The Annals of Statistics
2016, Vol. 44, No. 6, 2726–2755
DOI: 10.1214/16-AOS1444

CONVERGENCE RATES OF PARAMETER ESTIMATION FOR SOME WEAKLY IDENTIFIABLE FINITE MIXTURES[1]

BY NHAT HO AND XUANLONG NGUYEN

University of Michigan

We establish minimax lower bounds and maximum likelihood convergence rates of parameter estimation for mean-covariance multivariate Gaussian mixtures, shape-rate Gamma mixtures and some variants of finite mixture models, including the setting where the number of mixing components is bounded but unknown. These models belong to what we call "weakly identifiable" classes, which exhibit specific interactions among mixing parameters driven by the algebraic structures of the class of kernel densities and their partial derivatives. Accordingly, both the minimax bounds and the maximum likelihood parameter estimation rates in these models, obtained under some compactness conditions on the parameter space, are shown to be typically much slower than the usual $n^{-1/2}$ or $n^{-1/4}$ rates of convergence.

1. Introduction. Location-scale Gaussian mixtures are one of the most widely utilized modeling tools in statistics. Shape-rate Gamma mixtures are also a useful modeling choice for nonnegative valued data. Yet convergence behaviors of the parameters arising in these model classes remain largely open questions [10, 20, 22]. We seek to address these questions in this paper.

For finite mixtures of Gaussians, some facts are known when only one type of parameter varies (such as the mean/location or the variance/scale but not both). Specifically, if the number of mixing components generating the data is given, then the optimal rate of parameter estimation is the standard $n^{-1/2}$, where n is the sample size. If the number of mixing components is unknown but bounded by a known constant, then the convergence rate $n^{-1/4}$ for estimating the mixing distribution is achieved by a procedure established by Chen [9]. For multi-dimensional parameters, the $(\log n/n)^{1/4}$ rate of posterior concentration of the mixing distribution was established by Nguyen [23], under Wasserstein distance W_2. Ho and Nguyen [17] extended the results of [9] and [23] to a broader range of *strongly* identifiable models, which admit general rates for the mixing measure under maximum likelihood estimation (MLE): $(\log n/n)^{1/2}$ for exact-fitted mixtures under W_1 metric, and $(\log n/n)^{1/4}$ for over-fitted finite mixtures under W_2 metric.

Received February 2015; revised January 2016.

[1]Supported in part by Grants NSF CCF-1115769, NSF CAREER DMS-1351362 and CNS-1409303.

MSC2010 subject classifications. Primary 62F15, 62G05; secondary 62G20.

Key words and phrases. Mixture models, strong identifiability, weak identifiability, Wasserstein distances, minimax bounds, maximum likelihood estimation, system of polynomial equations.

Strong identifiablity and related notions, as studied by [9, 23] and several others (e.g., [21, 26]), refers to a linear independence condition on the class of kernel density functions and their first- and second-order partial derivatives with respect to the parameters. It is fruitful to delineate this condition further: first-order identifiability requires linear independence of the density functions and their first-order derivatives; second-order identifiability requires linear independence of the density functions and their partial derivatives up to the second order [17]. The classical identifiability condition—linear independence of the class of density functions—corresponds to zero-order identifiability. Gaussian mixtures with both the mean and covariance parameters varying are identifiable up to the first order, but *not* in the second order. Gamma mixtures are not identifiable even in the first order, despite being identifiable in the classical sense. In each of these examples, the violation of such identifiability conditions is due to a specific interaction among different parameters being present in the model class. Such interactions are driven by specific algebraic structures of the class of kernel densities and their partial derivatives. They can be succinctly expressed by certain partial differential equations satisfied by the kernel density function.

We shall informally refer to those finite mixture models *weakly identifiable* if they fail either the first- or second-order identifiability condition, but otherwise are identifiable in the classical sense. Most relevant existing works on the asymptotics of parameter estimation (e.g., [9, 17, 23]) concern only the settings of strong identifiability, and thus quite inapplicable to weakly identifiable classes. In fact, for such model classes the standard rates of convergence $n^{-1/2}$ and $n^{-1/4}$ (modulo a logarithmic term) no longer hold in general—the rates that we establish in this paper are nonstandard, and new. For instance, we shall show that for a location-scale Gaussian mixture where the number of mixing components is unknown and bounded by a constant, a minimax lower bound and the MLE convergence rate for estimating the mixing measure depend on how much we potentially overfit the model: the estimation rate is $n^{-1/8}$ under the 4th-order Wasserstein distance W_4, if overfitting by one extra component; $n^{-1/12}$ under the 6th-order Wasserstein distance W_6 if overfitting by two extra components. All these rates occur while the MLE convergence rate of the mixture density remains to be $n^{-1/2}$. Remarkably, for Gamma and some other mixtures, the minimax lower bound for estimating the mixing measure is shown to be worse than *any* polynomial rate of the form $n^{-1/r}$ even when the number of mixing components is known.

In the special case of overfitting location-scale Gaussian mixtures by one extra component, the poor convergence rate for parameter estimation has been noted before by several authors. Most notably, Chen and Chen [4] established the convergence rate $n^{-1/8}$ of the mixing distribution under a hypothesis testing for homogeneity. Kasahara and Shimotsu [18] also achieved the rate $n^{-1/8}$ of MLE of finite normal regression mixtures (overfitted by one more component) when parameters are reparameterized and mixing proportions are restricted to be bounded away from zero. We are not aware of existing work on Gamma mixtures.

1.1. *Main results for Gaussian mixtures.* Given an n-i.i.d. sample $X_1, \ldots, X_n$ generated according to a Gaussian mixture density $p_{G_0}(x) = \int f(x|\theta, \Sigma) G_0(\mathrm{d}\theta, \mathrm{d}\Sigma)$, where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$ has $k_0 \geq 1$ distinct support points. The class of Gaussian densities is denoted by $\{f(x|\theta, \Sigma), \theta \in \Theta \subset \mathbb{R}^d, \Sigma \in \Omega \subset S_d^{++}\}$, where S_d^{++} indicates the set of all symmetric positive definite matrices on $\mathbb{R}^{d \times d}$ and $d \geq 1$. Throughout this paper, Θ and Ω shall be restricted to be compact subsets where their precise formations are given in our main theorems. (We note that without these compactness conditions, the MLE of G_0 may not exist or be inconsistent.) Now, we shall fit a mixture of k Gaussian distributions using the n-sample, where $k \geq k_0 + 1$. Denote by $\mathcal{O}_k := \mathcal{O}_k(\Theta \times \Omega)$ the set of probability measures on $\Theta \times \Omega$ with at most k support points, $\mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta \times \Omega)$ the set of probability measures on $\Theta \times \Omega$ with exactly k_0 support points. In addition, given $c_0 \in [0, 1)$, define a subset of $\mathcal{O}_k$,

$$\mathcal{O}_{k,c_0} := \left\{ G = \sum_{i=1}^{k^*} p_i \delta_{(\theta_i, \Sigma_i)} \in \mathcal{O}_k : p_i \geq c_0 \ \forall 1 \leq i \leq k^* \right\}.$$

Let $\widehat{G}_n$ be an estimate of G_0. We seek to derive the rate of convergence of $\widehat{G}_n$ to G_0 under a number of settings. For evaluating the convergence of mixing measures, Wasserstein distances have proved to be a natural choice [23, 24]. Given two discrete probability measures $G = \sum_{i=1}^{k} p_i \delta_{(\theta_i, \Sigma_i)}$ and $G' = \sum_{i=1}^{k'} p_i' \delta_{(\theta_i', \Sigma_i')}$ on $\Theta \times \Omega$, recall that the sth ($s \geq 1$) order Wasserstein distance between G and G' takes the form [29]:

$$W_s(G, G') = \left(\inf \sum_{i,j} q_{ij} (\|\theta_i - \theta_j'\| + \|\Sigma_i - \Sigma_j'\|)^s \right)^{1/s},$$

where the infimum is taken over all couplings $\mathbf{q}$ between $\mathbf{p}$ and $\mathbf{p}'$, that is, $\mathbf{q} = (q_{ij})_{ij} \in [0, 1]^{k \times k'}$ such that $\sum_{i=1}^{k} q_{ij} = p_j'$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, \ldots, k$ and $j = 1, \ldots, k'$. In addition, $\|\cdot\|$ denotes either the ℓ_2 norm for elements in $\mathbb{R}^d$ or the entrywise ℓ_2 norm for matrices.

To see how a convergence rate in Wasserstein distance W_s is translated to that of the parameters, suppose that a sequence of mixing measures G_n tending to G_0 under W_s metric at a rate $\omega_n = o(1)$. If all G_n have the same number of atoms $k = k_0$ as that of G_0, then the set of atoms of G_n converge to the k_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_n under $\|\cdot\|$ metric. If G_n have varying $k_n \in [k_0, k]$ number of atoms, where k is a fixed upper bound, then a subsequence of G_n can be constructed so that each atom of G_0 is a limit point of a certain subset of atoms of G_n—the convergence to each such limit also happens at rate ω_n. Some atoms of G_n may have limit points that are not among G_0's atoms—the total mass associated with those "redundant" atoms of G_n must vanish at the generally faster rate ω_n^s.

For over-fitted Gaussian mixtures with both mean and variance varying, a main result of this paper is to show that the rate of convergence of the mixing measure is determined by the order of a set of polynomial equations, which we now describe precisely. Denote by $\overline{r} \geq 1$ the *minimum* value of $r \geq 1$ such that the following system of polynomial equations:

$$\sum_{j=1}^{k-k_0+1} \sum_{n_1,n_2} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \qquad \text{for each } \alpha = 1, \ldots, r \tag{1}$$

does *not* have any nontrivial solution for the unknowns $(a_j, b_j, c_j)_{j=1}^{k-k_0+1}$. The ranges of n_1, n_2 in the second sum are all natural pairs satisfying $n_1 + 2n_2 = \alpha$. A solution is considered nontrivial if all of c_j's are nonzeros, while at least one of the a_js is nonzero.

THEOREM 1.1 (Gaussian mixtures). *Let $L, \gamma, \underline{\lambda} < \overline{\lambda}$ be fixed positive numbers. Given $\Theta = [-a_n, a_n]^d$ where $a_n \leq L(\log n)^{\gamma}$, and Ω be a subset of S_d^{++} whose eigenvalues are bounded in an interval $[\underline{\lambda}, \overline{\lambda}]$.*

(a) (*Minimax lower bound.*) *For any $r < 2\overline{r}$,*

$$\inf_{\widehat{G}_n \in \mathcal{O}_k} \sup_{G \in \mathcal{O}_k \setminus \mathcal{O}_{k_0}} E_{p_G} W_1(\widehat{G}_n, G) \geq c_1 n^{-1/r}.$$

Here, the infimum is taken over all sequences of estimates $\widehat{G}_n$ ranging in $\mathcal{O}_k$, E_{p_G} denotes the expectation taken with respect to product measure with mixture density p_G^n, c_1 is a universal positive constant.

(b) (*Maximum likelihood estimation.*) *Let $c_0 = 0$ if $k - k_0 = 1$ or 2, and $c_0 > 0$ otherwise. Assume that $G_0 \in \mathcal{O}_{k,c_0}$ and let $\widehat{G}_n$ be the MLE ranging in $\mathcal{O}_{k,c_0}$. Then*

$$\mathbb{P}\big(W_{\overline{r}}(\widehat{G}_n, G_0) > C(\log n/n)^{1/(2\overline{r})}\big) \lesssim \exp(-c \log n).$$

Here, probability $\mathbb{P}$ is taken with respect to p_{G_0}. C, c are positive constants depending only on $d, L, \gamma, \underline{\lambda}, \overline{\lambda}, c_0$ and G_0.

Part (a) of Theorem 1.1 establishes a minimax lower bound for estimating mixing measure G under W_1 distance. Noting the general inequality $W_{\overline{r}} \geq W_1$, this lower bound obviously also holds for $W_{\overline{r}}$. In words, when the number of mixing components is unknown except that it lies in the interval $[k_0, k]$, then there is no method for estimating G at a rate better than $n^{-1/(2\overline{r})}$, uniformly for all $G \in \mathcal{O}_k \setminus \mathcal{O}_{k_0}$. The proof actually obtains something stronger: the lower bound holds uniformly for any fixed or suitably shrinking W_1 neighborhood in $\mathcal{O}_k$ of any $G_0 \in \mathcal{E}_{k_0}$. Part (b) of Theorem 1.1 establishes that, under the compactness of the parameter spaces Θ, Ω, the rate $n^{-1/(2\overline{r})}$ can be achieved, up to a logarithmic term $\log n$, by maximum likelihood estimation. We wish to emphasize that this is a pointwise convergence rate, that is, constant C depends on G_0. For a fixed G_0,

we do not know if the upper bound $n^{-1/(2\overline{r})}$ of the convergence rate for the MLE may still be improved without additional assumptions or not. As a consequence of part (a), the upper bound $n^{-1/(2\overline{r})}$ is sharp in the sense that it cannot be improved uniformly for any W_1 neighborhood for G_0.

