta \right\}.$$

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):  $\widehat{G}_N \in \arg \max_{G \in \mathcal{O}_K(\Theta)} \frac{1}{N} \sum_{n=1}^N \log(p_G(y_n | \mathbf{x}_n))$ .

**Practical Implication.** Practitioners can fit a single over-specified SGMoE with moderate  $K \geq K_0$ , 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

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	$p_{G_0}(y   \mathbf{x})$	$\exp(\omega_{0k}^0)$	$\omega_{1k}^0, b_k^0$	$\mathbf{a}_k^0, \sigma_k^0$
Exact-fit	$D_E$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$
Over-fit	$D_O$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2\bar{r}( \mathbb{A}_k )})$	$\mathcal{O}((\log N/N)^{1/\bar{r}( \mathbb{A}_k )})$
Merged	$D_{FRA}$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$

operator, aggregation path, fast pathwise rates, and DSC. The simulations in Section 4 illustrate parameter rates, path behavior, and model selection under clean and contaminated regimes. Then, we offer concluding remarks, limitations and future work in Section 5. Proof sketches appear at the end of Section 3; full proofs are deferred to the appendix.

**Notation.** Throughout the paper, for any natural number  $N \in \mathbb{N}$  we abbreviate  $\{1, 2, \dots, N\}$  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  $a_N \leq C b_N$  for all  $N \in \mathbb{N}$ . For a vector  $\mathbf{v} \in \mathbb{R}^D$ , set  $|\mathbf{v}| := v_1 + \dots + v_D$ , and let  $\|\mathbf{v}\|_p$  denote its  $p$ -norm; by default,  $\|\mathbf{v}\|$  refers to the 2-norm unless otherwise stated. We also use  $\|\mathbf{A}\|$  for the Frobenius norm of a matrix  $\mathbf{A} \in \mathbb{R}^{D \times D}$ . 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  $\mu$ , define  $D_{TV}(p, q) := \frac{1}{2} \int |p - q| d\mu$  as their Total Variation distance, while  $D_h^2(p, q) := \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\mu$  denotes the squared Hellinger distance between them. Let  $\Theta$  be the parameter space. Write  $\mathcal{E}_K(\Theta)$  for the collection of discrete probability measures on  $\Theta$  with exactly  $K$  atoms, and  $\mathcal{O}_K(\Theta) := \bigcup_{K' \leq K} \mathcal{E}_{K'}(\Theta)$  for those with at most  $K$  atoms. For a mixing measure  $G = \sum_{k=1}^K \pi_k \delta_{\theta_k}$ , we (slightly abusively) refer to each component  $\pi_k \delta_{\theta_k}$  as an ‘‘atom,’’ comprising both its weight  $\pi_k$  and parameter  $\theta_k$ . When clear from context, we drop  $\Theta$  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  $G = \sum_{k=1}^K \exp(\omega_{0k}) \delta_{(\omega_{1k}, \mathbf{a}_k, b_k, \sigma_k)}$  and the true  $G_0 =$

$$\sum_{k=1}^{K_0} \exp(\omega_{0k}^0) \delta_{(\omega_{1k}^0, \mathbf{a}_k^0, b_k^0, \sigma_k^0)}, \text{ define for } k \in [K_0]:$$

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

where we denote  $\boldsymbol{\theta}_\ell := (\boldsymbol{\omega}_{1\ell}, \mathbf{a}_\ell, b_\ell, \sigma_\ell)$ . We use the softmax-translation  $(t_0, \mathbf{t}_1)$  from identifiability (cf. Proposition 1 of Nguyen et al., 2023a) and the shorthand  $\Delta_{\mathbf{t}_1} \boldsymbol{\omega}_{1\ell k} := \boldsymbol{\omega}_{1\ell} - \boldsymbol{\omega}_{1k}^0 - \mathbf{t}_1$ ,  $\Delta \mathbf{a}_{\ell k} := \mathbf{a}_\ell - \mathbf{a}_k^0$ ,  $\Delta b_{\ell k} := b_\ell - b_k^0$ ,  $\Delta \sigma_{\ell k} := \sigma_\ell - \sigma_k^0$ . 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  $0 \leq |\ell_1| \leq r$ ,  $0 \leq \ell_2 \leq r - |\ell_1|$ ,  $|\ell_1| + \ell_2 \geq 1$ , the polynomial system

$$\sum_{j=1}^M \sum_{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{I}_{\ell_1, \ell_2}} \frac{p_{5j}^2 p_{1j}^{\alpha_1} p_{2j}^{\alpha_2} p_{3j}^{\alpha_3} p_{4j}^{\alpha_4}}{\alpha_1! \alpha_2! \alpha_3! \alpha_4!} = 0, \quad (3)$$

admits no non-trivial solution (all  $p_{5j} \neq 0$  and at least one  $p_{3j} \neq 0$ ). The ranges of  $\alpha_1, \alpha_2, \alpha_3, \alpha_4$  in the above sum satisfy  $\mathbb{I}_{\ell_1, \ell_2} = \{\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^D \times \mathbb{N}^D \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \ell_1, |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$ . 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:

$$\begin{aligned} D_O(G, G_0) &:= D_E(G, G_0) \\ &+ \inf_{t_0, \mathbf{t}_1} \sum_{k: |\mathbb{A}_k| > 1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}) \left( \|(\Delta_{\mathbf{t}_1} \boldsymbol{\omega}_{1\ell k}, \Delta b_{\ell k})\|^{\bar{r}(|\mathbb{A}_k|)} \right. \\ &\quad \left. + \|(\Delta \mathbf{a}_{\ell k}, \Delta \sigma_{\ell k})\|^{\bar{r}(|\mathbb{A}_k|)/2} \right), \end{aligned} \quad (4)$$

$$\begin{aligned} D_E(G, G_0) &:= \inf_{t_0, \mathbf{t}_1} \sum_{k=1}^{K_0} \left| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}) - \exp(\omega_{0k}^0 + t_0) \right| \\ &+ \sum_{k: |\mathbb{A}_k|=1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}) \|(\Delta_{\mathbf{t}_1} \boldsymbol{\omega}_{1\ell k}, \Delta \mathbf{a}_{\ell k}, \Delta b_{\ell k}, \Delta \sigma_{\ell k})\|. \end{aligned}$$

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  $\Theta$ ) s.t. the MLE  $\hat{G}_N$  of order  $K \geq K_0$  satisfies*

$$\mathbb{P}(D_O(\hat{G}_N, G_0) > C(\log N/N)^{1/2}) \lesssim e^{-c \log N}. \quad (5)$$

*Remarks.* (i) If  $K = K_0$  (all  $|\mathbb{A}_k| = 1$ ), then  $D_O = D_E$  and eq. (5) yields the exact-specified rate  $\mathbb{P}(D_E(\hat{G}_N, G_0) > C(\log N/N)^{1/2}) \lesssim e^{-c \log N}$ , implying parametric ( $N^{-1/2}$  up to logs) estimation of  $\exp(\omega_{0k}^0)$ ,  $\boldsymbol{\omega}_{1k}^0$  (up to translation),  $\mathbf{a}_k^0$ ,  $b_k^0$ ,  $\sigma_k^0$  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  $D_O(G, G_0)$  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

$$\begin{aligned} D_{FRA}(G, G_0) := D_O(G, G_0) \\ + \inf_{t_0, t_1} \sum_{k: |\mathbb{A}_k| > 1} \left( \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})(\Delta b_{\ell k}) \right\| \right. \\ + \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}) \right\| \\ + \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})[(\Delta b_{\ell k})^2 + (\Delta \sigma_{\ell k})] \right\| \\ + \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})[(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k})(\Delta b_{\ell k}) + (\Delta \mathbf{a}_{\ell k})] \right\| \\ \left. + \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell})(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k})(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k})^\top \right\| \right). \quad (6) \end{aligned}$$

**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  $|\mathbb{A}_k| = 1$ ,  $D_{FRA}$  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,  $D_{FRA}$  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 path-wise rates (Theorem 1) and height/likelihood controls (Theorems 2 and 3).

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

$$\begin{aligned} d(\exp(\omega_{0\ell_1}) \delta_{\theta_{\ell_1}}, \exp(\omega_{0\ell_2}) \delta_{\theta_{\ell_2}}) := \frac{\exp(\omega_{0\ell_1} + \omega_{0\ell_2})}{\exp(\omega_{0\ell_1}) + \exp(\omega_{0\ell_2})} \\ \times \left( \|(\boldsymbol{\omega}_{1\ell_1}, b_{\ell_1}) - (\boldsymbol{\omega}_{1\ell_2}, b_{\ell_2})\|^2 + \|(\mathbf{a}_{\ell_1}, \sigma_{\ell_1}) - (\mathbf{a}_{\ell_2}, \sigma_{\ell_2})\| \right). \quad (7) \end{aligned}$$

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

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

Then we define  $G^{(K-1)} = \exp (\omega_{0*}) \delta_{(\omega_{1*}, \mathbf{a}_*, b_*, \sigma_*)} + \sum_{k \neq i, j} \exp (\omega_{0k}) \delta_{(\omega_{1k}, \mathbf{a}_k, b_k, \sigma_k)}$ . 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 Aggregation Path: A Hierarchical View

**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  $\kappa$ -th level holding the atoms of  $G^{(\kappa)}$ ,  $\mathcal{E}$  storing the links between merged pairs across adjacent levels, and  $\mathcal{H} = (h^{(K)}, \dots, h^{(2)})$  with  $h^{(\kappa)} := \min \{d(\cdot, \cdot)\} \text{ over pairs in } G^{(\kappa)}$ . The quantity  $h^{(\kappa)}$  is the height between levels  $\kappa$  and  $\kappa - 1$ .

