Refined Convergence Rates for Maximum Likelihood Estimation under Finite Mixture Models

Tudor Manole ∗ Nhat Ho †

Department of Statistics and Data Science, Carnegie Mellon University ∗
Department of Statistics and Data Sciences, University of Texas, Austin †

February 12, 2022

Abstract

We revisit convergence rates for maximum likelihood estimation (MLE) under finite mixture models. The Wasserstein distance has become a standard loss function for the analysis of parameter estimation in these models, due in part to its ability to circumvent label switching and to accurately characterize the behaviour of fitted mixture components with vanishing weights. However, the Wasserstein metric is only able to capture the worst-case convergence rate among the remaining fitted mixture components. We demonstrate that when the log-likelihood function is penalized to discourage vanishing mixing weights, stronger loss functions can be derived to resolve this shortcoming of the Wasserstein distance. These new loss functions accurately capture the heterogeneity in convergence rates of fitted mixture components, and we use them to sharpen existing pointwise and uniform convergence rates in various classes of mixture models. In particular, these results imply that a subset of the components of the penalized MLE typically converge significantly faster than could have been anticipated from past work. We further show that some of these conclusions extend to the traditional MLE. Our theoretical findings are supported by a simulation study to illustrate these improved convergence rates.

1 Introduction

Finite mixture models form a celebrated tool for modelling heterogeneous data, and are used pervasively in the life and physical sciences [2, 21, 25]. The primary goal in many of these applications is to perform statistical inference for the mixture parameters. This raises the classical question of characterizing the optimal convergence rates for parameter estimation in finite mixture models. Though this topic has been the subject of considerable investigation in past literature, the aim of our work is to show how these existing results may be refined, through a careful choice of the loss function used in their analyses.

Mixture distributions do not enjoy the standard regularity conditions that are typically presumed in parametric models, such as non-degeneracy of the Fisher information. As a result, optimal rates of estimation in mixtures are strictly slower than the usual parametric rate of convergence. This observation dates back at least to the seminal work of [4], who analyzed univariate mixtures satisfying a regularity condition known as strong identifiability, which we formally define in Section 2.
below. A long line of recent work has further analyzed convergence rates in mixtures of general dimension, under varying degrees of strong identifiability. In particular, Nguyen [26] introduced the Wasserstein distance as a natural tool for metrizing convergence of parameters in finite mixtures, via their mixing measure. The Wasserstein metric was then used to analyze convergence rates for the maximum likelihood estimator (MLE) and related procedures, under various classes of finite mixture models [18, 17, 16, 19]. Moment-based estimators were also studied by [30, 8], and Bayesian estimators by [27, 14], to name a few.

A broad conclusion of these works is that slow convergence rates are pervasive to parameter estimation in finite mixture models. This observation contrasts the fact that the minimax rate of estimating the density of a finite mixture model is typically the standard parametric rate of convergence [12, 13, 8]. For example, Henrich et al. [16] show that the minimax rate for parameter estimation in a strongly identifiable mixture degrades exponentially as the number of components increases, when no separation conditions are placed on these components. This result suggests that the estimation of mixture parameters can be prohibitive, even when the number of components is moderate. On the other hand, practitioners have long been employing mixture models successfully, suggesting a discrepancy between practice and the worst-case rates suggested by the theory.

The goal of this paper is to revisit existing convergence rates for parameter estimation in finite mixture models, and to show that they may be refined by using stronger loss functions than the Wasserstein distance. We will argue that the Wasserstein distance is only able to capture the worst-case convergence rate among the estimated components of a mixture, and that in many cases, the vast majority of estimated component parameters may achieve considerably faster convergence rates than anticipated from prior work. Before describing these phenomena in further detail, we begin by formally introducing finite mixture models and related notions.

1.1 Problem Setting

**Finite Mixture Models.** Let \( \mathcal{F} = \{ f(x|\theta) : x \in \mathcal{X}, \theta \in \Theta \} \) be a given parametric family of density functions with respect to a dominating \( \sigma \)-finite measure \( \nu \). Here, we assume \( \mathcal{X} \subseteq \mathbb{R}^N \) for some \( N \geq 1 \), and \( \Theta \) is a parameter space which will either be a subset of the Euclidean space \( \mathbb{R}^d, d \geq 1 \), or of the set \( \mathbb{R}^d \times S^d_{++} \), where \( S^d_{++} \) denotes the cone of \( d \times d \) positive definite matrices. In either case, we shall always tacitly assume that \( \Theta \) is a compact set with nonempty interior. Let \( X_1, X_2, \ldots, X_n \) be an i.i.d. sample from a finite mixture model with \( k_0 \geq 1 \) components, whose density with respect to \( \nu \) is written as

\[
p_{G_0}(x) := \int f(x|\theta)dG_0(\theta) = \sum_{i=1}^{k_0} p^0_i f(x|\theta^0_i), \quad x \in \mathcal{X}.
\]

Here \( G_0 = \sum_{j=1}^{k_0} p^0_j \delta_{\theta^0_j} \) denotes an unknown mixing measure, where the \( p^0_j \geq 0 \) are called mixing proportions (or weights), satisfying \( \sum_{j=1}^{k_0} p^0_j = 1 \), and the \( \theta^0_j \in \Theta \) are called atoms, for \( j = 1, \ldots, k_0 \). When the mixing proportions are strictly positive and the atoms are distinct, we say \( G_0 \) has true
order $k_0$. More generally, any finitely-supported probability measure on $\Theta$ is called a mixing measure, and its support size is called its order. The set of mixing measures of order at most $k \geq 1$ is denoted $\mathcal{O}_k(\Theta)$, and we write $\mathcal{E}_k(\Theta) = \mathcal{O}_k(\Theta) \setminus \mathcal{O}_{k-1}(\Theta)$.

When dealing with parameter estimation in a finite mixture model, it is convenient to treat the mixing measure $G_0$ as the target of estimation, even if the main quantities of interest are the mixing proportions or atoms of $G_0$. Indeed, while the density $p_G$ is typically identifiable with respect to its mixing measure $G$, it is never identifiable with respect to the individual parameters of $G$, due to the possibility of label-switching. Throughout our work, we will consider both pointwise rates of estimating the mixing measure, that is, estimation rates which depend on the fixed mixing measure $G_0$, and uniform estimation rates, which hold uniformly over all mixing measures under consideration. We will always emphasize the latter setting by allowing $G_0 \equiv G_0^n$ to potentially depend on the sample size $n$.

**Maximum Likelihood Estimation.** Perhaps the most widely-used estimator of $G_0$ is the maximum likelihood estimator (MLE). We focus our analysis on estimators based on the MLE throughout this work, in part because they allow for a general theory of parameter estimation to be derived under minimal conditions on the family $\mathcal{F}$. Given an integer $k \geq 1$, the MLE of $G_0$ with order at most $k$ is given by

$$
\hat{G}_n = \sum_{i=1}^{\bar{k}_n} \hat{p}_i^n \hat{\delta}_{\hat{\theta}_i^n} = \arg\max_{G \in \mathcal{O}_k(\Theta)} \ell_n(G), \quad \text{where } \ell_n(G) = \sum_{i=1}^n \log p_G(X_i). \quad (1)
$$

Here, $\bar{k}_n \leq k$ denotes the fitted order of $\hat{G}_n$. We have defined the MLE with the general order $k$ to reflect the fact that true order $k_0$ of $G_0$ may be unknown. Notice that $\hat{G}_n$ is generally inconsistent if $k < k_0$, thus we shall always assume $k \geq k_0$. Our convergence rates will depend on the level of misspecification $k - k_0$.

In certain parts of our development, it will be technically convenient to ensure that the fitted mixing proportions of $\hat{G}_n$ do not vanish. While this can be achieved by constraining the maximum in equation (1), we will prefer to achieve this using a penalty on the likelihood function. Specifically, we follow [5] and define the penalized MLE of order at most $k$ by

$$
\hat{G}_n = \sum_{i=1}^{\tilde{k}_n} \hat{p}_i^n \hat{\delta}_{\hat{\theta}_i^n} = \arg\max_{G \in \mathcal{O}_k(\Theta)} \ell_n(G) + \xi_n \rho(G),
$$

where $\tilde{k}_n \leq k$ is the order of $\hat{G}_n$, $\xi_n \geq 0$ is a tuning parameter, and $\rho$ satisfies $\rho(G) \to -\infty$ as the smallest mixing weight of $G$ vanishes. For concreteness, we will use the penalty $\rho(G) = \sum_{j=1}^{k'} \log p_j^n$, where $k' \leq k$ denotes the order of $G = \sum_{j=1}^{k'} p_j^n \delta_{\theta_j^n}$. As discussed in Appendix C.1, with this choice of penalty, $\hat{G}_n$ may be numerically approximated using a simple modification of the EM algorithm.

In order to evaluate the risk of the estimators $\hat{G}_n$ and $\overline{G}_n$, we will require loss functions defined over $\mathcal{O}_k(\Theta)$. The most widely-used loss function appearing in past work is the Wasserstein distance, which we define next.
Wasserstein Distances. Let \( k, k' \geq 1 \), and set \( G = \sum_{i=1}^{k} p_i \delta_{\theta_i} \in \mathcal{O}_k(\Theta) \) and \( G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j} \in \mathcal{O}_{k'}(\Theta) \). Denote by \( \Pi(G, G') \) the set of joint probability mass functions \( q = (q_{ij} : i \in [k], j \in [k']) \) admitting marginal distributions equal to those of \( G \) and \( G' \), that is, \( \sum_{i=1}^{k} q_{ij} = \pi_j' \) and \( \sum_{j=1}^{k'} q_{ij} = \pi_i \), for all \( i \in [k] \) and \( j \in [k'] \). The Wasserstein distance of order \( r \geq 1 \) is defined by

\[
W_r(G, G') = \left( \inf_{q \in \Pi(G, G')} \frac{1}{r} \sum_{i=1}^{k} \sum_{j=1}^{k'} q_{ij} D^r(\theta, \theta'_j) \right)^{\frac{1}{r}},
\]

where \( D \) is a metric on \( \Theta \). When \( \Theta \subseteq \mathbb{R}^d \), we shall always assume that \( D = \| \cdot \| \) is induced by the Euclidean norm.

The use of Wasserstein distances in general dimension originated from the work of [26], and was partly motivated by its implication for the convergence of atoms, as we now recall. Let \( G_n \in \mathcal{O}_k(\Theta) \) be a sequence of mixing measures, and \( G_0 \in \mathcal{E}_{k_0}(\Theta) \). Then, if \( W_r(G_n, G_0) \leq \alpha_n \) for some \( \alpha_n \downarrow 0 \), there exists a subsequence of \( G_n \) such that every atom \( \theta_{0j} \) of \( G_0 \) is the limit point of at least one atom \( \theta_j^n \) of \( G_n \). Furthermore, the convergence rate of this fitted atom is \( D(\theta_j^n, \theta_{0j}) \lesssim \alpha_n \). When \( k > k_0 \), there may also be atoms \( \theta_{\ell}^n \) of \( G_n \) which do not converge to any atoms of \( G_0 \). It can be seen that their corresponding mixing proportions \( p_{\ell}^n \) must then vanish at the rate \( \alpha_n^n \). If we instead assume that the mixing proportions of \( G_n \) are bounded from below by a positive constant \( c_0 > 0 \), it must in fact hold that every atom of \( G_n \) converges to an atom of \( G_0 \) at rate \( \alpha_n \).

We note in particular that the Wasserstein distance can only induce the same convergence rate \( \alpha_n \) for those atoms of \( G_n \) which approach the atoms of \( G_0 \). In contrast, a key observation of our work is that maximum likelihood-based estimators have atoms which converge at distinct rates; such heterogeneous behaviour cannot be captured by the Wasserstein distance, and is the subject of this paper.

1.2 Contributions

Our goal is to provide sharper rates of convergence for parameter estimation in finite mixture models of various types. Our main technical contribution is the development of loss functions over the space of mixing measures, which are stronger than the Wasserstein distance, and which correctly characterize the heterogeneous convergence rates of the various mixture parameters in maximum likelihood-based estimators. To illustrate the refinements furnished by our theory, we consider the following example.

**Example 1** (Pointwise Convergence Rates for Strongly Identifiable Mixtures). Suppose \( \mathcal{F} \) is the location family of Gaussian densities with known variance. Furthermore, assume \( k = k_0 + 1 \). The works of [4, 18] show there exists a constant \( C(G_0) > 0 \) such that

\[
\text{EW}_2(G_n, G_0) \leq C(G_0)(\log n/n)^{1/4}.
\]

In particular, it follows that for every atom of \( G_0 \), there is at least one atom of \( G_n \) which converges.
to $G_0$ at the pointwise rate $(\log n/n)^{1/4}$. Equivalently, there exists an injection $u_n : [k_0] \to [k]$ such that
\[ \max_{1 \leq j \leq k_0} \mathbb{E} \| \tilde{\theta}_n(j) - \theta_0^j \| \leq C(G_0) \left( \frac{\log n}{n} \right)^{1/4}. \] (2)

In contrast, it will follow from our Theorem 4 below that there exists an injection $v_n : [k_0] \to [k]$ and a permutation $\sigma_n : [k_0] \to [k_0]$ such that
\[ \max_{1 \leq j \leq k_0 - 1} \mathbb{E} \| \tilde{\theta}_n(j) - \theta_0^{\sigma_n(j)} \| \leq C(G_0) \left( \frac{\log n}{n} \right)^{1/4}, \]
\[ \mathbb{E} \| \tilde{\theta}_n(k_0) - \theta_0^{\sigma_n(k_0)} \| \leq C(G_0) \left( \frac{\log n}{n} \right)^{1/4}. \]

This result shows that, ignoring polylogarithmic factors, all but two of the atoms of the overfitted MLE $\hat{G}_n$ achieve the parametric convergence rate. In contrast, equation (2) merely shows that these atoms converge at the slower rate $(\log n/n)^{1/4}$.

We will show that similar asymptotics hold for a broad family of strongly identifiable mixture models, and for general $k \geq k_0$, in Section 3.1. We further consider uniform convergence rates for such families in Section 4, as well as pointwise convergence rates for location-scale Gaussian mixture models (Section 3.2), which form an important example of weakly identifiable finite mixtures. We obtain these results by identifying distinct loss functions tailored to each of these three settings, which accurately capture the behaviour of individual fitted mixture parameters.

Our results highlight the underappreciated fact that the Wasserstein distance merely quantifies the worst-case convergence rate among the fitted parameters of a finite mixture; its use in past work may thus have painted an overly pessimistic picture of parameter estimation in these models. Though our primary emphasis is on such theoretical aspects, we will also discuss that certain loss functions developed in this work enjoy an improved computational complexity as compared to the Wasserstein distance, and may therefore be of practical significance in their own right.