The link of the estimation rate for location-scale Gaussian mixtures to the solvability of the system of polynomial equations (1) established by the above theorem is rather striking, as it describes precisely the hardness of parameter estimation in over-fitted situations. Determining the solvability of a system of polynomial equations is a basic question in (computational) algebraic geometry. For system (1), there does not seem to be an obvious answer as to the general value of $\overline{r}$. Since the number of variables in this system is $3(k - k_0 + 1)$, one expects that $\overline{r}$ keeps increasing as $k - k_0$ increases. Using a standard method of Groebner bases [2], we can show that for $k - k_0 = 1$ and 2, $\overline{r} = 4$ and 6, respectively. In addition if $k - k_0 \geq 3$, then $\overline{r} \geq 7$. Thus, the convergence rate of the mixing measure for Gaussian mixtures deteriorates rapidly as more extra components are included in the model. We expect, but do not have a proof, that the value $\overline{r}$ in the rate $n^{-1/2\overline{r}}$ tends to infinity as the number of redundant Gaussian components increases to infinity. We note several recent results at the other end of the rate spectrum: when the number of mixing components is unbounded (infinite), the convergence rate of the mixing measure under W_2 is shown to be $(\log n)^{-1/2}$ for the location Gaussian mixtures [3, 23]. This rate may also resonate with some classical results in the deconvolution literature (e.g., [12, 31]), but one should be reminded that these classical results are applicable to only location mixtures carrying smooth mixing densities. Interestingly, although the convergence rate of mixing measures in over-fitted finite mixtures may be poor, if one is interested in mixing proportions only, it follows from the previous discussion of Wasserstein distance $W_{\overline{r}}$ that the rate $(n^{-1/(2\overline{r})})^{\overline{r}} = n^{-1/2}$ is still achieved by the MLE. This rate is also obtained by a Bayesian estimation procedure studied by Rousseau and Mengersen [26].

1.2. *Results for other weakly identifiable classes.* We now briefly describe other model classes studied in this paper. Gamma densities represent an interesting instance: the Gamma density $f(x|a, b)$ has two positive parameters, a for shape and b for rate. This family is not identifiable in the first order. Moreover, we will show that there are particular combinations of the true parameter values which prevent the Gamma class from enjoying strong convergence properties. On the other hand, by excluding the measure-zero set of pathological cases of true mixing measures, the Gamma density class in fact can be shown to be strongly identifiable in both orders. Thus, this class is *almost* strongly identifiable, using the terminology of [1]. The generic/pathological dichotomy in the convergence behavior within the Gamma class is quite interesting: in the generic case of true mixing measures, the mixing measure can be estimated at the standard rate (i.e., $n^{-1/2}$ under W_1 for exact-fitted and $n^{-1/4}$ under W_2 for over-fitted mixtures). The pathological cases

are very unforgiving: even for exact-fitted mixtures, one can do no better than a logarithmic rate of convergence in a minimax sense.

Lest some readers wonder whether this unusually slow rate for the exact-fitted mixture setting can happen only in the measurably negligible (pathological) cases, we also introduce a location-extension of the exponential distribution, the location-exponential class: $f(x|\theta,\sigma) := \frac{1}{\sigma}\exp(-\frac{x-\theta}{\sigma})1(x > \theta)$. We show that the minimax lower bound for estimating the mixing measure in an location-exponentials is no faster than a logarithmic rate, even when the number of mixing component is known.

Practical implications. In theory, mixture models enjoy strong asymptotic properties as a black-box modeling device for density estimation; see [13, 14, 19, 25] and the references therein. In practice, the parameters specific to each mixing components may carry useful information about the heterogeneity among the underlying (latent) subpopulations. Thus, understanding the statistical efficiency of parameter estimation in mixture modeling is also relevant from a practical standpoint. Problematic convergence behaviors exhibited by widely utilized models such as Gaussian mixtures may have long been observed in practice, but a concrete theory has been largely unavailable. The results established in this paper present a cautionary tale about the limitation of Gaussian mixtures, when it comes to assessing the quality of parameter estimation, but only when the number of mixing components is unknown. Since a tendency in practice is to "over-fit" the mixture generously with many more extra mixing components, our theory warns against this because as we have shown, the convergence rate via standard methods such as MLE for subpopulation-specific parameters deteriorates rapidly with the number of redundant components. For Gamma and location-exponential distribution, our theory also paints wildly varied convergence behaviors within each model class, and thus a similarly extreme caution. We hope that the theoretical results obtained may hint at practically useful ways for determining benign scenarios and imposing helpful constraints when the mixture models enjoy strong identifiability properties and favorable convergence rates, and for identifying pathological scenarios where the practitioners would do well by avoiding them.

Paper organization. Section 2 is devoted to the proof of the results for Gaussian mixture models. Section 3 investigates Gamma mixtures and a location extension of exponential distribution. The theoretical bounds are illustrated via simulations in Section 4. Remaining proofs are given in Section 5 and in the supplemental material [16].

Notation. In addition to Wasserstein distances for mixing measures, we also utilize several familiar notions of distance for mixture densities, with respect to Lebesgue measure. They are total variation distance $V(p_G, p_{G'}) = \frac{1}{2}\int |p_G(x) - p_{G'}(x)|\,\mathrm{d}\mu(x)$ and Hellinger distance $h^2(p_G, p_{G'}) = \frac{1}{2}\int (\sqrt{p_G(x)} - \sqrt{p_{G'}(x)})^2\,\mathrm{d}\mu(x)$.

2. Proof of main results for Gaussian mixtures. This section is devoted to proving Theorem 1.1. This theorem addresses only over-fitted Gaussian mixtures, that is, when the true number of mixing components is bounded but otherwise unknown. If the number of mixing Gaussian components is known, it was already shown that the rate of estimating the mixing measure G is the standard rate $n^{-1/2}$ under W_1 metric [17]. This is due to the fact that the class of Gaussian densities with both mean and covariance parameters varying is identifiable in the first order. However, the Gaussian family is not identifiable in the second order—that is to say that the collection of Gaussian density functions and their partial derivatives up to the second order taken with respect to the mean and covariance parameters are *not* linearly independent. This can be seen by the following identity, which represents a partial differential equation satisfied by Gaussian density $f(x|\theta, \Sigma)$:

$$\frac{\partial^2 f}{\partial \theta^2}(x|\theta, \Sigma) = 2\frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma). \tag{2}$$

This identity, also noted previously by [4, 18], will play a fundamental role in our proof of Theorem 1.1.

2.1. *On the order $\overline{r}$.* Before proceeding to the proof of the theorem, let us briefly discuss some properties of $\overline{r}$ as defined in (1). This is a system of r polynomial equations with $3(k - k_0 + 1)$ unknowns. The condition $c_1, \ldots, c_{k-k_0+1} \neq 0$ is important. In fact, if $c_1 = 0$, then by choosing $a_1 \neq 0$, $a_i = 0$ for all $i = 2, \ldots, k - k_0 + 1$ and $b_j = 0$ for all $j = 1, \ldots, k - k_0 + 1$, we can check that system (1) is satisfied for all $\alpha \geq 1$. Therefore, without this condition, $\overline{r}$ does not exist.

To illustrate the possible values of $\overline{r}$, let us consider the case $k = k_0 + 1$, and let $r = 3$. System (1) reduces to the equations

$$\begin{aligned} &c_1^2 a_1 + c_2^2 a_2 = 0, \\ &\tfrac{1}{2}(c_1^2 a_1^2 + c_2^2 a_2^2) + c_1^2 b_1 + c_2^2 b_2 = 0, \\ &\tfrac{1}{3!}(c_1^2 a_1^3 + c_2^2 a_2^3) + c_1^2 a_1 b_1 + c_2^2 a_2 b_2 = 0. \end{aligned}$$

It is simple to see that a nontrivial solution exists, by choosing $c_2 = c_1 \neq 0$, $a_1 = 1$, $a_2 = -1$, $b_1 = b_2 = -1/2$. Hence, $\overline{r} \geq 4$. For $r = 4$, the system consists of the three equations given above, plus

$$\tfrac{1}{4!}(c_1^2 a_1^4 + c_2^2 a_2^4) + \tfrac{1}{2!}(c_1^2 a_1^2 b_1 + c_2^2 a_2^2 b_2) + \tfrac{1}{2!}(c_1^2 b_1^2 + c_2^2 b_2^2) = 0.$$

It will be shown in the sequel that this system has no nontrivial solution. Therefore, for $k = k_0 + 1$, we have $\overline{r} = 4$.

Determining the exact value of $\overline{r}$ in the general case appears quite challenging. For the specific value of $k - k_0$, one can find $\overline{r}$—there are well-developed methods in computational algebra for dealing with this type of polynomial equations, such as Groebner bases [2] and resultants [27]. Using the Groebner bases method, we shall show in Section 5 the following.

PROPOSITION 2.1. $\overline{r} = 4$ *if* $k = k_0 + 1$, $\overline{r} = 6$ *if* $k = k_0 + 2$. *If* $k \geq k_0 + 3$, *then* $\overline{r} \geq 7$.

2.2. *Discussion of conditions in Theorem* 1.1. The main conditions in the statement of Theorem 1.1 are concerned with compactness and boundedness of the mixture model's parameters, including the parameters of mixing components, and the parameters for mixing probabilities.

The parameters of mixing components lie in Ω and Θ. Compactness conditions for Ω and Θ are required for three reasons. First, the compactness of Ω is important in guaranteeing that the likelihood function is bounded. Indeed, if the smallest eigenvalue of the covariance parameter is not bounded below or the largest eigenvalue of the covariance parameter is not bounded above, the likelihood function will become unbounded [5, 11, 15]. Second, the compactness of Θ and Ω are also crucial in obtaining upper bounds of the (bracket) entropies that we need for Lemma 2.1. Such bounds yield convergence rate $n^{-1/2}$, up to logarithmic factor, for the convergence of mixture density p_G under Hellinger distance. Third, and most importantly, these compactness assumptions are required in establishing the lower bounds of Hellinger distance of mixture densities in terms of Wasserstein distance of mixing measures (cf. Proposition 2.2), thereby allowing us to translate the convergence rate of the mixture density into that of the corresponding mixing measure. Our proof technique hinges upon the compactness conditions. As pointed out by the referees, one may be able to relax somewhat the compactness assumptions by penalizing the likelihood function appropriately [7, 8]. While the first two issues discussed above may still be addressed, the third issue will require a substantially new proof technique; moreover, the rate of convergence will be likely different.

It is required in part (b) of the theorem that $\widehat{G}_n$ range in $\mathcal{O}_{k,c_0}$, where $c_0 > 0$ when $k - k_0 \geq 3$. This requirement is sufficient for establishing the bound in part (b) of Proposition 2.2. A consequence of this requirement is that it prevents the Fisher matrix at the masses from being degenerate [5, 6, 18]. As such, this condition is also crucial in obtaining the asymptotic distribution of parameter estimates. We note, however, that this requirement may not be necessary for the purpose of establishing rates of parameter estimation. In fact, when the Gaussian mixture is overfitted by at most two components, that is, $1 \leq k - k_0 \leq 2$, it will be demonstrated by Proposition 2.3 that this requirement can be removed (by letting $c_0 = 0$) without affecting the conclusion of the theorem.

2.3. *Sharp identifiability bounds*. A central ingredient in the proof of Theorem 1.1 are sharp inequalities which relate the distance of two Gaussian mixture densities to a Wasserstein distance between corresponding mixing measures. Let $V(p_G, p_{G_0})$ denote the variational distance, and $h(p_G, p_{G_0})$ the Hellinger distance of p_G and p_{G_0}. The order $\overline{r}$ enters the following bounds in an essential way.

PROPOSITION 2.2. *Let $\overline{r}$ be defined as above, and $G_0 \in \mathcal{E}_{k_0} \cap \mathcal{O}_{k_0,c_0}$ for some $c_0 > 0$.*

(a) *For any $1 \leq r < \overline{r}$, there holds*

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_k} \{h(p_G, p_{G_0})/W_1^r(G, G_0) : W_1(G, G_0) \leq \varepsilon\} = 0.$$

(b) *For any $G \in \mathcal{O}_{k,c_0}$ such that $W_{\overline{r}}(G, G_0)$ is sufficiently small, there holds*

$$h(p_G, p_{G_0}) \geq V(p_G, p_{G_0}) \gtrsim W_{\overline{r}}^{\overline{r}}(G, G_0) \geq W_1^{\overline{r}}(G, G_0).$$

The proof of this proposition is deferred to Section 5. We make several remarks:

(i) In part (a), the ratio h/W_1^r is set to ∞ if $W_1 = 0$. In part (b) and for the rest of the paper, the multiplying constant in $\gtrsim$ bound depends only on G_0.

(ii) Part (a) and part (b) together show that $W_{\overline{r}}^{\overline{r}}(G, G_0)$ is the sharp lower bound for the distance of mixture densities $V(p_G, p_{G_0})$. In particular, we cannot have $V \gtrsim W_1^r$ for any $r < \overline{r}$.

(iii) In part (b), G is restricted to a subset of $\mathcal{O}_k$, that is, set $\mathcal{O}_{k,c_0}$, which places a lower bound constraint on the mixing probability mass. This restriction seems to be an artifact of our proof technique. It can be removed completely with some extra hard work, at least for the case $k - k_0 \leq 2$, as follows.

PROPOSITION 2.3. *Let $k - k_0 = 1$ or 2. Fix $G_0 \in \mathcal{E}_{k_0}$. For any $G \in \mathcal{O}_k$ such that $W_{\overline{r}}(G, G_0)$ is sufficiently small, we have $V(p_G, p_{G_0}) \gtrsim W_{\overline{r}}^{\overline{r}}(G, G_0)$.*

The proof of Proposition 2.3 is deferred to [16]. Given the two propositions above, we can now complete the proof of Theorem 1.1.

