Beyond Black Box Densities: Parameter Learning for the Deviated Components

Dat Do
Department of Statistics
University of Michigan at Ann Arbor
Ann Arbor, MI 48109
dodat@umich.edu

Nhat Ho
Department of Statistics and Data Sciences
University of Texas at Austin
Austin, TX 78712
minhnhat@utexas.edu

XuanLong Nguyen
Department of Statistics
University of Michigan at Ann Arbor
Ann Arbor, MI 48109
xuanlong@umich.edu

Abstract

As we collect additional samples from a data population for which a known density function estimate may have been previously obtained by a black box method, the increased complexity of the data set may result in the true density being deviated from the known estimate by a mixture distribution. To learn about this phenomenon, we consider the *deviating mixture model* $(1-\lambda^*)h_0 + \lambda^*(\sum_{i=1}^k p_i^* f(x|\theta_i^*))$, where h_0 is a known density function, while the deviated proportion λ^* and latent mixing measure $G_* = \sum_{i=1}^k p_i^* \delta_{\theta_i^*}$ associated with the mixture distribution are unknown. Using a novel notion of distinguishability between the known density h_0 and the deviated mixture distribution, we establish rates of convergence for the maximum likelihood estimates of λ^* and G^* under Wasserstein metrics. Simulation studies are carried out to illustrate the theory.

1 Introduction

Most data-driven learning processes typically consist of an iteration of steps that involve model training and fine-tuning, with more data in-take leading to further model re-training and refinement. As more samples become available and exhibit more complex patterns, the initial model may be obsolete, risks being discarded, or absorbed into a richer class of models that adapt better to the increased complexity. It takes considerable resources to train complex models on a rich data population. Moreover, many successful models in modern real-world applications have become so complex that make them hard to properly evaluate and interpret; aside from the predictive performance they may as well be considered as black boxes. Nonetheless, as data populations evolve and so must the learning models, several desiderata remain worthy: the ability to adapt to new complexity while retaining aspects of old "wise" model, and the ability to interpret the changes.

In this paper we will investigate a class of complex models for density estimation that are receptive to *adaptation*, *reuse* and *interpretablity*: we posit that there is an existing distribution h_0 which may have been obtained a priori by some means for the data population of interest, e.g., via kernel density estimation (KDE) [22] or mixture models [20] or some modern black box methods, such as generative adversarial networks (GANs) [13, 1] or normalizing flows [9]. Nonetheless, as more samples become available and/or as the data population changes, it is possible that the true density may deviate from h_0. While h_0 is potentially difficult to explicate, it is the deviation from the known

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

h_0 that we wish to learn and interpret. We will use a mixture distribution to represent this deviation, leading to what we call a *deviating mixture model* for the underlying data population:

$$p_{\lambda^* G_*}(x) := (1-\lambda^*)h_0(x) + \lambda^* F(x, G_*), \tag{1}$$

for $x \in \mathbb{R}^d$, where $F(x, G_*) := \sum_{i=1}^{k_*} p_i^* f(x|\theta_i^*))$ represents a mixture distribution for the density components deviating from h_0. Such deviating components are from a known family of density function f. The unknown parameters for this model are the mixing proportion $\lambda^* \in [0,1]$, and the mixing measure $G_* = \sum_{i=1}^{k_*} p_i^* \delta_{\theta_i^*}$, where $k_* \geq 1$ number of *deviated* components. The choice of mixture distribution $F(x, G_*)$ allows us to express complex deviation from h_0, yet the overall model remains amenable to the interpretation of its parameters: λ_* represents the amount of deviation from the existing candidate h_0, while the mixing measure G_* represents heterogeneous patterns of the deviation. Because h_0 might be complex and trained with great computational resource to estimate the density of prior data population, it is reasonable to assume h_0 be known in the model (1). The primary contribution of this paper is a rigorous investigation into the rather challenging questions of identifiability and parameter learning rates that arise from a standard maximum likelihood estimation procedure.

Relations to existing works. This modeling framework owes its roots to several significant bodies of work in both statistics and machine learning literature. In classical statistics, a dominant approach to address the increased complexity of data populations is via hypothesis testing: one can test an alternative (possibly composite) hypothesis represented by a class of distributions against the null hypothesis represented by h_0. Due to the constraint for obtaining simple and theoretically valid test statistics in order to accept or reject the null hypothesis, the testing approaches were mostly restricted to simple choices of distribution for the null and alternative hypotheses [6, 10, 7, 4, 8]. More similar to (1) is the class of *contaminated mixture models* for density estimation: in this framework, the data are assumed to be sampled from a mixture of P_0 and Q where either P_0 or Q can be an unknown distribution that needs to be estimated. While this approach offers more flexibility in terms of modeling, it does not always guarantee the identifiability of the mixing weight or mixture components P_0, Q [25, 23, 19]. Without identifiability, it is virtually impossible to interpret the model parameters for the data domains. To avoid the identifiability issue, several researchers added the semi-parametric or parametric structures on P_0 and Q, such as P_0 and Q are mixture distributions [2, 12]. However, to the best of our knowledge, the convergence theory of these models remains poorly understood, except for some simple settings (see also [3, 5]). The main distinction between our modeling framework of deviating mixture models and the existing research on contaminated mixtures lies in our assumption that one of the mixture components, namely h_0 is known, allowing us to focus on the inference of the deviation from h_0, for which a considerable learning theory for the parameters of interest can be established and will be presented in this paper. Finally, estimating parameters of mixture distributions is an essential problem in mixture models. The convergence properties have been studied using identifiability notions and Wasserstein distances [21, 18, 15]. Our technical approach requires a generalization of the identifiability notion to take account of structural property of the existing component h_0, which helps to shed light on a considerably more complex convergence behavior of the deviated components.

Contributions. The primary contribution of this paper is a rich theory of *identifiability* and rates of convergence for *parameters* and density estimation that arise in the *deviating mixture model* (1), under various settings of the existing component h_0, and that of the deviating components (via f and G_*). Because the convergence of density estimation in Hellinger distance under the MLE procedure is well studied in [26], the bulk of our technical innovation lies in establishing a collection of *inverse bounds* which relate the Hellinger distance of densities in model (1) in terms of that of their parameters. To do that, we introduce a novel notion of *distinguishability* between h_0 and family of density f. The inverse bounds will be characterized under such distinguishability conditions (or the lack thereof). Our proof technique allows us to characterize different convergence rates of parameters in the deviating mixture model under distinguishable settings. It also gives rise to several new types of inverse bounds in partially distinguishable settings, where we may not have identifiability in our model. To the best of our knowledge, this is the first work in which such bounds are obtained in mixture modeling literature. Moreover, we will provide many examples to demonstrate the broad applicability of our theory, including cases where the existing component h_0 is obtained by a black box method (e.g., deep learning model) and a more traditional method (e.g., via KDE's or mixture models). By doing so, we are able to push the boundary of identifiability and learning theory of mixture models toward a larger class of modern machine learning models.

Organization. The remainder part of this paper is structured as follows. In Section 2, we review the MLE method and the identifiability conditions, where the notion of distinguishability is presented. In Section 3, the main results of inverse bounds and convergence rates for parameters estimation of model (1) are shown. In Section 4, multiple simulation experiments are carried out to support the theory. Finally, Section 5 is used to discuss and conclude. Proofs of all the results in the main text are deferred to the Supplementary Material.

Notation. We denote by $\mathcal{E}_k(\Theta) = \{\sum_{i=1}^{k} p_i f(x|\theta_i) : \sum_{i=1}^{k} p_i = 1, p_i > 0, \theta_i \in \Theta \,\forall 1 \leq i \leq k\}$ the family of mixtures with exactly k components and $\mathcal{O}_K(\Theta) = \{\sum_{i=1}^{K} p_i f(x|\theta_i) : \sum_{i=1}^{K} p_i = 1, p_i \geq 0, \theta_i \in \Theta \,\forall 1 \leq i \leq K\}$ the family of mixtures with no more than K components. $\mathcal{E}_{k,c_0}(\Theta) = \{\sum_{i=1}^{k} p_i f(x|\theta_i) : \sum_{i=1}^{k} p_i = 1, p_i \geq c_0, \theta_i \in \Theta \,\forall 1 \leq i \leq k, k \leq K\}$ is the family of mixtures with exactly K components and mixing proportions being bounded below by c_0, and $\mathcal{O}_{K,c_0}(\Theta) = \{\sum_{i=1}^{k'} p_i f(x|\theta_i) : \sum_{i=1}^{k'} p_i = 1, p_i \geq c_0, \theta_i \in \Theta \,\forall 1 \leq i \leq k', k' \leq K\}$. $\|\cdot\|_2$ is the usual l^2 norm for vectors in $\mathbb{R}^d$ and matrices in $\mathbb{R}^{d \times d}$. We write $g(x) \gtrsim h(x)$ if $g(x) > ch(x)$ for all x, where c is a constant does not depend on x (similar for $g(x) \lesssim h(x)$). For any $\lambda \in \mathbb{R}$ and $B \subset \mathbb{R}$, denote by $1_{\{\lambda \in B\}}$ the function that takes value 1 if $\lambda \in B$, and 0 otherwise. For any two densities p and q, we denote $h(p, q)$ by the Hellinger distance and $V(p, q)$ by the Total Variation distance between them.

2 Identifiability and distinguishability theory

The principal goal of the paper is to establish the efficiency of parameter learning for the deviating mixture model (1) via the standard maximum likelihood estimation (MLE) method. To achieve this goal, the parameters have to be identifiable to begin with. Thus, our theory builds on and extends a standard notion of identifiability of families of density $\{f(x|\theta) : \theta \in \Theta\}$ that has been considered in previous work [21, 15].