**Notation.** We denote by $O_{k,c_0}(\Theta)$ the set of mixing measures in $O_k(\Theta)$ with mixing weights bounded below by a constant $c_0 > 0$, and $\mathcal{E}_{k,c_0}(\Theta) = O_{k,c_0}(\Theta) \setminus O_{k-1}(\Theta)$. For any $n \geq 1$, we denote $[n] = \{1, 2, \ldots, n\}$. For any $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Given $(a_n)_{n \geq 1}, (b_n)_{n \geq 1} \subseteq \mathbb{R}_+$, we write $a_n \preceq b_n$ if there exists a universal constant $C > 0$, possibly depending on other problem parameters to be understood from context, such that $a_n \leq Cb_n$ for all $n \geq 1$. Furthermore, we write $a_n \asymp b_n$ when $a_n \preceq b_n \preceq a_n$. $C^\alpha(\Theta)$ denotes the Hölder space of regularity $\alpha > 0$ over $\Theta$ [11]. Given probability densities $p, q$ dominated by $\nu$, their squared Hellinger and Total Variation distances are denoted by $h^2(p, q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 d\nu$ and $V(p, q) = \frac{1}{2} \int |p - q| d\nu$. 

2 Preliminaries

2.1 Strong identifiability

We begin by recalling the strong identifiability condition for the parametric family $F$.

**Definition 1 (Strong Identifiability).** Let $r \geq 0$ be an integer. We say $F$ is $r$-strongly identifiable if $f(x|\cdot) \in C^r(\Theta)$ for $\nu$-almost every $x \in \mathcal{X}$, and for any $k \geq 1$, and $\theta_1, \ldots, \theta_k \in \Theta$, the following implication holds for all $\alpha^{(i)} \in \mathbb{R}$,

$$
\text{esssup}_{x \in \mathcal{X}} \left| \sum_{\ell=0}^{r} \sum |\eta| = \ell \sum_{i=1}^{k} \alpha^{(i)} \frac{\partial^{\lfloor \eta \rfloor}}{\partial \theta^{\eta}} f(x|\theta_i) \right| = 0 \implies \max_{|\eta| \leq r} \max_{1 \leq i \leq k} |\alpha^{(i)}| = 0.
$$

The notion of strong identifiability originates from the work of [4], and is stated here in a more general form due to [16, 18]. We refer to these references, as well as that of [20], for sufficient conditions under which the strong identifiability condition holds. For example, this condition is known to be satisfied for any finite $r \geq 1$ by the location Gaussian parametric family with known scale parameter, the Poisson family, and other common exponential families. Location-scale Gaussian densities form perhaps the most widely-used parametric family which fails to satisfy the $r$-strong identifiability condition for $r \geq 2$ [17], and we will treat this special case separately.

We will typically couple the strong identifiability condition with the following assumption on the modulus of continuity of the derivatives of $f(x|\cdot)$, up to order $r \geq 1$.

**A(r)** There exist $\Lambda, \delta > 0$ such that

$$
\text{esssup}_{x \in \mathcal{X}} \|f(x|\cdot)\|_{C^{r+\delta}(\Theta)} \leq \Lambda.
$$

Strong identifiability generalizes the condition of regular identifiability of the family $P_k(\Theta) = \{ p_G : G \in O_k(\Theta) \}$, and is a useful notion for deriving inequalities between Wasserstein-type distances over $O_k(\Theta)$ and statistical distances over $P_k(\Theta)$. Such bounds are at the heart of our proofs, and will allow us to derive parameter estimation rates from known convergence rates for maximum likelihood density estimation, to which we turn our attention next.

2.2 Convergence rates for maximum likelihood density estimators

In order to state a rate of convergence for the density estimators $p_{\hat{G}_n}$ and $p_{\hat{\sigma}_n}$, for instance under the Hellinger distance, we require a condition on the complexity of the class

$$
P_{k}^{1/2}(\Theta, \epsilon) = \left\{ \tilde{p}^{1/2}_G : G \in O_k(\Theta), h(\tilde{p}_G, p_{G_0}) \leq \epsilon \right\},
$$

where $\epsilon > 0$, and for any $G \in O_k(\Theta)$, we write $p_G = (p_G + p_{G_0})/2$. The definition of $P_{k}^{1/2}(\Theta, \epsilon)$ originates from [29], who place conditions on the convex combinations $\tilde{p}_G$ rather than $p_G$, as this
choice is guaranteed to place a non-negligible amount of probability mass over the support of \( p_{G_0} \).

The complexity of this class is measured through the bracketing entropy integral

\[
J_B(\epsilon, \mathcal{F}_k^{1/2}(\Theta, \epsilon), \nu) = \int_0^\epsilon \sqrt{H_B(u, \mathcal{F}_k^{1/2}(\Theta, u), \nu)} du \vee \epsilon,
\]

where \( H_B(\epsilon, \mathcal{P}, \nu) \) denotes the \( \epsilon \)-bracketing entropy of a set \( \mathcal{P} \subseteq L^2(\nu) \) with respect to the \( L^2(\nu) \) metric [29]. We shall assume that this quantity satisfies the following condition.

\( B(k) \) Given a universal constant \( J > 0 \), there exists a constant \( L > 0 \), possibly depending on \( d \) and \( k \), such that for all \( n \geq 1 \) and all \( \epsilon > L(\log n/n)^{1/2} \),

\[
J_B(\epsilon, \mathcal{F}_k^{1/2}(\Theta, \epsilon), \nu) \leq J \sqrt{n\epsilon^2}.
\]

We are now ready to state the following convergence rates.

**Theorem 2.** Given \( k \geq 1 \), assume condition \( B(k) \) holds.

(i) There exists a constant \( C > 0 \) depending only on \( d, k, \mathcal{F} \) such that for all \( n \geq 1 \),

\[
\sup_{G_0 \in \mathcal{O}_k(\Theta)} \mathbb{E}_{G_0} h(p_{G_n}, p_{G_0}) \leq C \sqrt{\log n \over n}.
\]

(ii) Furthermore, given \( c_0, c_1 > 0 \), if \( 0 \leq \xi_n \leq c_1 \log n \), then there exists a constant \( C' > 0 \) depending on \( d, k, c_0, c_1, \mathcal{F} \) such that for all \( n \geq 1 \),

\[
\sup_{G_0 \in \mathcal{O}_{k,c_0}(\Theta)} \mathbb{E}_{G_0} h(p_{\hat{G}_n}, p_{G_0}) \leq C' \log n \over \sqrt{n}.
\]

Theorem 2(ii) is a direct consequence of generic results for maximum likelihood density estimation (for instance, Theorem 7.4 of [29]). Its application to finite mixture models has previously been discussed by [18], who also argue that condition \( B(k) \) is satisfied by a broad collection of parametric families \( \mathcal{F} \), including the multivariate location-scale Gaussian and Student-\( t \) families. A version of Theorem 2(ii) is implicit in the work of [24], though with a stronger condition on the tuning parameter \( \xi_n \). We provide a self-contained proof of this result in Appendix A for completeness.

These results may also be used to show that the penalized MLE has nonvanishing mixing proportions.

**Proposition 3.** Let \( k \geq 1 \), \( c_0 \in (0, 1) \), and assume condition \( B(k) \) holds. Assume further that \( \xi_n \geq \log n \). Then, there exists a constant \( c > 1 \) depending on \( c_0, d, k, \mathcal{F} \) such that for all \( n \geq 1 \),

\[
\sup_{G_0 \in \mathcal{O}_{k,c_0}(\Theta)} \mathbb{P}_{G_0} \left( \min_{1 \leq j \leq k_n} \hat{p}_j \geq \frac{1}{c} \right) \leq \frac{c}{n}.
\]
In view of Proposition 3 and Theorem 2, we shall always tacitly assume that the tuning parameter $\xi_n$ is equal to $\log n$.

3 Pointwise convergence rates of the MLE

We first derive pointwise convergence rates for estimating a fixed mixing measure $G_0 \in \mathcal{E}_{k_0}(\Theta)$.

3.1 Strongly identifiable case

Assume the family $\mathcal{F}$ is twice strongly identifiable, with a compact parameter space $\Theta \subseteq \mathbb{R}^d$ admitting nonempty interior. We begin by defining a loss function on $O_k(\Theta)$ tailored to this setting. Given a mixing measure $G = \sum_{i=1}^{k'} p_i \delta_{\theta_i}$ of order $k' \leq k$, we partition its atoms into the following Voronoi cells, generated by the support of $G_0$,

$$A_j \equiv A_j(G) = \{ i \in [k'] : \| \theta_i - \theta_j^0 \| \leq \| \theta_i - \theta_{\ell}^0 \| \forall \ell \neq j \},$$

for all $j \in [k_0]$. We may then define the loss function

$$D(G,G_0) := \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \|^2 + \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \| + \sum_{j=1}^{k_0} \left| \sum_{i \in A_j} p_i - p_j^0 \right|. \quad (3)$$

Clearly, $D(G,G_0)=0$ if and only if $G=G_0$. Under this loss function, we obtain the following bound on the risk of $G_n$.

**Theorem 4.** Let $k \geq k_0$. Assume that the parametric family $\mathcal{F}$ is 2-strongly identifiable, and satisfies conditions $A(2)$ and $B(k)$. Then, there exists a constant $C(G_0) > 0$, depending on $G_0, d, k, \mathcal{F}$, such that

$$\mathbb{E} [D(G_n,G_0)] \leq C(G_0) \sqrt{\frac{\log n}{n}}.$$

The proof of Theorem 4 appears in Appendix A.3, where the main difficulty is to prove the following lower bound of the Hellinger distance in terms of $D$,

$$D(G,G_0) \leq C(G_0) h(p_G,p_{G_0}), \quad (4)$$

for any $G \in \mathcal{O}_k(\Theta)$. Using Theorem 2(i), the above bound directly leads to the stated convergence rate of $\tilde{G}_n$.

A few comments regarding Theorem 4 are in order. First, let $A_j^n = A_j(\tilde{G}_n)$ for all $j \in [k_0]$. The convergence rate $\sqrt{\log n/n}$ of $D(G_n,G_0)$ implies that for any index $j \in [k_0]$ such that $|A_j^n| = 1$, $\| \theta_i^n - \theta_j^0 \|$ and $|p_i^n - p_j^0|$ vanish at the near-parametric rate $\sqrt{\log n/n}$ for $i \in A_j^n$. Therefore, among the true
components which are only approximated by a single fitted component, the parameters of this fitted component converge as fast as if the order $k \geq k_0$ were not overspecified. In particular, in the exact-fitted setting $k = k_0$, we find that all fitted components and mixing proportions converge at the parametric rate, up to a polylogarithmic factor, which recovers Theorem 3.1 of [18]. Furthermore, when $k > k_0$, for any index $j \in [k_0]$ such that $|A^*_j| \geq 2$, $\sum_{i \in A^*_j} \hat{p}^n_i \|\theta^*_i - \theta^0_j\|_2$ and $|\sum_{i \in A^*_j} \hat{p}^n_i - p^0_j|$ decay at the rate $\sqrt{\log n/n}$. In particular, it follows that for every such $j$, there exists $i \in A^*_j$ such that $\theta^n_i$ converges to $\theta^0_j$ at the rate $(\log n/n)^{1/4}$, which is now markedly slower than the parametric rate. In contrast, the past works of [4, 26, 18] show that $\mathbb{E} W^2_2(G_n, G_0) \lesssim \sqrt{\log n/n}$, which implies a convergence rate no better than $(\log n/n)^{1/4}$ for all atoms of the MLE, rather than just those lying in a set $A^*_j$ with cardinality greater than one. These existing results painted a pessimistic picture of maximum likelihood estimation in overspecified mixtures—for example, they suggest that overspecifying the order $k_0$ merely by $k = k_0 + 1$ leads to poor convergence rates for each of the $k$ fitted atoms, whereas our work shows that at least $k_0 - 1$ fitted atoms enjoy considerably faster convergence rates.

Second, we can demonstrate that $D \gtrsim W^2_2$, and

$$\sup_{G \neq G_0} \frac{D(G, G_0)}{W^2_2(G, G_0)} = \infty.$$ 

See Lemma 14 in Appendix B for a formal statement. This shows that $D$ is a stronger loss function than the Wasserstein distance. In particular, we deduce that that Theorem 4 also implies the aforementioned convergence rate of $G_n$ under the Wasserstein distance.

Finally, the complexity of computing $D(G, G_0)$ is of the order of $O(k \times k_0)$. In contrast, computing $W^2_2(G, G_0)$ is equivalent to solving a linear programming problem, which has complexity no better than $O(k^3)$ [28]. Therefore, the loss function $D$ is computationally more efficient than the Wasserstein metric. This observation is significant because the Wasserstein distance has previously been used as a methodological tool for model selection in finite mixtures [14]. In these applications, the loss function $D$ provides an alternative to $W^2_2$ which is both statistically and computationally more efficient.

### 3.2 Weakly identifiable case: location-scale Gaussian mixtures

In this section, we study the convergence rate of the MLE when the model is not strongly identifiable in the second order. Location-scale Gaussian mixtures are a popular example of such models, as a result of the following equation:

$$\frac{\partial^2 f}{\partial \mu \partial \mu^\top}(x|\mu, \Sigma) = 2 \frac{\partial f}{\partial \Sigma}(x|\mu, \Sigma),$$

for all $x \in \mathbb{R}^d$ and $\theta = (\mu, \Sigma) \in \Theta$, where $\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$ denotes the family of location-scale Gaussian densities, with compact parameter space $\Theta \subseteq \mathbb{R}^d \times \mathbb{S}^{d-1}$. The absence of second
order identifiability in location-scale Gaussian mixtures leads to several challenges in studying the convergence rates of the MLE. To simplify our proofs, we will assume that all mixing measures have weights which are lower bounded by some small constant $c_0 > 0$. As a result, we only state a convergence rate for the penalized MLE $\hat{G}_n$, which indeed lies in the class $O_{k,c_0}(\Theta)$ with high probability, by Proposition 3. We would like to remark that constraints on the mixing weights are commonly assumed in past work on convergence rates for over-specified location-scale Gaussian mixtures [17], and are not a byproduct of our choice of loss function.

Proposition 2.2 in [17], together with Theorem 2 and Proposition 3, may be used to establish the following bound, for some constant $C(G_0) > 0$,

$$\mathbb{E}[W_{\tau(k-k_0+1)}(\hat{G}_n, G_0)] \leq C(G_0) \left( \frac{\log n}{\sqrt{n}} \right)^{\frac{1}{r(k-k_0+1)}},$$

where for any $k' \geq 2$, $\tau(k')$ is defined as the smallest integer $r$ such that the system of polynomial equations

$$\sum_{j=1}^{k'} \sum_{n_1,n_2} c_j^2 a_j^{n_1} b_j^{n_2} n_1! n_2! = 0, \text{ for each } \alpha = 1, \ldots, r$$

(6)

does not have any nontrivial solution for the unknown variables $(a_j, b_j, c_j)_{j=1}^{k'} \subseteq \mathbb{R}$. The range of $(n_1, n_2)$ in the second sum consist of all natural pairs satisfying the equation $n_1 + 2n_2 = \alpha$. A
solution to the above system is considered nontrivial if all variables \(c_j\) are non-zero, while at least one of the \(a_j\) is non-zero. For example, it was shown by [18] that \(r(2) = 4\) and \(r(3) = 6\).