2.4. *Proof of Theorem 1.1.* (a) The proof of this part follows from the same argument as that of Lemma 1 of [30] for establishing minimax lower bounds. Fix $r < \overline{r}$ and $G_0 \in \mathcal{E}_{k_0}$. Let $C_0 > 0$ be any fixed constant. According to part (a) of Proposition 2.2, for any sufficiently small $\varepsilon > 0$, there exists $G_0' \in \mathcal{O}_k$ such that $W_1(G_0, G_0') = 2\varepsilon$ and $h(p_{G_0}, p_{G_0'}) \leq C_0\varepsilon^r$. Take any sequence of estimates $\widehat{G}_n$ ranging in $\mathcal{O}_k$, we have

$$2 \max_{G \in \{G_0, G_0'\}} E_{p_G} W_1(\widehat{G}_n, G) \geq E_{p_{G_0}} W_1(\widehat{G}_n, G_0) + E_{p_{G_0'}} W_1(\widehat{G}_n, G_0'),$$

where $E_{p_{G_0}}$ (resp. $E_{p_{G_0'}}$) denotes the expectation taken with respect to the product measure with density $p_{G_0}^n$ ($p_{G_0'}^n$). By the triangle inequality, $W_1(\widehat{G}_n, G_0) + W_1(\widehat{G}_n, G_0') \geq W_1(G_0, G_0') = 2\varepsilon$. Thus,

$$E_{p_{G_0}} W_1(\widehat{G}_n, G_0) + E_{p_{G_0'}} W_1(\widehat{G}_n, G_0') \geq 2\varepsilon \inf_{f_1, f_2} (E_{p_{G_0}} f_1 + E_{p_{G_0'}} f_2),$$

where the infimum is taken over all measurable nonnegative functions f_1 and f_2 defined in terms of n arguments $X_1, \ldots, X_n$, subject to the constraint that $f_1 + f_2 = 1$. From the definition of the variational distance, the infimum value in the above expression is equal to $(1 - V(p^n_{G_0}, p^n_{G_0'}))$. Hence,

$$\max_{G \in \{G_0, G_0'\}} E_{p_G} W_1(\widehat{G}_n, G) \geq \varepsilon\big(1 - V\big(p^n_{G_0}, p^n_{G_0'}\big)\big).$$

Now, due to the general relationship between variational distance and Hellinger distance, that is, $V \leq h$, and by our construction that $h(p_{G_0}, p_{G_0'}) \leq C_0 \varepsilon^r$, we have

$$\begin{aligned} V\big(p^n_{G_0}, p^n_{G_0'}\big) &\leq h\big(p^n_{G_0}, p^n_{G_0'}\big) \\ &= \sqrt{1 - \big(1 - h^2(p_{G_0}, p_{G_0'})\big)^n} \\ &\leq \sqrt{1 - \big(1 - C_0^2 \varepsilon^{2r}\big)^n}. \end{aligned}$$

As a result,

$$\max_{G \in \{G_0, G_0'\}} E_{p_G} W_1(\widehat{G}_n, G) \geq \varepsilon\Big(1 - \sqrt{1 - \big(1 - C_0^2 \varepsilon^{2r}\big)^n}\Big).$$

By choosing $\varepsilon^{2r} = \frac{1}{C_0^2 n}$, the right-hand side of the above inequality is bounded below by $c_1 \varepsilon \asymp n^{-1/2r}$ for any $r < \overline{r}$ where c_1 is some positive universal constant. Noting that $G_0, G_0' \in \mathcal{O}_k \setminus \mathcal{O}_{k_0 - 1}$, this completes the proof for part (a).

(b) The proof follows from combining the result of part (b) of Proposition 2.2 with a standard result on convergence of density estimation via MLE, from [28]. To draw from the latter, we first recall some additional standard notation from the empirical process theory literature (which after this proof will unfortunately not be needed for the rest of the paper). Let $\Theta^* = \Theta \times \Omega$, $\mathcal{P}_k(\Theta^*) = \{p_G | G \in \mathcal{O}_k\}$. Let $N(\varepsilon, \mathcal{P}_k(\Theta^*), \|\cdot\|_\infty)$ denote the covering number of the metric space $(\mathcal{P}_k(\Theta^*), \|\cdot\|_\infty)$, and $H_B(\varepsilon, \mathcal{P}_k(\Theta^*), h)$ the bracketing entropy of $\mathcal{P}_k(\Theta^*)$ under Hellinger distance metric h. Put $\overline{P}_k(\Theta^*) = \{p_{(G+G_0)/2} : G \in \mathcal{O}_k\}$ and $\overline{\mathcal{P}}_k^{1/2}(\Theta^*) = \{f^{1/2} | f \in \overline{\mathcal{P}}_k(\Theta^*)\}$. For any $\delta > 0$, denote the intersection of a Hellinger ball centered at p_{G_0} and $\overline{\mathcal{P}}_k^{1/2}(\Theta^*)$ as

$$\overline{\mathcal{P}}_k^{1/2}\big(\Theta^*, \delta\big) = \big\{f^{1/2} \in \overline{\mathcal{P}}_k^{1/2}\big(\Theta^*\big) | h(f, p_{G_0}) \leq \delta\big\}.$$

The size of this set is captured by the entropy integral

$$\mathcal{J}_B\big(\delta, \overline{\mathcal{P}}_k^{1/2}\big(\Theta^*, \delta\big), \mu\big) = \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}\big(u, \overline{\mathcal{P}}_k^{1/2}\big(\Theta^*, u\big), \mu\big)\, \mathrm{d}u \vee \delta,$$

where μ denotes Lebesgue measure. Since $\overline{\mathcal{P}}_k^{1/2}(\Theta^*, u) \subset \overline{\mathcal{P}}_k^{1/2}(\Theta^*)$, for any $u > 0$,

$$H_B\big(u, \overline{\mathcal{P}}_k^{1/2}(\Theta^*, u), L_2(\mu)\big) \le H_B\big(u, \overline{\mathcal{P}}_k^{1/2}(\Theta^*), L_2(\mu)\big) = H_B\big(u/\sqrt{2}, \overline{\mathcal{P}}_k(\Theta^*), h\big), \tag{3}$$

where the identity is immediate from relationship between the Hellinger distance metric and $L_2(\mu)$.

Note that for any two mixing measures G_1, G_0, $p_{(G_1+G_0)/2} = (p_{G_1} + p_{G_0})/2$. Note also the fact that for any probability densities f_0, f_1, f_2 defined on the same space, $h^2((f_1 + f_0)/2, (f_2 + f_0)/2) \le h^2(f_1, f_2)/2$ (cf. Lemma 4.2 [28]). So, for any two mixing measures $G_1, G_2 \in \mathcal{O}_k$, we have

$$h^2(p_{(G_1+G_0)/2}, p_{(G_2+G_0)/2}) \le h^2(p_{G_1}, p_{G_2})/2.$$

This inequality yields $H_B(u/\sqrt{2}, \overline{\mathcal{P}}_k(\Theta^*), h) \le H_B(u, \mathcal{P}_k(\Theta^*), h)$. Combining with equation (3) we to obtain

$$H_B\big(u, \overline{\mathcal{P}}_k^{1/2}(\Theta^*, u), L_2(\mu)\big) \le H_B\big(u, \mathcal{P}_k(\Theta^*), h\big).$$

This inequality allows us to obtain an upper bound of the LHS in terms of a bound on the RHS. Specifically, we need the following.

LEMMA 2.1. *Suppose that* $\Theta^* = [-a, a]^d \times \Omega$, *where* Ω *is a subset of* S_d^{++} *whose eigenvalues are bounded in an interval* $[\underline{\lambda}, \overline{\lambda}]$, $a \le L(\log(1/\varepsilon))^\gamma$, $\gamma \ge 1/2$, $L > 0$. *Then for* $0 < \varepsilon < 1/2$,

$$\log N\big(\varepsilon, \mathcal{P}_k(\Theta^*), \|\cdot\|_\infty\big) \lesssim \log(1/\varepsilon), \tag{4}$$

$$H_B\big(\varepsilon, \mathcal{P}_k(\Theta^*), h\big) \lesssim \log(1/\varepsilon). \tag{5}$$

The proof of this lemma is an extension of the arguments in [14] to multivariate setting, and is deferred to [16]. Now we choose $L > 0$ and $\gamma_1 = \max\{1/2, \gamma\} \ge 1/2$ such that $a_n \le L(\log(n))^{\gamma_1}$. From Lemma 2.1, as long as $0 < u < 1/2$, we have

$$H_B\big(u, \overline{\mathcal{P}}_k^{1/2}(\Theta^*, u), L_2(\mu)\big) \le H_B\big(u, \mathcal{P}_k(\Theta^*), h\big) \lesssim \log(1/u). \tag{6}$$

Now we state the result of Theorem 7.4 of [28] adapted to the notation used in our paper.

THEOREM 2.1. *Take* $\Psi(\delta) \ge \mathcal{J}_B(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta^*, \delta), \mu)$ *in such a way that* $\Psi(\delta)/\delta^2$ *is a nonincreasing function of* δ. *Then, for a universal constant* c *and for*

$$\sqrt{n}\delta_n^2 \ge c\Psi(\delta_n),$$

we have for all $\delta \ge \delta_n$

$$P\big(h(p_{\widehat{G}_n}, p_{G_0}) > \delta\big) \le c\exp\left[-\frac{n\delta^2}{c^2}\right].$$

Based on the bracket entropy bound in (6), we can choose $\Psi(\delta) = \delta[\log(1/\delta)]^{1/2}$ for $\delta > 0$. Therefore, by choosing $\delta_n = O(\log n/n)^{1/2}$, we obtain $P(h(p_{\widehat{G}_n}, p_{G_0}) > \delta_n) \lesssim \exp(-c\log(n))$, where constant $c > 0$ depends only on $L, \gamma, \underline{\lambda}, \overline{\lambda}$. Combining this probability bound with part (b) of Proposition 2.2 completes the proof.

3. Gamma mixtures and location extensions. The Gamma family of densities takes the form $f(x|a,b) := \frac{b^a}{\Gamma(a)} x^{a-1}\exp(-bx)$ for $x > 0$, and 0 otherwise, where a, b are positive shape and rate parameters, respectively. The Gamma family is not identifiable in the first order when *both* shape and rate parameters vary—this is to say that the collection of Gamma density functions and their partial derivatives up to the first order taken with respect to the shape and rate parameters are *not* linearly independent. This can be seen by the following identity:

$$\frac{\partial f}{\partial b}(x|a,b) = \frac{a}{b} f(x|a,b) - \frac{a}{b} f(x|a+1,b). \tag{7}$$

Examining the identity in the above display shows that the violation of linear independence of the collection of Gamma density functions and its derivatives is due to certain combinations of the Gamma parameter values. This suggests that outside of these value combinations the Gamma densities may well be identifiable in the first order and even the second order. This observation leads to a remarkable consequence for Gamma mixtures, which display wildly distinct behaviors in two disjoint categories of the parameter values, which we call "generic cases" and "pathological cases."

Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(a_i^0, b_i^0)} \in \mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta)$ where $k_0 \geq 2$ and $\Theta \subset \mathbb{R}_+^2$. Assume that $a_i^0 \geq 1$ for all $1 \leq i \leq k_0$. To delineate the structure underying parameter values of G_0, we define

(A.1) Generic cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \neq \{1, 0\}$ for all $1 \leq i, j \leq k_0$.

(A.2) Pathological cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} = \{1, 0\}$ for some $1 \leq i, j \leq k_0$.

We have the following result under the exact-fitted setting of Gamma mixtures. Let $\widehat{G}_n \in \mathcal{E}_{k_0}$ denote the MLE estimate of G_0.

THEOREM 3.1 (Exact-fitted Gamma mixtures). *Given* $\Theta = [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}]$ *where* $\underline{a} \geq 1, \overline{a}, \underline{b}, \overline{b}$ *are given positive numbers.*

(a) Generic cases. *If the support points of* G_0 *satisfy assumption* (A.1), *then* $\mathbb{P}(W_1(\widehat{G}_n, G_0) > \delta_n) \lesssim \exp(-c\log n)$, *where* δ_n *is sufficiently large multiple of* $(\log n/n)^{1/2}$ *and* c *is positive constant depending only on* $\underline{a}, \overline{a}, \underline{b}, \overline{b}$.

(b) Pathological cases. *For any* $r \geq 2$,

$$\inf_{\widehat{G}_n \in \mathcal{E}_{k_0}} \sup_{G \in \mathcal{E}_{k_0}} E_{p_G} W_r(\widehat{G}_n, G) \gtrsim n^{-1/r}.$$

While the result of part (a) may seem "obvious" due to the standard rate $(\log n/n)^{1/2}$, this should be put in the context of the minimax lower bound of part (b), which shows that one cannot estimate the Gamma parameters efficiently uniformly over a W_1 neighborhood of G_0, when we do not know whether G_0 is pathological or not. As can be seen in the proof, the poor rate is due to the difficulty of distinguishing between the pathological and generic instances—no polynomial rate estimation method is possible.

Turning to the over-fitted Gamma mixture setting, as before let $G_0 \in \mathcal{E}_{k_0}$, while G varies in a larger subset of $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ for some given $k \geq k_0 + 1$. We have the following categories regarding the true G_0:

(A.3) Generic cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \notin \{\{1,0\},\{2,0\}\}$ for all $1 \leq i, j \leq k_0$.