When we represent  $\mathcal{T}(G)$  on a graph,  $h^{(\kappa)}$  is the height between  $\kappa$ -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:**  $G^{(\kappa)} = \sum_{k=1}^{\kappa} \exp (\omega_{0k}) \delta_{(\omega_{1k}, \mathbf{a}_k, b_k, \sigma_k)}$

- 1:  $(i, j) \leftarrow \arg \min_{\ell_1 \neq \ell_2 \in [\kappa]} d(\exp (\omega_{0\ell_1}) \delta_{\theta_{\ell_1}}, \exp (\omega_{0\ell_2}) \delta_{\theta_{\ell_2}})$
- 2: Compute  $(\omega_{0*}, \omega_{1*}, \mathbf{a}_*, b_*, \sigma_*)$  by eq. (8)
- 3: **return**  $G^{(\kappa-1)} = \exp (\omega_{0*}) \delta_{(\omega_{1*}, \mathbf{a}_*, b_*, \sigma_*)} + \sum_{k \neq i, j} \exp (\omega_{0k}) \delta_{\theta_k}$

---

**Algorithm 2** SGMoE Hierarchical Aggregation Path

**Require:**  $G^{(K)} = \sum_{k=1}^K \exp (\omega_{0k}) \delta_{(\omega_{1k}, \mathbf{a}_k, b_k, \sigma_k)}$

- 1: Initialize  $\mathcal{T}(G) = (\mathcal{V}, \mathcal{E}, \mathcal{H})$  with  $\mathcal{V}_K = \{\text{atoms of } G^{(K)}\}$ ,  $\mathcal{E} = \emptyset$
- 2: **for**  $\kappa = K, \dots, 2$  **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:      $h^{(\kappa)} \leftarrow \min d(\cdot, \cdot)$  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  $D_{\text{FRA}}(G^{(K)}, G_0) \rightarrow 0$ ,  $D_{\text{FRA}}(G^{(K)}, G_0) \gtrsim D_{\text{FRA}}(G^{(K-1)}, G_0) \gtrsim \dots \gtrsim D_{\text{FRA}}(G^{(K_0)}, G_0)$ , with constants depending only on  $G_0$ ,  $\Theta$ , 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*

$$\mathbb{P}(D_{\text{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) > C'_2 (\log N/N)^{1/2}) \lesssim e^{-c_2 \log N}, \quad (9)$$

$$\mathbb{P}(D_{\text{E}}(\widehat{G}_N^{(\kappa')}, G_0^{(\kappa')}) > C'_1 (\log N/N)^{1/2}) \lesssim e^{-c_1 \log N}.$$

### 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

$$\begin{aligned} h_N^{(\kappa)} := \min \left\{ d \left( \exp(\hat{\omega}_{0k_1}) \delta_{\hat{\theta}_{k_1}}, \exp(\hat{\omega}_{0k_2}) \delta_{\hat{\theta}_{k_2}} \right) \right. \\ \left. : k_1 \neq k_2, \text{ atoms of } \hat{G}_N^{(\kappa)} \right\}, \quad (10) \end{aligned}$$

and let  $h_0^{(\kappa')}$  be the analogous height on the true path. Then:

**Theorem 2** (Height control). *For all  $\kappa \in [K_0 + 1, K]$  and  $\kappa' \in [K_0]$ ,*

$$h_N^{(\kappa)} \lesssim (\log N/N)^{1/\bar{r}(\hat{G}_N)}, \quad |h_N^{(\kappa')} - h_0^{(\kappa')}| \lesssim (\log N/N)^{1/2}$$

with constants depending only on  $G_0$ ,  $\Theta$ , and  $\kappa$ .

**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  $\mathcal{L}(p_G) := \mathbb{E}_{(\mathbf{x}, y) \sim P_{G_0}} [\log p_G(y | \mathbf{x})]$ .

**Condition K.** There exist positive constants  $c_\alpha$  and  $c_\beta$  such that for all sufficiently small  $\epsilon$  and  $\theta_0, \theta \in \Theta$  such that  $\|\theta - \theta_0\| \leq \epsilon$ , we have  $\log f(\mathbf{x}, y | \theta) \geq (1 + c_\beta \epsilon) \log f(\mathbf{x}, y | \theta_0) - c_\alpha \epsilon$ .

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

### 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  $\kappa$ , define

$$DSC_N^{(\kappa)} := - \left( h_N^{(\kappa)} + \epsilon_N \bar{\ell}_N(p_{\hat{G}_N^{(\kappa)}}) \right),$$

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

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

**Theorem 4** (Consistency of model selection). *If  $K_0 \geq 2$ , then  $\hat{K}_N \rightarrow K_0$  in  $\mathbb{P}_{G_0}$ -probability as  $N \rightarrow \infty$ .*

**Interpretation.** Unlike pure likelihood criteria (AIC/BIC/ICL),  $DSC_N^{(\kappa)}$  also penalizes *structural closeness* through  $h_N^{(\kappa)}$ . 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  $\hat{G}_N \rightarrow G_0$ , 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

$$\begin{aligned} Q_N = & \left[ \sum_{k=1}^{K_0} \exp((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x} + \omega_{0k}^0 + t_0) \right] \\ & \times [p_{G_N}(y | \mathbf{x}) - p_{G_0}(y | \mathbf{x})], \end{aligned}$$

we analyze the sums over indices with  $|\mathbb{A}_k^N| > 1$  under  $1 \leq |\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k^N|)$ . For clarity, we also consider  $(\ell_1, \ell_2)$  with  $1 \leq |\ell_1| + \ell_2 \leq 2$ , 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  $D_{FRA}(G^{(K)}, G_0) \gtrsim D_{FRA}(G^{(K-1)}, G_0)$ . As  $D_{FRA}(G^{(K)}, G_0) \rightarrow 0$ , extract a sequence that satisfies  $(\mathbf{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \rightarrow (\mathbf{a}_k^0, b_k^0, \sigma_k^0)$  and there exist  $t_0 \in \mathbb{R}$ ,  $\mathbf{t}_1 \in \mathbb{R}^D$  with  $\sum_{\ell \in \mathbb{A}_k^N} \exp(\omega_{\ell k}^N) \rightarrow \exp(\omega_{0k}^0 + t_0)$  and  $\omega_{1\ell}^N \rightarrow \omega_{1k}^0 + \mathbf{t}_1$  for all  $\ell \in \mathbb{A}_k^N$ . The minimizing pair  $(\ell_1, \ell_2)$  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

$$\begin{aligned} & (\exp \omega_{0\ell_1}^N + \exp \omega_{0\ell_2}^N) \|(\Delta_{\mathbf{t}_1} \omega_{1*k}^N, \Delta b_{*k}^N)\|^{\bar{r}_k} \\ & \lesssim \sum_{j \in \{\ell_1, \ell_2\}} \exp \omega_{0j}^N \|(\Delta_{\mathbf{t}_1} \omega_{1jk}^N, \Delta b_{jk}^N)\|^{\bar{r}_k}, \\ & (\exp \omega_{0\ell_1}^N + \exp \omega_{0\ell_2}^N) \|(\Delta \mathbf{a}_{*k}^N, \Delta \sigma_{*k}^N)\|^{\bar{r}_k/2} \\ & \lesssim \sum_{j \in \{\ell_1, \ell_2\}} \exp \omega_{0j}^N \|(\Delta \mathbf{a}_{jk}^N, \Delta \sigma_{jk}^N)\|^{\bar{r}_k/2}, \end{aligned}$$

which yields the desired monotonicity.

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

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

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

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

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  $u(y|\mathbf{x}; \boldsymbol{\omega}_1, \mathbf{a}, b, \sigma) := \exp(\boldsymbol{\omega}_1^\top \mathbf{x}) \mathcal{N}(y|\mathbf{a}^\top \mathbf{x} + b, \sigma)$  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

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

For each experiment,  $N$  varies on a logarithmic grid from  $\log_{10}(N_{\min})$  to  $\log_{10}(N_{\max})$ , yielding  $N_{\text{num}}$  sizes in  $[N_{\min}, N_{\max}]$ . At each  $N$ , we generate  $N_{\text{rep}}$  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 softmax gate in eq. (1) is translation-invariant, we fix a baseline by setting  $\omega_{0K_0}^0 = 0$  and  $\omega_{1K_0}^0 = 0$ .

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

### 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  $N \in [10^2, 5 \times 10^4]$ ; 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

$$G_0(3) = e^{-2} \delta_{(3, 1, 0, 1)} + e^1 \delta_{(-3.5, 8, 7, 0.8)} + e^0 \delta_{(0, 3, 5, 0.6)}. \quad (11)$$

**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  $\epsilon$ -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 CONCLUSION

**Summary.** 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.

**Limitations.** Our analysis assumes linear softmax gates, Gaussian experts, compact  $\Theta$ , 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.

**Future work.** Empirically, the dendrogram remains

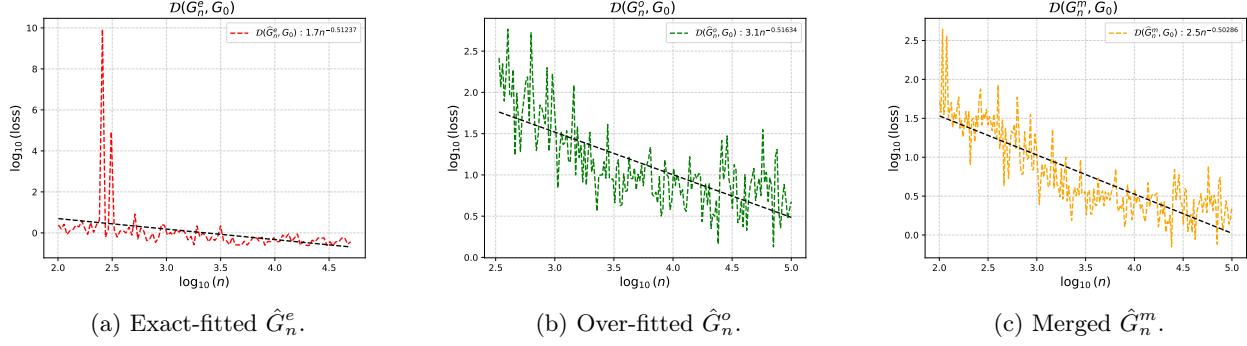


Figure 2: Convergence under three settings: (a) exact-fitted, (b) over-fitted, and (c) merged mixing measures.