The convergence rate \((\log n / \sqrt{n})^{1/r(k-k_0+1)}\) of \(\hat{G}_n\) indicates that the location and scale parameters of the penalized MLE converge to their population counterparts at this same slow rate. As before, this result does not precisely reflect the behavior of individual location and scale parameters in location-scale Gaussian mixtures, leading us to consider a stronger loss function than the Wasserstein distance. Given \(G = \sum_{i=1}^{k'} p_i \delta(\mu_i, \Sigma_i) \in \mathcal{E}_k(\Theta)\) for \(k' \leq k\), define the Voronoi cells \(A_j = A_j(G) = \{i \in [k'] : \|\mu_i - \mu_j^0\| + \|\Sigma_i - \Sigma_j^0\| \leq \|\mu_i - \mu_j^0\| + \|\Sigma_i - \Sigma_j^0\| \forall \ell \neq j\}\), for \(j \in [k_0]\), and set

\[
\mathcal{D}(G, G_0) := \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \left( \|\mu_i - \mu_j^0\| + \|\Sigma_i - \Sigma_j^0\| \right) + \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i \left( \|\mu_i - \mu_j^0\| / \|\Sigma_i - \Sigma_j^0\|^{r(|A_j|)} \right) + \sum_{j=1}^{k_0} \left| \sum_{i \in A_j} p_i - p_j^0 \right|
\]

It can be shown that \(\mathcal{D} \geq W_{\tau(k-k_0+1)}^r(G, G_0)\) and

\[
\sup_{G \in \mathcal{O}_k(\Theta)} \mathcal{D}(G, G_0) / W_{\tau(k-k_0+1)}^r(G, G_0) = \infty.
\]

The proof is similar to that of Lemma 14 in Appendix B; therefore, it is omitted. We deduce that \(\mathcal{D}\) is a stronger loss function than \(W_{\tau(k-k_0+1)}^r\). We bound the risk of the penalized MLE under \(\mathcal{D}\) as follows.

**Theorem 5.** Let \(\mathcal{F}\) denote the location-scale Gaussian density family with parameter space taking the form \(\Theta = [-a, a]^d \times \Omega\), where \(a > 0\) and \(\Omega\) is a compact subset of \(\mathbb{S}^{d-1}\) whose eigenvalues lie in a closed interval contained in \((0, \infty)\). Then, there exists a constant \(C(G_0) > 0\), depending only on \(G_0, k, d, \Theta\), such that

\[
\mathbb{E} \left[ \mathcal{D}(\hat{G}_n, G_0) \right] \leq C(G_0) \frac{\log n}{\sqrt{n}}.
\]

The proof of Theorem 5 appears in Appendix A.4. Recall that \(\hat{G}_n = \sum_{i=1}^{k_n} \hat{p}_i^n \delta(\hat{\mu}_i^n, \hat{\Sigma}_i^n)\), and write \(A_j^n = A_j(\hat{G}_n)\) for all \(j \in [k_0]\). Theorem 5 implies the following.

(i) Given \(j \in [k_0]\) such that \(|A_j^n| \geq 2\), we have, with probability tending to one,

\[
\|\hat{\mu}_i^n - \mu_j^0\| \lesssim (\log n / \sqrt{n})^{1/r(|A_j^n|)}, \quad \|\hat{\Sigma}_i^n - \Sigma_j^0\| \lesssim (\log n / \sqrt{n})^{2/r(|A_j^n|)}, \quad i \in A_j^n.
\]

In particular, the location parameters converge quadratically slower than the scale parameters.

(ii) On the other hand, for any index \(j \in [k_0]\) such that \(|A_j^n| = 1\) and for any \(i \in A_j^n\), we have with
probability tending to one,

$$\|\hat{\mu}_i^n - \mu_j^0\| \vee \|\Sigma_i^n - \Sigma_j^0\| \lec \log n / \sqrt{n}. \quad (7)$$

Hence, both location and scale parameters achieve the standard parametric rate up to a logarithmic factor. We refer to Figure 1(a) for an illustration of these convergence rates.

(iii) Notice that $|A^n_j| \leq \hat{k}_n - k_0 + 1$ for all $j \in [k_0]$. When equality is achieved for some $j$, there must be a single Voronoi cell with $\hat{k}_n - k_0 + 1$ elements, while the remaining cells have exactly one component. In this case, there are $k_0 - 1$ components of the penalized MLE which achieve the fast pointwise rate (7).

(iv) When $k = k_0 + 1$, there exists a unique index $j$ such that $A^n_j$ has at most two components, while the remaining Voronoi cells have exactly one component. Since $\tau(2) = 4$, this demonstrates that the two components having indices in $A_j$ have means converging at the slow rate $n^{-1/8}$, and covariances converging at the rate $n^{-1/4}$, up to polylogarithmic factors. These particular rates were already anticipated by the work of [3] when $k_0 = 1$. When $k_0 > 1$, our work shows that the remaining $k_0 - 1$ atoms of the penalized MLE converge at the fast rate (7).

(v) When $k = k_0 + 2$, there are two possible cases: either (a) there exists a unique index $j'$ such that $A^n_{j'}$ has at most three components while the remaining sets have exactly one component, or (b) there exist indices $j'_1$ and $j'_2$ such that $A^n_{j'_1}$ and $A^n_{j'_2}$ have at most two components while the remaining sets have exactly one component. Under case (a), since $\tau(3) = 6$, the means with indices in $A^n_{j'}$ converge at the rate $(\log n/n)^{1/12}$ while the remaining atoms of $\hat{G}_n$ converge at the parametric rate. Under case (b), the means with indices in $A^n_{j'_1} \cup A^n_{j'_2}$ converge at the $(\log n/n)^{1/8}$ rate while the remaining atoms converge at the rate $(\log n/n)^{1/2}$.

Finally, similar to the loss function $D$ in equation (3), we note that the complexity of evaluating $\overline{D}(G, G_0)$ for $G \in \mathcal{O}_k(\Theta)$ is of the order $O(k \times k_0)$, and is thus more computationally efficient than the Wasserstein metric.

4 Uniform convergence rates of the MLE

Thus far, we have derived pointwise convergence rates for the MLE or penalized MLE, which depend on the fixed mixing measure $G_0$. We next consider uniform rates of convergence, in which we allow the true mixing measure $G_0 \equiv G_0^n \in \mathcal{E}_{k_0}(\Theta)$ to vary with the sample size $n$, while converging to some limiting mixing measure $G_* = \sum_{i=1}^{k_*} p_i^* \delta_{\theta_i^*} \in \mathcal{E}_{k_*}(\Theta)$, of order $k_* \leq k_0 \leq k$. To simplify our proof technique, we will assume that $\Theta \subseteq \mathbb{R}$.

It is known that the optimal pointwise rate of estimation in a strongly identifiable mixture differs from the optimal uniform rate. Indeed, when $\mathcal{F}$ is $(k + k_0)$-strongly identifiable it can be inferred
from Theorem 6.3 in [16] that,
\[ \mathbb{E}[W_r(G_n, G_0^n)] \lesssim \left( \frac{\log n}{n} \right)^{1/2}, \tag{8} \]
where we fix \( r = k + k_0 - 2k_* + 1 \) throughout the remainder of this section. Furthermore, the above rate is minimax optimal up to a polylogarithmic factor, but is markedly slower than its pointwise analogue discussed in Section 3.1. It implies that the atoms of \( G_n \) with nonvanishing weights tend to those of \( G_0^n \) at this same slow rate. In contrast, we will show that the uniform convergence rates of individual components of the MLE can, in fact, be sharpened. Similarly to the previous subsection, however, our results will rely on the additional condition that the mixing proportions of \( G_0^n, G_* \) are uniformly bounded below by a small constant \( c_0 > 0 \). While this condition was not needed in the work of [16], we require it for our proof technique. As a result, we focus on deriving convergence rates for the penalized MLE \( \hat{G}_n \).

Given \( k' \in [k] \), let \( G = \sum_{i=1}^{k'} p_i \delta_{\theta_i} \in \mathcal{E}_k(\Theta) \) and \( G' = \sum_{i=1}^{k_0} p'_i \delta_{\theta'_i} \in \mathcal{E}_{k_0}(\Theta) \). We again partition the supports of these measures into Voronoi cells, which are now generated by the atoms of the measure \( G_* \) rather than \( G_0^n \):
\[ A_j(G) = \{ i \in [k'] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_{\ell}^*\| \forall \ell \neq j \} , \]
for all \( j \in [k_*] \). With this notation in place, we define the following loss function over \( \mathcal{O}_k(\Theta) \),
\[ \tilde{W}(G, G') := \inf_{q \in \Pi(G, G')} \left\{ \sum_{i=1}^{k_*} \sum_{(i,j) \in A_i(G) \times A_j(G')} q_{ij} |\theta_i - \theta_j'|^{A_i(G) + |A_i(G')| - 1} + \sum_{(i,j) \notin \bigcup_{i=1}^{k_*} A_i(G) \times A_j(G')} q_{ij} \right\}. \tag{9} \]
\( \tilde{W} \) may be viewed as a generalized optimal transport cost, whose ground cost depends on the measures \( G, G' \) via the exponent \(|A_i(G)| + |A_i(G')| - 1\). In the special case where \( k_* = 1 \), this exponent is given by \( k + k_0 - 1 \), and \( \tilde{W} \) is then equal to \( W_r^r \). On the other hand, when \( k_* > 1 \), it can be seen similarly as in previous subsections that,
\[ \tilde{W} \gtrsim W_r^r, \quad \text{and} \quad \sup_{G \neq G'} \frac{\tilde{W}(G, G')}{W_r^r(G, G')} = \infty. \tag{10} \]
Therefore, the loss function \( \tilde{W} \) is stronger than the one used by [16]. The main result of this section is the following convergence rate under \( \tilde{W} \).

**Theorem 6.** Let \( k \geq k_0 \geq k_* \) and \( c_0 > 0 \). Assume that \( G_* \in \mathcal{E}_{k_*,c_0}(\Theta) \) and \( G_0^n \in \mathcal{E}_{k_0,c_0}(\Theta) \) for all \( n \geq 1 \). Furthermore, assume that \( \mathcal{F} \) is \((k + k_0)\)-strongly identifiable, and satisfies conditions \( A(k + k_0) \) and \( B(k) \). Then, there exist constants \( C, \epsilon > 0 \), depending only on \( \mathcal{F}, k, c_0 \), such that for all
\( n \geq 1 \) satisfying \( \tilde{W}(G_0^n, G_*) \leq \epsilon \), we have

\[
\mathbb{E}[\tilde{W}((\hat{G}_n, G_0^n)) \leq C \log n \sqrt{n}.
\]

The proof of Theorem 6 appears in Appendix A.5. In view of equation (10) and the existing minimax lower bound of [16] under the Wasserstein distance, it can immediately be deduced that the convergence rate in Theorem 6 is minimax optimal, up to a logarithmic factor.

Theorem 6 may be interpreted similarly as in previous sections, thus we only provide an example. In the sequel, we ignore polylogarithmic factors. For all \( l \in [k_*] \), notice that

\[
|A_l(\hat{G}_n)| \leq k - k_* + 1, \quad |A_l(G_0^n)| \leq k_0 - k_* + 1.
\]

When these inequalities are both achieved by the same index \( \bar{l} \in [k_*] \), we find that for every \( i \in A_l(\hat{G}_n) \), there exists \( j \in A_l(G_0^n) \) such that, up to taking subsequences, the rate of [16] is achieved:

\[
|\hat{\theta}_i^n - \theta_j^0| \lesssim n^{-\frac{1}{2}}.
\]

However, the remaining \( k_* - 1 \) atoms of the penalized MLE converge uniformly at the parametric rate \( n^{-1/2} \), which could not have been deduced from equation (8). Furthermore, we emphasize that this setting—in which all redundant atoms of \( \hat{G}_n \) and \( G_0^n \) are concentrated near a single atom of \( G_* \)—is the only case where a subset of the atoms of \( \hat{G}_n \) achieve the worst-case rate predicted by [16]. Indeed, when the inequalities (8) are strict, rates faster than \( n^{-1/2r} \) are achieved by all atoms of \( \hat{G}_n \).

5 Discussion

The aim of our work has been to sharpen known convergence rates of the MLE for estimating individual parameters of a finite mixture model. Our key observation was that the Wasserstein distance, despite being an elegant tool for metrizing the space of mixing measures, is not well-suited to capturing the heterogeneous convergence behaviour of individual mixture parameters. We instead proposed new loss functions which achieve this goal. Our theoretical results are complemented by a simulation study, which is deferred to Appendix C.

Our analysis has focused on maximum likelihood-based estimators, whose computation involves the nonconvex optimization problem (1). Despite significant recent advances in the theoretical understanding of the EM algorithm for approximating the MLE in finite mixtures [1, 10, 22, 9], we make no claims that such approximations obey the asymptotics described in this paper, leaving open a potential gap between theory and practice. The method of moments provides a practical alternative to the MLE, which is minimax optimal for certain classes of finite mixture models under the Wasserstein distance [30, 8]. We leave open the question of characterizing the risk of moment-based estimators under the loss functions proposed in our work.
In Section 4, we obtained uniform convergence rates for strongly identifiable mixtures with mixing proportions bounded away from zero. We leave open the question of determining whether this constraint can be removed.

Finally, we derived both pointwise and uniform convergence rates for strongly identifiable mixtures, however we restricted our analysis of location-scale Gaussian mixtures to the pointwise case. Obtaining uniform convergence rates for such models remains an important open problem, which has not been studied beyond the special case of two component models [15, 23]. While this setting is beyond the scope of our work, we expect that considerations about the heterogeneity of parameter estimation, similar to those studied in this paper, would arise in such models as well.

References


16
Supplement to “Refined Convergence Rates for Maximum Likelihood Estimation under Finite Mixture Models”

In this supplementary material, we provide all proofs of results stated in the main text (Appendix A). We also state and prove certain results which were deferred from the main text (Appendix B), and provide a simulation study to illustrate the various convergence rates that were derived in this paper (Appendix C).

A Proofs

A.1 Proof of Theorem 2

Theorem 2(i) is an immediate consequence of Theorem 7.4 of [29], which provides a generic exponential inequality for the Hellinger loss of nonparametric maximum likelihood density estimators, under mere conditions on the bracketing integral $\mathcal{J}_B(\epsilon, \bar{F}_k^{1/2}, \Theta, \epsilon, \nu)$. The application of this result to finite mixture models has previously been discussed by [18, 17].

Theorem 2(ii) also follows by the same proof technique as Theorem 7.4 of [29], with modifications to account for the presence of the penalty in the definition of $\hat{G}_n$. An analogue of this result was previously proven by [24], though with different conditions on the tuning parameter $\xi_n$. For completeness, we provide a self-contained proof of Theorem 2(ii), under the conditions on $\xi_n$ required for our development.
As in [29], we shall reduce the problem to controlling the increments of the empirical process
\[ \nu_n(G) = \sqrt{n} \int \{p_{G_0} > 0\} \frac{1}{2} \log \frac{\bar{p}_G}{p_{G_0}} d(P_n - P_{G_0}), \]
where we recall that \( \bar{p}_G = (p_G + p_{G_0})/2 \), and we denote by \( P_G = \int p_G d\nu \) the distribution induced by \( p_G \), for any \( G \in \mathcal{O}_k(\Theta) \). Furthermore, \( P_n = (1/n) \sum_{i=1}^n \delta_{X_i} \) denotes the empirical measure. Our main technical tool will be the following special case of Theorem 5.11 [29].