(A.4) Pathological cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \in \{\{1,0\},\{2,0\}\}$ for some $1 \leq i, j \leq k_0$.

Additionally, for any $c_0 > 0$ and $l \geq 1$, define the following constrained set of $\mathcal{O}_l$:

$$\mathcal{O}_{l,c_0} = \Bigg\{ G = \sum_{i=1}^{k'} p_i \delta_{(a_i,b_i)} \Big| k' \leq k \text{ and } |a_i - a_j^0| \notin [1-c_0, 1+c_0] \cup [2-c_0, 2+c_0] \ \forall (i,j) \Bigg\}.$$

THEOREM 3.2 (Over-fitted Gamma mixtures). *Assume the same conditions on Θ as that of Theorem 3.1.*

(a) Generic cases. *If $G_0 \in \mathcal{O}_{k,c_0}$ and let $\widehat{G}_n \in \mathcal{O}_{k,c_0}$ be the MLE estimation of G_0, then $\mathbb{P}(W_2(\widehat{G}_n, G_0) > \delta_n) \lesssim \exp(-c \log n)$, where δ_n is sufficiently large multiple of $(\log n/n)^{1/4}$ and c is positive constant depending only on $c_0, \underline{a}, \overline{a}, \underline{b}, \overline{b}$.*

Moreover, the following minimax bound holds, for any $2 \leq r < 4$:

$$\inf_{\widehat{G}_n \in \mathcal{O}_{k,c_0}} \sup_{G \in \mathcal{O}_{k,c_0} \setminus \mathcal{O}_{k_0-1}} E_{p_G} W_r(\widehat{G}_n, G) \gtrsim n^{-1/r}.$$

(b) Pathological cases. *For any $r \geq 2$,*

$$\inf_{\widehat{G}_n \in \mathcal{O}_k} \sup_{G \in \mathcal{O}_k \setminus \mathcal{O}_{k_0-1}} E_{p_G} W_r(\widehat{G}_n, G) \gtrsim n^{-1/r}.$$

Part (a) shows that in the over-fitted setting, if the true G_0 falls in the generic cases, then the standard MLE method restricted to a suitable subset of $\mathcal{O}_k$ still yields the $(\log n/n)^{1/4}$ rate of convergence for the mixing measure. Outside of this category, however, one cannot hope to estimate G at any polynomial rate of convergence.

Not all is bad news for Gamma mixtures: since the pathological cases represent a Lebesgue measure zero set, Gamma mixtures can be viewed as almost strongly identifiable with the strong convergence properties for the parameter estimation.

Exponential location extension. Lest the reader think that pathological cases are rare, we introduce a location extension of the exponential distribution, for which there is no such generic/pathological dichotomy. With this family, the convergence behavior of the mixing parameters is always slow, even when the number of mixing components is known. The class of location-exponential distribution $\{f(x|\theta,\sigma), \theta \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ is defined as $f(x|\theta,\sigma) = \frac{1}{\sigma}\exp(-\frac{x-\theta}{\sigma}) \cdot 1_{\{x>\theta\}}$ for $x \in \mathbb{R}$. Direct calculation yields that

$$\frac{\partial f}{\partial \theta}(x|\theta,\sigma) = \frac{1}{\sigma} f(x|\theta,\sigma) \qquad \text{when } x \neq \theta. \tag{8}$$

Since this identity holds in general, the linear independence of the kernel densities f and their partial derivatives is clearly violated regardless of the true values of G_0. We shall state a result for the exact-fitted setting only. Let $\Theta = [-a, a]$ and $\Omega = [\underline{\sigma}, \overline{\sigma}]$ where $a, \underline{\sigma}, \overline{\sigma}$ are fixed positive constants.

THEOREM 3.3 (Exact-fitted location-exponential mixtures). *For any* $r \geq 2$,

$$\inf_{\widehat{G}_n \in \mathcal{E}_{k_0}} \sup_{G \in \mathcal{E}_{k_0}} E_{p_G} W_1(\widehat{G}_n, G) \gtrsim n^{-1/r}.$$

This is quite a surprising bound, especially considering this is a finite mixture model with the known number of mixing components k_0. Yet, one cannot hope to achieve a polynomial estimation rate uniformly over a neighborhood (in W_1) of any mixing measure G_0. As in the pathological cases of Gamma mixtures, the poor convergence behavior of parameter estimation is due to the interaction of mixing parameters θ and σ, which is induced by the algebraic structures of f and its partial derivatives. As can be observed from the proof, the algebraic structure makes it difficult to distinguish between mixing measures G carrying similar mixture densities.

4. Simulations. We illustrate via simulations the rich spectrum of convergence behaviors for weak identifiable classes. Both identifiability bounds $h \geq V \gtrsim W_r^r$, and the convergence behavior of the MLE are examined.

Weak identifiability bounds. We experiment with classes of Gaussian densities. The results for mixtures of location-scale Gaussian distributions are given in Figure 1. Simulation details are as follows. The true mixing measure G_0 has exactly $k_0 = 2$ support points with locations $\theta_1^0 = -2$, $\theta_2^0 = 4$, scales $\sigma_1^0 = 1$, $\sigma_2^0 = 2$ and $p_1^0 = 1/3$, $p_2^0 = 2/3$. 5000 random samples of discrete mixing measures $G \in \mathcal{E}_2$, 5000 samples of $G \in \mathcal{O}_3$ and another 5000 for $G \in \mathcal{O}_4$, where the support points are uniformly generated in $\Theta = [-10, 10]$ and $\Omega = [0.5, 5]$. Additionally, to illustrate the best lower bound W_4^4 when we over-fit by one point, we also generate sequence G in accordance with the construction of sequence G in the proof of part

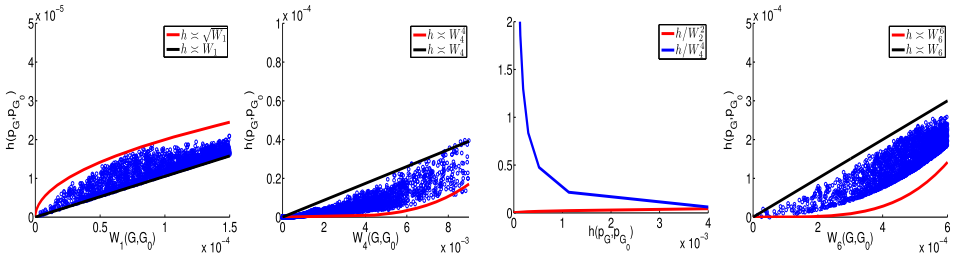

FIG. 1. *Location-scale Gaussian mixtures. From left to right*: (1) *Exact-fitted setting*; (2) *Over-fitted by one component*; (3) *Over-fitted by one component*; (4) *Over-fitted by two components*.

(a) of Proposition 2.2. The ratios h/W_2^2 and h/W_4^4 are plotted in the third panel of Figure 1 to verify that $h \gtrsim W_4^4$ holds, but $h \gtrsim W_2^2$ does not. It can be observed that both the lower bounds and upper bounds are in agreement with the theorems established earlier.

Convergence rates of MLE. First, we generate n-i.i.d. samples from a bivarite location-covariance Gaussian mixture with three components with an arbitrarily fixed choice of G_0. The true parameters for the mixing measure G_0 are: $\theta_1^0 = (0, 3), \theta_2^0 = (1, -4), \theta_3^0 = (5, 2)$, $\Sigma_1^0 = \begin{pmatrix} 4.2824 & 1.7324 \\ 1.7324 & 0.81759 \end{pmatrix}$, $\Sigma_2^0 = \begin{pmatrix} 1.75 & -1.25 \\ -1.25 & 1.75 \end{pmatrix}$, $\Sigma_3^0 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$, and $p_1^0 = 0.3$, $p_2^0 = 0.4$, $p_3^0 = 0.3$. MLE $\widehat{G}_n$ are obtained by the EM algorithm as we assume that the data come from a mixture of k Gaussians where $k \geq k_0 = 3$. See Figure 2 for a fixed choice of G_0. Wasserstein distances between $\widehat{G}_n$ and G_0 are plotted against increasing sample size n. The error bars were obtained by running the experiment 7 times for each n. These simulation results match quite well with the established rates and highlight that convergence slows down rapidly as $k - k_0$ increases.

We turn to mixtures of Gamma distributions. For generic cases, we generate n-i.i.d. samples from a Gamma mixture model that has exactly two mixing compo-

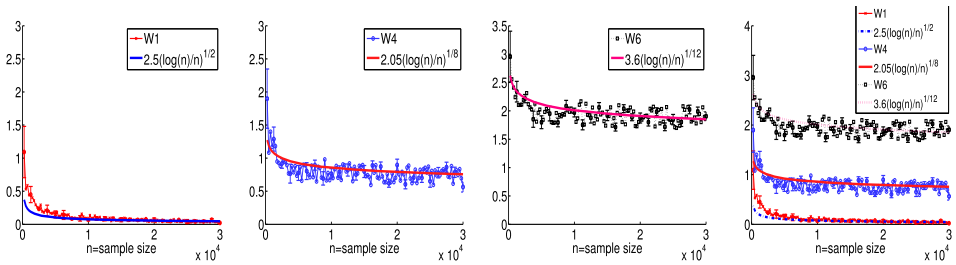

FIG. 2. *MLE rates for location-covariance mixtures of Gaussians. Left to right*: (1) *Exact-fitted*: $W_1 \asymp n^{-1/2}$. (2) *Over-fitted by one*: $W_4 \asymp n^{-1/8}$. (3) *Over-fitted by two*: $W_6 \asymp n^{-1/12}$. (4) *All previous plots are combined*.

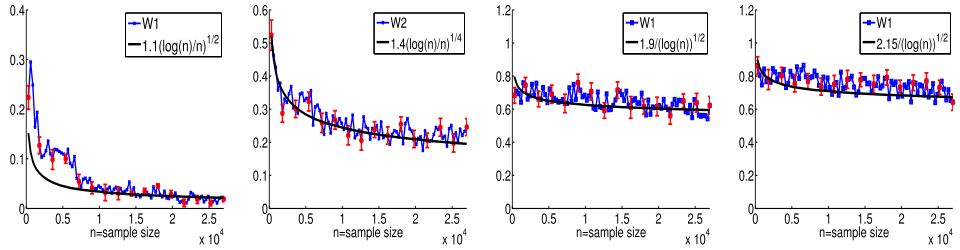

FIG. 3. *MLE rates for shape-rate mixtures of Gamma distributions. Left to right*: (1) *Generic/Exact-fitted*: $W_1(\widehat{G}_n, G_0) \asymp n^{-1/2}$. (2) *Generic/Over-fitted*: $W_2 \asymp n^{-1/4}$. (3) *Pathological/Exact-fitted*: $W_1 \approx 1/(\log n)^{1/2}$. (4) *Pathological/Over-fitted*: $W_1 \approx 1/(\log n)^{1/2}$.

nents. The true parameters for the mixing measure G_0 are: $a_1^0 = 8$, $a_2^0 = 2$, $b_1^0 = 3$, $b_2^0 = 4$, $p_1^0 = 1/3$, $p_2^0 = 2/3$. For pathological cases, everything else remains the same, except for our choice of G_0, for which we choose $a_1^0 = 8$, $a_2^0 = 7$, $b_1^0 = 3$, $b_2^0 = 3$, $p_1^0 = 1/3$, $p_2^0 = 2/3$.

It is remarkable to see the wild swing in behaviors within this same class. See Figure 3. Even for exact-fitted finite mixtures of Gamma, one can achieve very fast convergence rate of $n^{-1/2}$ in the generic case, or appear to be stagnant at a logarithmic rate if the true mixing measure G_0 belongs to the pathological category.

5. Proofs of other propositions and theorems.

5.1. *Proofs for over-fitted Gaussian mixtures.*

PROOF OF PROPOSITION 2.2. For the ease of exposition, we consider the setting of univariate location-scale Gaussian distributions, that is, both θ and $\Sigma = \sigma^2$ are scalars. The proof for general $d \geq 1$ is similar and omitted. Put $v = \sigma^2$, so we write $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, v_i^0)}$.

Step 1. For any sequence $G_n \in \mathcal{O}_k$, since k is finite, there is some $k^* \in [k_0, k]$ such that there exists a subsequence of G_n having exactly k^* support points. Denote $G_n = \sum_{i=1}^{k^*} p_i^n \delta_{(\theta_i^n, v_i^n)}$ (here, without loss of generality, we replace the whole sequence by its subsequence). Now if $G_n \to G_0$ in W_r, there exists a subsequence of G_n such that each support point (θ_i^0, σ_i^0) of G_0 is the limit of a subset of $s_i \geq 1$ support points of G_n. In general, there may also a subset of support points of G_n whose limits are not among the support points of G_0.

Note that with part (a), we shall construct one sequence of G_n to prove its conclusion. In our construction there are no constraints placed on p_i^n for all i. On the other hand, regarding part (b), we shall impose the constraint that $p_i^n \geq c_0$ for all i. Under this constraint, all the limit points of support points of G_n will be only those of G_0. To avoid notational cluttering, we replace the subsequence of G_n by

the whole sequence $\{G_n\}$. By re-labeling the support points, G_n can be expressed by

$$G_n = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{ij}^n, v_{ij}^n)}, \tag{9}$$

where $(\theta_{ij}^n, v_{ij}^n) \to (\theta_i^0, v_i^0)$, $\sum_{l=1}^{s_i} p_{il}^n \to p_i^0$ for all $i = 1, \ldots, k_0$ and $j = 1, \ldots, s_i$, where $s_1, \ldots, s_{k_0}$ are some natural constants less than k. All G_n have exactly the same $k^* = \sum s_i \leq k$ number of support points. This is the representation for G_n that we shall utilize in the proof of both part (a) and part (b).