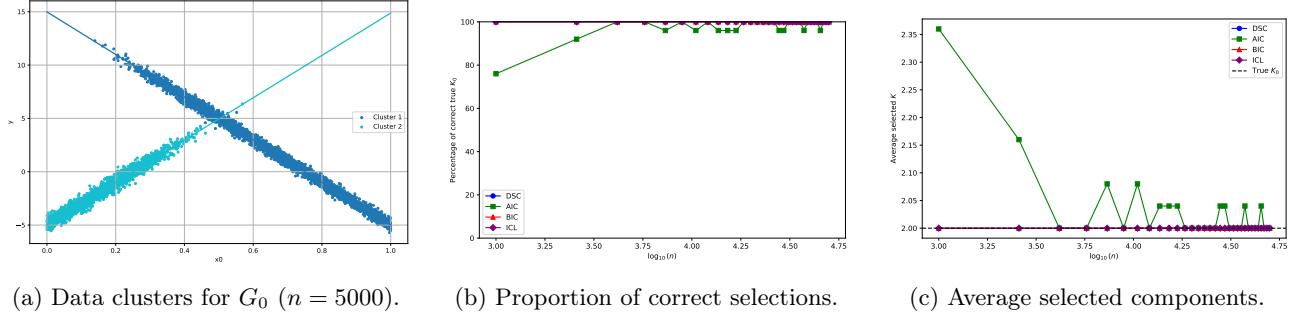


Figure 3: DSC vs. AIC, BIC, and ICL for selecting  $K_0 = 2$  of  $G_0$ .

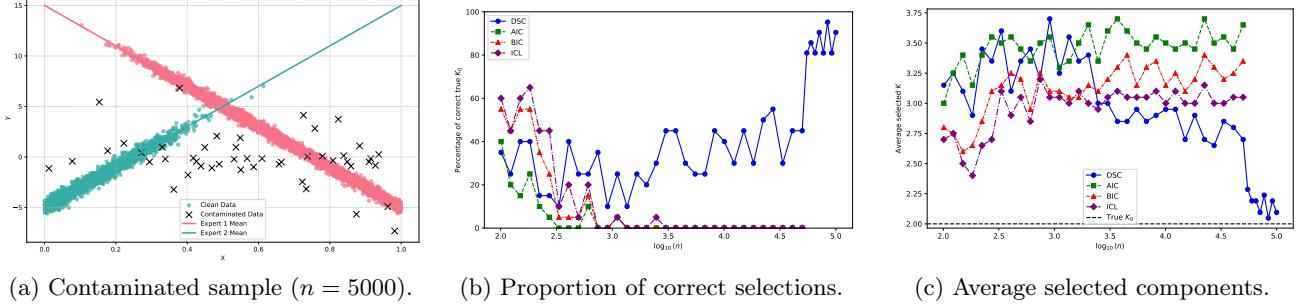


Figure 4: Model selection under  $\epsilon$ -contamination: DSC vs. AIC, BIC, and ICL with  $K_0 = 2$ . After AIC, BIC, and ICL fail to recover  $K_0$ , we evaluate DSC on 8 additional sample sizes between  $5.5 \times 10^4$  and  $10^5$ . The results show that DSC recovers  $K_0$  with a high proportion.

robust and recovers clustering structure under mild misspecification (Figure 4). A rigorous study of its asymptotic behavior under model misspecification is an important next step.

## Acknowledgments

TrungTin Nguyen and Christopher Drovandi acknowledge funding from the Australian Research Council Centre of Excellence for the Mathematical Analysis of Cellular Systems (CE230100001). Christopher Drovandi was also supported by an Australian Research Council Future Fellowship (FT210100260). 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 2.)
- Berrettini, M., Galimberti, G., Raciati, 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 page 14.)
- 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 page 14.)
- 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 8.)
- 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 8.)
- 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.)
- 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 3 and 9.)
- 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 page 2.)
- 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 2.)
- 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 2.)
- 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 page 2.)
- 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 sur-

- rogate 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, 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 30.)
- 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 pages 2 and 3.)
- 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 2.)
- Jiang, W. & Tanner, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9), 1253–1258. (Cited on page 21.)
- Jordan, M. I. & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2), 181–214. (Cited on page 2.)
- 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<sup>3</sup>ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*. (Cited on page 2.)
- 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., 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, 5, 7, 15, 16, 18, 21, 26, and 32.)
- Nguyen, H., Nguyen, T., Nguyen, K., & Ho, N. (2024b). 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., 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. (2025). 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. (2024c). 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 2.)
- 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 page 2.)
- 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 page 14.)
- 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 page 3.)
- van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press. (Cited on pages 8, 18, and 30.)
- Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society. (Cited on page 8.)
- Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer. (Cited on page 8.)
- 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.)

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

**Supplementary Organization.** This supplement has five parts. *First*, Section A provides a deeper illustration by applying our method to a real dataset. *Second*, Section B illustrates the Voronoi cells and the merge step used by our SGMoE aggregation path. *Third*, Section C details the main technical obstacles (softmax translation invariance, gate-expert PDE couplings, and algebraic cancellations). *Fourth*, Section D.1 expands the proof sketches for Lemma 1 and Theorems 1 to 3, showing how the Voronoi geometry drives the analysis. *Finally*, Section D presents the core proof ingredients and notational conventions used throughout the full arguments.

## A REAL DATA ILLUSTRATION

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 ( $L = 2$ ), which are *water use* (WU) and the proteins quantified under the *water deficit* (WD) condition, and 973 protein variables ( $D = 973$ ). 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.

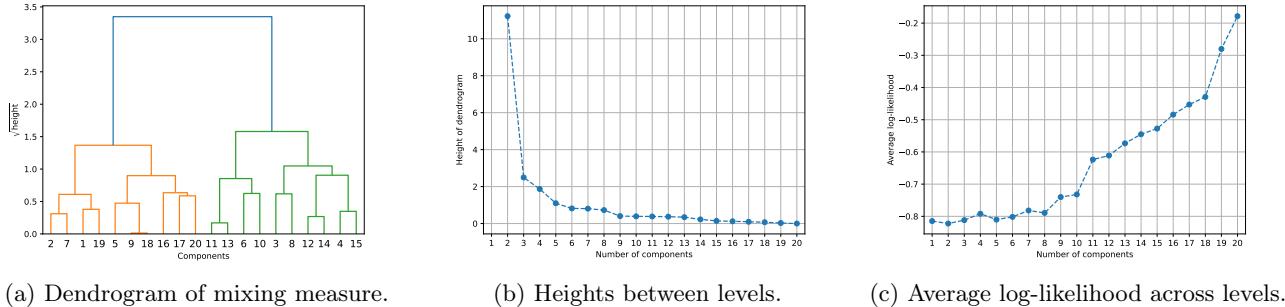


Figure 5: Dendrogram of mixing measure inferred from maize drought-responsive traits dataset.

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

## B ILLUSTRATION OF VORONOI CELLS AND MERGE STEPS FOR SGMoE

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

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

where  $\boldsymbol{\theta}_\ell := (\omega_{1\ell}, \mathbf{a}_\ell, b_\ell, \sigma_\ell)$ . We use the softmax translation  $(t_0, \mathbf{t}_1)$  from identifiability (cf. Proposition 1 of Nguyen et al., 2023a) and the shorthand  $\Delta_{\mathbf{t}_1} \omega_{1\ell} := \omega_{1\ell} - \omega_{1k}^0 - \mathbf{t}_1$ ,  $\Delta \mathbf{a}_{\ell k} := \mathbf{a}_\ell - \mathbf{a}_k^0$ ,  $\Delta b_{\ell k} := b_\ell - b_k^0$ ,  $\Delta \sigma_{\ell k} := \sigma_\ell - \sigma_k^0$ . 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 DFRA (eq. (6)), and enable our fast pathwise guarantees and sweep-free model selection via DSC.

## C 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:

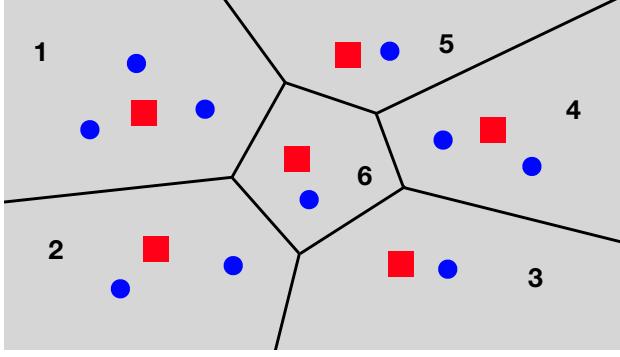
(i) *Softmax translation invariance.* Gating parameters are identifiable only up to common translations. Unlike covariate-independent gating functions, the softmax gate is invariant under simultaneous shifts of intercepts and slopes, which makes the parameterization non-unique. As a result, standard 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  $p_{\hat{G}_N}(y|\mathbf{x}) - p_{G_0}(y|\mathbf{x})$  into independent components. Moreover, the parameters of the softmax gating numerators and the Gaussian experts are intrinsically linked through explicit PDEs,

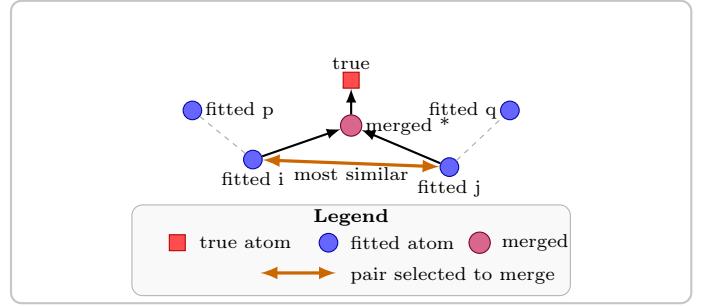
$$\frac{\partial^2 u}{\partial \boldsymbol{\omega}_1 \partial b} = \frac{\partial u}{\partial \mathbf{a}}, \quad \frac{\partial^2 u}{\partial b^2} = 2 \frac{\partial u}{\partial \sigma}, \quad (13)$$

where  $u(y|\mathbf{x}; \boldsymbol{\omega}, \mathbf{a}, b, \sigma) := \exp(\boldsymbol{\omega}_1^\top \mathbf{x}) \mathcal{N}(y|\mathbf{a}^\top \mathbf{x} + b, \sigma)$ . 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.