**Theorem 7** (Theorem 5.11 [29]). Let \( R > 0 \) and \( k \geq 1 \). Given \( G \subseteq \mathcal{O}_k(\Theta) \), let \( G_0 \in G \). Furthermore, given a universal constant \( C > 0 \), let \( a,C_1 > 0 \) be chosen such that
\[ a \leq C_1 \sqrt{n} R^2 \wedge 8 \sqrt{n} R, \]
and,
\[ a \geq \sqrt{C^2(C_1 + 1)} \left( \int_0^R \sqrt{H_B \left( \frac{u}{\sqrt{2}}, \{ p_G : G \in G, h(\bar{p}_G, p_0) \leq R \}, \nu \right) du \vee R \right), \]
Then,
\[ P \left\{ \sup_{G \in G \atop h(\bar{p}_G, p_{G_0}) \leq R} |\nu_n(G)| \geq a \right\} \leq C \exp \left( -\frac{a^2}{C^2(C_1 + 1)R^2} \right). \]

We are now in a position to prove the claim.

**Proof of Theorem 2(ii).** Let \( G_0 \in \mathcal{O}_{k,c_0}(\Theta) \). By a straightforward modification of Lemma 4.1 of [29], we have
\[ h^2(\bar{p}_{\hat{G}_n}, p_{G_0}) \leq \frac{1}{\sqrt{n}} \nu_n(\hat{G}_n) + \frac{\xi_n \rho(G_0)}{4n}. \]
Let \( u > \gamma_n = L \log n/\sqrt{n} \), where \( L \) is the constant in assumption B(k). In view of equation (13), and the fact that \( h^2(p_G, p_{G_0}) \leq 4h(\bar{p}_G, p_{G_0}) \) for all \( G \in \mathcal{O}_k(\Theta) \) (cf. Lemma 4.2 of [29]), we have
\[ P \left\{ h(\bar{p}_{\hat{G}_n}, p_{G_0}) > u \right\} \leq P \left\{ h(\bar{p}_{\hat{G}_n}, p_{G_0}) > u/4 \right\} \]
\[ \leq P \left\{ \sup_{G \in \mathcal{O}_{k,c_0}(\Theta) \atop h(\bar{p}_G, p_0) > u/4} n^{-\frac{1}{2}} \nu_n(G) + \frac{\xi_n \rho(G_0)}{4n} - h^2(\bar{p}_G, p_{G_0}) \geq 0 \right\}. \]

18
Let \( S = \min\{ s : 2^{s+1}u/4 > 1 \} \). Then,

\[
\mathbb{P}\left\{ \sup_{G \in \mathcal{O}_{k,c_0}(\Theta)} n^{-\frac{1}{2}} \nu_n(G) + \frac{\xi_n \rho(G_0)}{4n} - h^2(\hat{p}_G, p_0) \geq 0 \right\}
\]

\[
\leq \sum_{s=0}^{S} \mathbb{P}\left\{ \sup_{G \in \mathcal{O}_{k,c_0}(\Theta)} \nu_n(G) \geq \sqrt{n} 2^{2s} \left( \frac{u}{4} \right)^2 - \frac{\xi_n \rho(G_0)}{4\sqrt{n}} \right\}.
\]

We have thus reduced the problem to that of bounding the supremum of the empirical process \( \nu_n \), for which we shall invoke Theorem 7. Let \( R = 2^{s+1}u, C_1 = 15, C_0 = C^2(C_1 + 1) \) and

\[ a = \sqrt{n} 2^{2s} \left( \frac{u}{4} \right)^2 - \frac{\xi_n \rho(G_0)}{4\sqrt{n}}. \]

It can be directly verified that condition (11) holds for all \( s = 0, \ldots, S \). To further show that condition (12) holds, note that

\[
\int_0^{2^{s+1}u} \sqrt{H_B \left( \frac{t}{\sqrt{2}}, \mathcal{T}_k^{1/2} \left( \Theta, 2^{s+1}u/4 \right), \nu \right)} dt \leq 2^{s+1}u
\]

\[
\leq \sqrt{2} \int_0^{2^{s+1}u} \sqrt{H_B \left( t, \mathcal{T}_k^{1/2} \left( \Theta, 2^{s+1}u/4 \right), \nu \right)} dt \leq 2J \sqrt{n} 2^{2s+1}u^2,
\]

where we invoked condition \( \mathbf{B}(k) \). Now, notice that \( \rho(G_0) \) is bounded above by a universal constant \( L_0 > 0 \) depending only on \( k, c_0 \), irrespective of the choice of \( G_0 \in \mathcal{O}_{k,c_0}(\Theta) \). Furthermore, we have \( \sqrt{n} \gamma_n^2 \asymp (\log n)^2 / \sqrt{n} \), and \( \xi_n / \sqrt{n} \asymp \log n / \sqrt{n} \), thus for all \( u > \gamma_n \), the second term in the definition of \( a \) is of lower order than the first. Deduce that there exists a constant \( N > 0 \), depending only on \( L_0, c_1, k \) such that for all \( n \geq N \),

\[ a \geq \frac{1}{2} \sqrt{n} 2^{2s} (u/4)^2 = \sqrt{n} 2^{2s-5} u^2 \geq \sqrt{C_0} \cdot (2J \sqrt{n} 2^{2s+1}u^2), \]

for a sufficiently small choice of the universal constant \( J > 0 \). We may therefore invoke Theorem 7,
to deduce that for all \( n \geq N \),

\[
\mathbb{P}\left\{ h(p_{\tilde{G}_n}, p_{G_0}) > u \right\} \leq \sum_{s=0}^{S} \mathbb{P}\left\{ \sup_{O_{k,c_0}(\Theta)} \nu_n(G) \geq \sqrt{n} 2^{2s-5} u^2 \right\}
\leq C \sum_{s=0}^{\infty} \exp\left\{ \frac{1}{16C^2 2^{2s+2} \gamma_n^2} \left[ \sqrt{n} 2^{2s-5} u^2 \right]^2 \right\}
\leq C \sum_{s=0}^{\infty} \exp\left\{ \frac{n 2^{2s-16} u^2}{C^2} \right\}
\leq c \exp(-nu^2/c),
\]

for a large enough constant \( c > 0 \). It follows that, for all \( n \geq N \),

\[
\mathbb{E} h(p_{\tilde{G}_n}, p_{G_0}) = \int_0^{\infty} \mathbb{P}(h(p_{\tilde{G}_n}, p_{G_0}) \geq u) du \leq \gamma_n + c \int_{\gamma_n}^{\infty} \exp\left\{ \frac{-nu^2}{c} \right\} du \leq c' \gamma_n,
\]

for another universal constant \( c' > 0 \). Since the Hellinger distance is bounded above by 1, it is clear that the above display holds for all \( n \geq 1 \), up to modifying the constant \( c' \) in terms of \( N \). Furthermore, the above calculation is clearly uniform in the \( G_0 \) under consideration, so the claim follows.

\[ \square \]

### A.2 Proof of Proposition 3

We shall require a bound on the log-likelihood ratio statistic based on the MLE \( \widetilde{G}_n \). Such a bound is implicit in the proof of Theorem 7.4 of [29]. Specifically, the following can be deduced from their Corollary 7.5.

**Proposition 8** (Corollary 7.5 [29]). Assume that condition \( B(k) \) holds. Then, given \( k \geq 1 \), there exists a constant \( C > 0 \) depending on \( k, d \) and \( \mathcal{F} \), such that for all \( u \geq L(\log n/n)^{1/2} \),

\[
\sup_{G_0 \in O_{k, c_0}(\Theta)} \mathbb{P}_{G_0} \left( \int \log \frac{p_{\tilde{G}_n}}{p_{G_0}} dP_n \geq u^2 \right) \leq C \exp\left( -\frac{nu^2}{C^2} \right).
\]

Let \( G_0 \in O_{k,c_0}(\Theta) \). After possibly replacing \( C \) by \( C \vee L \), apply Proposition 8 with \( u = C \sqrt{\log n/n} \) to deduce that

\[
\ell_n(\widetilde{G}_n) - \ell_n(G_0) \leq C^2 \log n,
\]

with probability at least \( 1 - C/n \). Now, by definition of the penalized MLE \( \hat{G}_n \) and of the non-
penalized MLE $\tilde{G}_n$, we have
\[
0 \leq \left[ \ell_n(\tilde{G}_n) - \ell_n(G_0) \right] + \xi_n \left[ \rho(\tilde{G}_n) - \rho(G_0) \right] \leq \left[ \ell_n(\tilde{G}_n) - \ell_n(G_0) \right] + \xi_n \left[ \rho(\tilde{G}_n) - \rho(G_0) \right] \\
\leq C^2 \log n + \xi_n \left[ \rho(\tilde{G}_n) - \rho(G_0) \right],
\]
with probability at least $1 - C/n$. Therefore, since $\xi_n \geq \log n$, we obtain
\[
\rho(\tilde{G}_n) \geq -C^2 + \rho(G_0) \geq -C^2 + k_0 \log c_0 = -C_1,
\]
where $C_1 = C^2 + k_0 \log(1/c_0) > 0$. By definition of $\rho$, it must follow that
\[
\hat{p}_i \geq \exp(-C_1), \quad i = 1, \ldots, \hat{k}_n,
\]
with probability at least $1 - C/n$. The claim follows with $c = \exp(C_1) \lor C$. \hfill \qed

A.3 Proof of Theorem 4

The claim will follow from the following result, relating the discrepancy $D$ to the Total Variation distance between the corresponding densities $p_G$ and $p_{G_0}$.

Lemma 9. Assume the same conditions as Theorem 4. Then, there exists a constant $c > 0$ depending on $G_0, d, k, F$, such that for any $G \in \mathcal{O}_k(\Theta)$,
\[
V(p_G, p_{G_0}) \geq c D(G, G_0). \tag{14}
\]

Recall that we have assumed condition $B(k)$. Therefore, by combining Lemma 9 with Theorem 2(i) and the well-known inequality $V \leq h$, we deduce that
\[
\mathbb{E} D(\tilde{G}_n, G) \lesssim \mathbb{E} V(p_{\tilde{G}_n}, p_{G_0}) \leq \mathbb{E} h(p_{\tilde{G}_n}, p_{G_0}) \lesssim \sqrt{\frac{\log n}{n}},
\]
as claimed. It thus remains to prove Lemma 9.

Proof of Lemma 9. Our proof proceeds using a similar argument as that of [18], though with key differences to account for our choice of loss function. We will prove that
\[
\lim_{\delta \to 0} \inf_{G \in \mathcal{O}_k(\Theta)} \frac{V(p_G, p_{G_0})}{D(G, G_0)} > 0. \tag{15}
\]
This implies a local version of the claim, namely that there exist constants $\delta_0, C > 0$ such that for all $G \in \mathcal{O}_k(G_0; \delta_0)$,
\[
D(G, G_0) \leq C V(p_G, p_{G_0}). \tag{16}
\]
We begin by showing how this local inequality leads to the claim, and we will then prove equation (15). Taking equation (15) for granted, it suffices to prove

\[
\inf_{G \in \mathcal{O}_k(\Theta)} \frac{V(p_G, p_{G_0})}{D(G, G_0)} > 0. \tag{17}
\]

Suppose by way of a contradiction that the above display does not hold. Then, there exists a sequence of mixing measures \(G_n \in \mathcal{O}_k(\Theta)\) with \(D(G_n, G_0) \geq \delta_0\) such that \(\frac{V(p_{G_n}, p_{G_0})}{D(G_n, G_0)} \to 0\). Since the parametric family \(\mathcal{F}\) is assumed to be 2-strongly identifiable, the model \(\{p_G : G \in \mathcal{O}_k(\Theta)\}\) is identifiable, thus the map

\[
(G, G') \in \mathcal{O}_k(\Theta) \times \mathcal{O}_k(\Theta) \mapsto V(p_G, p_{G'})
\]

defines a metric on \(\mathcal{O}_k(\Theta)\). Since this metric is bounded, the sequence \(\{G_n\}\) admits a converging subsequence, under the above metric, to some mixing measure \(\bar{G} \in \mathcal{O}_k(\Theta)\). For ease of exposition, we replace this subsequence by the entire sequence \(G_n\) in what follows, thus we have \(V(p_{G_n}, p_{\bar{G}}) \to 0\).

Now, notice that \(D(\bar{G}, G_0) \geq \delta_0\) by definition of \(G_n\), whence \(V(p_{G_n}, p_{G_0}) \to 0\). Combining these facts leads to \(V(p_{G_0}, p_{\bar{G}}) = 0\), and hence \(\bar{G} = G_0\). This contradicts the assumption \(D(\bar{G}, G_0) > 0\), and hence proves equation (17).

It remains to prove the local inequality (15). We again assume by way of a contradiction that there exists a sequence of mixing measures \(G_n = \sum_{j=1}^{k_n} p_j^n \delta_{\theta^n_j} \in \mathcal{O}_k(\Theta)\) such that \(D(G_n, G_0) \to 0\) but

\[
\frac{V(p_{G_n}, p_{G_0})}{D(G_n, G_0)} \to 0, \quad n \to \infty. \tag{18}
\]

Define

\[
\mathcal{A}_j^n = \mathcal{A}_j(G_n) = \{i \in \{1, \ldots, k_n\} : \|\theta_i^n - \theta_j^0\| \leq \|\theta_i^n - \theta_\ell^n\| \quad \forall \ell \neq j\}, \quad j = 1, \ldots, k_0.
\]

Since \(k_n \leq k\) for all \(n\), there exists a subsequence of \(G_n\) such that \(k_n\) does not change with \(n\). Therefore, up to replacing \(G_n\) by a subsequence, we may assume that \(k_n = k' \leq k\) for all \(n\). Similarly, since there are only a finite number of distinct sets \(\mathcal{A}_1^n \times \cdots \times \mathcal{A}_{k_0}^n\) over the range of \(n \geq 1\), we may assume without loss of generality that \(\mathcal{A}_j = \mathcal{A}_j^n\) does not change with \(n\), for all \(j = 1, \ldots, k_0\). Now, consider the decomposition

\[
p_{G_n}(x) - p_{G_0}(x) = \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i \left( f(x|\theta_i^n) - f(x|\theta_j^0) \right) + \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \left( f(x|\theta_i^n) - f(x|\theta_j^0) \right) + \sum_{j=1}^{k_0} (p_j^n - p_j^0) f(x|\theta_j^0) + A_{n,1}(x) + A_{n,2}(x) + B_n(x).
\]

\[
A_{n,1}(x) + A_{n,2}(x) + B_n(x).
\]

22
By a Taylor expansion to second order, notice that

\[ A_{n,1}(x) = \sum_{j: |A_j|>1} \sum_{i \in A_j} p_i \left[ (\theta_i^n - \theta_j^n)^\top \frac{\partial f}{\partial \theta}(x|\theta_0^0) + \frac{1}{2} (\theta_i^n - \theta_j^n)^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_0^0)(\theta_i^n - \theta_j^n) \right] + R_{n,1}(x) \]

where \( R_{n,1}(x) \) is a Taylor remainder satisfying

\[ \|R_{n,1}\|_\infty \lesssim \sum_{j: |A_j|>1} \sum_{i \in A_j} p_i \|\theta_i^n - \theta_j^n\|^{2+\gamma}, \] (19)

for some \( \gamma > 0 \), due to condition A(2). Furthermore, by a Taylor expansion to first order, we also have

\[ A_{n,2}(x) = \sum_{j: |A_j|=1} \sum_{i \in A_j} p_i (\theta_i^n - \theta_j^n)^\top \frac{\partial f}{\partial \theta}(x|\theta_0^0) + R_{n,2}(x), \]

where, again, the Taylor remainder \( R_{n,2} \) satisfies

\[ \|R_{n,2}\|_\infty \lesssim \sum_{j: |A_j|=1} \sum_{i \in A_j} p_i \|\theta_i^n - \theta_j^n\|^{1+\gamma}, \] (20)