Step 2. For any $x \in \mathbb{R}$,

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \big(f(x|\theta_{ij}^n, v_{ij}^n) - f(x|\theta_i^0, v_i^0)\big) \\ &\quad + \sum_{i=1}^{k_0} \big(p_{i\cdot}^n - p_i^0\big) f(x|\theta_i^0, v_i^0), \end{aligned}$$

where $p_{i\cdot}^n := \sum_{j=1}^{s_i} p_{ij}^n$. For any $r \geq 1$, integer $N \geq r$ and $x \in \mathbb{R}$, by means of Taylor's expansion up to the order N, we obtain

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \sum_{|\alpha|=1}^{N} \big(\Delta\theta_{ij}^n\big)^{\alpha_1} \big(\Delta v_{ij}^n\big)^{\alpha_2} \frac{D^{|\alpha|} f(x|\theta_i^0, v_i^0)}{\alpha!} \\ &\quad + A_1(x) + R_1(x). \end{aligned} \tag{10}$$

Here, $\alpha = (\alpha_1, \alpha_2)$, $|\alpha| = \alpha_1 + \alpha_2$, $\alpha! = \alpha_1!\alpha_2!$, $\Delta\theta_{ij}^n = \theta_{ij}^n - \theta_i^0$, $\Delta v_{ij}^n = v_{ij}^n - v_i^0$. Additionally, $A_1(x) = \sum_{i=1}^{k_0} (p_{i\cdot}^n - p_i^0) f(x|\theta_i^0, v_i^0)$, and $R_1(x) = O(\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta\theta_{ij}^n|^{N+\delta} + |\Delta v_{ij}^n|^{N+\delta}))$ for some positive constant $\delta > 0$.

Step 3. Enter the key identity (2): $\frac{\partial^2 f}{\partial \theta^2}(x|\theta, v) = 2\frac{\partial f}{\partial v}(x|\theta, v)$ for all x. This entails, for any natural orders n_1, n_2, that $\frac{\partial^{n_1+n_2} f}{\partial \theta^{n_1} \partial v^{n_2}}(x|\theta, v) = \frac{1}{2^{n_2}} \frac{\partial^{n_1+2n_2} f}{\partial \theta^{n_1+2n_2}}(x|\theta, v)$. Thus, by converting all derivatives to those taken with respect to only θ, we may rewrite (10) as

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \sum_{\alpha \geq 1} \sum_{n_1, n_2} \frac{(\Delta\theta_{ij}^n)^{n_1} (\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_i^0, v_i^0) \\ &\quad + A_1(x) + R_1(x) \\ &:= A_1(x) + B_1(x) + R_1(x), \end{aligned} \tag{11}$$

where n_1, n_2 in the sum satisfy $n_1 + 2n_2 = \alpha$, $n_1 + n_2 \leq N$.

Step 4. We proceed to proving part (a) of the proposition. From the definition of $\overline{r}$, by setting $r = \overline{r} - 1$, there exist nontrivial solutions $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ for the

system of equations (1). Construct a sequence of probability measures $G_n \in \mathcal{O}_k$ under the representation given by equation (9) as follows:

$$\theta_{1j}^n = \theta_1^0 + \frac{a_j^*}{n}, \qquad v_{1j}^n = v_1^0 + \frac{2b_j^*}{n^2}, \qquad p_{1j}^n = \frac{p_1^0 (c_j^*)^2}{\sum_{j=1}^{k-k_0+1} (c_j^*)^2},$$
$$\text{for all } j = 1, \ldots, k - k_0 + 1,$$

and $\theta_{i1}^n = \theta_i^0$, $v_{i1}^n = v_i^0$, $p_{i1}^n = p_i^0$ for all $i = 2, \ldots, k_0$. (I.e., we set $k^* = k$, $s_1 = k - k_0 + 1$, $s_i = 1$ for all $2 \leq i \leq k_0$.) Note that b_j^* may be negative, but we are guaranteed that $v_{1j}^n > 0$ for sufficiently large n. It is easy to verify that $W_1(G_n, G_0) = \sum_{i=1}^{k-k_0+1} p_{1i}^n (\frac{|a_i^*|}{n} + \frac{2|b_i^*|}{n^2}) \asymp \frac{1}{n}$, because at least one of the a_i^* is non-zero.

Step 5. Select $N = \overline{r}$ in equation (11). By our construction of G_n, clearly $A_1(x) = 0$. Moreover,

$$\begin{aligned} B_1(x) &= \sum_{i=1}^{k-k_0+1} p_{1i}^n \sum_{\alpha=1}^{\overline{r}-1} \sum_{n_1,n_2} \frac{(\Delta\theta_{1i}^n)^{n_1} (\Delta v_{1i}^n)^{n_2}}{2^{n_2} n_1! n_2!} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_1^0, v_1^0) \\ &\quad + \sum_{i=1}^{k-k_0+1} p_{1i}^n \sum_{\alpha=\overline{r}}^{2\overline{r}} \sum_{n_1,n_2} \frac{(\Delta\theta_{1i}^n)^{n_1} (\Delta v_{1i}^n)^{n_2}}{2^{n_2} n_1! n_2!} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_1^0, v_1^0) \\ &:= \sum_{\alpha=1}^{\overline{r}-1} B_{\alpha n} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_1^0, v_1^0) + \sum_{\alpha \geq \overline{r}} C_{\alpha n} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_1^0, v_1^0). \end{aligned}$$

In the above display, for each $\alpha \geq \overline{r}$, observe that $C_{\alpha n} = O(n^{-\alpha})$. Moreover, for each $1 \leq \alpha \leq \overline{r} - 1$,

$$B_{\alpha n} = \frac{1}{n^\alpha \sum_{i=1}^{k-k_0+1} (c_i^*)^2} \sum_{i=1}^{k-k_0+1} (c_i^*)^2 \sum_{n_1+2n_2=\alpha} \frac{(a_i^*)^{n_1} (b_i^*)^{n_2}}{n_1! n_2!} = 0,$$

because $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ form a nontrivial solution to system (1).

Step 6. We arrive at an upper bound for the Hellinger distance of mixture densities:

$$\begin{aligned} h^2(p_{G_n}, p_{G_0}) &\leq \frac{1}{2p_1^0} \int_{\mathbb{R}} \frac{(p_{G_n}(x) - p_{G_0}(x))^2}{f(x|\theta_1^0, v_1^0)} \, \mathrm{d}x \\ &\lesssim \int_{\mathbb{R}} \left(\left(\sum_{\alpha=\overline{r}}^{2\overline{r}} C_{\alpha n}^2 \left(\frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_1^0, v_1^0) \right)^2 + R_1^2(x) \right) \Big/ f(x|\theta_1^0, v_1^0) \right) \mathrm{d}x. \end{aligned}$$

For Gaussian densities, it can be verified that $(\frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_1^0, v_1^0))^2 / f(x|\theta_1^0, v_1^0)$ is integrable for all $1 \leq \alpha \leq 2\overline{r}$. So, $h^2(p_{G_n}, p_{G_0}) \leq O(n^{-2\overline{r}}) + \int R_1^2(x)/f(x|\theta_1^0, v_1^0) \, \mathrm{d}x$.

Turning to the Taylor remainder $R_1(x)$, note that

$$|R_1(x)| \lesssim \sum_{i=1}^{k-k_0+1} \sum_{|\beta|=\overline{r}+1} \frac{(\overline{r}+1)}{\beta!} |\Delta\theta_{1i}^n|^{\beta_1} |\Delta v_{1i}^n|^{\beta_2} \times \int_0^1 (1-t)^{\overline{r}} \left| \frac{\partial^{\overline{r}+1} f}{\partial\theta^{\beta_1}\, \partial v^{\beta_2}} \big(x|\theta_1^0 + t\Delta\theta_{1i}^n, v_1^0 + t\Delta v_{1i}^n\big) \right| \mathrm{d}t.$$

Now, $(\Delta\theta_{1i}^n)^{\beta_1}(\Delta v_{1i}^n)^{\beta_2} \asymp n^{-\beta_1-2\beta_2} = o(n^{-2\overline{r}})$. In addition, as n is sufficiently large, we have for all $|\beta| = \overline{r}+1$ that

$$\sup_{t\in[0,1]} \int_{x\in\mathbb{R}} \left(\frac{\partial^{\overline{r}+1} f}{\partial\theta^{\beta_1}\, \partial v^{\beta_2}} \big(x|\theta_1^0 + t\Delta\theta_{1i}^n, v_1^0 + t\Delta v_{1i}^n\big) \right)^2 \Big/ f\big(x|\theta_1^0, v_1^0\big)\, \mathrm{d}x < \infty.$$

It follows that $h(p_{G_n}, p_{G_0}) = O(n^{-\overline{r}})$. As noted above, $W_1(G_n, G_0) \asymp n^{-1}$, so the claim of part (a) is established.

Step 7. Turning to part (b) of Proposition 2.2, it suffices to show that

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{O}_{k,c_0}} \Big\{ \sup_{x\in\mathcal{X}} \big|p_G(x) - p_{G_0}(x)\big| / W_{\overline{r}}^{\overline{r}}(G, G_0) : W_{\overline{r}}(G, G_0) \le \varepsilon \Big\} > 0. \tag{12}$$

Then one can arrive at the proposition's claim by passing through an argument using Fatou's lemma (cf. proof of Theorem 1 of [23] or step 4 in the proof of Theorem 3.1 of [17]). Suppose that (12) does not hold. Then we can find a sequence of probability measures $G_n \in \mathcal{O}_{k,c_0}$ that are represented by equation (9), such that $W_{\overline{r}}^{\overline{r}}(G_n, G_0) \to 0$ and $\sup_x |p_{G_n}(x) - p_{G_0}(x)| / W_{\overline{r}}^{\overline{r}}(G_n, G_0) \to 0$. Define

$$D_n := d(G_n, G_0) := \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \big(|\Delta\theta_{ij}^n|^{\overline{r}} + |\Delta v_{ij}^n|^{\overline{r}}\big) + \sum_{i=1}^{k_0} |p_{i\cdot}^n - p_i^0|.$$

It is easy to see that $W_{\overline{r}}^{\overline{r}}(G_n, G_0) \lesssim D_n$, since D_n is the multiple of the $W_{\overline{r}}^{\overline{r}}$ cost of moving mass from G_n to G_0 by a (possibly) nonoptimal coupling. So, for all $x \in \mathbb{R}$, $(p_{G_n}(x) - p_{G_0}(x))/D_n \to 0$. Combining this fact with (11), where $N = \overline{r}$, we obtain

$$\big(A_1(x) + B_1(x) + R_1(x)\big)/D_n \to 0. \tag{13}$$

We have $R_1(x)/D_n = o(1)$ as $n \to \infty$.

Step 8. $A_1(x)/D_n$ and $B_1(x)/D_n$ are the linear combination of elements of $\frac{\partial^\alpha f}{\partial\theta^\alpha}(x|\theta, v)$ where $\alpha = n_1 + 2n_2$ and $n_1 + n_2 \le \overline{r}$. Note that the natural order α ranges in $[0, 2\overline{r}]$. Let $E_\alpha(\theta, v)$ denote the corresponding coefficient of $\frac{\partial^\alpha f}{\partial\theta^\alpha}(x|\theta, v)$. Extracting from (11), for $\alpha = 0$, $E_0(\theta_i^0, v_i^0) = (p_{i\cdot}^n - p_i^0)/D_n$. For $\alpha \ge 1$,

$$E_\alpha(\theta_i^0, v_i^0) = \left[\sum_{j=1}^{s_i} p_{ij}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2\le\overline{r}}} \frac{(\Delta\theta_{ij}^n)^{n_1} (\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \right] \Big/ D_n.$$

In the remainder of this proof step, we shall show that as $n \to \infty$, at least one of the coefficients $E_\alpha(\theta_i^0, v_i^0)$ must not vanish. Suppose this is not the case, that is, $E_\alpha(\theta_i^0, v_i^0) \to 0$ for all $i = 1, \ldots, k_0$ and $0 \le \alpha \le 2\overline{r}$ as $n \to \infty$. By taking the summation of all $|E_0(\theta_i^0, v_i^0)|$, we get $\sum_{i=1}^{k_0} |p_{i\cdot}^n - p_i^0|/D_n \to 0$ as $n \to \infty$. As a consequence, we obtain

$$\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \left(|\Delta\theta_{ij}^n|^{\overline{r}} + |\Delta v_{ij}^n|^{\overline{r}}\right)/D_n \to 1 \qquad \text{as } n \to \infty.$$

Hence, we can find an index $i^* \in \{1, 2, \ldots, k_0\}$ such that as $n \to \infty$

$$\sum_{j=1}^{s_{i^*}} p_{i^*j}^n \left(|\Delta\theta_{i^*j}^n|^{\overline{r}} + |\Delta v_{i^*j}^n|^{\overline{r}}\right)/D_n \not\to 0.$$

Without loss of generality, we assume that $i^* = 1$. Accordingly,

$$\begin{aligned} F_\alpha(\theta_1^0, v_1^0) &:= \frac{D_n E_\alpha(\theta_1^0, \sigma_1^0)}{\sum_{j=1}^{s_1} p_{1j}^n (|\Delta\theta_{1j}^n|^{\overline{r}} + |\Delta v_{1j}^n|^{\overline{r}})} \\ &= \left(\sum_{j=1}^{s_1} p_{1j}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \le \overline{r}}} \frac{(\Delta\theta_{1j}^n)^{n_1} (\Delta v_{1j}^n)^{n_2}}{2^{n_2} n_1! n_2!}\right) \\ &\quad \Big/ \left(\sum_{j=1}^{s_1} p_{1j}^n \left(|\Delta\theta_{1j}^n|^{\overline{r}} + |\Delta v_{1j}^n|^{\overline{r}}\right)\right) \\ &\to 0. \end{aligned}$$

If $s_1 = 1$ then $F_1(\theta_1^0, v_1^0)$ and $F_{2\overline{r}}(\theta_1^0, v_1^0)$ yield

$$|\Delta\theta_{11}^n|^{\overline{r}}/\left(|\Delta\theta_{11}^n|^{\overline{r}} + |\Delta v_{11}^n|^{\overline{r}}\right), |\Delta v_{11}^n|^{\overline{r}}/\left(|\Delta\theta_{11}^n|^{\overline{r}} + |\Delta v_{11}^n|^{\overline{r}}\right) \to 0,$$

which is a contradiction. As a consequence, $s_1 \ge 2$.