(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  $\{\theta_k^0\}$ . Blue circles are fitted atoms  $\{\theta_\ell\}$ . 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

$$\text{fitted } i: \exp(\omega_{0i})\delta_{\theta_i}, \quad \text{fitted } j: \exp(\omega_{0j})\delta_{\theta_j}, \quad \text{merged } *: \exp(\omega_{0*})\delta_{\theta_*}.$$

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

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

(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.

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.

## D 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  $[N] := \{1, 2, \dots, N\}$ . 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  $\mathbf{v} \in \mathbb{R}^D$  and any multi-index  $\mathbf{p} \in \mathbb{N}^D$ , set  $|\mathbf{p}| := p_1 + \dots + p_D$ ,  $\mathbf{v}^\mathbf{p} := v_1^{p_1} v_2^{p_2} \dots v_D^{p_D}$ ,  $\mathbf{p}! := p_1! p_2! \dots p_D!$ , and let  $\|\mathbf{v}\|_p$  denote its  $p$ -norm; by default,  $\|\mathbf{v}\|$  refers to the 2-norm unless otherwise stated. We also use  $\|\mathbf{A}\|$  for the Frobenius norm of a matrix  $\mathbf{A} \in \mathbb{R}^{D \times D}$ . For any set  $\mathbb{S}$ ,  $|\mathbb{S}|$  denotes its cardinality. For two probability density functions  $p$  and  $q$  with respect to the Lebesgue measure  $\mu$ , define  $D_{\text{TV}}(p, q) := \frac{1}{2} \int |p - q| d\mu$  as their total variation distance, while  $D_h^2(p, q) := \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\mu$  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

$$\frac{\partial^{|\boldsymbol{\alpha}|}}{\partial^{\boldsymbol{\alpha}} \boldsymbol{\mu}} f(\boldsymbol{\mu}) := \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial \mu_1^{\alpha_1} \partial \mu_2^{\alpha_2} \cdots \partial \mu_D^{\alpha_D}} f(\boldsymbol{\mu}).$$

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

## D.1 Proof Sketches

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

**Why the  $D_{FRA}$  loss in eq. (6)?** When  $\widehat{G}_N \rightarrow G_0$  with  $K > K_0$ , some Voronoi cells  $\mathbb{A}_k$  are multi-covered. The slow directions in  $D_O$  (with exponents  $\bar{r}(|\mathbb{A}_k|)$ ) arise from these cells.  $D_{FRA}$  augments  $D_O$  with first-order *merged-moment* block-sums that vanish when a cell behaves as a single aggregate. Thus  $D_{FRA}$  is simultaneously (i) exact-fit consistent, it reduces to  $D_E$  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

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

we analyze the sums over indices with  $|\mathbb{A}_k| > 1$  under  $1 \leq |\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|)$ ; for clarity, we also isolate the case  $1 \leq |\ell_1| + \ell_2 \leq 2$ , corresponding to  $|\mathbb{A}_k| = 1$ . This leads directly to the merge operator and the aggregation path.

### D.1.1 Proof Sketch of Lemma 1

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

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

The minimizing pair  $(i, j)$  of  $\mathbf{d}$  must lie in the same cell  $\mathbb{A}_k$ . Let the merged atom be  $\exp(\omega_{0*})\delta_{(\boldsymbol{\omega}_{1*}, \mathbf{a}_*, b_*, \sigma_*)}$  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

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

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

### D.1.2 Proof Sketch of Theorem 1

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

$$\mathbb{E}_{\mathbf{x}} [\text{DTV}(p_G(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x}))] \geq C D_{FRA}(G, G_0). \quad (14)$$

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

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

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

$$\mathbf{x}^{\ell_1} \exp((\boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}) \frac{\partial^{\ell_2}}{\partial h_1^{\ell_2}} \mathcal{N}(y | \mathbf{a}_k^{0\top} \mathbf{x} + b_k^0, \sigma_k^0), \quad 1 \leq |\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|),$$

with coefficients that are precisely the atomwise sums appearing in  $D_{FRA}$  (up to constants). If eq. (14) failed, all these coefficients would have to vanish at a rate faster than  $D_{FRA}(G, G_0)$ , 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. \(2023a\)](#),  $\mathbb{E}_{\mathbf{x}}[D_h^2(p_{\widehat{G}_N}(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x}))] = \mathcal{O}_{\mathbb{P}}((\log N/N)^{1/2})$ . Using the inequality  $D_{TV} \leq \sqrt{2} D_h^{2^{1/2}}$  and eq. (14) with  $G = \widehat{G}_N$ , we obtain

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

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

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

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

### D.1.3 Proof Sketch of Theorem 2

For  $\kappa \in [K_0 + 1, K]$ , the height  $h_N^{(\kappa)}$  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

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

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

### D.1.4 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  $\mathcal{L}(p_G) = \mathbb{E}_{(\mathbf{x}, y) \sim P_{G_0}} [\log p_G(y|\mathbf{x})]$ . Under Condition K, a local Lipschitz/curvature argument yields

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

For  $\kappa \geq K_0$ , combine the inverse bound  $\mathbb{E}[D_{TV}] \gtrsim D_{FRA}$  with  $D_{FRA}(\widehat{G}_N^{(\kappa)}, G_0) = \mathcal{O}_{\mathbb{P}}((\log N/N)^{1/2})$  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  $\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa')}}) \rightarrow \mathcal{L}(p_{G_0^{(\kappa')}})$  in probability.

## D.2 Proof of Lemma 1

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

Assume that  $G_N := G_N^{(K)} = \sum_{k=1}^K \exp(\omega_{0k}^N) \delta_{(\boldsymbol{\omega}_{1k}^N, \mathbf{a}_k^N, b_k^N, \sigma_k^N)} \in \mathcal{E}_K$  varies so that  $D_{FRA}(G_N, G_0) \rightarrow 0$ . 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  $(\mathbf{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \rightarrow (\mathbf{a}_k^0, b_k^0, \sigma_k^0)$ , and there exist  $t_0 \in \mathbb{R}$  and  $\mathbf{t}_1 \in \mathbb{R}^D$  such that  $\sum_{\ell \in \mathbb{A}_k^N} \exp(\omega_{0\ell}^N) \rightarrow \exp(\omega_{0k}^0 + t_0)$  and  $\boldsymbol{\omega}_{1\ell}^N \rightarrow \boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1$  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  $(\ell_1, \ell_2)$  must belong to a common  $\mathbb{A}_k$ . Indeed, for every pair  $(\ell_1, \ell_2)$  in a common  $\mathbb{A}_k$ , since  $(\mathbf{a}_{\ell_1}^N, b_{\ell_1}^N, \sigma_{\ell_1}^N) \rightarrow (\mathbf{a}_k^0, b_k^0, \sigma_k^0)$  and  $\boldsymbol{\omega}_{1\ell_1}^N \rightarrow \boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1$ , and  $(\mathbf{a}_{\ell_2}^N, b_{\ell_2}^N, \sigma_{\ell_2}^N) \rightarrow (\mathbf{a}_k^0, b_k^0, \sigma_k^0)$  and  $\boldsymbol{\omega}_{1\ell_2}^N \rightarrow \boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1$ , we have

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

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

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

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

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

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

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

We prove that  $\mathcal{D}(G_N^{(K-1)}, G_0, t_0, \mathbf{t}_1) \lesssim \mathcal{D}(G_N^{(K)}, G_0, t_0, \mathbf{t}_1)$ . Let the merging pair of indices  $(\ell_1, \ell_2)$  in the Voronoi cell  $\mathbb{A}_k$ , then  $|\mathbb{A}_k| > 1$  and the merged atom is  $\exp(\omega_{0*}^N) \delta_{(\boldsymbol{\omega}_{1*}^N, \mathbf{a}_*^N, b_*^N, \sigma_*^N)}$ , i.e,

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

Hence, we have that

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

It follows that the term

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

Similarly, we can show that

$$\begin{aligned}
 \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N) \right\| &= \left\| \sum_{\ell \in \mathbb{A}_k \setminus \{\ell_1, \ell_2\}} \exp(\omega_{0\ell}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N) + \exp(\omega_{0*}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1*k}^N) \right\|, \\
 \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)[(\Delta b_{\ell k}^N)^2 + (\Delta \sigma_{\ell k}^N)] \right\| &= \left\| \sum_{\ell \in \mathbb{A}_k \setminus \{\ell_1, \ell_2\}} \exp(\omega_{0\ell}^N)[(\Delta b_{\ell k}^N)^2 + (\Delta \sigma_{\ell k}^N)] \right. \\
 &\quad \left. + \exp(\omega_{0*}^N)[(\Delta b_{*k}^N)^2 + (\Delta \sigma_{*k}^N)] \right\|, \\
 \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)[(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N) + (\Delta \boldsymbol{a}_{\ell k}^N)] \right\| &= \left\| \sum_{\ell \in \mathbb{A}_k \setminus \{\ell_1, \ell_2\}} \exp(\omega_{0\ell}^N)[(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta b_{\ell k}^N) + (\Delta \boldsymbol{a}_{\ell k}^N)] \right. \\
 &\quad \left. + \exp(\omega_{0*}^N)[(\Delta_{t_1} \boldsymbol{\omega}_{1*k}^N)(\Delta b_{*k}^N) + (\Delta \boldsymbol{a}_{*k}^N)] \right\|, \\
 \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N)^\top \right\| &= \left\| \sum_{\ell \in \mathbb{A}_k \setminus \{\ell_1, \ell_2\}} \exp(\omega_{0\ell}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1\ell k}^N)^\top \right. \\
 &\quad \left. + \exp(\omega_{0*}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1*k}^N)(\Delta_{t_1} \boldsymbol{\omega}_{1*k}^N)^\top \right\|.
 \end{aligned}$$

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

$$\begin{aligned}
 \alpha_i &:= \exp(\omega_{0\ell_i}^N - \omega_{0*}^N), \quad i = 1, 2, \\
 \mathbf{x}_i &:= (\boldsymbol{\omega}_{1\ell_i}^N - \boldsymbol{\omega}_{1k}^0 - \mathbf{t}_1, \quad b_{\ell_i}^N - b_k^0) = (\Delta_{t_1} \boldsymbol{\omega}_{1\ell_i}^N, \quad \Delta b_{\ell_i k}^N), \quad i = 1, 2.
 \end{aligned}$$

Note that  $\alpha_1, \alpha_2 \in (0, 1)$  and  $\alpha_1 + \alpha_2 = 1$  and by the definition of the merged atom eq. (8), we have the convex combination identity  $(\Delta_{t_1} \boldsymbol{\omega}_{1*k}^N, \Delta b_{*k}^N) = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2$ .