Definition 2.1. The family $\{f(x|\theta), \theta \in \Theta\}$ (or in short, f) is identifiable in the order r, for some $r \geq 1$, if $f(x|\theta)$ is differentiable up to the order r in θ and the following holds:

A1. For any $k \geq 1$, given k different elements $\theta_1, \ldots, \theta_k \in \Theta$, if we have $\alpha_\eta^{(i)}$ such that for almost all x

$$\sum_{l=0}^{r} \sum_{|\eta|=l} \sum_{i=1}^{k} \alpha_\eta^{(i)} \frac{\partial^{|\eta|} f}{\partial \theta^\eta}(x|\theta_i) = 0$$

then $\alpha_\eta^{(i)} = 0$ for all $1 \leq i \leq k$ and $|\eta| \leq r$.

Many commonly used families f for mixture modeling satisfy the first order identifiability condition, including location-scale Gaussian distributions, e.g., $f(x|\theta) = N(x|\mu, \sigma^2)$ where μ and σ^2 represent the mean (location) and variance (scale) parameters, and location-scale Student's t-distributions. In model (1), however, due to the presence of the existing component h_0, the deviated mixture components need to be *distinguishable* from h_0. This motivates a more general notion of identifiability, namely, *distinguishability* that we now define. This condition specifies a property jointly for both the existing component h_0 and the family of density functions f that make up the deviated components.

Definition 2.2. For any natural numbers $k, r \geq 1$, we say that the family of density functions $\{f(\cdot|\theta), \theta \in \Theta\}$ with complexity level k (or in short, (f, k)) is distinguishable up to the order r from h_0 if the following holds:

A2. For any k distinct components $\theta_1, \ldots, \theta_k$, if we have real coefficients $\alpha_\eta^{(i)}$ for $0 \leq i \leq k$ such that

$$\alpha^{(0)} h_0(x) + \sum_{l=0}^{r} \sum_{|\eta|=l} \sum_{i=1}^{k} \alpha_\eta^{(i)} \frac{\partial^{|\eta|} f}{\partial \theta^\eta}(x|\theta_i) = 0,$$

for almost surely $x \in \mathcal{X}$, then $\alpha^{(0)} = \alpha_\eta^{(i)} = 0$ for $1 \leq i \leq k$ and $|\eta| \leq r$.

We observe that the identifiable condition is a direct consequence of the corresponding distinguishable condition. A simple but non-trivial example of the distinguishability condition can be derived directly from the definitions.

Example 2.3. (a) When $h_0(x) = \sum_{i=1}^{k_0} p_i^0 f(x|\theta_i^0)$ for some given weights $(p_1^0, \ldots, p_{k_0}^0)$ and parameters $(\theta_1^0, \ldots, \theta_{k_0}^0)$ where $k_0 \geq 1$, then (f, k) is distinguishable in the order r from h_0 as long as $k < k_0$ and the family of density f is identifiable in the order r.
(b) Given the choice of h_0 in (a), (f, k) is not distinguishable in the order r from h_0 when $k \geq k_0$.

More significantly, we can establish a broad class of h_0 and families f for which distinguishability holds. This is exemplified by the following theorem, where f represents a family of location or location-scale Gaussian kernels, and h_0 is subject to a relatively weak condition.

Theorem 2.4. *(a) Suppose that $-\log h_0(x) \gtrsim \|x\|_2^{\beta_1}$ or $-\log h_0(x) \lesssim \|x\|_2^{\beta_2}$ for all $\|x\|_2 > x_0$, for some $x_0 > 0$, $\beta_1 > 2$, and $\beta_2 < 2$. Then, for f being family of location-scale Gaussian and any $k > 0$, (f, k) is distinguishable from h_0 up to the first order, where the derivatives in Assumption A2 are taken with respect to both location and scale parameters, and (f, k) is also distinguishable from h_0 up to any order, where the derivatives in Assumption A2 are taken only with location parameters.*

(b) Suppose that h_0 is the pdf of a pushforward measure of $N(0, I_d)$ by a piecewise linear function with a finite and positive number of breakpoints. Then, the same conclusions as in part (a) hold.

The proof of Theorem 2.4 is in Appendix C.1, where the main proof technique is carefully examining the tail densities of h_0 and f at infinity. Note that in part (a), h_0 can be a pdf of any distribution possessing a lighter or heavier tail than Gaussian distributions, and in part (b), h_0 represents the pushforward of a Gaussian distribution by any piecewise linear function (recall that family of piecewise linear functions is dense in the Banach space of continuous functions with compact support). In the sequel we shall demonstrate several examples of interest that are applicable to Theorem 2.4 where h_0 may have been estimated by some popular "black box" methods.

Kernel based representation. Suppose that h_0 was obtained from a m- sample $Y_1, \ldots, Y_m \in \mathbb{R}^d$ by a classical kernel density estimation (KDE) method [22] or a RKHS-based method [24], so that

$$h_0(x) = \frac{1}{m} \sum_{j=1}^{m} k_\sigma(x, Y_j) \quad \forall x \in \mathbb{R}^d, \tag{2}$$

where k_σ is a kernel function with bandwidth σ. Popular choices of kernels include the Gaussian kernel $k_\sigma(x, x') = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$ and the multivariate Student's kernel $k_\sigma(x, x') = \left(\frac{1}{\sqrt{\pi}\sigma}\right)^d \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)} \left(1 + \frac{\|x-x'\|_2^2}{\nu\sigma^2}\right)^{-\frac{\nu+d}{2}}$. The corresponding distinguishability guarantee is as follows.

Corollary 2.5. *Suppose h_0 is defined by Eq. (2), where k_σ is Gaussian kernel and $m > K$, or k_σ is the multivariate Student's kernel. Then, for f being family of location-scale Gaussian, (f, K) is distinguishable from h_0 up to the first order, where the derivatives in Assumption A2 are taken with respect to both location and scale parameters, and (f, K) is also distinguishable from h_0 up to any order, where the derivatives in Assumption A2 are taken only with location parameters.*

In application, it is common that the condition $m > K$ is satisfied. It is also matches with the scenario that we consider in the paper, where h_0 is already trained using a big data set, and there is a small number of deviated components.

Neural networks. Deep neural networks represent a powerful, albeit black box, approximation device for constructing rich classes of distribution for generative models [13, 1]. Accordingly, h_0 is the pdf function of a Gaussian distribution being push-forwarded by a map T, which is represented by a neural network (NN). Suppose that the NN representing T has a positive and finite number of layers L, and so

$$T(x) = a(W_L a(W_{L-1}(\ldots a(W_1 x + b_1)) + b_{L-1}) + b_L), \tag{3}$$

where $W_1, \ldots, W_L \in \mathbb{R}^{d \times d}$ are the weights and $b_1, \ldots, b_L \in \mathbb{R}^d$ are the biases. The activation function a is chosen to be rectified linear unit (ReLU) function defined by $a(x) = \max\{x, 0\}$, and is applied elementwise to any vector in $\mathbb{R}^d$. The corresponding guarantee on the distinguishability condition is as follows.

Corollary 2.6. *Suppose that h_0 is the pdf a pushforward measure of $N(0, I_d)$ by a map T defined by Eq. (3). Then, for f being family of location-scale Gaussian and any $k > 0$, (f, k) is distinguishable* *from h_0 up to the first order, where the derivatives in Assumption A2 are taken with respect to both location and scale parameters, and (f, k) is also distinguishable from h_0 up to any order, where the derivatives in Assumption A2 are taken only with location parameters.*

3 Convergence rates of density estimation

In this section, we first establish the rate of density estimation for the deviating mixture models in Section 3.1. We then describe a general procedure to obtain the convergence rate of parameter estimation based on that of density estimation via inverse bounds in Section 3.2. Finally, we provide comprehensive inverse bounds under several settings of the deviating mixture models in Section 3.3.

3.1 MLE for deviating mixture model

Given n i.i.d. sample $X_1, X_2, \ldots, X_n$ from $p_{\lambda^* G_*}$ as in model (1), where G_* has k_* components, we want to estimate λ^* and G_* from the data. We refer to the problem as in *exact-fitted setting* if k_* is known, and we refer to it as in *over-fitted setting* if k_* is unknown but is known to be bounded by some number K. We denote the MLE for exact-fitted setting by

$$\widehat{\lambda}_n, \widehat{G}_n \in \operatorname*{arg\,max}_{\lambda \in [0,1], G \in \mathcal{E}_{k_*}(\Theta)} \sum_{i=1}^{n} \log(p_{\lambda G}(X_i)),$$

and for the over-fitted setting, we replace $\mathcal{E}_{k_*}(\Theta)$ in the equation above by $\mathcal{O}_K(\Theta)$, where $K \geq k_*$.

In order to state a rate of convergence for the density estimators $p_{\hat{G}_n}$ under the Hellinger distance h [26], we need a condition on the complexity of the function class

$$\overline{\mathcal{P}}_k^{1/2}(\Theta, \epsilon) = \left\{ \bar{p}_{\lambda G}^{1/2} : G \in \mathcal{O}_k(\Theta),\ h(\bar{p}_{\lambda G}, p_{\lambda^* G_*}) \leq \epsilon \right\}, \tag{4}$$

where for any $G \in \mathcal{O}_K(\Theta)$, we write $\bar{p}_{\lambda G} = (p_{\lambda G} + p_{\lambda^* G_*})/2$. The definition of $\overline{\mathcal{P}}_k(\Theta, \epsilon)$ originates from [26]. We measure the complexity of this class through the bracketing entropy integral

$$\mathcal{J}_B(\epsilon, \overline{\mathcal{P}}_k^{1/2}(\Theta, \epsilon), \nu) = \int_{\epsilon^2/2^{13}}^{\epsilon} \sqrt{\log N_B(u, \overline{\mathcal{P}}_k^{1/2}(\Theta, \epsilon), \nu)}\, du \vee \epsilon, \tag{5}$$

where $N_B(\epsilon, X, \eta)$ denotes the ϵ-bracketing number of a metric space (X, η) and ν is the Lebesgue measure. We require the following assumption.