Denote $\overline{p}_n = \max_{1 \le j \le s_1} \{p_{1j}^n\}$, $\overline{M}_n = \max\{|\Delta\theta_{11}^n|, \ldots, |\Delta\theta_{1s_1}^n|, |\Delta v_{11}^n|^{1/2}, \ldots, |\Delta v_{1s_1}^n|^{1/2}\}$. Since $0 < p_{1j}^n/\overline{p}_n \le 1$ for all $1 \le j \le s_1$, by a subsequence argument, there exist $c_j^2 := \lim_{n\to\infty} p_{1j}^n/\overline{p}_n$ for all $j = 1, \ldots, s_1$. Similarly, define $a_j := \lim_{n\to\infty} \Delta\theta_{1j}^n/\overline{M}_n$, and $2b_j := \lim_{n\to\infty} \Delta v_{1j}^n/\overline{M}_n^2$ for each $j = 1, \ldots, s_1$. By the constraints of $\mathcal{O}_{k,c_0}$, $p_{1j}^n \ge c_0$, so all of c_j^2 differ from 0 and at least one of them equals to 1. Likewise, at least one element of $(a_j, b_j)_{j=1}^{s_1}$ equal to -1 or 1. Now, for each $\alpha = 1, \ldots, \overline{r}$, divide both the numerator and denominator of $F_\alpha(\theta_1^0, v_1^0)$ by $\overline{p}_n$ and then $\overline{M}_n^\alpha$ and let $n \to \infty$, we obtain the following system of polynomial equations:

$$\sum_{j=1}^{s_1} \sum_{n_1+2n_2=\alpha} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \qquad \text{for each } \alpha = 1, \ldots, \overline{r}.$$

Since $s_1 \geq 2$, we get $\overline{r} \geq 4$. If $a_i = 0$ for all $1 \leq i \leq s_1$ then by choosing $\alpha = 4$, we obtain $\sum_{j=1}^{s_1} c_j^2 b_j^2 = 0$. However, it demonstrates that $b_i = 0$ for all $1 \leq i \leq s_1$—a contradiction to the fact that at least one element of $(a_i, b_i)_{i=1}^{s_1}$ is different from 0. Therefore, at least one element of $(a_i)_{i=1}^{s_1}$ is not equal to 0. Observe that $s_i \leq k - k_0 + 1$ (because the number of distinct atoms of G_n is $\sum_{i=1}^{k_0} s_i \leq k$ and all $s_i \geq 1$). Thus, the existence of nontrivial solutions for the system of equations given in the above display entails the existence of nontrivial solutions for system of equations (1). This contradicts with the definition of $\overline{r}$. Therefore, our hypothesis that all coefficients $E_\alpha(\theta_i^0, v_i^0)$ vanish does not hold—there must be at least one coefficient which does not converge to 0 as $n \to \infty$.

Step 9. Let m_n be the maximum of the absolute values of $E_\alpha(\theta_i^0, v_i^0)$ where $0 \leq \alpha \leq 2\overline{r}$, $1 \leq i \leq k_0$ and $d_n = 1/m_n$. Since $m_n \not\to 0$ as $n \to \infty$, d_n is uniformly bounded above for all n. As $d_n |E_\alpha(\theta_i^0, v_i^0)| \leq 1$, we have $d_n E_\alpha(\theta_i^0, v_i^0) \to \beta_{i\alpha}$ for all $0 \leq \alpha \leq 2\overline{r}$, $1 \leq i \leq k_0$ where at least one of $\beta_{i\alpha}$ differs from 0. Incorporating these limits to equation (13), we obtain that for all $x \in \mathbb{R}$,

$$\bigl(p_{G_n}(x) - p_{G_0}(x)\bigr)/D_n \to \sum_{i=1}^{k_0} \sum_{\alpha=0}^{2\overline{r}} \beta_{i\alpha} \frac{\partial^\alpha f}{\partial \theta^\alpha}\bigl(x|\theta_i^0, v_i^0\bigr) = 0.$$

By direct calculation, we can rewrite the above equation as

$$\sum_{i=1}^{k_0} \Biggl(\sum_{j=1}^{2\overline{r}+1} \gamma_{ij} \bigl(x - \theta_i^0\bigr)^{j-1} \Biggr) \exp\Biggl(-\frac{(x - \theta_i^0)^2}{2v_i^0} \Biggr) = 0 \qquad \text{for all } x \in \mathbb{R}, \tag{14}$$

where γ_{ij} for odd j are linear combinations of $\beta_{i(2l_1)}$, for $(j-1)/2 \leq l_1 \leq \overline{r}$, such that all of the coefficients are functions of v_i^0 differing from 0. For even j, γ_{ij} are linear combinations of $\beta_{i(2l_2+1)}$, for $j/2 \leq l_2 \leq \overline{r}$, such that all of the coefficients are functions of v_i^0 differing from 0. Now, without loss of generality, we assume that $v_1^0 \leq v_2^0 \leq \cdots \leq v_{k_0}^0$. Denote $\overline{i} \in [1, k_0]$ to be the minimum index i such that $v_i^0 = v_{k_0}^0$. It implies that $v_{\overline{i}}^0 = v_{\overline{i}+1}^0 = \cdots = v_{k_0}^0$. Therefore, θ_i^0 are pairwise different as $\overline{i} \leq i \leq k_0$. Now, let call $\underline{i} = \arg\max_{\overline{i} \leq i \leq k_0} \theta_i^0$. Multiply both sides of (14) with $\exp[(x - \theta_{\underline{i}}^0)^2/2v_{\underline{i}}^0]$ and let $x \to +\infty$, then we can check that

$$\sum_{j=1}^{2\overline{r}+1} \gamma_{\underline{i}j} \bigl(x - \theta_{\underline{i}}^0\bigr)^{j-1} \to 0,$$

which only happens when $\gamma_{\underline{i}j} = 0$ for all $1 \leq j \leq 2\overline{r} + 1$. Employing the same argument to the remained indices, we obtain $\gamma_{ij} = 0$ for all $i = 1, \ldots, k_0$, $j = 1, \ldots, 2\overline{r} + 1$. This entails that $\beta_{i\alpha} = 0$ for all $i = 1, \ldots, k_0$, $\alpha = 0, \ldots, 2\overline{r}$—a contradiction. Thus, we achieve the conclusion of (12). □

PROOF OF PROPOSITION 2.1. Our proof is based on the Groebner basis method for determining solutions for a system of polynomial equations. (i) For

the case $k - k_0 = 1$, the system (1) when $r = 4$ can be written as

$$c_1^2 a_1 + c_2^2 a_2 = 0, \tag{15}$$

$$\tfrac{1}{2}(c_1^2 a_1^2 + c_2^2 a_2^2) + c_1^2 b_1 + c_2^2 b_2 = 0, \tag{16}$$

$$\tfrac{1}{3!}(c_1^2 a_1^3 + c_2^2 a_2^3) + c_1^2 a_1 b_1 + c_2^2 a_2 b_2 = 0, \tag{17}$$

$$\tfrac{1}{4!}(c_1^2 a_1^4 + c_2^2 a_2^4) + \tfrac{1}{2!}(c_1^2 a_1^2 b_1 + c_2^2 a_2^2 b_2) + \tfrac{1}{2!}(c_1^2 b_1^2 + c_2^2 b_2^2) = 0. \tag{18}$$

Suppose that the above system has a nontrivial solution. If $c_1 a_1 = 0$, then equation (15) implies $c_2 a_2 = 0$. Since $c_1, c_2 \neq 0$, we have $a_1 = a_2 = 0$. This violates the constraint that one of a_1, a_2 is non-zero. Hence, $c_1 a_1, c_2 a_2 \neq 0$. Divide both sides of (15), (16), (17), (18) by $c_1^2 a_1$, $c_1^2 a_1^2$, $c_1^2 a_1^3$, $c_1^2 a_1^4$, respectively, we obtain the following system of polynomial equations:

$$\begin{aligned}
&1 + x^2 a = 0,\\
&1 + x^2 a^2 + 2(b + x^2 c) = 0,\\
&1 + x^2 a^3 + 6(b + x^2 ac) = 0,\\
&1 + x^2 a^4 + 12(b + x^2 a^2 c) + 12(b^2 + x^2 c^2) = 0,
\end{aligned}$$

where $x = c_2/c_1$, $a = a_2/a_1$, $b = b_1/a_1$, $c = b_2/a_1$. By taking the lexicographical order $a \succ b \succ c \succ x$, the Groebner basis of the above system contains $x^6 + 2x^4 + 2x^2 + 1 > 0$ for all $x \in \mathbb{R}$. Therefore, the above system of polynomial equations does not have real solutions. As a consequence, the original system of polynomial equations does not have a nontrivial solution, which means that $\overline{r} \leq 4$. However, we have already shown that as $r = 3$, equation (1) has a nontrivial solution. Therefore, $\overline{r} = 4$.

(ii) The case $k - k_0 = 2$. System (1) when $r = 6$ takes the form

$$\sum_{i=1}^{3} c_i^2 a_i = 0, \tag{19}$$

$$\frac{1}{2}\sum_{i=1}^{3} c_i^2 a_i^2 + \sum_{i=1}^{3} c_i^2 b_i = 0, \tag{20}$$

$$\frac{1}{6}\sum_{i=1}^{3} c_i^2 a_i^3 + \frac{1}{2}\sum_{i=1}^{3} c_i^2 a_i b_i = 0, \tag{21}$$

$$\frac{1}{24}\sum_{i=1}^{3} c_i^2 a_i^4 + \frac{1}{2}\sum_{i=1}^{3} c_i^2 a_i^2 b_i + \frac{1}{2}\sum_{i=1}^{3} c_i^2 b_i^2 = 0, \tag{22}$$

$$\frac{1}{120}\sum_{i=1}^{3} c_i^2 a_i^5 + \frac{1}{6}\sum_{i=1}^{3} c_i^2 a_i^3 b_i + \frac{1}{2}\sum_{i=1}^{3} c_i^2 a_i b_i^2 = 0, \tag{23}$$

$$\frac{1}{720}\sum_{i=1}^{3} c_i^2 a_i^6 + \frac{1}{24}\sum_{i=1}^{3} c_i^2 a_i^4 b_i + \frac{1}{4}\sum_{i=1}^{3} c_i^2 a_i^2 b_i^2 + \frac{1}{6}\sum_{i=1}^{3} c_i^2 b_i^3 = 0. \tag{24}$$

Nontrivial solution constraints require that $c_1, c_2, c_3 \neq 0$ and without loss of generality, $a_1 \neq 0$. Dividing both sides of the six equations above by $c_1^2 a_1$, $c_1^2 a_1^2$, $c_1^2 a_1^3$, $c_1^2 a_1^4$, $c_1^2 a_1^5$, $c_1^2 a_1^6$, respectively, we obtain

$$\begin{aligned}
&1 + x^2 a + y^2 b = 0,\\
&\tfrac{1}{2}(1 + x^2 a^2 + y^2 b^2) + c + x^2 d + y^2 e = 0,\\
&\tfrac{1}{3}(1 + x^2 a^3 + y^2 b^3) + c + x^2 ad + y^2 be = 0,\\
&\tfrac{1}{12}(1 + x^2 a^4 + y^2 b^4) + c + x^2 a^2 d + y^2 b^2 e + c^2 + x^2 d^2 + y^2 e^2 = 0,\\
&\tfrac{1}{60}(1 + x^2 a^5 + y^2 b^5) + \tfrac{1}{3}(c + x^2 a^3 d + y^2 b^3 e) + c^2 + x^2 a d^2 + y^2 b e^2 = 0,\\
&\tfrac{1}{360}(1 + x^2 a^6 + y^2 b^6) + \tfrac{1}{12}(c + x^2 a^4 d + y^2 b^4 e) + \tfrac{1}{2}(c^2 + x^2 a^3 d + y^2 b^3 e)\\
&\qquad + \tfrac{1}{3}(c^3 + x^2 d^3 + y^2 e^3) = 0,
\end{aligned}$$

where $x = c_2/c_1$, $y = c_3/c_1$, $a = a_2/a_1$, $b = a_3/a_1$, $c = b_1/a_1^2$, $d = b_2/a_1^2$, $e = b_3/a_1^2$. By taking the lexicographical order $a \succ b \succ c \succ d \succ x \succ y$, we can verify that the Groebner basis of the above system of polynomial equations contains a polynomial in terms of x^2, y^2 with all of the positive coefficient numbers, which cannot be 0 when $x, y \in \mathbb{R}$. Therefore, the original system of polynomial equations does not have a nontrivial solution. It follows that $\overline{r} \leq 6$.