Let \( D_n = D(G_n, G_0) \). By equations (19)–(20) and the definition of \( D \), we deduce that \( \|R_{n,\ell}\|_\infty / D_n = o(1) \) for \( \ell = 1, 2 \). Therefore, we have uniformly in \( x \in \mathcal{X} \) that,

\[ \frac{p_{G_n}(x) - p_{G_0}(x)}{D_n} \asymp \frac{A_{n,1}(x) + A_{n,2}(x) + B_n(x)}{D_n}. \]

Notice that the ratio \((A_{n,1}(x) + A_{n,2}(x) + B_n(x))/D_n\) is a linear combination of \( f(x|\theta_0^0) \) and its first two partial derivatives, with coefficients not depending on \( x \). We claim that at least one of these coefficients does not tend to zero as \( n \to \infty \). Indeed, suppose by way of a contradiction that this is not the case. Then, in particular the coefficients corresponding to the second derivatives in \( A_{n,1}/D_n \) and the coefficients corresponding to the first derivatives in \( A_{n,2}/D_n \) must vanish, and their absolute sum must vanish, implying the following display,

\[ \frac{1}{D_n} \left[ \sum_{j: |A_j|>1} \sum_{i \in A_j} p_i \|\theta_i^n - \theta_j^n\|^2 + \sum_{j: |A_j|=1} \sum_{i \in A_j} p_i \|\theta_i^n - \theta_j^n\| \right] \to 0. \]

The definition of \( D_n \) then implies that

\[ \frac{\sum_{j=1}^{k_n} |p_j - p_j^0|}{D_n} \to 1. \]

We deduce that at least one coefficient in the linear combination \( B_n(x)/D_n \) does not tend to zero, which is a contradiction. Thus, there indeed exists at least one coefficient in the linear combinations \( A_{n,\ell}(x)/D_n, B_n(x)/D_n, \ell = 1, 2 \), which does not vanish. Let \( m_n \) denote the greatest absolute value
of these nonzero coefficients, and set $d_n = 1/m_n$. Then, there must exist scalars $\alpha_i \in \mathbb{R}$ and vectors $\beta_j, \nu_j \in \mathbb{R}^d, j = 1, \ldots, k_0$, not all of which are zero, such that for all $x \in \mathcal{X}$,

$$
\frac{d_n A_{n,1}(x)}{D_n} + \frac{d_n A_{n,2}(x)}{D_n} \to \sum_{j=1}^{k_0} \left[ \beta_j^\top \frac{\partial f}{\partial \theta}(x|\theta_j^0) + \nu_j^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^0)\nu_j \right] \tag{21}
$$

On the other hand, the assumption (18) and the fact that $d_n$ are uniformly bounded implies that

$$
d_n \frac{V(p_{G_n}, p_{G_0})}{D_n} = \int d_n \frac{A_{n,1}(x) + A_{n,2}(x) + B_n(x)}{D_n} dx \to 0.
$$

By Fatou’s Lemma combined with equation (21), it follows that for almost all $x \in \mathcal{X}$,

$$
\sum_{j=1}^{k_0} \left[ \alpha_j f(x|\theta_j^0) + \beta_j^\top \frac{\partial f}{\partial \theta}(x|\theta_j^0) + \nu_j^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^0)\nu_j \right] = 0.
$$

Since the coefficients $\alpha_j, \beta_j, \nu_j$ are not all zero, the above display contradicts the second-order strong identifiability assumption on the parametric family $\mathcal{F}$. It follows that equation (18) could not have held, whence the claim (15) is proved. This completes the proof. \hfill \square

A.4 Proof of Theorem 5

We will prove Theorem 5 as a consequence of the following upper bound of $\mathcal{D}$ by the Total Variation distance.

Lemma 10. Assume the same conditions as Theorem 5, and let $c_0 \in (0, \min_{1 \leq j \leq k_0} p_{0_j})$. Then, there exists $C > 0$, depending on $G_0, c_0, d, k, \Theta$ such that for all $G \in \mathcal{O}_{k,c_0}(\Theta)$,

$$
V(p_{G}, p_{G_0}) \geq C \mathcal{D}(G, G_0).
\tag{22}
$$

Before proving Lemma 10, we show how it leads to the claim. Under the conditions of Theorem 5 regarding the parameter space $\Theta$, it follows from Lemma 2.1 of [18] (see also [13]) that the location-scale Gaussian density family $\mathcal{F}$ satisfies

$$
H_B(\epsilon, \mathcal{P}^{1/2}_k(\Theta, \epsilon), \nu) \leq C_1 \log(1/\epsilon), \quad \epsilon > 0,
$$

for a constant $C_1 > 0$ depending on $d, k, \Theta$. Given $L > 0$, it follows that for all $\epsilon \geq L(\log n/n)^{1/2}$,

$$
\mathcal{J}_B(\epsilon, \mathcal{P}^{1/2}_k(\Theta, \epsilon), \nu) \leq C_1 \epsilon \sqrt{\log(1/\epsilon)} \leq C_1 \sqrt{n} \left( \frac{\epsilon}{\sqrt{n}} \right) \sqrt{\log(1/\epsilon)} \leq \frac{C_1 \sqrt{n} \epsilon^2}{L}.
$$

24
Condition $B(k)$ is then satisfied by choosing $L = C_1/J$, thus we may apply Theorem 2 and Proposition 3 in what follows.

By Proposition 3, there is an event $A_n$ and a constant $c > 1$ such that $\mathbb{P}(A_n^c) \leq c/n$ and $\tilde{G}_n^c \geq 1/c$ for all $i = 1, \ldots, \hat{k}_n$. In particular, letting $c_0 = \min\{p_{ij}^0 : j \in [k_0]\} \wedge c^{-1}$, we have $\hat{G}_n \in \mathcal{O}_{k,c_0}(\Theta)$ over the event $A_n$. Therefore, by Lemma 11 and the fact that $\mathcal{D}$ is bounded by a constant depending only on diam($\Theta$), $k$ we arrive at

$$\mathbb{E}[\mathcal{D}(\hat{G}_n, G_n^0)] = \mathbb{E}[\mathcal{D}(\hat{G}_n, G_n^0)_{I_{A_n}}] + \mathbb{E}[\mathcal{D}(\hat{G}_n, G_n^0)_{I_{A_n}}]$$

$$\lesssim \mathbb{E}[h(p_{\hat{G}_n}, p_{G_0})_{I_{A_n}}] + \mathbb{P}(A_n^c) \approx \log n/\sqrt{n} + 1/n \lesssim \log n/\sqrt{n},$$

where we used the inequality $V \leq h$ and we invoked the Hellinger rate of convergence of $p_{\hat{G}_n}$, given in Theorem 2(ii). The claim follows; it thus remains to prove Lemma 10.

**Proof of Lemma 10.** We will prove the following local version of the claim: there exists $\delta_0 > 0$ such that

$$\inf_{G \in \mathcal{O}_{k,c_0}(\Theta)} \frac{V(p_G, p_{G_0})}{\mathcal{D}(G, G_0)} < 0. \quad (23)$$

The above local statement directly leads to the claim by the same argument as in the beginning of the proof of Lemma 9, and we therefore omit it. Our proof follows along similar lines as the proof of Proposition 2.2 of [17], though with key modifications to account for our distinct loss function. We proceed with the following steps.

**Step 1: Setup.** To prove inequality (23), assume by way of a contradiction that it does not hold. Then, there exists a sequence of mixing measures $G_n = \sum_{i=1}^{k_n} p_{ij}^0 \delta_{(\mu_i^0, \Sigma_i^0)}$ with $p_{ij}^0 \geq c_0$ for all $j \in [k_n]$, such that $D_n := \mathcal{D}(G_n, G_0) \to 0$ and $V(p_{G_n}, p_{G_0}) / D_n \to 0$. Furthermore, since $k_n \leq k$ for all $n \geq 1$, there exists a subsequence of $G_n$ admitting a fixed number of atoms $k_n = k' \leq k$. For notational simplicity, replace $G_n$ by such a subsequence throughout the sequel.

Define the Voronoi diagram

$$A_n^j = \{1 \leq i \leq k : \|\mu_i^0 - \mu_j^0\| + \|\Sigma_i^0 - \Sigma_j^0\| \leq \|\mu_i^0 - \mu_j^0\| + \|\Sigma_i^0 - \Sigma_j^0\|, \forall l \neq i\}, \quad i = 1, \ldots, k'.$$

By the same argument as in the proof of Lemma 9, we may assume, up to taking a further subsequence of $G_n$, that the sets $A_j \equiv A_j^0$ do not change with $n$ for all $j = 1, \ldots, k_0$ and all $n \geq 1$. Furthermore, we note that, since the mixing proportions of $G_n$ are bounded below by $c_0$, the fact that $D_n \to 0$ implies

$$\sup_{i \in A_j} \left[\|\mu_i^n - \mu_j^0\| + \|\Sigma_i^n - \Sigma_j^0\|\right] \to 0, \quad j = 1, \ldots, k_0.$$

Throughout what follows, we write the coordinates of $\mu_j^0$ and $\Sigma_j^0$ as $\mu_j^0 = (\mu_{j,1}, \ldots, \mu_{j,d})$ and $\Sigma_j^0 = (\Sigma_{j,uv})_{u,v=1}^d$, for all $j = 1, \ldots, k_0$, and similarly for $\mu_i^n, \Sigma_i^n$, $i = 1, \ldots, k$. We also write for simplicity $\theta_i^n = (\mu_i^n, \Sigma_i^n)$ and $\theta_j^0 = (\mu_j^0, \Sigma_j^0)$ for all $j = 1, \ldots, k_0$ and $i = 1, \ldots, k'$. 

25
**Step 2: Taylor Expansions.** Similarly to the proof of Lemma 9, consider the following representation

\[ p_{G_n}(x) - p_{G_0}(x) = \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i^n \left( f(x|\theta_i^n) - f(x|\theta_j^0) \right) + \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i^n \left( f(x|\theta_i^n) - f(x|\theta_j^0) \right) + \sum_{j=1}^{k_0} (p_j^n - p_j^0) f(x|\theta_j^0) \]

:= \tilde{A}_n(x) + B_n(x) + C_n(x).

By repeated Taylor expansions to order \(\bar{r}(|A_j|)\) for all \(j = 1, \ldots, k_0\), we obtain

\[ \tilde{A}_n(x) = \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i^n \sum_{\alpha,\beta} \frac{1}{\alpha!\beta!} (\mu_i^n - \mu_j^0)^\alpha (\Sigma_i^n - \Sigma_j^0)^\beta \frac{\partial^{\alpha+\beta}}{\partial \mu^\alpha \partial \Sigma^\beta} (x|\theta_i^0) + R_{n,1}(x) =: A_n(x) + R_{n,1}(x), \]

where the third summation in the above display is over all multi-indices \(\alpha \in \mathbb{N}^d\) and \(\beta \in \mathbb{N}^{d \times d}\) satisfying \(1 \leq |\alpha| + |\beta| := \sum_{i=1}^d \alpha_i + \sum_{l,s=1}^d \beta_{ls} \leq \bar{r}(|A_j|)\). Above, we write \(\alpha! = \prod_{l=1}^d \alpha_l!\) and \(\beta! = \prod_{l,s=1}^d \beta_{ls}!\). Furthermore, \(R_{n,1}\) is a Taylor remainder which satisfies

\[ \|R_{n,1}\|_\infty \leq \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i^n \left[ \|\mu_i^n - \mu_j^0\|^r(\bar{r}(|A_j|) + \gamma) + \|\Sigma_i^n - \Sigma_j^0\|^r(\bar{r}(|A_j|) + \gamma) \right], \]

for some constant \(\gamma > 0\). Now, recall the key PDE (5), which implies that for any multi-indices \(\alpha \in \mathbb{N}^d\) and \(\beta \in \mathbb{N}^{d \times d}\),

\[ \frac{\partial^{\alpha+\beta}}{\partial \mu^\alpha \partial \Sigma^\beta} f = \frac{1}{2|\beta|!} \frac{\partial^{\alpha+2|\beta|}}{\partial \mu^{\alpha+\tau_0(\alpha,\beta)}} f, \]

where we denote by \(\tau_0(\alpha, \beta) \in \mathbb{N}^d\) the multi-index with coordinates \(\alpha_v + \sum_{u=1}^d (\beta_{uv} + \beta_{vu}), v = 1, \ldots, d\). Notice that we may then write for all \(x \in \mathbb{R}^d\),

\[ A_n(x) = \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i^n \sum_{1 \leq |\alpha| + |\beta| \leq \bar{r}(|A_j|)} \frac{1}{2|\beta|!} (\mu_i^n - \mu_j^0)^\alpha (\Sigma_i^n - \Sigma_j^0)^\beta \frac{\partial^{\alpha+2|\beta|}}{\partial \mu^{\alpha+\tau_0(\alpha,\beta)}} (x|\theta_j^0) \]

\[ = \sum_{j:|A_j|>1} \sum_{|\tau| = 1} 2^{\bar{r}(|A_j|)} a_{\tau,j} \frac{\partial |\tau|}{\partial \mu^\tau} (x|\theta_j^0), \]

where for all \(\tau \in \mathbb{N}^d\), we write

\[ a_{\tau,j} = \sum_{1 \leq |\alpha| + |\beta| \leq \bar{r}(|A_j|)} \frac{1}{\tau_0(\alpha,\beta)!} p_i^n (\mu_i^n - \mu_j^0)^\alpha (\Sigma_i^n - \Sigma_j^0)^\beta. \]
Furthermore, by a first-order Taylor expansion in the definition of $\bar{B}_n$, we obtain

$$
\bar{B}_n(x) = \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \left\{ (\mu_i^n - \mu_j^0)^\top \frac{\partial f}{\partial \mu_j^n} (x|\theta_j^0) + \text{tr} \left[ \frac{\partial f}{\partial \Sigma} (x|\theta_j^0)^\top (\Sigma_i^n - \Sigma_j^0) \right] \right\} + R_{n,2}(x)
$$

$$
=: B_n(x) + R_{n,2}(x),
$$

where $R_{n,2}$ is a Taylor remainder which satisfies,

$$
\|R_{n,2}\|_{\infty} \lesssim \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i^n \left[ \|\mu_i^n - \mu_j^0\|^{1+\gamma} + \|\Sigma_i^n - \Sigma_j^0\|^{1+\gamma} \right].
$$

Similarly as the term $A_n$, we may explicitly rewrite $B_n$ as a linear combination of the first- and second-order partial derivatives of the density $f$ with respect to $\mu$,

$$
B_n(x) = \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \left\{ (\mu_i^n - \mu_j^0)^\top \frac{\partial f}{\partial \mu_j^n} (x|\theta_j^0) + \frac{1}{2} \text{tr} \left[ \frac{\partial f}{\partial \mu_j^n \partial \mu_j^n} (x|\theta_j^0)^\top (\Sigma_i^n - \Sigma_j^0) \right] \right\}
$$

$$
= \sum_{j:|A_j|=1} \sum_{|\kappa|=1} 2 b_{j,\kappa} \frac{\partial |\kappa| f}{\partial \mu^\kappa}(x|\theta_j^0),
$$

where

$$
b_{\kappa,j} = \sum_{\alpha,\beta:\alpha+|\beta|=1} \frac{1}{2!|\beta|} p_i^n (\mu_i^n - \mu_j^0)^\alpha (\Sigma_i^n - \Sigma_j^0)^\beta.
$$

Notice that the conditions on the remainder terms $R_{n,1}, R_{n,2}$ together with the definition of $D_n$ readily imply that, uniformly in $x \in \mathbb{R}^d$,

$$
\frac{p_{G_n}(x) - p_{G_0}(x)}{D_n} \asymp \frac{A_n(x) + B_n(x) + C_n(x)}{D_n}.
$$

Letting $c_j = \bar{p}_j^n - p_j^0$, it can be seen that the right-hand side of the above display is a linear combination of partial derivatives of $f$ with respect to $\mu$, with coefficients $a_{\tau,j}/D_n, b_{n,j}/D_n, c_j/D_n$, $j = 1, \ldots, k_0$, where $\tau$ and $\kappa$ vary over the aforementioned ranges. In the next step, we will show that not all of these coefficients decay to zero.