A3. Given a universal constant $J > 0$, there exists $N > 0$, possibly depending on Θ and k, such that for all $n \geq N$ and all $\epsilon > (\log n/n)^{1/2}$,

$$\mathcal{J}_B(\epsilon, \overline{\mathcal{P}}_k^{1/2}(\Theta, \epsilon), \nu) \leq J\sqrt{n}\epsilon^2.$$

Theorem 3.1. *Assume that Assumption A3 holds, and let $k \geq 1$. There exists a constant $C > 0$ depending only on Θ, k such that for all $n \geq 1$,*

$$\sup_{G_* \in \mathcal{O}_k(\Theta), \lambda^* \in [0,1]} \mathbb{E}_{\lambda^*, G_*} h(p_{\widehat{\lambda}_n \widehat{G}_n}, p_{\lambda^* G_*}) \leq C\sqrt{\log n/n}.$$

Therefore, in order to get convergence rate for density functions based on MLE procedure, we only need to check assumption A3. This assumption holds true for a wide range class of parametric model [26]. For our model, we give an example that it holds when h_0 has an exponential tail (satisfied for KDE's and Neural networks above) and f is location-scale Gaussian distribution.

Proposition 3.2. *Suppose f is location-scale Gaussian family and $\Theta = [-a, a]^d \times \Omega$, where Ω is a subset of S_d^{++} whose eigenvalues are bounded in $[\underline{\lambda}, \overline{\lambda}]$, $a, \underline{\lambda}, \overline{\lambda} > 0$, and h_0 is bounded with tail $-\log h_0(x) \gtrsim \|x\|_2^{\beta}$ for some $\beta > 0$. Then, the family of densities $\{p_{\lambda G} : \lambda \in [0, 1], G \in \mathcal{O}_k(\Theta)\}$ satisfies assumption A3.*

3.2 Parameter learning rates of deviated components

The core of this paper lies in establishing a collection of *inverse bounds*, provided that some distinguishability condition developed in Section 2 holds. The inverse bounds basically say that a small distance between $p_{\lambda G}$ and $p_{\lambda^* G_*}$ under the total variation distance entails that (λ, G) and (λ^*, G_*)

are similar under appropriate distances, where (λ^*, G_*) is fixed. To this end, we employ Wasserstein metrics [27] and their extensions.

Wasserstein distances. Wasserstein distances are natural and useful for assessing the convergence of latent mixing measures in mixture models [21, 16, 14]. Given two measures $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$ on a space Θ endowed with a metric ρ, the Wasserstein metric of order $r \geq 1$ is:

$$W_r(G, G') = [\inf_q \sum_{i,j} q_{ij} \rho^r(\theta_i, \theta'_j)]^{1/r},$$

where the infimum is taken over all joint distribution on $[1, \dots, k] \times [1, \dots, k']$ such that $\sum_i q_{ij} = p'_j, \sum_j q_{ij} = p_i$. Note that if G_n is a sequence of discrete measures that converges to G in a Wasserstein distance, then for every atom of G, there is a subset of atoms of G_n converges to it. Therefore, the convergence in Wasserstein metrics implies convergence of parameters in mixture models. In this paper, space Θ is often chosen to be a compact subset of $\mathbb{R}^d$ and ρ is the usual l^2 distance. In the case of location-scale Gaussian mixtures, space Θ is a compact subset of $\mathbb{R}^d \times S_d^{++}$, where S_d^{++} is the set of positive definite and symmetric matrices in $\mathbb{R}^{d \times d}$, and for every $(\mu, \Sigma), (\mu', \Sigma') \in \Theta$, the distance ρ is defined by $\rho((\mu, \Sigma), (\mu', \Sigma')) = \|\mu - \mu'\|_2 + \|\Sigma - \Sigma'\|_2$.

From inverse bounds to parameter learning rates. Suppose that some distinguishablity condition is satisfied, then we will establish an inverse bound providing a guarantee that a small distance between $p_{\lambda^* G_*}$ and $p_{\lambda G}$ entails a small distance between λ and λ^* and between G and G_*. More concretely, define a divergence between two measures λG and $\lambda^* G_*$ via

$$\overline{W}_r(\lambda G, \lambda^* G_*) := |\lambda - \lambda^*| + (\lambda + \lambda^*) W_r^r(G, G_*).$$

for all $r \geq 1$, and the inverse bounds will have the form that $V(p_{\lambda G}, p_{\lambda^*, G_*}) \gtrsim \overline{W}_r(\lambda G, \lambda^* G_*)$, for some r that depends on the level of distinguishable level of the model. Since total variational distance is upper bounded by Hellinger distance, if Assumption A3. holds, then combining the aforementioned inverse bound with Theorem 3.1 we immediately obtain

$$\mathbb{E}_{\lambda^*, G_*} \overline{W}_r(\widehat{\lambda}_n \widehat{G}_n, \lambda^* G_*) \leq C \sqrt{\frac{\log n}{n}}.$$

This further implies that the convergence rate of $\hat{\lambda}_n$ to λ^* is of order $(\log(n)/n)^{1/2}$ and the convergence rate of $W_r(\hat{G}_n, G_*)$ to 0 is of order $(\log(n)/n)^{1/2r}$.

3.3 Inverse bounds in distinguishable setting

We shall establish inverse bounds provided a distinguishability condition for model (1) holds under either exact-fitted and over-fitted settings regarding the true number of components k_*.

Theorem 3.3. *Assume that k_* is known and (f, k_*) is distinguishable in the first order from h_0. Then, for any $G \in \mathcal{E}_{k_*}(\Theta)$, there exist positive constant C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following holds:*

(a) When $\lambda^* = 0$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \lambda$.

(b) When $\lambda^* \in (0, 1]$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \overline{W}_1(\lambda G, \lambda^* G_*)$.

We now present a proof sketch for Theorem 3.3. It is a combination of the Taylor expansion around the true parameters and the Fatou's lemma; the proof technique for the remaining results also shares similar spirit as that of Theorem 3.3. Detailed proof of Theorem 3.3 is deferred to the Appendix.

Proof sketch for part (b): Suppose that the bound is not correct, so there exists a sequence $\lambda_n \in (0, 1]$ and $G_n \in \mathcal{E}_{k_*}(\Theta)$ such that $V(p_{\lambda^* G_*}, p_{\lambda_n G_n}) / \overline{W}_1(\lambda^* G_*, \lambda_n G_n) \to 0$. Because of the compactness of the parameter space, by extracting a subsequence if necessary, we can assume $\lambda_n \to \lambda', G_n \xrightarrow{W_1} G'$. If $(\lambda', G') \neq (\lambda^*, G_*)$, we have $\overline{W}_1(\lambda^* G_*, \lambda_n G_n) \to \overline{W}_1(\lambda^* G_*, \lambda' G') \neq 0$. It indicates that $V(p_{\lambda^* G_*}, p_{\lambda_n G_n}) \to 0$, which leads to $p_{\lambda^* G_*} = p_{\lambda' G'}$. It contradicts to the distinguishable condition when $(\lambda', G') \neq (\lambda^*, G_*)$).

Otherwise, we have $\lambda_n \to \lambda^*, G_n \to G_*$, and can present $G_n = \sum_{i=1}^{k_*} p_i^n \delta_{\theta_i^n}$ and $G_* = \sum_{i=1}^{k_*} p_i^* \delta_{\theta_i^*}$ such that $p_i^n \to p^*, \theta_i^n \to \theta_i^*$. Because of these limits and by Taylor expansion, we can arrange the difference $(p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x)) / \overline{W}_1(\lambda_n G_n, \lambda^* G_*)$ in terms of a linear combination of

$h_0(x), f(x|\theta_i^*), \frac{\partial}{\partial\theta} f(x|\theta_i^*)$ such that at least one coefficient is different from 0. By Fatou's lemma, $0 = \dfrac{\liminf V(p_{\lambda_n G_n}, p_{\lambda^* G_*})}{\overline{W}_1(\lambda_n G_n, \lambda^* G_*)} dx \geq \displaystyle\int \left|\liminf \frac{p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x))}{\overline{W}_1(\lambda_n G_n, \lambda^* G_*)}\right| dx$, which equals to the absolute integral of the linear combination above. Hence, there exists a non-trivial linear combination of $h_0(x), f(x|\theta_i^*), \frac{\partial}{\partial\theta} f(x|\theta_i^*)$ that equals 0, which contradict to the distinguishability condition. Therefore, we complete the proof.

In application, the true number of components k_* might not be known and we often fit the model (1) with $G \in \mathcal{O}_K(\Theta)$ for some large $K \geq k_*$. The next result shows that similar bounds can also be established in this case, where we require distinguishability of f and h_0 in a higher order.

Theorem 3.4. *Assume that k_* is unknown and strictly upper bounded by a given K. Assume additionally that (f, K) is distinguishable in second order from h_0. Then, for any $G \in \mathcal{O}_K(\Theta)$, there exist positive constant C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following holds:*

(a) When $\lambda^* = 0$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \lambda$.

(b) When $\lambda^* \in (0, 1]$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \overline{W}_2(\lambda G, \lambda^* G_*)$.

Thanks to the distinguishability up to second order, no matter how large the number of over-fitted components K is, we always get the $\overline{W}_2$ lower bound for the total variation distances. Proof of this theorem shares the same spirit with what of Theorem 3.3. The difference here is when we overfit G_* with some $\hat{G}$, there are some atoms of $\hat{G}$ that converges to the same atom of G_*, which requires us to do Taylor expansion up to second order and explain the higher order of Wasserstein distance here. Next, we relax the assumption of Theorem 3.4 by working on the setting where f is not second order identifiable. This is an instance of the so-called *weakly identifiable* setting — One popular example of weakly identifiable f is location-scale Gaussian distribution, which admits the partial differential equation (PDE) structure $\dfrac{\partial^2 f}{\partial \mu^2}(x|\mu, \Sigma) = 2\dfrac{\partial f}{\partial \Sigma}(x|\mu, \Sigma)$, for all $x \in \mathbb{R}^d$ where $f(x|\mu, \Sigma)$ stands for location-scale Gaussian density function with location μ and covariance Σ. In order to illustrate the result of our bound for that weak identifiability setting of f, we specifically consider f to be location-scale Gaussian distribution. In this case, the parameter space Θ is a compact subset of $\mathbb{R}^d \times S_d^{++}$, where S_d^{++} is the set of positive definite and symmetric matrices in $\mathbb{R}^{d \times d}$ equipped with the usual Frobenius norm. To put our result in context, we shall adopt a notion used in analyzing the convergence rate of parameter estimation in location-scale Gaussian mixtures in [16]. For any $k \geq 1$, let $\overline{r}(k)$ be the minimum value of r such that the following system of polynomial equations:

$$\sum_{j=1}^{k+1} \sum_{n_1, n_2} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \text{ for each } \alpha = 1, \ldots, r, \tag{6}$$

does not have any nontrivial solution for the unknown variables $(a_j, b_j, c_j)_{j=1}^{k+1}$, where the ranges of n_1 and 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 of variables c_j are non-zeroes, while at least one of the a_j is non-zero. Some examples of known values of $\overline{r}$ are $\overline{r}(1) = 4$ and $\overline{r}(2) = 6$, and $\overline{r}(k) \geq 7$ for all $k \geq 3$. Using this notion, we can characterize the convergence of parameters of model (1) for the location-scale Gaussian family via the following theorem for inverse bounds.