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  $m = \bar{r}_k$ ):

$$\|\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2\|^{\bar{r}_k} \leq \alpha_1 \|\mathbf{x}_1\|^{\bar{r}_k} + \alpha_2 \|\mathbf{x}_2\|^{\bar{r}_k}.$$

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

$$\begin{aligned}
 (\exp(\omega_{0\ell_1}^N) + \exp(\omega_{0\ell_2}^N)) \|\Delta_{t_1} \boldsymbol{\omega}_{1*k}^N, \Delta b_{*k}^N\|^{\bar{r}_k} &= \exp(\omega_{0*}^N) \|\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2\|^{\bar{r}_k} \\
 &\leq \exp(\omega_{0*}^N) (\alpha_1 \|\mathbf{x}_1\|^{\bar{r}_k} + \alpha_2 \|\mathbf{x}_2\|^{\bar{r}_k}) \\
 &= \exp(\omega_{0\ell_1}^N) \|\mathbf{x}_1\|^{\bar{r}_k} + \exp(\omega_{0\ell_2}^N) \|\mathbf{x}_2\|^{\bar{r}_k} \\
 &= \exp(\omega_{0\ell_1}^N) \|(\Delta_{t_1} \boldsymbol{\omega}_{1\ell_1}^N, \Delta b_{\ell_1 k}^N)\|^{\bar{r}_k} + \exp(\omega_{0\ell_2}^N) \|(\Delta_{t_1} \boldsymbol{\omega}_{1\ell_2}^N, \Delta b_{\ell_2 k}^N)\|^{\bar{r}_k}.
 \end{aligned}$$

Analogously, we can show that

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

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

$$\begin{aligned} \exp(\omega_{0\ell_1}^N) \left( \|(\Delta \mathbf{t}_1 \omega_{1\ell_1 k}^N, \Delta b_{\ell_1 k}^N)\|^{\bar{r}_k} + \|(\Delta \mathbf{a}_{\ell_1 k}^N, \Delta \sigma_{\ell_1 k}^N)\|^{\bar{r}_k/2} \right) + \exp(\omega_{0\ell_2}^N) \left( \|(\Delta \mathbf{t}_1 \omega_{1\ell_2 k}^N, \Delta b_{\ell_2 k}^N)\|^{\bar{r}_k} + \|(\Delta \mathbf{a}_{\ell_2 k}^N, \Delta \sigma_{\ell_2 k}^N)\|^{\bar{r}_k/2} \right) \\ \gtrsim (\exp(\omega_{0\ell_1}^N) + \exp(\omega_{0\ell_2}^N)) \left( \|(\Delta \mathbf{t}_1 \omega_{1*k}^N, \Delta b_{*k}^N)\|^{\bar{r}_k} + \|(\Delta \mathbf{a}_{*k}^N, \Delta \sigma_{*k}^N)\|^{\bar{r}_k/2} \right). \end{aligned}$$

Hence

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

and therefore

$$D_{\text{FRA}}(G_N^{(K-1)}, G_0) \lesssim D_{\text{FRA}}(G_N^{(K)}, G_0).$$

### D.3 Proof of Theorem 1

First of all, we study the convergence rate of the MLE  $\hat{G}_N \in \mathcal{E}_K$  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  $G = \sum_{k=1}^K \exp(\omega_{0k}) \delta_{(\omega_{1k}, \mathbf{a}_k, b_k, \sigma_k)}$  and  $G' = \sum_{k=1}^{K'} \exp(\omega'_{0k}) \delta_{(\omega'_{1k}, \mathbf{a}'_k, b'_k, \sigma'_k)}$ , if we have  $p_G(y|\mathbf{x}) = p_{G'}(y|\mathbf{x})$  for almost surely  $(\mathbf{x}, y)$ , then it follows that  $K = K'$  and  $G \equiv G'_{t_0, \mathbf{t}_1}$  where  $G'_{t_0, \mathbf{t}_1} := \sum_{k=1}^{K'} \exp(\omega'_{0k} + t_0) \delta_{(\omega'_{1k} + \mathbf{t}_1, \mathbf{a}'_k, b'_k, \sigma'_k)}$  for some  $t_0 \in \mathbb{R}$  and  $\mathbf{t}_1 \in \mathbb{R}^D$ .*

The identifiability of the softmax gating Gaussian mixture of experts guarantees that the MLE  $\hat{G}_N$  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  $p_{\hat{G}_N}(y|\mathbf{x})$  to the true conditional density  $p_{G_0}(y|\mathbf{x})$ , which lays an important foundation for the study of MLE's convergence rate.

**Fact 4** ([Nguyen et al., 2023a](#), Proposition 2). *The density estimation  $p_{\hat{G}_N}(y|\mathbf{x})$  converges to the true density  $p_{G_0}(y|\mathbf{x})$  under the Hellinger distance  $D_h^2(\cdot, \cdot)$  at the following rate:*

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

That is,

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

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  $p_{\hat{G}_N}(y|\mathbf{x})$  to the true one  $p_{G_0}(y|\mathbf{x})$  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  $D_{\text{FRA}}$ .

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

$$\mathbb{E}_{\mathbf{x}}[D_h^2(p_G(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x}))] \geq C \cdot D_{\text{FRA}}(G, G_0),$$

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

$$\mathbb{P}(D_{\text{FRA}}(\hat{G}_N, G_0) > C'(\log(N)/N)^{1/2}) \lesssim \exp(-c \log N), \quad (15)$$

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  $\Theta$  such that, for any  $G \in \mathcal{O}_K$ ,

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

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):

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

Assume that the inequality in eq. (17) does not hold true, there exists a sequence of mixing measures  $G_N := \sum_{k=1}^{K_N} \exp(\omega_{0k}^N) \delta_{(\omega_{1k}^N, \mathbf{a}_k^N, b_k^N, \sigma_k^N)} \in \mathcal{O}_K$  such that

$$\begin{aligned} \mathbb{E}_{\mathbf{x}} [\text{D}_{\text{TV}}(p_{G_N}(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x}))] / \text{D}_{\text{FRA}}(G_N, G_0) &\rightarrow 0, \\ \text{D}_{\text{FRA}}(G_N, G_0) &\rightarrow 0, \end{aligned}$$

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  $\text{D}_{\text{FRA}}(G_N, G_0) \rightarrow 0$ , we have  $(\mathbf{a}_\ell^N, b_\ell^N, \sigma_\ell^N) \rightarrow (\mathbf{a}_\ell^0, b_\ell^0, \sigma_\ell^0)$  for any  $\ell \in \mathbb{A}_k$  as  $N \rightarrow \infty$ . Furthermore, there exist  $t_0 \in \mathbb{R}$  and  $\mathbf{t}_1 \in \mathbb{R}^D$  such that  $\sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \rightarrow \exp(\omega_{0k}^0 + t_0)$  and  $\omega_{1\ell}^N \rightarrow \omega_{1k}^0 + \mathbf{t}_1$  as  $N$  approaches infinity for any  $\ell \in \mathbb{A}_k$  and  $k \in [K_0]$ . It suggests that we can upper bound  $\text{D}_{\text{FRA}}$  as  $\text{D}_{\text{FRA}}(G_N, G_0) \leq \text{D}_{\text{V}}(G_N, G_0)$ , where

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

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

### Step 1: Density Decomposition

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

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

where we denote  $u(y|\mathbf{x}; \omega_1, \mathbf{a}, b, \sigma) := \exp(\omega_1^\top \mathbf{x}) \mathcal{N}(y|\mathbf{a}^\top \mathbf{x} + b, \sigma)$  and  $v(y|\mathbf{x}; \omega_1) := \exp(\omega_1^\top \mathbf{x}) p_{G_N}(y|\mathbf{x})$ .

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

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

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:

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

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

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

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

$$\frac{\partial^2 u}{\partial \mathbf{a} \partial b} = \frac{\partial u}{\partial \mathbf{a}}, \quad \frac{\partial^2 u}{\partial b^2} = 2 \frac{\partial u}{\partial \sigma},$$

we have

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

Hence

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

It follows that

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

where  $\mathbb{I}_{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2} = \{\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \alpha_3, \alpha_4) \in \mathbb{N}^D \times \mathbb{N}^D \times \mathbb{N} \times \mathbb{N} : \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = \boldsymbol{\ell}_1, |\boldsymbol{\alpha}_2| + \alpha_3 + 2\alpha_4 = \boldsymbol{\ell}_2\}$ .