When $r = 5$, we retain the first five equations in the system described in the above display. By choosing $x = y = 1$, under lexicographical order $a \succ b \succ c \succ d \succ e$, we can verify that the Groebner basis contains a polynomial of e with roots $e = \pm\sqrt{2}/3$ or $e = (-3 \pm \sqrt{2})/6$ while a, b, c, d can be uniquely determined by e. Thus, system of polynomial equations (1) has a nontrivial solution. It follows that $\overline{r} = 6$.

(iii) For the case $k - k_0 \geq 3$, we choose $c_1 = c_2 = \cdots = c_{k-k_0+1} = 1$, $a_i = b_i = 0$ for all $4 \leq i \leq k - k_0 + 1$. Additionally, take $a_1 = a_2 = 1$. Now, by choosing $r = 6$ in system (1), we can check by the Groebner basis that this system of polynomial equations has a nontrivial solution. As a result, $\overline{r} \geq 7$. $\square$

5.2. *Mixture of Gamma distributions and location-exponential distributions.*

PROOF OF THEOREM 3.1. The proof of this theorem proceeds in the same manner as that of Theorem 1.1. Therefore, it suffices to prove the following.

PROPOSITION 5.1 (Bounds for exact-fitted Gamma mixtures).

(a) (*Generic cases*) *Assume that the support points of* G_0 *satisfy assumption* (A.1). *Then for* $G \in \mathcal{E}_{k_0}$ *and* $W_1(G, G_0)$ *sufficiently small, we have*

$$V(p_G, p_{G_0}) \gtrsim W_1(G, G_0).$$

(b) (*Pathological cases*) *If the support points of* G_0 *satisfy assumption* (A.2), *then for any* $r \geq 1$

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{E}_{k_0}} \{V(p_G, p_{G_0}) / W_r^r(G, G_0) : W_r(G, G_0) \leq \varepsilon\} = 0.$$

PROOF. (a) For the range of generic parameter values of G_0, we shall show that the first-order identifiability still holds for Gamma mixtures, so that the conclusion can be drawn immediately from Theorem 3.1 of [17]. It suffices to show that for any $\alpha_{ij} \in \mathbb{R}$ $(1 \leq i \leq 3, 1 \leq j \leq k_0)$ such that for almost sure $x > 0$

$$\sum_{i=1}^{k_0} \alpha_{1i} f(x|a_i^0, b_i^0) + \alpha_{2i} \frac{\partial f}{\partial a}(x|a_i^0, b_i^0) + \alpha_{3i} \frac{\partial f}{\partial b}(x|a_i^0, b_i^0) = 0 \tag{25}$$

then $\alpha_{ij} = 0$ for all i, j. Equation (25) is rewritten as

$$\sum_{i=1}^{k_0} (\beta_{1i} x^{a_i^0 - 1} + \beta_{2i} (\log x) x^{a_i^0 - 1} + \beta_{3i} x^{a_i^0}) \exp(-b_i^0 x) = 0, \tag{26}$$

where $\beta_{1i} = \alpha_{1i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} + \alpha_{2i} \frac{(b_i^0)^{a_i^0} (\log(b_i^0) - \psi(a_i^0))}{\Gamma(a_i^0)} + \alpha_{3i} \frac{a_i^0 (b_i^0)^{a_i^0 - 1}}{\Gamma(a_i^0)}$, $\beta_{2i} = \alpha_{2i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$, and $\beta_{3i} = -\alpha_{3i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$. Without loss of generality, we assume that $b_1^0 \leq b_2^0 \leq \cdots \leq b_{k_0}^0$. Denote $\bar{i}$ to be the maximum index i such that $b_i^0 = b_1^0$. Then we have that $a_1^0, \ldots, a_{\bar{i}}^0$ are pairwise different. Multiply both sides of (26) with $\exp(b_{\bar{i}}^0 x)$ and let $x \to +\infty$, we obtain

$$\sum_{i=1}^{\bar{i}} \beta_{1i} x^{a_i^0 - 1} + \beta_{2i} (\log x) x^{a_i^0 - 1} + \beta_{3i} x^{a_i^0} \to 0.$$

Since $|a_i^0 - a_j^0| \neq 1$ and $a_i^0 \geq 1$ for all $1 \leq i, j \leq \bar{i}$, the above result implies that $\beta_{1i} = \beta_{2i} = \beta_{3i} = 0$ for all $1 \leq i \leq \bar{i}$ or equivalently $\alpha_{1i} = \alpha_{2i} = \alpha_{3i}$ for all $1 \leq i \leq \bar{i}$. Repeat the same argument for the remained indices, we obtain $\alpha_{1i} = \alpha_{2i} = \alpha_{3i} = 0$ for all $1 \leq i \leq k_0$. This completes the proof.

(b) Without loss of generality, we assume that $\{|a_2^0 - a_1^0|, |b_2^0 - b_1^0|\} = \{1, 0\}$. In particular, $b_1^0 = b_2^0$ and assume $a_2^0 = a_1^0 - 1$. We construct the following sequence of measures: $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{(a_i^n, b_i^n)}$, where $a_i^n = a_i^0$ for all $1 \leq i \leq k_0$, $b_1^n = b_1^0$, $b_2^n = b_1^0 (1 + \frac{1}{a_2^0 (n p_2^0 - 1)})$, $b_i^n = b_i^0$ for all $3 \leq i \leq k_0$, $p_1^n = p_1^0 + 1/n$, $p_2^n = p_2^0 - 1/n$, $p_i^n = p_i^0$ for all $3 \leq i \leq k_0$. We can check that $W_r^r(G_n, G_0) \asymp 1/n + (p_2^0 -$

$1/n)|b_2^n - b_1^0|^r \asymp n^{-1}$ as $n \to \infty$. For any natural order $r \geq 1$, by applying Taylor's expansion up to the $([r]+1)$th-order, we obtain

$$p_{G_n}(x) - p_{G_0}(x) = \sum_{i=1}^{k_0} p_i^n \big(f(x|a_i^n, b_i^n) - f(x|a_i^0, b_i^0)\big) + \big(p_i^n - p_i^0\big) f(x|a_i^0, b_i^0)$$

$$= \big(p_1^n - p_1^0\big) f(x|a_1^0, b_1^0) + \big(p_2^n - p_2^0\big) f(x|a_2^0, b_2^0) + \sum_{j=1}^{[r]+1} p_2^n \frac{(b_2^n - b_2^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) + R_n(x). \tag{27}$$

The Taylor expansion remainder $|R_n(x)| = O(p_2^n |b_2^n - b_2^0|^{[r]+1+\delta})$ for some $\delta > 0$ due to $a_2^0 \geq 1$. Therefore, $R_n(x) = o(W_r^r(G_n, G_0))$ as $n \to \infty$. For the choice of p_2^n, b_2^n, we can check that as $j \geq 2$, $p_2^n (b_2^n - b_2^0)^j = o(W_r^r(G_n, G_0))$. Now, we can rewrite (27) as

$$p_{G_n}(x) - p_{G_0}(x) = A_n x^{a_2^0} \exp\big(-b_1^0 x\big) + B_n x^{a_2^0 - 1} \exp\big(-b_1^0 x\big) + \sum_{j=2}^{[r]+1} p_2^n \frac{(b_2^n - b_2^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) + R_n(x),$$

where we have $A_n = \frac{(b_1^0)^{a_1^0}}{\Gamma(a_1^0)}(p_1^n - p_1^0) - \frac{(b_1^0)^{a_2^0}}{\Gamma(a_2^0)} p_2^n (b_2^n - b_1^0) = 0$ and similarly $B_n = \frac{(b_1)^{a_2^0}}{\Gamma(a_2^0)}(p_2^n - p_2^0) + \frac{a_2^0 (b_1^0)^{a_2^0 - 1}}{\Gamma(a_2^0)} p_2^n (b_2^n - b_1^0) = 0$ for all n. Since $a_2^0 \geq 1$, $|\frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0)|$ is bounded for all $2 \leq j \leq r+1$. It follows that $\sup_{x>0} |p_{G_n}(x) - p_{G_0}(x)| = O(n^{-2})$. Observe that

$$V(p_{G_n}, p_{G_0}) = 2 \int_{p_{G_n}(x) < p_{G_0}(x)} \big(p_{G_0}(x) - p_{G_n}(x)\big)\, \mathrm{d}(x) \leq 2 \int_{x \in (0, a_2^0/b_1^0)} \big|p_{G_n}(x) - p_{G_0}(x)\big|\, \mathrm{d}x.$$

As a consequence $V(p_{G_n}, p_{G_0}) = O(n^{-1/2})$ so for any $r \geq 1$, $V(p_{G_n}, p_{G_0}) = o(W_r^r(G_n, G_0))$ as $n \to \infty$. □ □

PROOF OF THEOREM 3.2. As in the proof of Theorem 3.1, it is sufficient to prove the following.

PROPOSITION 5.2 (Bounds for over-fitted Gamma mixtures).

(a) (*Generic cases.*) *Assume that we have* $G_0 \in \mathcal{O}_{k,c_0}$. *Then, for* $G \in \mathcal{O}_{k,c_0}$ *and* $W_2(G, G_0)$ *sufficiently small, we obtain*

$$V(p_G, p_{G_0}) \gtrsim W_2^2(G, G_0).$$

(b) (*Pathological cases.*) *Assume that the support points of* G_0 *satisfy assumption* (A.4), *then for any* $r \geq 1$,

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_k} \{V(p_G, p_{G_0}) / W_r^r(G, G_0) : W_r(G, G_0) \leq \varepsilon\} = 0.$$

PROOF. (a) As in step 7 in the proof of Proposition 2.2, it suffices to show that

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{O}_{k,c_0}} \left\{ \sup_{x \in \mathcal{X}} |p_G(x) - p_{G_0}(x)| / W_2^2(G, G_0) : W_2(G, G_0) \leq \varepsilon \right\} > 0. \tag{28}$$

Suppose this does not hold, by repeating the arguments of step 1 of Proposition 2.2, there is a sequence $G_n = \sum_{i=1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(a_{ij}^n, b_{ij}^n)} \to G_0 = \sum_{i=1}^{k_0+m} p_i^0 \delta_{(a_i^0, b_i^0)}$ such that $(a_{ij}^n, b_{ij}^n) \to (a_i^0, b_i^0)$ for all $1 \leq i \leq k_0 + m$ where (a_i^0, b_i^0) are limit points that lie outside the support points of G_0 as $k_0 + 1 \leq i \leq k_0 + m$. Additionally, $p_i^0 = 0$ as $k_0 + 1 \leq i \leq k_0 + m$. Invoke the Taylor expansion up to the second order and assume that all of the coefficients corresponding to the first and second derivatives with respect to the parameters go to 0. Use the same argument as that of step 8 in Proposition 2.2, and by summing up all the coefficients of second derivative, we obtain the contradiction. Now, by proceeding in the same way as that of step 9 in Proposition 2.2, as we let $n \to \infty$, we have for almost every x,

$$\begin{aligned}
\frac{p_{G_n}(x) - p_{G_0}(x)}{d(G_n, G_0)} \to & \sum_{i=1}^{k_0+m} \left\{ \alpha_{1i} f(x|a_i^0, b_i^0) + \alpha_{2i} \frac{\partial f}{\partial a}(x|a_i^0, b_i^0) \right. \\
& + \alpha_{3i} \frac{\partial f}{\partial b}(x|a_i^0, b_i^0) + \sum_{j=1}^{s_i} \alpha_{4ij}^2 \frac{\partial^2 f}{\partial a^2}(x|a_i^0, b_i^0) \\
& \left. + \sum_{j=1}^{s_i} \alpha_{5ij}^2 \frac{\partial^2 f}{\partial b^2}(x|a_i^0, b_i^0) + 2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij} \frac{\partial^2 f}{\partial a \, \partial b}(x|a_i^0, b_i^0) \right\} \\
= & \ 0,
\end{aligned}$$

where at least one of $\alpha_{1i}, \alpha_{2i}, \alpha_{3i}, \sum_{j=1}^{s_i} \alpha_{4ij}^2, \sum_{j=1}^{s_i} \alpha_{5ij}^2, 2\sum_{j=1}^{s_i} \alpha_{4ij}\alpha_{5ij}$ is nonzero. We can rewrite the above equation as

$$\begin{aligned}
& \sum_{i=1}^{k_0+m} \{\beta_{1i} x^{a_i^0 - 1} + \beta_{2i} x^{a_i^0} + \beta_{3i} x^{a_i^0 + 1} + \beta_{4i} (\log x) x^{a_i^0 - 1} \\
& \qquad + \beta_{5i} (\log x)^2 x^{a_i^0 - 1} + \beta_{6i} (\log x) x^{a_i^0} \} e^{-b_i^0 x} = 0,
\end{aligned} \tag{29}$$

where $\beta_{1i} = \alpha_{1i} \frac{b_i^0}{\Gamma(a_i^0)} + \beta_i^0 \frac{\partial}{\partial a}\left(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}\right) + \alpha_{3i} \frac{a_i^0 (b_i^0)^{a_i^0 - 1}}{\Gamma(a_i^0)} + \sum_{j=1}^{s_i} \alpha_{4ij}^2 \frac{\partial}{\partial a^2}\left(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}\right) + \sum_{j=1}^{s_i} \alpha_{5ij}^2 \frac{a_i^0 (a_i^0 - 1)(b_i^0)^{a_i^0 - 2}}{\Gamma(a_i^0)} + 2\sum_{j=1}^{s_i} \alpha_{4ij}\alpha_{5ij} \frac{\partial}{\partial a}\left(\frac{a_i^0 (b_i^0)^{a_i^0 - 1}}{\Gamma(a_i^0)}\right)$, $\beta_{2i} = -\alpha_{3i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} +$