Theorem 3.5. *Assume that $G^* \in \mathcal{E}_{k^*, c_0}(\Theta)$, and k_* is unknown and strictly upper bounded by a given K. In addition, f is location-scale Gaussian distribution and (f, K) with varied location, fixed variance parameters is distinguishable in any order from h_0. Then, for any $G \in \mathcal{O}_{K, c_0}(\Theta)$, there exist positive constant C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following holds:*

(a) When $\lambda^* = 0$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \lambda$.

(b) When $\lambda^* \in (0, 1]$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \overline{W}_{\overline{r}(K - k_*)}(\lambda G, \lambda^* G_*)$.

The proof technique of this result involves doing Taylor expansion of both location and scale parameter up to order $\overline{r}$, then utilize the heat equation $\dfrac{\partial f}{\partial \Sigma}(x|\mu, \Sigma) = \dfrac{1}{2}\dfrac{\partial^2 f}{\partial \mu^2}(x|\mu, \Sigma)$ to compress this expression into linear combination of h_0 and derivatives of $f(x|\mu, \Sigma)$ with respect to μ only. This allows us to use the condition in this theorem to imply a contradiction, and gives rise to Eq. (6).

3.4 Inverse Bounds in Partially Distinguishable Setting

What happens if the distinguishability condition required by Def. 2.2 no longer holds generally? Recall in Example 2.3 (b) that this situation is not uncommon, specifically when

$$h_0(x) = f(x; G_0) = \sum_{i=1}^{k_0} p_i^0 f(x|\theta_i^0), \tag{7}$$

where $G_0 := \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$. In some specific cases of this setting, in fact, we fail to attain distinguishability, and the model may not even be identifiable in the classical sense, i.e. $p_{\lambda G} = p_{\lambda^* G_*}$ does not guarantee to have $\lambda G = \lambda^* G_*$. Since h_0 is the pdf of a mixture distribution — a popular choice for modeling complex forms of probability densities given its amenability to interpretation compared to black box type models — it is of interest to study the implication of parameter estimation for the deviated components in this setting, provided that the distinguisability condition may be at least partially achieved in some suitable sense. As we shall see, our theory demands a more refined analysis. To facilitate the presentation, denote $\mathcal{A} := \{1 \leq i \leq k_* : \theta_i^* \in \{\theta_1^0, \ldots, \theta_{k_0}^0\}\}$. Also, set $\bar{k} := |\mathcal{A}|$, which stands for the cardinal of the set $\mathcal{A}$. Our results will be divided into three separate regimes of $\bar{k}$ and λ^*: (i) $\lambda^* = 0$, (ii) $\bar{k} < k_0$ and $\lambda^* \in (0, 1]$, and (iii) $\bar{k} = k_0$ and $\lambda^* \in (0, 1]$. We only choose to present results of the second regime (ii) in the main text because of limited space and because of its representativeness as it shows all the intriguing behaviours of the model in this partially distinguishable setting. The first and third regime are deferred to Appendix A.

3.4.1 Regime B: $\bar{k} < k_0$ and $\lambda^* \in (0, 1]$

First, we consider the exactly-specified setting of model (1), namely, k_* is known. When $\bar{k} < k_0$, we can check that we still have dishtinguishability of h_0 and linear combinations of $\{f(x|\theta_i^*)\}_{i=1}^{k_*}$ and its derivatives. Therefore, as long as f is first order identifiable, one can invoke the proof of Theorem 3.3 to establish the same lower bound $V(p_{\lambda G}, p_{\lambda^* G_*})$ in terms of $\overline{W}_r(\lambda G, \lambda^* G_*)$ for some $r \geq 1$. Thus, our focus in this subsection is the settings when k_* is unknown.

Over-fitted setting with strongly identifiable f. Moving to the over-fitted settings of model setup (1), i.e., k_* is unknown and strictly upper bounded by a given K, as long as $K \geq k_0$, (f, K) is not distinguishable from h_0. Therefore, the results of Theorem 3.3 are not always applicable to the setting when $K \geq k_0$. Besides, in the over-fitted setting, the identifiability of model (1) no longer holds. Indeed, for any $\lambda > \lambda^*$, if we take

$$\overline{G}_*(\lambda) = (1 - \lambda^*/\lambda)\, G_0 + (\lambda^*/\lambda) G_*, \tag{8}$$

then $p_{\lambda^* G_*} = p_{\lambda \overline{G}_*(\lambda)}$. We present this pathological behavior in the following result.

Theorem 3.6. *Assume that h_0 takes the form (7) and $\bar{k} < k_0$. Besides that, $K \geq k_0$ and f is second order identifiable. Then, for any $G \in \mathcal{O}_K(\Theta)$, there exist positive constants C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following hold:*

(a) If $K \leq k_* + k_0 - \bar{k} - 1$, then $V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_1 \overline{W}_2(\lambda G, \lambda^* G_*)$,

(b) If $K \geq k_* + k_0 - \bar{k}$, then

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \left(1_{\{\lambda \leq \lambda^*\}} \overline{W}_2(\lambda G, \lambda^* G_*) + 1_{\{\lambda > \lambda^*\}} W_2^2(G, \overline{G}_*(\lambda)) \right).$$

(c) As a special case, if $K = k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_3 1_{\{\lambda > \lambda^* + \delta\}} W_1(G, \overline{G}_*(\lambda)),$$

for all $\delta > 0$, where C_3 depends on $\lambda^*, G_*, h_0, \Theta, \delta$.

As we can see, the magnitude of λ compared to λ^* will decide the solution of (λ, G) to the identifiable equation $p_{\lambda G} = p_{\lambda^* G^*}$, therefore lead to different lower bounds such in part (b) of the theorem. In particular, if $\lambda \leq \lambda^*$, the solution is $(\lambda, G) = (\lambda^*, G_*)$, and for any $\lambda > \lambda^*$, the solution is $G = \overline{G}_*(\lambda)$ given in Eq. (8). Specifically, when λ is strictly larger than λ^* by some amount $\delta > 0$, then the latter case is well separated from the former, and we have an exact-fitted result when $K = k_0 + k_* - \bar{k}$.

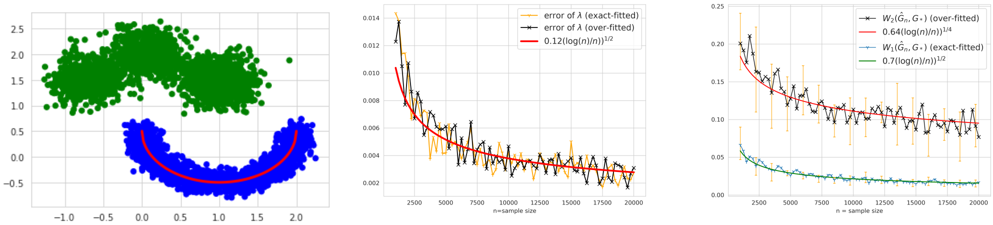
4 Experiments

We now would like to demonstrate the convergence rates in Section 3 via two synthetic experiments: one for distinguishable setting and one for partially distinguishable setting. For the partially distinguishable one, the experiments are in Appendix B.

Distinguishable setting. We conduct an experiment where the original data distribution comes from an uniform distribution on a curve (half circle) in $\mathbb{R}^2$ convoluted with Gaussian noises (red curve and blue points in Fig. 1(a)), and train a Normalizing Flow neural network [11] (Masked Autoregressive architecture) with 5 layers to get a good density estimation h_0 for this dataset. Then we assume that there are new data coming in, and the original distribution h_0 is deviated by a mixture of distributions in the location Gaussian family $f(x|\theta)$. So the true generating density now is

$$p_{\lambda^* G_*}(x) = (1-\lambda^*)h_0(x) + \lambda^* \sum_{i=1}^{3} p_i^* f(x|\theta_i^*), \tag{9}$$

where $\lambda^* = 0.5, G_* = \sum_{i=1}^{3} p_i^* \delta_{\theta_i^*}$, where $p_1^* = 0.3, p_2^* = 0.3, p_3^* = 0.4, \theta_1^* = (-0.7, 1.5), \theta_2^* = (0.1, 2.0), \theta_3^* = (1.0, 1.5)$. Samples from the deviated component are green points in Fig. 1(a). It can be seen from Proposition 2.4(a) that h_0 is distinguishable with family f. For each n, we simulate n data points from true model (9), estimate $\hat{\lambda}_n, \hat{G}_n$ by the EM algorithm (it is possible because Normalizing Flows provides exact density computation), and measure its convergence to the true λ^*, G_*. We conduct 16 replications for each sample size. The average error estimations with a 75% error bar can be seen in Fig. 1. The W_1 error in the exact-fitted case is of order $(\log(n)/n)^{1/2}$ and W_2 error in the over-fitted case is of order $(\log(n)/n)^{1/4}$. Meanwhile, thanks to the distinguishability, the estimation errors in both cases of λ are all of the order $(\log(n)/n)^{1/2}$. These simulation results are matched with the theoretical results found in Theorem 3.3 and Theorem 3.4. From the result, we see that the deviating mixture model successfully learns the deviated components and reuses the pre-trained black box model h_0, which helps to reduce computational costs.