**Step 3: Nonvanishing Coefficients.** Assume by way of a contradiction that all coefficients
\[a_{r,j}/D_n, b_{n,j}/D_n, c_{j}/D_n\] tend to zero. Define the following quantities,

\[
D_{n,1} = \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i^n \left\{ ||\mu_i^n - \mu_j^0||^{r(|A_j|)} + ||(\Sigma_{i,uu}^n - \Sigma_{j,uu}^0)_{1 \leq u \leq d}||^{r(|A_j|)/2} \right\}
\]
\[
D_{n,2} = \sum_{j:|A_j|>1} \sum_{i \in A_j} p_i^n ||(\Sigma_{i,uu}^n - \Sigma_{j,uu}^0)_{1 \leq u \neq v \leq d}||^{r(|A_j|)/2}
\]
\[
D_{n,3} = \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i^n \left( ||\mu_i^n - \mu_j^0|| + ||\Sigma_i^n - \Sigma_j^0|| \right)
\]
\[
D_{n,4} = \sum_{j=1}^{k_0} |\tilde{p}_i^n - p_j^0|.
\]

Note that there must exist \(1 \leq i \leq 4\) such that \(D_{n,i}/D_n \not\to 0\). We will consider four cases according to which of the terms \(D_{n,i}\) dominates \(D_n\)

**Case 3.1:** \(D_{n,1}/D_n \not\to 0\). In this case, it must hold that for some indices \(1 \leq j \leq k_0\) and \(1 \leq u \leq d\) such that \(\tilde{D}_{n,1}/D_n \not\to 0\), where

\[
\tilde{D}_{n,1} = \sum_{i \in A_j} p_i^n \left[ ||\mu_{j,uu}^n - \mu_j^0||^{r(|A_j|)} + ||\Sigma_{j,uu}^n - \Sigma_j^0||^{r(|A_j|)/2} \right]
\]

Fix such \(j\) and assume \(u = 1\) without loss of generality, throughout the rest of this Case. It follows by assumption that \(a_{r,j}/\tilde{D}_{n,1} \to 0\) for all \(1 \leq |\tau| \leq \tilde{r}(|A_j|)\). In particular, this property holds for all \(\tau\) such that \(\tau_l = 0\) for \(l = 2, \ldots, d\). Notice that \(\tau = \tau_0(\alpha, \beta)\) takes the latter form if and only if \(\alpha_l = \beta_{1l} = \beta_{sl} = 0\) for all \(l, s = 2, \ldots, d\). Therefore, taking the sum over such multi-indices leads to the limit

\[
\frac{1}{\tilde{D}_{n,1}} \sum_{i \in A_j} \sum_{\alpha_1, \beta_1} \frac{p_i^n}{2^{\alpha_1} 1!^{\beta_1}} (\mu_{j,11}^n - \mu_{j,11}^0)^{\alpha_1} (\Sigma_{j,11}^n - \Sigma_{j,11}^0)^{\beta_1} \to 0, \quad \tau_1 = 1, \ldots, \tilde{r}(|A_j|). \tag{25}
\]

Now, define

\[
\bar{m}_n = \max_{i \in A_j} p_i^n, \quad \bar{M}_n = \max\{||\mu_{i,11}^n - \mu_{j,11}^0||, ||\Sigma_{i,11}^n - \Sigma_{j,11}^0||^{1/2} : i \in A_j\}.
\]

For any \(i \in A_j, p_i^n/\bar{m}_n\) forms a bounded sequence of positive real numbers. Therefore, up to replacing it by a subsequence, it admits a nonnegative limit which we denote by \(z_i^2 = \lim_{n \to \infty} p_i^n/\bar{m}_n\). We similarly define \(x_i = \lim_{n \to \infty} (\mu_{i,11}^n - \mu_{j,11}^0)/\bar{M}_n\), and \(y_i = \lim_{n \to \infty} (\Sigma_{i,11}^n - \Sigma_{j,11}^0)/2\bar{M}_n^2\). We note that, since \(p_i^n \geq c_0\) due to the definition of \(O_{k,c_0}(\Theta)\), at least one of the real numbers \(z_i\) is equal to 1, and similarly, at least one of each of the \(x_i\) and \(y_i\) is equal to 1 or \(-1\). Furthermore, \(\tilde{D}_{n,1}/(\bar{m}_n \bar{M}_n) \not\to 0\) for any \(\tau_1 = 1, \ldots, \tilde{r}(|A_j|)\). We may then divide the numerator and denominator in equation (25)
by $\overline{M}_n^\tau_i m_i$ and take $n \to \infty$, to obtain the following system of polynomial equations

$$
\sum_{i \in A_j} \sum_{\alpha_1+2\beta_1=\tau_1} \frac{z_i^2 x_i^{\alpha_1} y_i^{\beta_1}}{\alpha_1! \beta_1!} = 0, \quad \tau_1 = 1, \ldots, \tilde{r}(|A_j|).
$$

By definition of $\tilde{r}(|A_j|)$, this system cannot have any nontrivial solutions, which is a contradiction.

**Case 3.2:** $D_{n,2}/D_n \not\to 0$. In this case, there must instead exist indices $1 \leq j \leq k_0$ and $1 \leq u \neq v \leq d$ for which $D_{n,2}/D_{n,2} \not\to 0$, where

$$
\tilde{D}_{n,2} = \sum_{i \in A_j} p_i^n |\sum_{\tau_0(\alpha,\beta)=\tau} \frac{1}{2^{\beta_1}|\alpha_1!\beta_1!} p_i^n (\mu_i^n - \mu_j^0)^\alpha (\Sigma_i^n - \Sigma_j^0)^\beta |^{\tilde{r}(|A_j|)/2}.
$$

Without loss of generality, we assume $u = 1$ and $v = 2$. Similarly to the previous case, we have by assumption that $a_{\tau,j}/D_{n,2} \to 0$ for all $1 \leq |\tau| \leq \tilde{r}(|A_j|)$. In particular, this property holds for the value $\tau = (2, 2, 0, \ldots, 0)$, where we note that this choice of $\tau$ is allowable because $|A_j| \geq 2$, hence $\tilde{r}(|A_j|) \geq 4 = |\tau|$. Therefore,

$$
\frac{1}{D_{n,2}} \sum_{\alpha_1,\beta_1} \sum_{\tau_0(\alpha,\beta)=\tau} \frac{1}{2^{\beta_1}|\alpha_1!\beta_1!} p_i^n (\mu_i^n - \mu_j^0)^\alpha (\Sigma_i^n - \Sigma_j^0)^\beta \to 0.
$$

Since Case 3.1 does not hold, we have $D_{n,1}/D_n \to 0$. Therefore, under the assumption of Case 3.2, any term in the above summation with $\alpha_l > 0$ and $\beta_l > 0$ ($l = 1, 2$) vanishes, and the preceding display thus reduces to

$$
\frac{1}{D_{n,2}} \sum_{i \in A_j} p_i^n (\Sigma_{i,12} - \Sigma_{j,12})^2 \to 0.
$$

By definition of $\tilde{D}_{n,2}$, this is a contradiction, thus Case 3.2 could not have held.

**Case 3.3:** $D_{n,3}/D_n \not\to 0$. By assumption, the coefficients $b_{\kappa,j}/D_n$ vanish for all multi-indices $\kappa \in \mathbb{N}^d$ satisfying $|\kappa| \in \{1, 2\}$, and all $j = 1, \ldots, k_0$. Therefore, their absolute sum also vanishes, implying

$$
\frac{1}{D_n} \sum_{j:|A_j|=1} \sum_{|\kappa|=1} |b_{\kappa,j}| = \frac{1}{D_n} \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i^n \left( \|\mu_i^n - \mu_j^0\|_1 + \frac{1}{2} \|\Sigma_i^n - \Sigma_j^0\|_1 \right) \to 0.
$$

The assumption of Case 3.3, together with the topological equivalence of the norms $\|\cdot\|_1$ and $\|\cdot\|_2$, then implies

$$
1 = \frac{1}{D_{n,3}} \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i^n \left( \|\mu_i^n - \mu_j^0\| + \|\Sigma_i^n - \Sigma_j^0\| \right) \to 0,
$$

which is a clear contradiction.

**Case 3.4:** $D_{n,4}/D_n \not\to 0$. In this case, it is clear that the coefficients $c_j/D_{n,4} \not\to 0$, whence $c_j/D_n \not\to 0$, for all $j = 1, \ldots, k_0$, and we immediately obtain a contradiction.
We have thus shown that each of Cases 3.1-3.4 lead to a contradiction. We conclude that at least one of the coefficients $a_{\tau,j}/D_n, b_{\kappa,j}/D_n, c_j/D_n$ does not tend to zero.

**Step 4: Reduction to Location-Gaussian Strong Identifiability.** Let $m_n$ denote the maximum of the absolute values of the coefficients $a_{\tau,j}/D_n, b_{\kappa,j}/D_n, c_j/D_n$, and set $d_n = 1/m_n$. Similarly as in the proof of Lemma 9, there exist real numbers $\zeta_{\tau,j}, \xi_{\kappa,j}, \nu_j$ not all zero such that

$$
\frac{d_n A_n(x)}{D_n} \rightarrow \sum_{j:|A_j|>1} 2^{|A_j|} \sum_{|\tau|=1} \zeta_{\tau,j} \frac{\partial |\tau| f(x|\theta_0^j)}{\partial \mu^\tau} \frac{\partial |\tau| f(x|\theta_0^j)}{\partial \mu^\tau},
$$

$$
\frac{d_n B_n(x)}{D_n} \rightarrow \sum_{j:|A_j|=1} 2^{2^{|A_j|}} \sum_{|\tau|=1} \xi_{\kappa,j} \frac{\partial |\kappa| f(x|\theta_0^j)}{\partial \mu^\kappa},
$$

$$
\frac{d_n C_n(x)}{D_n} \rightarrow \sum_{j=1}^{k_0} \nu_j f(x|\theta_0^j).
$$

Furthermore, by Step 3, $\sup_{n \geq 1} d_n < \infty$, and by the assumption $V(p_{G_n}, p_{G_0})/D_n \rightarrow 0$, we arrive at

$$
d_n \frac{V(p_{G_n}, p_{G_0})}{D_n} \propto \int d_n \frac{A_n(x) + B_n(x) + C_n(x)}{D_n} dx \rightarrow 0.
$$

By Fatou’s Lemma, the integrand of the above display vanishes for almost all $x \in \mathbb{R}$, whence

$$
\sum_{j:|A_j|>1} 2^{2^{|A_j|}} \sum_{|\tau|=1} \zeta_{\tau,j} \frac{\partial |\tau| f(x|\theta_0^j)}{\partial \mu^\tau} + \sum_{j:|A_j|=1} 2^{2^{|A_j|}} \sum_{|\tau|=1} \xi_{\kappa,j} \frac{\partial |\kappa| f(x|\theta_0^j)}{\partial \mu^\kappa} + \sum_{j=1}^{k_0} \nu_j f(x|\theta_0^j) = 0.
$$

The strong identifiability of the location-Gaussian family now implies that the coefficients $\zeta_{\tau,j}, \xi_{\kappa,j}, \nu_j$ are all zero, which is a contradiction. The claim follows. □

### A.5 Proof of Theorem 6

For any mixing measure $G \in \mathcal{O}_k(\Theta)$, let $F(x, G) = \int_p p_G(x) d\nu(x)$ denote the CDF of $p_G$. Similar to previous subsections, the proof will follow from the following key inequality relating $\hat{W}$ to a statistical distance, which we take to be the Kolmogorov-Smirnov distance by analogy with [16].

**Lemma 11.** Under the same conditions as Theorem 6, there exist $C, \epsilon_0 > 0$ depending on $c_0, \mathcal{F}$, and $G_*$ such that

$$
\|F(\cdot, G) - F(\cdot, G')\|_{\infty} \geq C\hat{W}(G, G'),
$$

for any $G \in \mathcal{O}_{k, c_0}(\Theta)$ and $G' \in \mathcal{E}_{k_0, c_0}(\Theta)$ such that $\hat{W}(G, G_*) \vee \hat{W}(G', G_*) \leq \epsilon_0$.

Taking Lemma 11 for granted, notice that

$$
\|F(\cdot, \hat{G}_n) - F(\cdot, G_0^{\hat{G}_n})\|_{\infty} \leq h(p_{G_0^{\hat{G}_n}}, p_{G_0^{\hat{G}_n}}).
$$
Furthermore, under the conditions of Theorem 6, we may apply Proposition 3 to deduce that there is an event $A_n$ and a constant $c > 1$ such that $\mathbb{P}(A'_n) \leq c/n$ and $\hat{p}_i^n \geq 1/c$ for all $i \in [k_0]$. As in the proof of Theorem 5, we may therefore set $c'_0 = c_0 \wedge c^{-1}$ and deduce that $\hat{G}_n \in \mathcal{O}_{k,c'_0}(\Theta)$ over the event $A_n$. Therefore, by Lemma 11 and Theorem 2(ii),

$$\mathbb{E}\hat{W}(\hat{G}_n,G_0^n) = \mathbb{E}\left[\hat{W}(\hat{G}_n,G_0^n)I_{A_n}\right] + \mathbb{E}\left[\hat{W}(\hat{G}_n,G_0^n)I_{A_n^c}\right] \lesssim \mathbb{E}\left[h(p_{G_n},p_{G_0^n})I_{A_n}\right] + 1/n \lesssim \log n/\sqrt{n}.$$ 

This proves the claim; it thus remains to prove the key Lemma 11.

**Proof of Lemma 11.** The proof of Lemma 11 is a refinement of the proof of Theorem 6.3 in [16] where we carefully consider the behavior of individual mixing components and weights of the mixing measures involved. Notice that in the special case $k^* = 1$, the loss function $\hat{W}$ is equal to $W_{k+k_0-1}$, and the claim can be deduced identically as in [16]. We therefore assume $k^* \geq 2$ throughout the sequel.