Similarly, we can decompose  $A_{N,2}$  by the first-order Taylor expansion as

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

where

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

Analogously,  $B_N$  can be rewritten as

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

where

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

Therefore,  $Q_N$  can be represented as

$$\begin{aligned} Q_N &= \sum_{k=1}^{K_0} \sum_{|\ell_1|+|\ell_2|=1}^{2\bar{r}(|\mathbb{A}_k|)} T_{\ell_1, \ell_2}^N(k) \cdot \mathbf{x}^{\ell_1} \exp((\boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}) \frac{\partial^{\ell_2} \mathcal{N}}{\partial h_1^{\ell_2}} (y|\mathbf{a}_k^{0\top} \mathbf{x} + b_k^0, \sigma_k^0) \\ &\quad + \sum_{k=1}^{K_0} \sum_{|\gamma|=1}^{1+\mathbf{1}\{|\mathbb{A}_k|>1\}} S_\gamma^N(k) \cdot \mathbf{x}^\gamma \exp((\boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}) p_{G_N}(y|\mathbf{x}) + \sum_{\rho=1}^4 R_\rho^N(\mathbf{x}, y) \\ &\quad + \sum_{k=1}^{K_0} \left( \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) - \exp(\omega_{0k}^0 + t_0) \right) [u(y|\mathbf{x}; \boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1, \mathbf{a}_k^0, b_k^0, \sigma_k^0) - v(y|\mathbf{x}; \boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1)] \\ &= \sum_{k=1}^{K_0} \sum_{|\ell_1|+|\ell_2|=0}^{2\bar{r}(|\mathbb{A}_k|)} T_{\ell_1, \ell_2}^N(k) \cdot \mathbf{x}^{\ell_1} \exp((\boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}) \frac{\partial^{\ell_2} \mathcal{N}}{\partial h_1^{\ell_2}} (y|\mathbf{a}_k^{0\top} \mathbf{x} + b_k^0, \sigma_k^0) \\ &\quad + \sum_{k=1}^{K_0} \sum_{|\gamma|=0}^{1+\mathbf{1}\{|\mathbb{A}_k|>1\}} S_\gamma^N(k) \cdot \mathbf{x}^\gamma \exp((\boldsymbol{\omega}_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}) p_{G_N}(y|\mathbf{x}) + \sum_{\rho=1}^4 R_\rho^N(\mathbf{x}, y), \end{aligned} \tag{18}$$

with coefficients  $T_{\ell_1, \ell_2}^N(k)$  and  $S_\gamma^N(k)$  are defined for any  $k \in [K_0]$ ,  $0 \leq |\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|)$  and  $0 \leq |\gamma| \leq 2$  as

$$\begin{aligned} T_{\ell_1, \ell_2}^N(k) &= \begin{cases} \sum_{\ell \in \mathbb{A}_k} \sum_{\alpha \in \mathbb{I}_{\ell_1, \ell_2}} \frac{\exp(\omega_{0\ell}^N)}{2^{\alpha_4} \alpha!} (\Delta_{\mathbf{t}_1} \boldsymbol{\omega}_{1\ell k}^N)^{\alpha_1} (\Delta \mathbf{a}_{\ell k}^N)^{\alpha_2} (\Delta b_{\ell k}^N)^{\alpha_3} (\Delta \sigma_{\ell k}^N)^{\alpha_4}, & (\ell_1, \ell_2) \neq (0_D, 0), \\ \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) - \exp(\omega_{0k}^0 + t_0), & (\ell_1, \ell_2) = (0_D, 0), \end{cases} \\ S_\gamma^N(k) &= \begin{cases} - \sum_{\ell \in \mathbb{A}_k} \frac{\exp(\omega_{0\ell}^N)}{\gamma!} (\Delta_{\mathbf{t}_1} \boldsymbol{\omega}_{1\ell k}^N)^\gamma, & |\gamma| \neq 0, \\ - \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) + \exp(\omega_{0k}^0 + t_0), & |\gamma| = 0. \end{cases} \end{aligned}$$

## Step 2: Non-vanishing coefficients

Next, we will show that not all the quantities  $T_{\ell_1, \ell_2}^N(k)/D_V(G_N, G_0)$  and  $S_\gamma^N(k)/D_V(G_N, G_0)$  go to 0 as  $N \rightarrow \infty$ . We assume that all of them go to 0 as  $N \rightarrow \infty$ . Then, by assumption  $T_{0_D, 0}^N(k)/D_V(G_N, G_0) \rightarrow 0$ , we have

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

For any  $k$  such that  $|\mathbb{A}_k| = 1$ , consider all  $(|\ell_1|, \ell_2)$  implying  $1 \leq |\ell_1| + \ell_2 \leq 2$ , we have  $T_{\ell_1, \ell_2}^N(k) / D_V(G_N, G_0) \rightarrow 0$  for all  $k$  such that  $|\mathbb{A}_k| = 1$ . Hence

$$\frac{1}{D_V(G_N, G_0)} \left( \sum_{k: |\mathbb{A}_k|=1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) \|(\Delta_{t_1} \omega_{1\ell k}^N, \Delta \mathbf{a}_{\ell k}^N, \Delta b_{\ell k}^N, \Delta \sigma_{\ell k}^N)\| \right) \rightarrow 0. \quad (20)$$

Next, we consider  $k$  such that  $|\mathbb{A}_k| > 1$  and  $(|\ell_1|, \ell_2)$  such that  $1 \leq |\ell_1| + \ell_2 \leq 2$ :

- For  $(|\ell_1|, \ell_2) = (0, 1)$ , then

$$\frac{1}{D_V(G_N, G_0)} \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) (\Delta b_{\ell k}^N) \right\| \rightarrow 0.$$

- For  $(|\ell_1|, \ell_2) = (1, 0)$ , then

$$\frac{1}{D_V(G_N, G_0)} \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) (\Delta_{t_1} \omega_{1\ell k}^N) \right\| \rightarrow 0.$$

- For  $(|\ell_1|, \ell_2) = (1, 1)$ , then

$$\frac{1}{D_V(G_N, G_0)} \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) [(\Delta_{t_1} \omega_{1\ell k}^N)(\Delta b_{\ell k}^N) + (\Delta \mathbf{a}_{\ell k}^N)] \right\| \rightarrow 0.$$

- For  $(|\ell_1|, \ell_2) = (0, 2)$ , then

$$\frac{1}{D_V(G_N, G_0)} \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) [(\Delta b_{\ell k}^N)^2 + (\Delta \sigma_{\ell k}^N)] \right\| \rightarrow 0.$$

- For  $(|\ell_1|, \ell_2) = (2, 0)$ , then

$$\frac{1}{D_V(G_N, G_0)} \left\| \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}^N) (\Delta_{t_1} \omega_{1\ell k}^N) (\Delta_{t_1} \omega_{1\ell k}^N)^\top \right\| \rightarrow 0.$$

Combining the above limit and the formulation of  $D_{FRA}(G_N, G_0)$  together, it follows that

$$\frac{1}{D_V(G_N, G_0)} \cdot \sum_{k: |\mathbb{A}_k| > 1} \sum_{\ell \in \mathbb{A}_k} \exp(\omega_{0\ell}) \left( \|(\Delta_{t_1} \omega_{1\ell k}^N, \Delta b_{\ell k}^N)\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \mathbf{a}_{\ell k}^N, \Delta \sigma_{\ell k}^N)\|^{\bar{r}(|\mathbb{A}_k|)/2} \right) \not\rightarrow 0$$

which implies that there exists some index  $k^* \in [K_0]$  such that  $|\mathbb{A}_{k^*}| > 1$  and

$$\frac{1}{D_V(G_N, G_0)} \cdot \sum_{\ell \in \mathbb{A}_{k^*}} \exp(\omega_{0\ell}) \left( \|(\Delta_{vt_1} \omega_{1\ell k^*}^N, \Delta b_{\ell k^*}^N)\|^{\bar{r}(|\mathbb{A}_k|)} + \|(\Delta \mathbf{a}_{\ell k^*}^N, \Delta \sigma_{\ell k^*}^N)\|^{\bar{r}(|\mathbb{A}_k|)/2} \right) \not\rightarrow 0$$

for all  $\mathbf{t}_1 \in \mathbb{R}^D$ . WLOG, we assume that  $k^* = 1$ . For  $(\ell_1, \ell_2) \in \mathbb{N}^D \times \mathbb{N}$  such that  $1 \leq |\ell_1| + \ell_2 \leq \bar{r}(|\mathbb{A}_1|)$ , we have  $T_{\ell_1, \ell_2}^N(1) / D_V(G_N, G_0) \rightarrow 0$  as  $N \rightarrow \infty$ . Thus, by dividing this ratio and the left hand side of the above equation and let  $\mathbf{t}_1 = 0$ , we have

$$\frac{\sum_{\ell \in \mathbb{A}_1} \sum_{\alpha \in \mathbb{I}_{\ell_1, \ell_2}} \frac{\exp(\omega_{0\ell}^N)}{2^{\alpha_4} \alpha!} (\Delta_{t_1} \omega_{1\ell 1}^N)^{\alpha_1} (\Delta \mathbf{a}_{\ell 1}^N)^{\alpha_2} (\Delta b_{\ell 1}^N)^{\alpha_3} (\Delta \sigma_{\ell 1}^N)^{\alpha_4}}{\sum_{\ell \in \mathbb{A}_1} \exp(\omega_{0\ell}^N) \left( \|(\Delta_{t_1} \omega_{1\ell 1}^N, \Delta b_{\ell 1}^N)\|^{\bar{r}(|\mathbb{A}_1|)} + \|(\Delta \mathbf{a}_{\ell 1}^N, \Delta \sigma_{\ell 1}^N)\|^{\bar{r}(|\mathbb{A}_1|)/2} \right)} \rightarrow 0 \quad (21)$$

for all  $(\ell_1, \ell_2)$  such that  $1 \leq |\ell_1| + \ell_2 \leq \bar{r}(|\mathbb{A}_1|)$ .

Let us define  $\bar{M}_N := \max \left\{ \|\Delta_{t_1} \omega_{1\ell 1}^N\|, \|\Delta \mathbf{a}_{\ell 1}^N\|^{1/2}, |\Delta b_{\ell 1}^N|, |\Delta \sigma_{\ell 1}^N|^{1/2} : \ell \in \mathbb{A}_1 \right\}$  and  $\bar{\omega}_N := \max_{\ell \in \mathbb{A}_1} \exp(\omega_{0\ell}^N)$ . Since the sequence  $\exp(\omega_{0\ell}^N) / \bar{\omega}_N$  is bounded, we can replace it by its subsequence that has a positive limit  $p_{5\ell}^2 := \lim_{N \rightarrow \infty} \exp(\omega_{0\ell}^N) / \bar{\omega}_N$ . Hence, at least one among  $p_{5\ell}^2$ , for  $\ell \in \mathbb{A}_1$ , equals 1.