(a) Synthetic data set

(b) Convergence rates of $\hat{\lambda}_n$

(c) Convergence rates of $W(\hat{G}_n, G_*)$

Figure 1: Convergence rates for parameter estimation in the distinguishable case.

5 Discussion

In this work, we have presented the deviating mixture model and studied its parameter learning rates under MLE procedure. With a novel notion of distinguishability between distributions, we are able to prove inverse bounds for our model under several distinguishability settings, which allow us to deduce the parameter learning rates from the convergence rate of density functions. The distinguishability condition is shown to be satisfied for multiple families of distributions including those that come from black box models.

We now discuss practical implication of the theory. The deviating mixture model is designed to capture the deviated mixture components, and learning its parameters can reveal meaningful information about subpopulations in the data. When there is distinguishability in the model, our theory implies that we can learn the deviated proportion with the parametric rate and deviated components with a rate depending on the identifiablity of f. However, our theory does not support employing the deviating mixture model when the existing distribution h_0 itself is a mixture distributions in family f and possesses parameters similar to deviated part, as the learning rate can be slow, and the deviated proportion estimator may not converge to the true value. Asymptotically, when h_0 is estimated using a very complex model (eg. a wide and deep neural network) and somehow approximates a mixture of f, and/or the signal from deviating components is low, then the provided learning rates in the paper, while still the same with respect to sample size n, may deteriorate from a large multiplicative constant that depends on h_0, λ^*, and G_*.

We believe that this work is the first attempt in the effort of understanding a broader class of mixture models combining with black box models, and interpreting the learned model parameters. There is room for future work going forward. From a theoretical viewpoint, one may be interested in establishing minimax lower bounds for the learning behavior of the deviating mixture model, or show uniform inverse bounds for the model when λ^* and G_* are considered as signals that will change with samples. From a modeling viewpoint, it is worthwhile to explore mixtures of black box models and develop a suitable notion of identifiability and inverse bounds so that the learning process is efficient.

Acknowledgements

Nhat Ho acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning. Long Nguyen is partially supported by NSF grant DMS-2015361.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34, 2006.

[3] Cristina Butucea and Pierre Vandekerkhove. Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics*, 41(1):227–239, 2014.

[4] T. Cai, X. J. Jeng, and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 2011.

[5] T Tony Cai, Jiashun Jin, and Mark G Low. Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6):2421–2449, 2007.

[6] H. Chen and J. Chen. Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, 13:351–365, 2003.

[7] J. Chen and P. Li. Hypothesis test for normal mixture models: the em approach. *Annals of Statistics*, 37:2523–2542, 2009.

[8] J. Chen, P. Li, and Y. Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107:1096–1105, 2012.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[10] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32, 2004.

[11] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. nflows: normalizing flows in PyTorch, November 2020.

[12] Sébastien Gadat, Jonas Kahn, Clément Marteau, and Cathy Maugis-Rabusseau. Parameter recovery in two-component contamination mixtures: The l^2 strategy. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 1391–1418. Institut Henri Poincaré, 2020.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Aritra Guha, Nhat Ho, and XuanLong Nguyen. On posterior contraction of parameters and interpretability in bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188, 2021.

[15] Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46(6A):2844–2870, 2018.

[16] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016.

[17] N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307, 2016.

[18] Nhat Ho and XuanLong Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, 2019.

[19] J. Katz-Samuels, G. Blanchard, and C. Scott. Decontamination of mutual contamination models. *Journal of Machine Learning Research*, 20, 2019.

[20] Geoffrey J. McLachlan and David Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York, 2000.

[21] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013.

[22] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[23] Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.

[24] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[25] C. Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, 2015.

[26] Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[27] C. Villani. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemtical Sciences]*. Springer, Berlin, 2009.

[28] Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] It can be seen in the source code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 1, 2, 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The experiments are run on CPU's only.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [No] We do not use any existing assets.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We do not use crowdsourcing/conducted research with human subjects
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

Supplement for "Beyond Black Box Densities: Parameter Learning for the Deviated Components"

In the supplementary material, we collect proofs and results deferred from the main text. Section A provides remaining results for the partially distinguishable case. Section B presents the simulation studies that demonstrates the results in the partially distinguishable case. Section C contains proofs of results in Section 2, and Section D contains proofs of Section 3.

A Additional results

In this appendix, we provide theory for the inverse bounds in partially distinguishable setting when $\bar{k} = k_0$ and $\lambda^* \in (0, 1]$.

A.1 Regime A: $\lambda^* = 0$.

Theorem A.1. *Assume that h_0 takes the form* (7) *and $\lambda^* = 0$. Then, there exist positive constants C_1 and C_2 depending only on h_0, Θ such that the following holds:*

(a) (exact-fitted) If f is first order identifiable, then for any $G \in \mathcal{E}_{k_0}(\Theta)$

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_1 \lambda W_1(G, G_0).$$

(b) (over-fitted) If f is second order identifiable, then for any $G \in \mathcal{O}_K(\Theta)$ that $K > k_0$

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \lambda W_2^2(G, G_0).$$

(c) (over-fitted and weakly identifiable) If f is location-scale Gaussian distribution and we further assume that $G_* \in \mathcal{E}_{k_*, c_0}(\Theta)$, then for any $G \in \mathcal{O}_{K, c_0}(\Theta)$ that $K > k_0$, there exists C_3 depends on h_0, Θ_0, c_0 such that

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_3 \lambda W_{\overline{r}(K-k_*)}^{\overline{r}(K-k_*)}(G, G_0).$$

We may also "underfit" the deviated components by imposing $G \in \mathcal{O}_K(\Theta)$ such that $K < k_0$. In that case, because of having less atoms, $p_{\lambda G}$ is $K-$distinguishable with h_0 and the result in Theorem 3.3 applies.

A.2 Regime B: $\bar{k} < k_0$ and $\lambda^* \in (0, 1]$

We recall Theorem 3.6 in the main text, together with a similar theorem on weak identifiable family (Theorem A.3), and then provide some additional comments on the results.

Theorem A.2. *Assume that h_0 takes the form* (7) *and $\bar{k} < k_0$. Besides that, $K \geq k_0$ and f is second order identifiable. Then, for any $G \in \mathcal{O}_K(\Theta)$, there exist positive constants C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following hold:*

(a) If $K \leq k_* + k_0 - \bar{k} - 1$, then $V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_1 \overline{W}_2(\lambda G, \lambda^* G_*)$,

(b) If $K \geq k_* + k_0 - \bar{k}$, then

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \left(1_{\{\lambda \leq \lambda^*\}} \overline{W}_2(\lambda G, \lambda^* G_*) + 1_{\{\lambda > \lambda^*\}} W_2^2(G, \overline{G}_*(\lambda)) \right).$$

(c) As a special case, if $K = k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_3 1_{\{\lambda > \lambda^* + \delta\}} W_1(G, \overline{G}_*(\lambda)),$$

for all $\delta > 0$, where C_3 depends on $\lambda^*, G_*, h_0, \Theta, \delta$.

We can view $p_{\lambda G}$ as a mixture distributions with latent mixing measures $\widehat{G} = (1 - \lambda) \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0} + \sum_{i=1}^{K} p_i \delta_{\theta_i}$ having at most $K + k_0$ elements, while $p_{\lambda^* G_*}$ as a mixture with latent measure $\widehat{G}_* = \sum_{i=1}^{\bar{k}} \left[(1-\lambda^*) p_i^0 + \lambda^* p_i^* \right] \delta_{\theta_i^0} + \sum_{i=\bar{k}+1}^{k_0} (1-\lambda^*) p_i^0 \delta_{\theta_i^0} + \sum_{i=\bar{k}+1}^{k_*} \lambda^* p_i^* \delta_{\theta_i^*}$ having exactly $k_0 + k_* - \bar{k}$

elements. Because $k_0 + k_* - \bar{k} < K + k_0$, a direct application of Theorem 3.2 in [17] gives us $V(p_{\lambda^*,G_*}, p_{\lambda,G}) \gtrsim W_2^2(\widehat{G}_*, \widehat{G})$. But this bound is not as tight as what in Theorem 3.6(c), since $W_1 \gtrsim W_2^2$. The bounds established in the theorem are possible as we carefully explore the structure of $\widehat{G}_*$ and $\widehat{G}$.

Over-fitted setting with weakly identifiable f. Similar to Theorem 3.5, when f is the location-scale Gaussian, the weak identifiability can worsen the power of the bound in the over-fitted case.

Theorem A.3. *Assume that h_0 takes the form (7). Besides that, $K \geq k_0$ and f is location-scale Gaussian distribution. Then, for any $\lambda \in [0, 1]$ and $G \in \mathcal{O}_{K,c_0}(\Theta)$ for some $c_0 > 0$, there exist positive constants C_1, C_2, C_3, C_4 depending only on $\lambda^*, G_*, G_0, \Theta$ (C_3 and C_4 also depend on δ) such that the following holds:*

(a) When $K \leq k_* + k_0 - \bar{k} - 1$, then $V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_1 \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*)$.