To prove inequality (26), we assume that it does not hold. Therefore, there exist sequences $G_n \in \mathcal{O}_{k,c_0}(\Theta), G'_n \in \mathcal{E}_{k,c_0}(\Theta)$ such that $\hat{W}(G_n,G_n) \to 0$, $\hat{W}(G_n,G_n) \to 0$, and $||F(\cdot,G_n) - F(\cdot,G'_n)||/\hat{W}(G_n,G'_n) \to 0$ as $n \to \infty$. Similarly to the proof of Theorem 9, we can find subsequences of $G_n,G'_n$ such that $A_j(G_n), A_j(G'_n)$ do not change with $n \geq 1$, for all $1 \leq j \leq k_0$. Without loss of generality, we therefore assume that $A_j = A_j(G_n)$ and $A'_j = A_j(G'_n)$ are constant with $n \geq 1$, for all $j \in [k_0]$. Furthermore, up to taking subsequences once again, we may assume that $G_n$ has exact order $\hat{k} \leq k$ for all $n \geq 1$, and we denote $G_n = \sum_{i=1}^{\hat{k}} p_i^n \delta_{\theta_i^n}$ and $G'_n = \sum_{i=1}^{k_0} p_i^n \delta_{(\theta_i^n)'}. Now, define

$$(\omega_i^n, \nu_i^n) = \begin{cases} (p_i^n, \theta_i^n), & 1 \leq i \leq \hat{k} \\ (-p_{i-\hat{k}}^n, (\theta_{i-\hat{k}}^n)'), & \hat{k} + 1 \leq i \leq \hat{k} + k_0. \end{cases}$$

and let $B_j = A'_j + \hat{k} = \{i + \hat{k} : i \in A'_j\}$. Based on this notation, we may rewrite $\hat{W}(G_n,G'_n)$ as follows:

$$\hat{W}(G_n,G'_n) = \inf_{q \in \Pi(G_n,G'_n)} \left\{ \sum_{i=1}^{\hat{k}} \sum_{(i,j) \in A_i \times B_j} q_{i(j-\hat{k})} |\nu_i^n - \nu_j^n| + \sum_{(i,j) \notin \cup_{i=1}^{\hat{k}} A_i \times B_j} q_{i(j-\hat{k})} \right\}.$$ 

From Lemma 7.1 in [16], we can find a finite number $S$ of scaling sequences $0 \equiv \tau_0(n) < \tau_1(n) < \ldots < \tau_S(n) \equiv 1$, where $\tau_s(n) = o(\tau_{s+1}(n))$, such that for any $j,j' \in \{1,2,\ldots,\hat{k} + k_0\}$, we can find a unique integer $s(j,j') \in \{0,1,\ldots,S\}$ satisfying $|\nu_j^n - \nu_j'^n| < \tau_{s(j,j')}(n)$. In the sequel, we shall sometimes omit the dependence on $n$ in the preceding notation. It can be inferred from its definition that $s(\cdot,\cdot)$ defines an ultrametric on the set $\{1,2,\ldots,\hat{k} + k_0\}$. As in [16], this allows us to construct a coarse graining tree over the set of balls in $\{1,\ldots,\hat{k} + k_0\}$ relative to the metric $s$. In the interest of being self-contained, we recall their definition as follows.

**Definition 2** (Definition 7.2 [16]). The coarse-graining tree $T$ is the collection of distinct balls $J = \{i \in 1,\ldots,\hat{k} + k_0\}$ such that $s(i,j) \leq s$, called nodes, for $j = 1,\ldots,\hat{k} + k_0$ and $s = 0,\ldots,S$. Moreover,

- The root of $T$ is $J_{\text{root}} = \{1,\ldots,\hat{k} + k_0\}$. 

31
• $J^\uparrow \in \mathcal{T}$ is called the parent of a node $J \in \mathcal{T}$ if the following implication holds for all $I \in \mathcal{T}$,

$$(J \subseteq I \subseteq J^\uparrow, I \in \mathcal{T}) \implies I = J.$$ 

• The set of children of a node $J \in \mathcal{T}$ is $\text{Child}(J) = \{I \in \mathcal{T} : I^\uparrow = J\}$.

• The set of descendants of a node $J \in \mathcal{T}$ is $\text{Desc}(J) = \{I \in \mathcal{T} : I^\uparrow \subseteq J\}$.

• The diameter of a node $J \in \mathcal{T}$ is $s(J) = \max_{j,j' \in J} s(j,j')$.

Since $k_* \geq 2$, it is a straightforward consequence of these definitions that the cardinality of $\text{Child}(J_{\text{root}})$ is exactly $k_*$, and we shall write $\text{Child}(J_{\text{root}}) = \{J_1, \ldots, J_{k_*}\}$. Furthermore, note that

$$J_l = A_l \cup B_l, \quad l = 1, \ldots, k_*.$$  

(27)

Now, let $\pi_J = \sum_{j \in J} \omega^n_j$ and $\tau_J = \tau_s(J)(n)$, for all $J \in \mathcal{T}$. We claim that the following key asymptotic equivalence holds.

**Lemma 12.** We have,

$$\tilde{W}(G_n, G'_n) \asymp \max \left\{ \max_{1 \leq l \leq k_*} \max_{J \in \text{Desc}(J_l)} |\pi_J| \tau_{J^\uparrow} J_l, \max_{1 \leq l \leq k_*} |\pi_{J_f}| \right\}. $$

(28)

The proof of Lemma A.5.1 is deferred to Section A.5.1. We next show how this Lemma may be used to lower bound the expansion of $F(\cdot, G_n)$ around $F(\cdot, G'_n)$. We begin with the following result, which is a simplified statement of Lemma 7.4 of [16]. In the sequel, for any node $J \in \mathcal{T}$, let $\nu_J$ denote an arbitrary but fixed element of $\{\nu^n_j : j \in J\}$.

**Lemma 13** (Lemma 7.4 [16]). For every $l = 1, \ldots, k_*$, there exists a vector $a_l = (a_l(p))_{0 \leq p \leq k_*}$ and a remainder $R_l(x)$ such that for all $x \in \mathbb{R}$,

$$\sum_{j \in J_l} \omega_j F(x, \nu^n_J) = \sum_{p=0}^{k_*} a_l(p) \tau_{J_l}^p F(p, \nu_{J_l}) + R_l(x),$$

Furthermore, the following assertions hold.

(i) We have $a_l(0) = \pi_{J_l}$, and,

$$\|a_l\| \asymp \max_{0 \leq p \leq |J_l| - 1} |a_l(p)| \gtrsim \max_{J \in \text{Desc}(J_l)} |\pi_J| \left( \frac{\tau_{J_l}}{\tau_{J_f}} \right)^{|J_l| - 1}.$$ 

(ii) We have, $\|R_l\|_\infty = o(\|a_l\| \tau_{J_l}^{k_*})$. 

32
By Lemma 13, we have for all $x \in \mathbb{R}$,

$$F(x, G_n) - F(x, G'_n) = \sum_{l=1}^{k_*} \sum_{j \in J_l} \omega_j F(x, \nu^n_j) = \sum_{l=1}^{k_*} \sum_{p=0}^{k_0} a_l(p) \tau^p_{J_l} F(p)(x, \nu_{J_l}) + \sum_{l=1}^{k_*} R_l(x).$$

Let $M_{n,l} = \max_{0 \leq p \leq |J_l|-1} |a_l(p)| \tau^p_{J_l}$ for any $l = 1, \ldots, k_*$, and let $M_n = \max_{1 \leq l \leq k_*} M_{n,l}$. By Lemma 13(i), we have

$$M_{n,l} \geq |a_l(0)| = |\pi_{J_l}|,$$

and additionally,

$$M_{n,l} \geq \max_{J \in \text{Desc}(J_l)} |\pi_J| \left( \frac{|J_l|-1}{\tau^p_{J_l}} \right) \min_{0 \leq p \leq |J_l|-1} \tau^p_{J_l} = \max_{J \in \text{Desc}(J_l)} |\pi_J|^{(|J_l|-1)}.$$

Let $D_n = \widetilde{W}(G_n, G'_n)$. By Lemma 12 and equations (29)–(30), we deduce that $M_n/D_n \geq 1$.Additionally, by Lemma 13(ii), we have $\|R_n\|_\infty = o(M_n)$. Therefore, setting $d_n = D_n/M_n$, we obtain that there exist finite real numbers $\alpha_{lp} \in \mathbb{R}$, not all of which are zero, such that,

$$\left\| d_n \frac{F(\cdot, G_n) - F(\cdot, G'_n)}{D_n} - \sum_{l=1}^{k_*} \sum_{p=0}^{k_0} \alpha_{lp} F(p)(\cdot, \theta^*_l) \right\|_\infty \to 0.$$

On the other hand, since $d_n \leq 1$, we have by assumption that $d_n\|F(\cdot, G_n) - F(\cdot, G'_n)\|_\infty/D_n \to 0$, thus we must obtain

$$\left\| \sum_{l=1}^{k_*} \sum_{p=0}^{k_0} \alpha_{lp} F(p)(\cdot, \theta^*_l) \right\|_\infty = 0.$$

By the strong identifiability condition of order $k + k_0$, it must follow that $\alpha_{lp} = 0$ for all $l = 1, \ldots, k_*$ and $p = 0, \ldots, k + k_0$, which is a contradiction. The claim thus follows.

### A.5.1 Proof of Lemma 12.

We first prove the lower bound of equation (28). For any coupling $q \in \Pi(G_n, G'_n)$ and for any $J, J' \in \mathcal{T}$, we denote

$$W(J, J'; q) = \sum_{l=1}^{k_*} \sum_{(i,j) \in (A_l \cap J) \times (B_l \cap J')} q_{(i,j)} p^n_{i} - \nu^n_{j} |A_l| + |B_l| - 1 + \sum_{(i,j) \in \mathcal{M}(J, J') \setminus \bigcup_{l=1}^{k_*} (A_l \cap J) \times (B_l \cap J')} q_{(i,j)},$$

where $\mathcal{M}(J, J') = (J \cap \{1, \ldots, \bar{k}\}) \times (J' \cap \{\bar{k} + 1, \ldots, \bar{k} + k_*\})$. From the above definition, we obtain that $\widetilde{W}(G_n, G'_n) = \inf_{q \in \Pi(G_n, G'_n)} W(J_{\text{root}}, J_{\text{root}}; q)$. Now, for any coupling $q$ between $G_n$ and $G'_n$, and for any node $J$ in the tree $\mathcal{T}$, we obtain that

$$W(J_{\text{root}}, J_{\text{root}}; q) \geq W(J, J'; q) + W(J', J; q).$$
Since $|v_i^n - v_j^n| \geq \tau_{J^\uparrow}$ for any $(i, j) \in J \times J^c$ or $(i, j) \in J^c \times J$, it follows that

$$
W(J, J^c; q) + W(J^c, J; q) \geq \sum_{l=1}^{k_0} \left[ \sum_{(i, j) \in (A_l \cap J) \times (B_l \cap J)} q_i(j-k) + \sum_{(i, j) \in (A_l \cap J^c) \times (B_l \cap J)} q_i(j-k) \right] + \sum_{(i, j) \in M(J, J^c) \cup \bigcup_{l=1}^{k_0} (A_l \cap J) \times (B_l \cap J^c)} q_i(j-k) + \sum_{(i, j) \in M(J^c, J) \cup \bigcup_{l=1}^{k_0} (A_l \cap J^c) \times (B_l \cap J)} q_i(j-k) := C,
$$

(31)

There are two settings of node $J$:

**Case 1:** $J \in \text{Child}(J_{\text{root}})$. In this case, $J = J_l$ for some $l \in [k_s]$. We deduce from equation (27) that $A_l \cap J^c = B_l \cap J^c = \emptyset$. Therefore, from equation (31), we obtain that

$$
C = \sum_{(i, j) \in M(J, J^c)} q_i(j-k) + \sum_{(i, j) \in M(J^c, J)} q_i(j-k) \geq \left| \sum_{(i, j) \in M(J \cup J^c)} q_i(j-k) - \sum_{(i, j) \in M(J \cup J^c, J^c)} q_i(j-k) \right| = |\pi_J|,
$$

(32)

**Case 2:** $J \in \text{Desc}(J_l)$ for some $l \in [k_s]$. Under this case, we can verify that

$$
C \geq \left[ \sum_{(i, j) \in (A_l \cap J) \times (B_l \cap J^c)} q_i(j-k) + \sum_{(i, j) \in (A_l \cap J^c) \times (B_l \cap J)} q_i(j-k) + \sum_{(i, j) \in M(J, J^c) \cup \bigcup_{l=1}^{k_0} (A_l \cap J) \times (B_l \cap J^c)} q_i(j-k) + \sum_{(i, j) \in M(J^c, J) \cup \bigcup_{l=1}^{k_0} (A_l \cap J^c) \times (B_l \cap J)} q_i(j-k) \right] + \sum_{l=1}^{k_0} \left| A_l \cup B_l \right| \geq \left| \pi_J \right| |A_l \cup B_l| - 1.
$$

(33)

Combining the results of equations (31), (32), and (33), we obtain the lower bound that

$$
\overline{W}(G_n, G'_n) \geq \max \left\{ \max_{1 \leq l \leq k_s} \left| \pi_{J_l} \right| \left| A_l \cup B_l \right| - 1, \max_{1 \leq l \leq k_s} \left| \pi_{J_l} \right| \right\}.
$$

Therefore, to obtain the conclusion of claim (28), it remains to verify the upper bound of $\overline{W}(G_n, G'_n)$ in that claim. Based on Lemma B.2 of [16], we can construct a coupling $\tilde{q}$ between $G_n$ and $G'_n$ such that for any node $J \in T$, we have

$$
\sum_{l=1}^{k_0} \sum_{(i, j) \in (A_l \cap J) \times (B_l \cap J)} \tilde{q}_i(j-k) = \min \{p_J, p'_J\},
$$

where $p_J = \sum_{i \in J} p_{i}^{n}$ and $p'_J = \sum_{i \in J} (p_{i-k}^{n})'$. Given the coupling $\tilde{q}$, we first
prove that for any node $J$ that is a descendant of $\mathcal{J}^l$ or equal to $\mathcal{J}^l$ for some $l \in [k_*]$, then

$$W(J, J; \bar{q}) \lesssim \max_{K \in \text{Desc}(J)} |\pi_K|^{|A_l|+|B_l|-1}. \quad (34)$$

We prove the inequality (34) by induction. When $J$ is an end node of $\mathcal{J}^l$, $W(J, J; \bar{q}) = 0$; therefore, inequality (34) holds true. We assume that this inequality holds for any node $K$ which is a child of a given node $J$. We now proceed to show that this inequality also holds for $J$. In fact, we have the following identity:

$$W(J, J; \bar{q}) = \sum_{K \in \text{Child}(J)} \left( W(K, K; \bar{q}) + \sum_{K' \neq K; K' \in \text{Child}(J)} W(K, K'; \bar{q}) \right).$$

Note that, for any $K$ and $K'$ that are children of node $J$, we have

$$W(K, K'; \bar{q}) = \sum_{(i, j) \in (A_l \cap K) \times (B_l \cap K')} \bar{q}_{i(j-k)} \nu_i^n - \nu_j^n |A_l|+|B_l|-1.$$ 

From the induction hypothesis, we obtain that $W(K, K; \bar{q}) \lesssim \max_{K' \in \text{Desc}(J)} |\pi_K|^{|A_l|+|B_l|-1}$. Furthermore, for any $K' \neq K$ and $K' \in \text{Child}(J)$, we find that

$$W(K, K'; \bar{q}) \lesssim \sum_{(i, j) \in (A_l \cap K) \times (B_l \cap K')} \bar{q}_{i(j-k)} \tau_j |A_l|+|B_l|-1 \lesssim |\pi_K|^{|A_l|+|B_l|-1}.$$ 