Similarly, we also define

$$\begin{aligned} (\Delta_{\mathbf{t}_1} \omega_{1\ell_1}^N) / \bar{M}_N &\rightarrow p_{1\ell}, (\Delta \mathbf{a}_{\ell_1}^N) / \bar{M}_N \rightarrow p_{2\ell}, \\ (\Delta b_{\ell_1}^N) / \bar{M}_N &\rightarrow p_{3\ell}, (\Delta \sigma_{\ell_1}^N) / [2\bar{M}_N] \rightarrow p_{4\ell}. \end{aligned}$$

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  $\bar{\omega}_N \bar{M}_N^{\ell_1 + \ell_2}$ , and then achieve the following system of polynomial equations:

$$\sum_{\ell \in \mathbb{A}_1} \sum_{\alpha \in \mathbb{I}_{\ell_1, \ell_2}} \frac{1}{\alpha!} \cdot p_{5\ell}^2 p_{1\ell}^{\alpha_1} p_{2\ell}^{\alpha_2} p_{3\ell}^{\alpha_3} p_{4\ell}^{\alpha_4} = 0$$

for all  $(\ell_1, \ell_2) \in \mathbb{N}^D \times \mathbb{N}$  such that  $1 \leq |\ell_1| + \ell_2 \leq \bar{r}(|\mathbb{A}_1|)$ . However, based on the definition of  $\bar{r}(|\mathbb{A}_1|)$ , the above system has no non-trivial solutions, which is a contradiction. Thus, not all the quantities  $T_{\ell_1, \ell_2}^N(k) / \text{Dv}(G_N, G_0)$  and  $S_\gamma^N(k) / \text{Dv}(G_N, G_0)$  go to 0 as  $N \rightarrow \infty$ .

### 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  $1/m_N \not\rightarrow \infty$ . Then, by applying the Fatou's lemma, we obtain that

$$\lim_{N \rightarrow \infty} \frac{\mathbb{E}_{\mathbf{x}} [\text{D}_{\text{TV}}(p_G(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x}))]}{m_N \cdot \text{Dv}(G_N, G_0)} \geq \int \liminf_{N \rightarrow \infty} \frac{|p_G(y|\mathbf{x}), p_{G_0}(y|\mathbf{x})|}{2m_N \cdot \text{Dv}(G_N, G_0)} \, d(\mathbf{x}, y). \quad (22)$$

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  $(\mathbf{x}, y)$ . Hence, we get that  $Q_N / [m_N \text{Dv}(G_N, G_0)] \rightarrow 0$  as  $N \rightarrow \infty$  for almost surely  $(\mathbf{x}, y)$ . It follows from the decomposition of  $Q_N$  in eq. (18) that

$$\begin{aligned} &\sum_{k=1}^{K_0} \sum_{|\ell_1|+|\ell_2|=0}^{2\bar{r}(|\mathbb{A}_k|)} \tau_{\ell_1, \ell_2}(k) \cdot \mathbf{x}^{\ell_1} \exp\left((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}\right) \frac{\partial^{\ell_2} \mathcal{N}}{\partial h_1^{\ell_2}} \left(y | (\mathbf{a}_k^0)^\top \mathbf{x} + b_k^0, \sigma_k^0\right) \\ &+ \sum_{k=1}^{K_0} \sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathbb{A}_k|>1\}}} \xi_\gamma(j) \cdot \mathbf{x}^\gamma \exp\left((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}\right) p_{G_0}(y|\mathbf{x}) = 0, \end{aligned}$$

for almost surely  $(\mathbf{x}, y)$ , where  $\tau_{\ell_1, \ell_2}(k)$  and  $\xi_\gamma(k)$  denote the limits of  $T_{\ell_1, \ell_2}^N(k) / [m_N \text{Dv}(G_N, G_0)]$  and  $S_\gamma^N(k) / [m_N \text{Dv}(G_N, G_0)]$  as  $N \rightarrow \infty$ , respectively, for all  $k \in [K_0], 0 \leq |\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|)$  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:

$$\begin{aligned} \mathcal{W} := &\left\{ \mathbf{x}^{\ell_1} \exp\left((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}\right) \frac{\partial^{\ell_2} \mathcal{N}}{\partial h_1^{\ell_2}} \left(y | (\mathbf{a}_k^0)^\top \mathbf{x} + b_k^0, \sigma_k^0\right) : k \in [K_0], 0 \leq |\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|) \right\} \\ &\cup \left\{ \mathbf{x}^\gamma \exp\left((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}\right) p_{G_0}(y|\mathbf{x}) : k \in [K_0], 0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathbb{A}_k|>1\}} \right\}. \end{aligned}$$

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 independent w.r.t  $\mathbf{x}$  and  $y$ , where  $\mathcal{W}_1$  is denoted as follows:*

$$\begin{aligned} \mathcal{W}_1 := &\left\{ \mathbf{x}^{\ell_1} \exp\left((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}\right) \frac{\partial^{\ell_2} \mathcal{N}}{\partial h_1^{\ell_2}} \left(y | (\mathbf{a}_k^0)^\top \mathbf{x} + b_k^0, \sigma_k^0\right) : k \in [K_0], 0 \leq |\ell_1| + \ell_2 \leq 2 \right\} \\ &\cup \left\{ \mathbf{x}^\gamma \exp\left((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x}\right) p_{G_0}(y|\mathbf{x}) : k \in [K_0], 0 \leq |\gamma| \leq 1 \right\}, \end{aligned}$$

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

$$\tau_{\ell_1, \ell_2}(k) = \xi_\gamma(k) = 0$$

for all  $k \in [K_0]$ ,  $0 \leq 2|\ell_1| + \ell_2 \leq 2\bar{r}(|\mathbb{A}_k|)$  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:

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

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

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

Then, we get that  $\mathbb{E}_{\boldsymbol{x}} [D_{TV}(p_{G'_N}(\cdot | \boldsymbol{x}), p_{G_0}(\cdot | \boldsymbol{x}))] \rightarrow 0$  as  $N \rightarrow \infty$ . Since the set  $\Theta$  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  $D_{FRA}(G', G_0) > \varepsilon'$ . Then, by the Fatou's lemma, we get

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

It follows that

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

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  $G' = \sum_{k=1}^{K_0} \exp(\omega_{0\nu(k)}^0 + t_0) \delta_{(\omega_{1\nu(k)}^0 + \mathbf{t}_1, \mathbf{a}_{\nu(k)}^0, b_{\nu(k)}^0, \sigma_{\nu(k)}^0)}$  for some  $(t_0, \mathbf{t}_1) \in \mathbb{R} \times \mathbb{R}^D$ , where  $\nu$  is some permutation of the set  $\{1, 2, \dots, K_0\}$ . It follows that  $D_{FRA}(G', G_0) = 0$ , which contradicts the hypothesis  $D_{FRA}(G', G_0) > \varepsilon' > 0$ . Hence, we obtain the inequality in eq. (16).  $\square$

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

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

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  $D_{FRA}(\widehat{G}_N^{(\kappa')}, G_0)$ , we obtain that  $D_{FRA}(\widehat{G}_N^{(K_0)}, G_0) = D_E(\widehat{G}_N^{(K_0)}, G_0)$ . Hence, by Lemma 1, we get the convergence rate

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

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

$$\begin{aligned} |\exp(\omega_{0k}^N) - \exp(\omega_{0k}^0 + t_0)| &\lesssim \left( \frac{\log N}{N} \right)^{1/2}, \\ \|(\Delta_{\mathbf{t}_1} \omega_{1k}^N, \Delta \mathbf{a}_k^N, \Delta b_k^N, \Delta \sigma_k^N)\| &\lesssim \left( \frac{\log N}{N} \right)^{1/2}, \end{aligned}$$

for every  $k \in [K_0]$ , where  $\Delta_{t_1}^N \omega_{1k}^N := \omega_{1k}^N - \omega_{1k}^0 - t_1$ ,  $\Delta \mathbf{a}_k^N := \mathbf{a}_k^N - \mathbf{a}_k^0$ ,  $\Delta b_k^N := b_k^N - b_k^0$  and  $\Delta \sigma_k^N := \sigma_k^N - \sigma_k^0$ . This implies that for every  $(i, j) \in [K_0]^2$ , by the triangle inequality, we have

$$\begin{aligned} \left| \|(\omega_{1i}^N - \omega_{1j}^N, b_i^N - b_j^N)\| - \|(\omega_{1i}^0 - \omega_{1j}^0, b_i^0 - b_j^0)\| \right| &\leq \|(\omega_{1i}^N - \omega_{1j}^N - \omega_{1i}^0 + \omega_{1j}^0, b_i^N - b_j^N - b_i^0 + b_j^0)\| \\ &\leq \|(\omega_{1i}^N - \omega_{1i}^0 - t_1, b_i^N - b_i^0)\| + \|(\omega_{1j}^N - \omega_{1j}^0 - t_1, b_j^N - b_j^0)\| \\ &\lesssim \left( \frac{\log N}{N} \right)^{1/2}. \end{aligned}$$

Similarly, we have

$$\left| \|(\mathbf{a}_i^N - \mathbf{a}_j^N, \sigma_i^N - \sigma_j^N)\| - \|(\mathbf{a}_i^0 - \mathbf{a}_j^0, \sigma_i^0 - \sigma_j^0)\| \right| \lesssim \left( \frac{\log N}{N} \right)^{1/2}.$$

Hence, we obtain that

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

Hence, on  $A_N$ , the optimal choice of indices  $(\ell_1, \ell_2)$  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  $\exp(\omega_{0*}^N) \delta_{(\omega_{1*}^N, \mathbf{a}_*^N, b_*^N, \sigma_*^N)}$  and  $\exp(\omega_{0*}^0) \delta_{(\omega_{1*}^0, \mathbf{a}_*^0, b_*^0, \sigma_*^0)}$  denoted as follows:

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

and

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

After merging, we also have

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

and

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

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

#### D.4 Proof of Theorem 2

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

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

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

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

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

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

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

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

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.