(b) When $K \geq k_* + k_0 - \bar{k}$, then

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_2 \bigg(1_{\{\lambda \leq \lambda^*\}} \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*) + 1_{\{\lambda > \lambda^*\}} W_{\overline{r}(K-k_*)}^{\overline{r}(K-k_*)}(G, \overline{G}_*(\lambda)) \bigg).$$

(c) For $\delta > 0$, when $K = k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_3 1_{\{\lambda > \lambda^* + \delta\}} W_1(G, \overline{G}_*(\lambda)),$$

and when $K > k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_4 1_{\{\lambda > \lambda^* + \delta\}} W_{\overline{r}(K-k_0-k_*+\bar{k})}^{\overline{r}(K-k_0\ -k_*+\bar{k})}(G, \overline{G}_*(\lambda)).$$

In this theorem, we once again observe the pathological behavior of the lower bound by Wasserstein distances caused by the unidentifiability of the model (1). In part (c), when there is a well separation between two region of solutions of equation $p_{\lambda G} = p_{\lambda^* G_*}$, we can improve the order of Wasserstein distances for both exact-fitted case and over-fitted case. In application, if $\hat{\lambda}_n$ and $\hat{G}_n$ are the MLE of model (1) estimated by n i.i.d. data, then the convergence of $(\hat{\lambda}_n, \hat{G}_n)$ depends on the limit of $\hat{\lambda}_n$ (or its subsequence) comparing to λ^*. If $K = k_0 + k_* - \overline{k}$, any subsequence of $(\hat{\lambda}_n)$ having limit greater than λ^* can achieve W_1 convergence rate of the distance between $\hat{G}_n$ and $\overline{G}_*(\hat{\lambda}_n)$. If $K > k_0 + k_* - \overline{k}$, any subsequence of $(\hat{\lambda}_n)$ having limit greater than λ^* can achieve $W_{\overline{r}(K-k_0-k_*+\bar{k})}^{\overline{r}(K-k_0-k_*+\bar{k})}$ convergence rate of the distance between $\hat{G}_n$ and $\overline{G}_*(\hat{\lambda}_n)$, where $\overline{r}(K - k_0 - k_* + \bar{k})$ is smaller than $\overline{r}(K - k_*)$ in part (b).

A.3 Regime C: $\bar{k} = k_0$ and $\lambda^* \in (0, 1]$.

When $\bar{k} = k_0$, (f, k_*) and (f, K) are not distinguishable from h_0. It indicates that the results of Theorem 3.3 are no longer applicable to this setting. If $G^* = G_0$, the setting goes back to the case $\lambda^* = 0$ and it is already considered, so from this section, we assume that $G_* \neq G_0$. To streamline the argument, we further denote a few more notations. As $\bar{k} = k_0$, we can rewrite G_* as follows:

$$G_* = \sum_{i=1}^{k_0} p_i^* \delta_{\theta_i^0} + \sum_{i=k_0+1}^{k_*} p_i^* \delta_{\theta_i^*}. \tag{10}$$

Because of the non-identifiability, the lower bound of $V(p_{\lambda G}, p_{\lambda^* G_*})$ must be inspected carefully based on the magnitude of mixing proportions of $p_{\lambda G}$ compared to what of $p_{\lambda^* G_*}$. To serve this purpose, we denote

$$\mathcal{B} := \{\lambda \in [0, 1] : (\lambda^* - \lambda) p_i^0 \leq \lambda^* p_i^* \ \forall \ 1 \leq i \leq k_0\},$$
$$\mathcal{I}(\lambda) := \{1 \leq i \leq k_0 : (\lambda^* - \lambda) p_i^0 > \lambda^* p_i^*\}.$$

For any $\lambda \in [0, 1]$, we say that the set $\mathcal{I}(\lambda)$ is *ratio-independent* if and only if $|\mathcal{I}(\lambda)| = 1$ or $p_i/p_i^* = p_j/p_j^*$ for all $i, j \in \mathcal{I}(\lambda)$ when $|\mathcal{I}(\lambda)| \geq 2$. Moreover, we define

$$\widetilde{G}_*(\lambda) := \frac{1}{\mathcal{S}(\mathcal{I}(\lambda))} \bigg(\sum_{i \in \mathcal{I}(\lambda)^c} \Big[p_i^* \lambda^* + (\lambda - \lambda^*) \, p_i^0 \Big] \delta_{\theta_i^0}$$
$$+ \lambda^* \sum_{i=k_0+1}^{k_*} p_i^* \delta_{\theta_i^*} \bigg), \tag{11}$$

where $\mathcal{S}(\mathcal{I}(\lambda)) := \sum_{i \in \mathcal{I}(\lambda)^c} \left[p_i^* \lambda^* + (\lambda - \lambda^*) \, p_i^0 \right] + \lambda^* \sum_{i=k_0+1}^{k} p_i^*$. In the case $\mathcal{I}(\lambda)$ is ratio-independent, the identifiable equation $p_{\lambda G} = p_{\lambda^* G_*}$ attains a solution $G = \widetilde{G}_*(\lambda)$ as in equation (11). Hence, in the following, we need to divide λ into several regimes to specify the lower bound for $V(p_{\lambda G}, p_{\lambda^* G_*})$ based on appropriate distances of (λ, G) and (λ^*, G_*).

Setting with second order identifiable f: We first consider the setting when f is second order identifiable and the model setup (1) is over-fitted. The following result demonstrates that under different settings of λ and $\mathcal{I}(\lambda)$, the lower bound of $V(p_{\lambda G}, p_{\lambda^* G_*})$ in terms of its corresponding parameters (λ, G) and (λ^*, G_*) can be very different.

Theorem A.4. *Assume that h_0 takes the form* (7) *and $\overline{k} = k_0$. Besides that, f is second order identifiable. Then, for any $\lambda \in [0,1]$ and $G \in \mathcal{O}_K(\Theta)$ that $K \geq k_*$, there exist positive constants C_1 and C_2 depending only on $\lambda^*, G_*, G_0, \Theta$ such that the following holds:*

(a) If $\mathcal{I}(\lambda)$ is not ratio-independent, then

$$V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \left[1_{\{\lambda \in \mathcal{B}^c\}} + 1_{\{\lambda \in \mathcal{B}\}} W_2^2(G, \overline{G}_*(\lambda)) \right]. \tag{12}$$

(b) If $\mathcal{I}(\lambda)$ is ratio-independent, then

$$\begin{aligned} V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \Bigg[1_{\{\lambda \in \mathcal{B}^c\}} \Bigg(\sum_{i \in \mathcal{I}(\lambda)} \bigg[(\lambda^* - \lambda) p_i^0 \\ - \lambda^* p_i^* \bigg] + \mathcal{S}(\mathcal{I}(\lambda)) W_2^2(G, \widetilde{G}_*(\lambda)) \Bigg) \\ + 1_{\{\lambda \in \mathcal{B}\}} W_2^2(G, \overline{G}_*(\lambda)) \Bigg]. \end{aligned} \tag{13}$$

We can see that when $\lambda \in \mathcal{B}^c$ and $\mathcal{I}(\lambda)$ is not ratio-independent, the bound in equation (12) shows that $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1$. It is due to the fact that $(\lambda^* - \lambda) p_i^0 - \lambda^* p_i^*$ cannot be simultaneously arbitrarily small as $i \in \mathcal{I}(\lambda)$. On the other hand, these terms can become very small at the same time when $\mathcal{I}(\lambda)$ is ratio-independent. It implies that $V(p_{\lambda^* G_*}, p_{\lambda G})$ can become arbitrarily close to 0 under this setting of $\mathcal{I}(\lambda)$. It explains the difference of bounds between two settings of $\mathcal{I}(\lambda)$.

Setting with weakly identifiable f: Finally, we consider the settings of model setup (1) when f is weakly identifiable. We specifically choose f to be location-scale Gaussian distribution and study the lower bounds of $V(p_{\lambda G}, p_{\lambda^* G_*})$ in terms of their parameters.

Theorem A.5. *Assume that h_0 takes the form* (7) *and $\overline{k} = k_0$. Besides that, f is location-scale Gaussian distribution. Then, for $\tilde{k} := \max\{k_* - k_0, 1\}$, and for any $\lambda \in [0,1]$ and $G \in \mathcal{O}_{K,c_0}(\Theta)$ for some $K \geq k_*$ and $c_0 > 0$, there exist positive constants C_1 and C_2 depending only on $\lambda^*, G_*, G_0, \Theta$ such that on $\lambda^*, G_*, G_0, \Theta$ such that*

(a) If $\mathcal{I}(\lambda)$ is not ratio-independent, then

$$\begin{aligned} V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \Bigg[1_{\{\lambda \in \mathcal{B}^c\}} \\ + 1_{\{\lambda \in \mathcal{B}\}} W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \bar{G}_*(\lambda)) \Bigg]. \end{aligned} \tag{14}$$

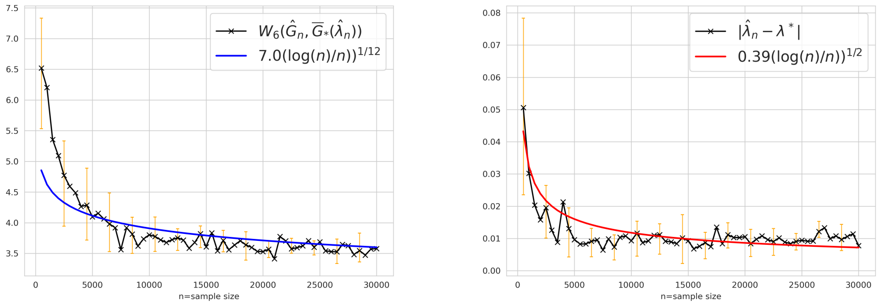
(b) If $\mathcal{I}(\lambda)$ is ratio-independent, then

$$\begin{aligned} V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \Bigg[1_{\{\lambda \in \mathcal{B}^c\}} \Bigg(\sum_{i \in \mathcal{I}(\lambda)} \bigg[(\lambda^* - \lambda) p_i^0 \\ - \lambda^* p_i^* \bigg] + \mathcal{S}(\mathcal{I}(\lambda)) W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \widetilde{G}_*(\lambda)) \Bigg) \\ + 1_{\{\lambda \in \mathcal{B}\}} W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \bar{G}_*(\lambda)) \Bigg]. \end{aligned} \tag{15}$$

B Additional Experiment

We provide a simulation experiment with partially distinguishable setting in this section to demonstrate the theoretical results in Section 3.4.