$2\sum_{j=1}^{s_i}\alpha_{5ij}^2\frac{a_i^0(b_i^0)^{a_i^0-1}}{\Gamma(a_i^0)}+2\sum_{j=1}^{s_i}\alpha_{4ij}\alpha_{5ij}\frac{\partial}{\partial a}(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)})$, $\beta_{3i}=\sum_{j=1}^{s_i}\alpha_{5ij}^2\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$, $\beta_{4i}=\alpha_{2i}\frac{(b_i^0)}{\Gamma(a_i^0)}+2\sum_{j=1}^{s_i}\alpha_{4ij}^2\frac{\partial}{\partial a}(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)})+2\sum_{j=1}^{s_i}\alpha_{4ij}\alpha_{5ij}\frac{a_i^0(b_i^0)^{a_i^0-1}}{\Gamma(a_i^0)}$, $\beta_{5i}=\sum_{j=1}^{s_i}\alpha_{4ij}^2\times\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$, and $\beta_{6i}=-2\sum_{j=1}^{s_i}\alpha_{4ij}\alpha_{5ij}\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$. Using the same argument as that of the proof of part (a) of Proposition 5.1, by multiplying both sides of the above equation with $\exp(b_{\bar{i}}^0x)$ and let $x\to+\infty$, we obtain

$$\sum_{i=1}^{\bar{i}}\{\beta_{1i}x^{a_i^0-1}+\beta_{2i}x^{a_i^0}+\beta_{3i}x^{a_i^0+1}+\beta_{4i}(\log x)x^{a_i^0-1}+\beta_{5i}(\log x)^2x^{a_i^0-1}+\beta_{6i}(\log x)x^{a_i^0}\}\to 0.$$

By the constraints of $\mathcal{O}_{k,c_0}$, we have $|a_i^0-a_j^0|\notin\{1,2\}$ for all $1\le i,j\le k_0+m$. Therefore, this limit yields $\beta_{1i}=\beta_{2i}=\beta_{3i}=\beta_{4i}=\beta_{5i}=\beta_{6i}=0$ for all $1\le i\le\bar{i}$ or equivalently $\alpha_{1ij}=\alpha_{2ij}=\alpha_{3ij}=\alpha_{4ij}=\alpha_{5ij}=0$ for all $1\le i\le\bar{i}, 1\le j\le s_i$. The same argument for the remaining indices yields $\alpha_{1ij}=\alpha_{2ij}=\alpha_{3ij}=\alpha_{4ij}=\alpha_{5ij}=0$ for all $1\le i\le k_0+m, 1\le j\le s_i$, which leads to contradiction. This completes the proof.

(b) If there exists (i,j) such that $\{|a_i^0-a_j^0|,|b_i^0-b_j^0|\}\equiv\{1,0\}$, then we can use the same way of construction as that of part (b) of Proposition 5.1. Now, the only case of interest is when we have some (i,j) such that $\{|a_i^0-a_j^0|,|b_j^0-b_j^0|\}\equiv\{2,0\}$. Without loss of generality, assume that $a_2^0=a_1^0-2$. We construct the sequence $G_n=\sum_{i=1}^{k_0+1}p_i^n\delta_{(a_i^n,b_i^n)}$ as $a_1^n=a_1^0, a_2^n=a_3^n=a_2^0, a_i^n=a_{i-1}^0$ for all $4\le i\le k_0+1$, $b_1^n=b_1^0, b_2^n-b_1^0=b_1^0-b_3^n=\frac{b_1^0}{a_2^0n}$, $b_i^n=b_{i-1}^0$ for all $4\le i\le k_0+1$, $p_1^n=p_1^0-c_n$, $p_2^n=\frac{p_2^0}{2}+\frac{1}{2}(c_n+\frac{1}{n})$, $p_3^n=\frac{p_2^0}{2}+\frac{1}{2}(c_n-\frac{1}{n})$, $p_i^n=p_{i-1}^0$ for all $4\le i\le k_0+1$ where $c_n=\frac{(a_2^0+1)p_2^0}{(2n^2-1)a_2^0-1}$. Now, we can check that for any $r\ge1$, $W_r^r(G_n,G_0)\gtrsim c_n+\frac{1}{n^r}$. As $r\ge2$, by means of Taylor's expansions up to the $([r]+1)$th order, we obtain

$$\begin{aligned}p_{G_n}(x)-p_{G_0}(x)&=(p_1^n-p_1^0)f(x|a_1^0,b_1^0)+\Bigg(\sum_{i=2}^{3}p_i^n-p_2^0\Bigg)f(x|a_2^0,b_2^0)\\&\quad+\sum_{j=1}^{[r]+1}\frac{\sum_{i=2}^{3}p_i^n(b_i^n-b_i^0)^j}{j!}\frac{\partial^j f}{\partial b^j}(x|a_2^0,b_2^0)+R_n(x),\end{aligned}\tag{30}$$

where $R_n(x)$ is the remainder term and, therefore, $|R_n(x)|/W_r^r(G_n,G_0)\to0$. We can check that as $j\ge3$, $\sum_{i=2}^{3}p_i^n(b_i^n-b_i^0)^j/W_r^r(G_n,G_0)\to0$ as $n\to\infty$. Addi-

tionally, direct computation demonstrates that

$$(p_1^n - p_1^0) f(x|a_1^0, b_1^0) + \left(\sum_{i=2}^{3} p_i^n - p_2^0\right) f(x|a_2^0, b_2^0) + \sum_{j=1}^{2} \frac{\sum_{i=2}^{3} p_i^n (b_i^n - b_i^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) = 0.$$

The rest of the proof proceeds in the same way as that of Proposition 5.1 part (b). □

PROOF OF THEOREM 3.3. It suffices to demonstrate the following bound.

PROPOSITION 5.3 (Location-exponential mixtures). *For any* $r \geq 1$,

$$\lim_{\varepsilon \to 0} \inf_{G \in \mathcal{E}_{k_0}} \{V(p_G, p_{G_0}) / W_1^r(G, G_0) : W_1(G, G_0) \leq \varepsilon\} = 0.$$

PROOF. Choose the sequence $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{(\theta_i^n, \sigma_i^n)}$ such that $\sigma_i^n = \sigma_i^0$ for all $1 \leq i \leq k_0$, $(p_i^n, \theta_i^n) = (p_i^0, \theta_i^0)$ for all $3 \leq i \leq k_0$. The parameters $p_1^n, p_2^n, \theta_1^n, \theta_2^n$ are to be determined. With this construction of G_n, we obtain $W_1(G_n, G_0) \asymp |p_1^n - p_1^0| + |p_2^n - p_2^0| + p_1^0|\theta_1^n - \theta_1^0| + p_2^0|\theta_2^n - \theta_2^0|$. Now, for any $x \notin \{\theta_1^0, \theta_2^0\}$ and for any $r \geq 1$, taking the Taylor expansion with respect to θ up to the $([r]+1)$th order, we obtain

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{2} p_i^0 (f(x|\theta_i^n, \sigma_i^0) - f(x|\theta_i^0, \sigma_i^0)) + (p_i^n - p_i^0) f(x|\theta_i^n, \sigma_i^0) \\ &= \sum_{i=1}^{2} (p_i^n - p_i^0) f(x|\theta_i^n, \sigma_i^0) \\ &\quad - p_i^0 \left[\sum_{j=1}^{[r]+1} \frac{(\theta_i^0 - \theta_i^n)^j}{j!} \frac{\partial^j f}{\partial \theta^j}(x|\theta_i^n, \sigma_i^0)\right] + R(x) \\ &= \sum_{i=1}^{2} \left[(p_i^n - p_i^0) - p_i^0 \sum_{j=1}^{[r]+1} \frac{(\theta_i^0 - \theta_i^n)^j}{j!(\sigma_i^0)^j}\right] f(x|\theta_i^n, \sigma_i^0) + R(x), \end{aligned}$$

where the last inequality is due to the identity (8) and $R(x)$ is the remainder of Taylor's expansion. Note that

$$\sup_{x \notin \{\theta_1^0, \theta_2^0\}} |R(x)| / W_1^r(G_n, G_0) \leq \sum_{i=1}^{2} O(|\theta_i^n - \theta_i^0|^{[r]+1+\delta}) / |\theta_i^n - \theta_i^0|^r \to 0.$$

Now, we choose $p_1^n = p_1^0 + 1/n$, $p_2^n = p_2^0 - 1/n$, which means $p_1^n + p_2^n = p_1^0 + p_2^0$ and $p_1^n \to p_1^0$, $p_2^n \to p_2^0$. As $p_i^0/j!(\sigma_i^0)^j$ are fixed positive constants for all $1 \leq j \leq [r] + 1$. It is clear that there exists sequences θ_1^n and θ_2^n such that for both $i = 1$ and $i = 2$, $\theta_i^n - \theta_i^0 \to 0$, the identity $p_i^0 \sum_{j=1}^{[r]+1} \frac{(\theta_i^0 - \theta_i^n)^j}{j!(\sigma_i^0)^j} = p_i^n - p_i^0$ holds for all n (sufficiently large). With these choices of p_1^n, p_2^n, θ_1^n, θ_2^n, we have

$$\sup_{x \notin \{\theta_1^0, \theta_2^0\}} \bigl|p_{G_n}(x) - p_{G_0}(x)\bigr| / W_1^r(G_n, G_0) = \sup_{x \notin \{\theta_1^0, \theta_2^0\}} \bigl|R(x)\bigr| / W_1^r(G_n, G_0) \to 0.$$

The rest of the proof proceeds in the same way as that of Proposition 5.1, part (b). □

□

SUPPLEMENTARY MATERIAL

Supplement to "Convergence rates of parameter estimation for some weakly identifiable finite mixtures" (DOI: 10.1214/16-AOS1444SUPP; .pdf). In this supplemental material, we present the proofs of several remaining technical lemmas.

REFERENCES

[1] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

[2] BUCHBERGER, B. (1965). An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. Ph.D. thesis, Johannes Kepler Univ. Linz.

[3] CAILLERIE, C., CHAZAL, F., DEDECKER, J. and MICHEL, B. (2011). Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.* **5** 1394–1423. MR2851684

[4] CHEN, H. and CHEN, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statist. Sinica* **13** 351–365. MR1977730

[5] CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Ann. Statist.* **37** 2523–2542. MR2543701

[6] CHEN, J., LI, P. and FU, Y. (2012). Inference on the order of a normal mixture. *J. Amer. Statist. Assoc.* **107** 1096–1105. MR3010897

[7] CHEN, J. and TAN, X. (2009). Inference for multivariate normal mixtures. *J. Multivariate Anal.* **100** 1367–1383. MR2514135

[8] CHEN, J., TAN, X. and ZHANG, R. (2008). Inference for normal mixtures in mean and variance. *Statist. Sinica* **18** 443–465. MR2432278

[9] CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665

[10] DASGUPTA, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York. MR2664452

[11] DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474. MR0254956

[12] FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. MR1126324

[13] GENOVESE, C. R. and WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127. MR1810921

[14] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

[15] HATHAWAY, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13** 795–800. MR0790575

[16] HO, N. and NGUYEN, X. (2016). Supplement to "Convergence rates of parameter estimation for some weakly identifiable finite mixtures." DOI:10.1214/16-AOS1444SUPP.

[17] HO, N. and NGUYEN, X. (2016). On strong identifiability and optimal rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** 271–307.

[18] KASAHARA, H. and SHIMOTSU, K. (2015). Testing the number of components in normal mixture regression models. *J. Amer. Statist. Assoc.* **110** 1632–1645.

[19] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885

[20] LINDSAY, B. (1995). *Mixture Models*: *Theory, Geometry and Applications*. IMS, Hayward, CA.

[21] LIU, X. and SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31** 807–832. MR1994731

[22] MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models*: *Inference and Applications to Clustering*. *Statistics*: *Textbooks and Monographs* **84**. Dekker, New York. MR0926484

[23] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422

[24] NGUYEN, X. (2016). Borrowing strength in hierarchical Bayes: Convergence of the Dirichlet base measure. *Bernoulli* **22** 1535–1571.

[25] ROUSSEAU, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* **38** 146–180. MR2589319

[26] ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454

[27] STURMFELS, B. (2002). *Solving Systems of Polynomial Equations*. *CBMS Regional Conference Series in Mathematics* **97**. Amer. Math. Soc., Providence, RI. MR1925796

[28] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[29] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454

[30] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York.

[31] ZHANG, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18** 806–831. MR1056338

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109-1107
USA
E-MAIL: minhnhat@umich.edu
xuanlong@umich.edu