Collecting the above results, we arrive at $W(J, J; \bar{q}) \lesssim \max_{K \in \text{Desc}(J)} |\pi_K|^{|A_l|+|B_l|-1}$. Therefore, inequality (34) is proved for any node $J$ that is a descendant of $\mathcal{J}^l$ or equal to $\mathcal{J}^l$ for some $l \in [k_*]$. Now, we proceed to prove the following inequality

$$W(\mathcal{J}_{\text{root}}, \mathcal{J}_{\text{root}}; \bar{q}) \lesssim \max \left\{ \max_{1 \leq l \leq k_*} \frac{\max_{J \in \text{Desc}(\mathcal{J}^l)} |\pi_J|^{|A_l|+|B_l|-1}}{\max_{1 \leq l \leq k_*} |\pi_{\mathcal{J}^l}|} \right\} \cdot (35)$$

In fact, we have

$$W(\mathcal{J}_{\text{root}}, \mathcal{J}_{\text{root}}; \bar{q}) = \sum_{l=1}^{k_*} \left( W(\mathcal{J}^l, \mathcal{J}^l; \bar{q}) + \sum_{l' \neq l} W(\mathcal{J}^l, \mathcal{J}^{l'}; \bar{q}) \right).$$

From inequality (34), we obtain that $W(\mathcal{J}^l, \mathcal{J}^l; \bar{q}) \lesssim \max_{J \in \text{Desc}(\mathcal{J}^l)} |\pi_J|^{|A_l|+|B_l|-1}$ for any $l \in [k_*]$. Furthermore, for any $l' \neq l$, we find that

$$W(\mathcal{J}^l, \mathcal{J}^{l'}; \bar{q}) = \sum_{(i, j) \in \mathcal{M}(\mathcal{J}^l, \mathcal{J}^{l'}) \setminus \bigcup_{l=1}^{k_*} (A_l \cap \mathcal{J}^l) \times (B_l \cap \mathcal{J}^{l'})} \bar{q}_{i(j-k)} \lesssim |\pi_{\mathcal{J}^l}| = |\pi_{\mathcal{J}^{l'}}|.$$ 

Putting the above results together, we obtain the conclusion of inequality (35). Since $\tilde{W}(G_n, G'_n) \leq W(\mathcal{J}_{\text{root}}, \mathcal{J}_{\text{root}}; \bar{q})$, we reach the conclusion of claim (28). \qed
B Additional Results

In this appendix, we state and prove the following result which was deferred from the main text.

**Lemma 14.** Let $\Theta \subseteq \mathbb{R}^d$ be a compact set with nonempty interior.

(a) Let $\Delta = 1 \vee \text{diam}(\Theta) < \infty$ and $G_0 \in \mathcal{E}_{k_0}(\Theta)$. Then, for any $G \in \mathcal{O}_k(\Theta)$, we have

\[ \mathcal{D}(G, G_0) \geq \frac{1}{\Delta^2} W_2^2(G, G_0). \]

(b) Assume the mixing measure $G_0 \in \mathcal{E}_{k_0}(\Theta)$ admits a support point $\theta_0$ lying in the interior of $\Theta$. Then,

\[ \sup_{G \in \mathcal{O}_k(\Theta) \atop G \neq G_0} \frac{\mathcal{D}(G, G_0)}{W_2^2(G, G_0)} = \infty. \]

**Proof.** By Lemma B.2 of [16], there exists a coupling $\bar{q} \in \Pi(G, G_0)$ such that

\[ \sum_{i \in A_j} \bar{q}_{ij} = p_j^0 \land \sum_{i \in A_j} p_i, \quad j = 1, \ldots, k_0. \]

Using the above display and the marginal constraints in the definition of a coupling, we obtain

\[ W_2^2(G, G_0) \leq \sum_{i=1}^k \sum_{j=1}^{k_0} \bar{q}_{ij} \| \theta_i - \theta_j^0 \|^2 \]

\[ \leq \sum_{j=1}^{k_0} \sum_{i \in A_j} \bar{q}_{ij} \| \theta_i - \theta_j^0 \|^2 + \Delta^2 \sum_{j=1}^{k_0} \sum_{i \notin A_j} \bar{q}_{ij} \]

\[ = \sum_{j=1}^{k_0} \sum_{i \in A_j} \bar{q}_{ij} \| \theta_i - \theta_j^0 \|^2 + \Delta^2 \sum_{j=1}^{k_0} \left[ p_j^0 - \sum_{i \in A_j} \bar{q}_{ij} \right] \]

\[ \leq \sum_{j=1}^{k_0} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \|^2 + \Delta^2 \sum_{j=1}^{k_0} \left[ p_j^0 - \sum_{i \in A_j} p_i \right] \]

\[ \leq \Delta \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \|^2 + \sum_{j:|A_j| \geq 2} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \|^2 + \Delta^2 \sum_{j=1}^{k_0} \left[ p_j^0 - \sum_{i \in A_j} p_i \right] \]

\[ \leq \Delta^2 \left\{ \sum_{j:|A_j|=1} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \|^2 + \sum_{j:|A_j| \geq 2} \sum_{i \in A_j} p_i \| \theta_i - \theta_j^0 \|^2 + \sum_{j=1}^{k_0} \left[ p_j^0 - \sum_{i \in A_j} p_i \right] \right\} \]

\[ = \Delta^2 \mathcal{D}(G, G_0), \]
since $\Delta \geq 1$. This proves part (a). To prove part (b), recall that $G_0 = \sum_{j=1}^{k_0} p_j^0 \delta_{\theta_j^0}$ admits a support point lying in the interior of $\Theta$. Without loss of generality, we assume this support point is $\theta_1^0$. Therefore, there exists $\epsilon_0 > 0$ such that for all $\epsilon \in (0, \epsilon_0)$, $\theta_0^\epsilon := (1 + \epsilon)\theta_1^0 \in \Theta$. Define the mixing measure

$$G_\epsilon = p_1^0 \delta_{\theta_0^\epsilon} + \sum_{j=2}^{k_0} p_j^0 \delta_{\theta_j^0} \in O_{k_0}(\Theta) \subseteq O_k(\Theta).$$

Clearly, we may also choose $\epsilon_0$ small enough such that $\theta_0^\epsilon \in A_1(G_\epsilon)$ for all $\epsilon \in (0, \epsilon_0)$. Thus, $|A_j(G_\epsilon)| = 1$ for every $j = 1, \ldots, k_0$. By equation (36), we therefore have

$$W_2^2(G_\epsilon, G_0) \leq p_1^0 \epsilon.$$

On the other hand, using again the fact that $|A_j(G_\epsilon)| = 1$ for each $j = 1, \ldots, k_0$, we have

$$\mathcal{D}(G_\epsilon, G_0) = p_1^0 \epsilon.$$

We deduce that

$$\sup_{G \in O_k(\Theta)} \frac{\mathcal{D}(G, G_0)}{W_2^2(G, G_0)} \geq \sup_{\epsilon \in (0, \epsilon_0)} \frac{\mathcal{D}(G_\epsilon, G_0)}{W_2^2(G_\epsilon, G_0)} \geq \sup_{\epsilon \in (0, \epsilon_0)} \frac{1}{\epsilon} = \infty,$$

as claimed. \qed

C Simulation Study

We perform a simulation study to illustrate the convergence rates of the penalized MLE given in Sections 3 and 4. All simulations hereafter were performed in Python 3.7 on a standard Unix machine, and we provide further numerical details in Appendix C.1.

We consider three models A–C, which respectively correspond to the settings described in Sections 3.1, 3.2, and 4. In each case, we choose the kernel density $f$ to be the $d$-dimensional Gaussian density, and we generate observations from the Gaussian mixture density,

$$p_{G_0}(x) = \sum_{j=1}^{k_0} \pi_j^0 \frac{\exp \left\{-\frac{1}{2}(x - \mu_j^0)^\top (\Sigma_j^0)^{-1}(x - \mu_j^0) \right\}}{\sqrt{\det(2\pi \Sigma_j^0)}},$$

where $x \in \mathbb{R}^d$. The models are defined as follows.

**Model A.** We treat the scale parameters as equal and known, and set

$$\Sigma_1^0 = \ldots = \Sigma_{k_0}^0 = .01 I_d,$$

with $d = 2$ and $k_0 = 2$. The resulting location-Gaussian family of densities is strongly identifiable [4,
Thus the result of Theorem 4 applies to this family. We set the location parameters and mixing proportions as follows,

$$\theta_1^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \theta_2^0 = \begin{pmatrix} .2 \\ .2 \end{pmatrix}, \quad \pi_1^0 = \pi_2^0 = \frac{1}{2}.$$

**Model B.** We next consider a two-dimensional Gaussian mixture model with \(k_0 = 3\) components, however we now treat both location and scale parameters as unknown. Define,

$$\mu_1^0 = \begin{pmatrix} 0 \\ .3 \end{pmatrix}, \quad \mu_2^0 = \begin{pmatrix} .1 \\ -.4 \end{pmatrix}, \quad \mu_3^0 = \begin{pmatrix} .5 \\ .2 \end{pmatrix},$$

$$\Sigma_1^0 = \begin{pmatrix} .042824 & .017324 \\ .017324 & .081759 \end{pmatrix}, \quad \Sigma_2^0 = \begin{pmatrix} .0175 & -.0125 \\ -.0125 & .0175 \end{pmatrix},$$

$$\Sigma_3^0 = \begin{pmatrix} .01 & -.0125 \\ -.0125 & .0175 \end{pmatrix}, \quad \pi_1^0 = \frac{1}{3}, \quad \pi_2^0 = \frac{1}{4}, \quad \pi_3^0 = \frac{1}{3}.$$

The above parameters are taken from the simulation study of [17], up to rescaling. This model falls within the setting of Theorem 5.

**Model C.** We again consider a location-Gaussian family as in Model A, but now with parameters \(G_0 \equiv G_0^0\) depending on the sample size \(n\). We set the scale parameters as in equation (38) with \(d = 1\). Furthermore, we consider two distinct submodels, depending on the true number of components \(k_0\). Our definitions depend on the sequence \(\epsilon_n = n^{-\frac{1}{4k_0-6}}\).

- When \(k_0 = 3\), we set

  $$\mu_{1,n}^0 = 0, \quad \mu_{2,n}^0 = .2 + \epsilon_n, \quad \mu_{3,n}^0 = .2 + 4\epsilon_n.$$

- When \(k_0 = 4\), we retain the above parameters and additionally define

  $$\mu_{4,n}^0 = .2 - 1.5\epsilon_n.$$

In both cases, the mixing proportions are chosen such that the resulting mixtures are balanced. These models correspond to the setting described in Section 4, relative to the limiting mixing measure

$$G_* = \frac{1}{2} \delta_0 + \frac{1}{2} \delta_2, \quad k_* = 2.$$

For each model, we generate 20 samples of size \(n\), for 100 different choices of \(n\) between \(10^2\) and \(10^5\). For each sample, we compute the penalized MLE \(\hat{G}_n\) with respect to the tuning parameter \(\xi_n = \log n\), and with respect to a number of components \(k\). For the fixed Models A–B, we choose \(k \in \{k_0 + 1, k_0 + 2\}\), whereas for the varying Model C, we choose \(k = k_0 \in \{k^* + 1, k^* + 2\}\).
Figure 2: Log-log scale plots for the simulation results under Models A–C. For each model and sample size $n$, we compute the estimator $\hat{G}_n$ on 20 independent samples of size $n$. Its average discrepancy from the true mixing measure is plotted in blue, with error bars representing two empirical standard deviations. We additionally plot, in orange, the fitted linear regression line of these points, obtained using the method of least squares.
We now report additional numerical details. We implement the penalized MLE \( \hat{C} \).

### C.1 Numerical Specifications

Parameters in finite mixture models may converge as fast as the parametric rate. Models typically enjoy heterogeneous rates of convergence. In particular, a subset of the estimated parameters is likely to converge at the faster rate, while others may converge at the approximate rate. Under Figure 1(a), a similar conclusion holds true, but now two possibilities arise: either \( \alpha_n = n^{-1/6} \) and \( \beta_n = n^{-1/2} \), or \( \alpha_n = \beta_n = n^{-1/4} \). In contrast, past literature on mixture models only implies that the worst of these rates (i.e. \( n^{-1/4} \) for Model A and \( n^{-1/6} \) for Model C) hold for all three fitted parameters. The main contribution of our work was to show that such results are overly pessimistic, and that the fitted parameters of finite mixture models typically enjoy heterogeneous rates of convergence. In particular, a subset of the estimated parameters in finite mixture models may converge as fast as the parametric rate.

**Algorithm 1** Modified EM Algorithm.

Starting values \( \Psi^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)}, \pi_1^{(0)}, \ldots, \pi_k^{(0)}) \); i.i.d. sample \( X_1, \ldots, X_n \); tuning parameter \( \xi_n = \log n \); maximum number of iterations \( T > 0 \); convergence criterion \( \epsilon > 0 \).

repeat \( \| \Psi^{(t)} - \Psi^{(t-1)} \| \leq \epsilon \) or \( t \geq T \).

\[
\begin{align*}
\pi_j^{(t+1)} & = \frac{\sum_{i=1}^n w_{ij}^{(t)} + \xi_n}{n + k \xi_n}, \\
\mu_j^{(t+1)} & = \frac{\sum_{i=1}^n w_{ij}^{(t)} X_i}{\sum_{i=1}^n w_{ij}^{(t)}}, \\
\Sigma_j^{(t+1)} & = \frac{\sum_{i=1}^n w_{ij}^{(t)} (X_i - \mu_j^{(t)}) (X_i - \mu_j^{(t)})^\top}{\sum_{i=1}^n w_{ij}^{(t)}}, \\
\Psi^{(t+1)} & = (\theta_1^{(t)} , \ldots, \theta_k^{(t)}, \pi_1^{(t)}, \ldots, \pi_k^{(t)}),
\end{align*}
\]

Compute \( w_{ij}^{(t+1)} \) using Algorithm 1 as written for Model B, whereas for Models A and C, we omitted the update on line 8 for the scale parameters, and simply held them fixed to their true values.

We chose the convergence criteria \( \epsilon = 10^{-8} \) and \( T = 2,000 \). The tuning parameter for the penalty is taken to be \( \xi_n = \log n \), in accordance to our theoretical results in Sections 3–4. We made no attempt at further tuning the leading constant of 1 in the definition of \( \xi_n \), since this parameter...
has previously been reported to have a negligible impact on the fitted mixture parameters (see for
instance [24] and references therein).

Since our purpose is to illustrate theoretical properties of the estimator $\hat{G}_n$, we initialized the EM
algorithm favorably. In particular, for any given $k$ and $k_0$, and for each replication, we randomly
partitioned the set $\{1, \ldots, k\}$ into $k_0$ index sets $I_1, \ldots, I_{k_0}$, each containing at least one point. We
then sampled $\theta^{(0)}_j$ (resp. $\Sigma^{(0)}_j$) from a Gaussian distribution with vanishing covariance, centered at
$\theta^{0}_\ell$ (resp. $\Sigma^{0}_\ell$), where $\ell$ is the unique index such that $j \in I_\ell$. 

41