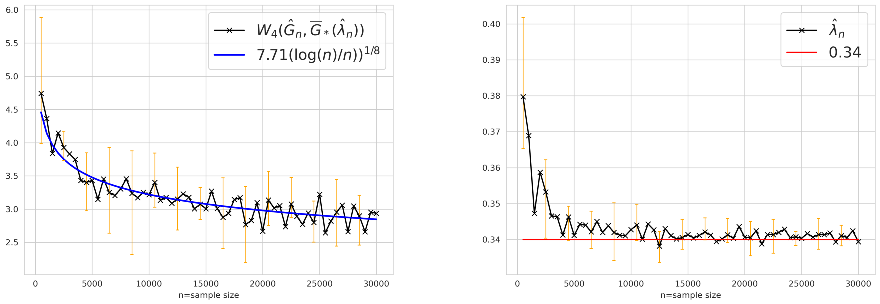
Partially distinguishable setting. Consider the partial distinguishable case as in Theorem A.3 with weakly identifiable f, we will conduct an experiment to distinguish two regimes in part (b) and (c) of the theorem, which are $\lambda > \lambda^*$ and $\lambda \leq \lambda^*$. We simulate n data from the true data generating model (1), where $p_1^0 = 0.4, p_2^0 = 0.6, p_1^* = 1, \lambda^* = 0.3, \mu_1^0 = \mu_1^* = (-2, 3), \Sigma_1^0 = \Sigma_1^* = \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}, \mu_2^0 = (1, -4), \Sigma_2^0 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$. In this case, $k_* = 1, k_0 = 2, \overline{k} = 1, k_* + k_0 - \overline{k} = 2$ and we will fit the data with model $p_{\lambda G}$, where G has 3 atoms. The MLE $(\hat{\lambda}_n, \hat{G}_n)$ is found by the EM algorithm. In the regime $\hat{\lambda}_n < \lambda^*$, we see that $\hat{\lambda}_n \to \lambda^*$ in the parametric rate and the convergence of $\hat{G}_n$ to G_* is of order $(\log(n)/n)^{2\overline{r}(K-k_*)} = (\log(n)/n)^{12}$ (Fig. 2). When $\hat{\lambda}_n > \lambda^*$, because of the indistinguishability of the model, we do not expect $\hat{\lambda}_n \to \lambda^*$ but the Wasserstein distance between $\hat{G}_n$ and $\overline{G}_*(\hat{\lambda}_n)$ converges to 0 with the rate $(\log(n)/n)^{2\overline{r}(2)} = (\log(n)/n)^{1/8}$. The simulation study matches with this result, where $\hat{\lambda}_n$ converges to some number greater than λ^*, and the rate that $W_4(G, \overline{G}_*(\hat{\lambda}_n))$ converges to 0 is of order $(\log(n)/n)^{1/8}$ (Fig. 3).



(a) Convergence rates of $W_6(\hat{G}_n, G_*)$

(b) Convergence rates of $|\hat{\lambda}_n - \lambda^*|$

Figure 2: Parameter learning rates in regime $\lambda \leq \lambda^*$.



(a) Convergence rates of $W_4(\hat{G}_n, G_*)$

(b) Limit of $\hat{\lambda}_n$

Figure 3: Parameter learning rates in regime $\lambda > \lambda^*$.

C Proofs of Section 2

C.1 Proof of Theorem 2.4

(a) We first prove that h_0 is distinguishable with (f, k) up to first order with any k and f being location-scale Gaussian family, i.e., if there exists $\lambda, \alpha_j \in \mathbb{R}, \beta_j \in \mathbb{R}^d$, symmetric matrices $\gamma_i \in \mathbb{R}^{d\times d}$, $\theta_j \in \mathbb{R}^d$, and positive definite symmetric $\Sigma_j \in \mathbb{R}^{d\times d}$ for $j = 1, \dots, k$ such that

$$\lambda h_0(x) + \sum_{j=1}^{k} \alpha_j f(x|\theta_j, \Sigma_j) + \sum_{j=1}^{k} \beta_j^T \frac{\partial f}{\partial \theta}(x|\theta_j, \Sigma_j) + \operatorname{tr}\left(\frac{\partial f}{\partial \Sigma}(x|\theta_j, \Sigma_j)^T \gamma_j\right) = 0,$$

then $\lambda = \alpha_j = \beta_j = \gamma_j = 0$ for all $j = 1, \dots, k$, where $f(x|\theta, \Sigma)$ is the density evaluated at x of Gaussian distribution with mean θ and covariance Σ and $(\theta_j, \Sigma_j)_{j=1}^k$ are pairwise different. Suppose there exists such $(\lambda, \alpha_j, \beta_j, \gamma_j)_{j=1}^k$. We borrow a technique from [17, 28], where we find a one-dimensional space to project $x \in \mathbb{R}^d$ onto and work with the order of means and variances in that space to show that the solution must be trivial. Calculating the first derivatives of f gives

$$\lambda h_0(x) + \sum_{j=1}^{k} \left(\alpha_j' + (\beta_j')^T(x - \theta_j) + (x - \theta_j)^T \gamma_j^{-1}(x - \theta_j)\right) e^{-\frac{1}{2}(x-\theta_j)^T \Sigma_j^{-1}(x-\theta_j)} = 0, \quad (16)$$

where

$$\alpha_j' = \frac{2\alpha_j - \operatorname{tr}(\Sigma_j^{-1}\gamma_j)}{2\pi^{d/2}|\Sigma_j|^{1/2}}, \quad \beta_j' = \frac{2}{\pi^{d/2}|\Sigma_j|^{1/2}}\Sigma_j^{-1}\beta_j, \quad \gamma_j' = \frac{1}{\pi^{d/2}|\Sigma_j|^{1/2}}\Sigma_j^{-1}\gamma_j\Sigma_j^{-1},$$

for all $j = 1, \dots, k$. If all the covariance matrices are equal, i.e., $\Sigma_1 = \dots = \Sigma_k$, then $(\theta_j)_{j=1}^k$ are pairwise different. Denote by $\delta_{ij} = \theta_i - \theta_j$, then for any $x' \notin \cup_{1\le i\le j\le k}\{u \in \mathbb{R}^d : \delta_{ij}^T u = 0\}$, we have $(x')^T\theta_1, \dots, (x')^T\theta_k$ are distinct. Otherwise, if (without loss of generality) there are $\Sigma_1, \dots, \Sigma_m$ different matrices among $\Sigma_1, \dots, \Sigma_k$, then for every $x' \notin \cup_{1\le i\le j\le m}\{u \in \mathbb{R}^d : u^T(\Sigma_i - \Sigma_j)u = 0\}$, we have $(x')^T\Sigma_1(x'), \dots, (x')^T\Sigma_m(x')$ are distinct. In both cases, we find a finite collection of hyperplanes and cones such that for every x' not belongs to any set of this collection, we have $((x')^T\theta_1, (x')^T\Sigma_1(x')), \dots, ((x')^T\theta_k, (x')^T\Sigma_k(x'))$ are pairwise different. Note that because the union of these collection of $(d-1)$ dimensional manifolds can not be $\mathbb{R}^d$, such a non-zero x' exists. Now we only consider x belongs to the one-dimensional linear space spanned by this x', i.e., $x = y(x')$, where $y \in \mathbb{R}$. Denote by

$$a_j = (x')^T\gamma_j' x',\ b_j = [(\beta_j')^T - 2\theta_j^T\gamma_j']x',\ c_j = \theta_j^T\gamma_j'\gamma_j - (\beta_j')^T\theta_j,$$
$$d_j = (x')^T\Sigma_j^{-1}x',\ e_j = (x')^T\Sigma_j^{-1}\theta_j',\ f_i = \theta_j^T\Sigma_j^{-1}\theta_j,$$

for $j = 1, \dots, k$, we proved that $((d_j, e_j))_{j=1}^k$ are distinct. Equation (16) implies that

$$\lambda h_0(yx') + \sum_{j=1}^{k}(\alpha_j' + a_j y^2 + b_j y + c_j)\exp(d_j y^2 + e_j y + f_j) = 0. \quad (17)$$

Case 1. If $-\log h_0(x) \gtrsim \|x\|_2^{\beta_1}$ for some $\beta_1 > 2$ and for all $\|x\|_2 > x_0$, we have $h_0(x) \lesssim \exp^{-\|x\|_2^{\beta_1}}$. Choose $d_{i_1} = \max_{1\le i\le k} d_k$ and $e_{i_2} = \max\{e_j : d_j = d_{j_1}\}$. Because h_0 has a lighter tail than Gaussian and

$$d_j y^2 + e_j y + f_j < d_{i_2}y^2 + e_{i_2}y + f_{i_2}, \quad \forall j \neq i_2,$$

for all y large enough, divide both sides of (17) by $\exp(d_{i_2}y^2 + e_{i_2}y + f_{i_2})$ and let $y \to \infty$, we have $a_{i_2} = b_{i_2} = 0$. It implies that $(x')^T\gamma_{i_2}'x' = [(\beta_{i_2}')^T - 2\theta_{i_2}^T\gamma_{i_2}']x' = 0$. If $\gamma_{i_2}' \neq 0$ then we can further choose x' outside a cone such that $(x')^T\gamma_{i_2}'x' \neq 0$. Hence, $\gamma_{i_2} = 0$, which implies $(\beta_{i_2}')^T(x') = 0$. If $\beta_{i_2} \neq 0$ then we can further choose x' outside a hyperplane such that $(\beta_{i_2}')^T(x') \neq 0$. Hence, in any case, we can argue so that $\beta_{i_2}' = \theta_{i_2}' = 0$. Put it back to (17), we also have $\alpha_{i_2}' = 0$. Therefore, $\alpha_{i_2} = \beta_{i_2} = \gamma_{i_2} = 0$. Repeat the same argument, notice that the tail of h_0 is lighter than any Gaussian distribution, we have $\alpha_j = \beta_j = \gamma_j = 0$ for all $j = 1, \dots, k$. It finally leads to $\lambda = 0$. Hence, we have the distinguishability of h_0 with family of location-scale Gaussians up to first order.

Case 2. If $-\log h_0(x) \lesssim \|x\|_2^{\beta_2}$ for some $\beta_2 < 2$ and for all $\|x\|_2 > x_0$. We have $p(x|\theta_j, \Sigma_j)/h_0(x) \to 0$ as $x \to \infty$ for all $j = 1, \dots, k$. Therefore, dividing both sides of (16) by $h_0(x)$ and let $x \to \infty$ by some direction, we have $\lambda = 0$. Now proceed to argue similar to Case 1, we also have the distinguishability of h_0 with family of location-scale Gaussians up to first order.