### D.5 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  $\mathbf{x}_1, \dots, \mathbf{x}_N \sim P_{G_0}$ . 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

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

then

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

**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

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

where  $\Pi(\mathbf{p}, \mathbf{p}')$  is the set of all couplings between  $\mathbf{p} = (p_1, \dots, p_K)$  and  $\mathbf{p}' = (p'_1, \dots, p'_{K'})$ , i.e.,  $\Pi(\mathbf{p}, \mathbf{p}') = \left\{ \mathbf{q} \in \mathbb{R}_+^{K \times K'} : \sum_{k=1}^K q_{k\ell} = p'_\ell, \sum_{l=1}^{K'} q_{k\ell} = p_k, \forall k \in [K], \ell \in [K'] \right\}$ . 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  $W_r(G, G_0) \rightarrow 0$ , we obtain that

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

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  $\phi$ . 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  $p_G(y|\mathbf{x})$  and  $p_{G'}(y|\mathbf{x})$  with the Wasserstein metric between  $G$  and  $G'$  following:

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

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:

$$\begin{aligned}\bar{\ell}_N(p_G) &= \frac{1}{N} \sum_{n=1}^N \log p_G(y_n | \mathbf{x}_n) =: P_N \log p_G, \\ \mathcal{L}(p_G) &= \mathbb{E}_{(\mathbf{x}, y) \sim P_{G_0}} [\log p_G(y | \mathbf{x})] = \int \log p_G(y | \mathbf{x}) dP_{G_0}(\mathbf{x}, y) =: P_{G_0} \log p_G,\end{aligned}$$

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  $(\mathbf{x}, y)$  is then constructed by first sampling  $\mathbf{x} \sim P_{\mathbf{x}}$  and then  $y | \mathbf{x} \sim p_{G_0}(y | \mathbf{x})$ .

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

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

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

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

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  $D_{\text{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) \lesssim (\log N/N)^{1/2}$ , and obviously we have

$$\inf_{t_0, t_1} W_{\bar{r}(\widehat{G}_N)}(\widehat{G}_N^{(\kappa)}, G_{0, t_0, t_1}) \leq D_{\text{FRA}}(\widehat{G}_N^{(\kappa)}, G_0),$$

where  $G_{0, t_0, t_1} = \sum_{k=1}^{K_0} \exp(\omega_{0k}^0 + t_0) \delta_{(\omega_{1k}^0 + t_1, \mathbf{a}_k^0, b_k^0, \sigma_k^0)}$ , so there exists a constant  $D$  such that

$$\mathbb{P}_{G_0} \left( \inf_{t_0, t_1} W_{\bar{r}(\widehat{G}_N)}(\widehat{G}_N^{(\kappa)}, G_{0, t_0, t_1}) \leq D \left( \frac{\log N}{N} \right)^{1/2\bar{r}(\widehat{G}_N)} \right) \geq 1 - c_1 N^{-c_2}, \quad \kappa \in [K_0, K].$$

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  $\mathbb{E}_{\mathbf{x}}(D_h^2(p_G(\cdot | \mathbf{x}), p_{G'}(\cdot | \mathbf{x})) \lesssim W_2(G, G')$  and  $p_{G_0} =$

$p_{G_{0,t_0,t_1}}, \forall (t_0, t_1) \in \mathbb{R} \times \mathbb{R}^D$ , we also have

$$\begin{aligned} & \mathbb{P}_{G_0} \left( \mathbb{E}(\text{D}_h^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x})) \leq D \left( \frac{\log N}{N} \right)^{1/2\bar{r}(\widehat{G}_N)} \right) \\ &= \mathbb{P}_{G_0} \left( \inf_{t_0, t_1} \mathbb{E}_{\mathbf{x}}(\text{D}_h^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\mathbf{x}), p_{G_{0,t_0,t_1}}(\cdot|\mathbf{x})) \leq D \left( \frac{\log N}{N} \right)^{1/2\bar{r}(\widehat{G}_N)} \right) \\ &\geq \mathbb{P}_{G_0} \left( \inf_{t_0, t_1} W_{\bar{r}(\widehat{G}_N)}(\widehat{G}_N^{(\kappa)}, G_{0,t_0,t_1}) \leq D \left( \frac{\log N}{N} \right)^{1/2\bar{r}(\widehat{G}_N)} \right) \geq 1 - c_1 N^{-c_2}, \quad \kappa \in [K_0, K]. \end{aligned}$$

Let  $\mathcal{P}_K(\Theta) := \{p_G(y|\mathbf{x}) : G \in \mathcal{O}_K(\Theta)\}$  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  $a \leq \sqrt{N}R^2 \leq \sqrt{NR}$  and

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

By Fact 6, we get

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

Combining with the bound on Hellinger distance, we have

$$\begin{aligned} & \mathbb{P}_{G_0} \left( \left| (P_N - P_{G_0}) \log \frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} \right| \geq D \frac{\log^{\alpha+1/2} N}{N^{\alpha+1/2}} \right) \\ &\leq \mathbb{P}_{G_0} \left( \mathbb{E}_{\mathbf{x}}(\text{D}_h^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x})) \geq D(\log N/N)^\alpha \right) \\ &\quad + \mathbb{P}_{G_0} \left( \left| (P_N - P_{G_0}) \log \frac{\bar{p}_{\widehat{G}_N^{(\kappa)}}}{p_{G_0}} \right| \geq D \frac{\log^{\alpha+1/2} N}{N^{\alpha+1/2}}, \mathbb{E}_{\mathbf{x}}(\text{D}_h^2(p_{\widehat{G}_N^{(\kappa)}}(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x})) \leq D(\log N/N)^\alpha \right) \\ &\leq c_1 N^{-c_2} + \mathbb{P}_{G_0} \left( \sup_{\mathbb{E}_{\mathbf{x}}(\text{D}_h^2(p_G(\cdot|\mathbf{x}), p_{G_0}(\cdot|\mathbf{x})) \leq D(\log N/N)^\alpha} \left| \sqrt{N}(P_N - P_{G_0}) \log \frac{\bar{p}_G}{p_{G_0}} \right| \geq D \frac{\log^{\alpha+1/2} N}{N^\alpha} \right) \\ &\leq c'_1 N^{-c_2}. \end{aligned}$$

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

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

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

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

Assume that  $\widehat{G}_N^{(K_0)} = \sum_{k=1}^{K_0} \exp(\omega_{0k}^N) \delta_{(\omega_{1k}^N, \mathbf{a}_k^N, b_k^N, \sigma_k^N)}$ , since  $D_E(\widehat{G}_N^{(K_0)}, G_0) \rightarrow 0$  as  $N \rightarrow \infty$ , 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  $\mathbf{t}_1 \in \mathbb{R}^D$  independent of  $N$  such that  $\exp(\omega_{0k}^N) \rightarrow \exp(\omega_{0k}^0 + t_0)$  and  $\omega_{1k}^N \rightarrow \omega_{1k}^0 + \mathbf{t}_1$  as  $N \rightarrow \infty$  for all  $k \in [K_0]$ . By the definition of  $D_E(\widehat{G}_N^{(K_0)}, G_0)$ , we get for large  $N$  enough

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

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

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

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

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

$$\begin{aligned} \exp(\omega_{0k}^N) &\geq (1 - c_p \epsilon_N) \exp(\omega_{0k}^0 + t_0), & \forall k \in [K_0], \\ \sum_{k=1}^{K_0} \exp((\omega_{1k}^0 + \mathbf{t}_1)^\top \mathbf{x} + \omega_{0k}^0 + t_0) &\geq (1 - c_q \epsilon_N) \sum_{k=1}^{K_0} \exp((\omega_{1k}^N)^\top \mathbf{x} + \omega_{0k}^N), & \forall k \in [K_0]. \end{aligned}$$

Hence, we have

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

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

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

Therefore, we have

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

Hence

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

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

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

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

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

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

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

approach 1 when  $N \rightarrow \infty$ , where  $A_N$  is defined in Section D.3. 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  $|\log p_G(y | \mathbf{x})| \leq m(y | \mathbf{x})$  for a measurable function  $m$  for all  $G \in \mathcal{O}_\kappa$ , we can use uniform law of large number to get that

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

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

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

We know that  $\log p_{\widehat{G}_N^{(\kappa)}} \rightarrow \log p_{G_0^{(\kappa)}}$  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  $f(\mathbf{x}, y|\boldsymbol{\theta}) := \exp(\boldsymbol{\omega}_1^\top \mathbf{x}) \mathcal{N}(y|\mathbf{a}^\top \mathbf{x} + b, \sigma)$ .

**Lemma 2.** *The condition K is satisfied for  $f(\mathbf{x}, y|\boldsymbol{\theta}) := \exp(\boldsymbol{\omega}_1^\top \mathbf{x}) \mathcal{N}(y|\mathbf{a}^\top \mathbf{x} + b, \sigma)$ , where  $\boldsymbol{\theta} = (\boldsymbol{\omega}_1, \mathbf{a}, b, \sigma) \in \mathbb{R}^D \times \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}$  and  $\mathcal{X}$  are bounded as from the initial setup, and the eigenvalues of  $\sigma$  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  $\boldsymbol{\theta}^0 = (\boldsymbol{\omega}_1^0, \mathbf{a}^0, b^0, \sigma^0)$ , by the equivalence of the norm, we can consider the cases where  $\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_1^0\|, \|\mathbf{a} - \mathbf{a}^0\|, \|b - b^0\|, \|\sigma - \sigma^0\| \leq \epsilon$ . We aim to show that for sufficiently small  $\epsilon$ , there exist  $c_\alpha, c_\beta > 0$  such that

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

which is equivalent to

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

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

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

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

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

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

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

So that

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

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

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

Let  $u := y - (\mathbf{a}^0)^\top \mathbf{x} - b^0$ ,  $\Delta u := (\mathbf{a}^0)^\top \mathbf{x} + b^0 - [\mathbf{a}^\top \mathbf{x} + b]$ , using the boundedness of  $\sigma$ , there exist  $c_\sigma$  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

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

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  $c_\alpha$  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  $(1 + c_\beta \epsilon) (y - (\mathbf{a}^0)^\top \mathbf{x} - b^0)^\top (\sigma^0)^{-1} (y - (\mathbf{a}^0)^\top \mathbf{x} - b^0) - (y - \mathbf{a}^\top \mathbf{x} - b)^\top (\sigma)^{-1} (y - \mathbf{a}^\top \mathbf{x} - b) + c_\alpha \epsilon \geq 0$ .

Therefore, we complete the proof.  $\square$

## D.6 Proof of Theorem 4

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

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

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

and in the proof of Theorem 3, we get

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

Then we have

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

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