Now we proceed to prove that h_0 is distinguishable with (f, k) up to the any order, for f being family of location Gaussian and any $k > 0$. Arguing similar to above, we only need to work on one-dimensional space. Suppose that there exists λ, $(c_{i,j})_{i=1,\dots,k,j=1,\dots,r}$ such that

$$\lambda h_0(x) + \sum_{i=1}^{k}\sum_{j=0}^{r} c_{i,j}\frac{\partial^j f}{\partial \theta^j}(x|\theta_i, v_i) = 0, \tag{18}$$

where $f(\cdot|\theta, v)$ is the density function of normal distribution with mean θ and variance v, and $(\theta_1, v_1), \dots, (\theta_k, v_k)$ are distinct. We need to prove that $\lambda = c_{i,j} = 0$ for all $i = 1, \dots, k, j = 1, \dots, r$. Calculating the partial derivatives of f, we have

$$\lambda h_0(x) + \sum_{i=1}^{k}\left(\sum_{j=0}^{r}\gamma_{i,j}(x-\theta_i)^j\right)\exp\left(-\frac{(x-\theta_i)^2}{2v_i}\right) = 0, \tag{19}$$

such that $\gamma_{i,j}$ for odd j are linear combination of $(c_{i,l})$ with odd $l \le j$, $\gamma_{i,j}$ for even j are linear combination of $(c_{i,l})$ with even $l \le j$, and one can prove (for example, by induction) that $\gamma_{i,j} = 0 \forall j$ is equivalent to $c_{i,j} = 0 \forall j$. Now we can argue similar to Case 1 and Case 2 above to get the contradiction, with the notice that polynomials grow slower than exponential functions.

(b) Let T be a piecewise linear function with a positive finite number of breakpoints and h_0 is the density function of $N(0, I_d)$ being pushforwarded by T. Argue similar to above, we only need to prove the result in one-dimensional case. In order to prove the distinguishable of h_0 with mixtures of location Gaussians family or mixtures of location-scale Gaussians family, it all boils down to prove that if there exists $\lambda \in \mathbb{R}$ and polynomials $Q_1(x), Q_2(x), \dots, Q_k(x)$ such that

$$\lambda h_0(x) + \sum_{i=1}^{k} Q_i(x) f(x|\theta_i, v_i^2) = 0, \tag{20}$$

where $(\theta_1, v_1^2), \dots, (\theta_k, v_k^2)$ are distinct, then $\lambda = Q_1(x) = \dots = Q_k(x) = 0$. We will prove this by induction in k. Consider the case $k = 1$, we have

$$\lambda h_0(x) + Q_1(x) f(x|\theta_1, v_1^2) = 0. \tag{21}$$

Because T has finite number of break points, there exists some x_0 large enough so that for all $x > x_0$, T is a linear one-to-one function between $[x_0, \infty)$ and its image. Denote by $T(x) = ax + b$ when $x > x_0$. We can argue that $a \neq 0$, because otherwise the distribution of h_0 will has an atom, which directly leads to distinguishability between h_0 and mixtures of Gaussians. Then, $h_0(x) = f(x|b, a^2)$ and we have

$$\lambda f(x|b, a^2) + Q_1(x) f(x|\theta_1, v_1^2) = 0.$$

Argue similar to part (a), if $(b, a^2) \neq (\theta_1, v_1^2)$, we have $\lambda = Q_1(x) = 0$, which implies the distinguishability. Otherwise, we have $b = \theta_1, a^2 = v_1^2$, and $Q_1(x) = -\lambda$ for all $x \in \mathbb{R}$. We can rewrite (21) as

$$h_0(x) - f(x|\theta_1, v_1^2) = 0.$$

Because h_0 is $N(0, 1)$ being pushforwarded by a piecewise linear function, we can write $\mathbb{R}$ as a partition $(-\infty, c_1], (c_1, c_2], \dots, [c_m, \infty)$ such that each semi-open interval is image of some linear functions of T. Consider a semi-open interval $(c_i, c_{i+i}]$ being image of $T_j(z) = a_j z + b_j$ for $j = 1, \dots, h$, by the change of variable formula for many-to-one map, we have

$$0 = h_0(x) - f(x|\theta_1, v_1^2) = \sum_{j=1}^{h} f(x|b_j, a_j^2) - f(x|\theta_1, v_1^2), \tag{22}$$

for all $x \in (c_i, c_{i+i}]$. Applying Lemma C.1, we have equation (22) is true for all $x \in \mathbb{R}$. Hence, by integrating both side, we get $h = 1$, and then $b_1 = \theta_1, a_1^2 = v_1^2$. Because this is true for all semi-open intervals $(c_i, c_{i+i}]$, we have $T(x) = a_1 x + b_1$ for all $x \in \mathbb{R}$, which is contradict to our assumption that T is non-linear.

Suppose that our inductive hypothesis is correct for $k = n$, now we proceed to prove it is true for $k = n + 1$. If there exists $\lambda \in \mathbb{R}$ and polynomials $Q_1(x), Q_2(x), \ldots, Q_{n+1}(x)$ such that

$$\lambda h_0(x) + \sum_{i=1}^{n+1} Q_i(x) f(x|\theta_i, v_i^2) = 0, \tag{23}$$

where $(\theta_1, v_1), \ldots, (\theta_{n+1}, v_{n+1}^2)$ are distinct. Without loss of generality, assume that $v_1^2 = \max_{1 \le i \le n+1} v_k^2$ and $\theta_1 = \max\{\theta_j : v_j^2 = v_1^2\}$. Because T has finite number of break points, there exists some x_0 large enough so that for all $x > x_0$, T is a linear one-to-one function between $[x_0, \infty)$ and its image. Denote by $T(x) = ax + b$ when $x > x_0$. We have

$$\lambda f(x|b, a^2) + \sum_{i=1}^{n+1} Q_i(x) f(x|\theta_i, v_i^2) = 0, \quad \forall\, x > x_0. \tag{24}$$

If $a^2 > v_1^2$ or $a^2 = v_1^2, b > \theta_1$, divide both sides of equation (24) by $\exp((x-b)/2a^2)$ and let $x \to \infty$, we have $\lambda = 0$ and the conclusion follows from the identifiability of Gaussians family.

If $v_1^2 > a^2$ or $v_1^2 = a^2, \theta_1 > b$, divide both sides of equation (24) by $\exp((x-\theta_1)/2v_1^2)$ and let $x \to \infty$, we have $Q_1(x) = 0$. The problem is back to the case $k = n$ and is proved using the inductive hypothesis.

If $a^2 = v_1^2, b = \theta_1$, divide both sides of equation (24) by $\exp((x-b)/2a^2)$ and let $x \to \infty$, we have $Q_1(x) = -\lambda$ for all $x \in \mathbb{R}$. Hence for x large enough,

$$\sum_{i=2}^{n+1} Q_i(x) f(x|\theta_i, v_i^2) = 0,$$

which implies $Q_2(x) = \cdots = Q_{n+1}(x) = 0$. The problem is back to the case $k = 1$ and is proved using the inductive hypothesis.

The following lemma presents the local identifiability of location-scale Gaussians mixtures.

Lemma C.1. *Denote by $f(\cdot|\theta, \sigma^2)$ the density function of Gaussian distribution with mean θ and variance σ^2. For all $a < b$ and pairs $\{(\theta_i, \sigma_i^2)\}_{i=1}^k$, if there exists $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$ such that*

$$\alpha_1 f(x|\theta_1, \sigma_1^2) + \cdots + \alpha_k f(x|\theta_k, \sigma_k^2) = 0$$

for all $x \in [a, b]$, then

$$\alpha_1 f(x|\theta_1, \sigma_1^2) + \cdots + \alpha_k f(x|\theta_k, \sigma_k^2) = 0, \tag{25}$$

for all $x \in \mathbb{R}$.

Proof. Step 1. (Centralize and normalize coefficients). Suppose that there exists $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$ such that

$$\alpha_1 f(x|\theta_1, \sigma_1^2) + \cdots + \alpha_k f(x|\theta_k, \sigma_k^2) = 0$$

for all $x \in [a, b]$. Denote by $\theta_i' = \theta_i - \dfrac{a+b}{2}$ for all $i = 1, \ldots, k$, then

$$\alpha_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\theta_1')^2}{2\sigma_1^2}\right) + \cdots + \alpha_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\theta_k')^2}{2\sigma_k^2}\right) = 0, \tag{26}$$

for all $x \in [-\frac{b-a}{2}, \frac{b-a}{2}]$. Denote by $\sigma_{i_1} = \min\{\alpha_1, \ldots, \alpha_k\}$. Multiple both sides of (26) by $\exp(-\frac{x^2}{\sigma_{i_1}^2})$, and denote by $s_i^2 = \frac{1}{\sigma_{i_1}^2} - \frac{1}{2\sigma_i^2}, m_i = \theta_i'/\sigma_i^2, \beta_i = \dfrac{1}{\sqrt{2\pi}\sigma_i} \exp(-(\theta_i')^2/2\sigma_i^2)$ for all $i = 1, \ldots, k$, we have

$$\beta_1 \exp\left(s_1^2 x^2 + m_1 x\right) + \cdots + \beta_k \exp\left(s_k^2 x^2 + m_k x\right) = 0, \tag{27}$$

for all $x \in [-\frac{b-a}{2}, \frac{b-a}{2}]$.

Step 2. (Use properties of Laplace transformation). The left-hand side of equation (27) is the Laplace transformation of $\sum_{i=1}^k \beta_i f(x|m_i, s_i^2)$ and is identical to 0 in an open set around 0. Hence

$$\sum_{i=1}^k \beta_i f(x|m_i, s_i^2) = 0,$$

for all $x \in \mathbb{R}$. It implies that

$$\beta_1 \exp\left(s_1^2 x^2 + m_1 x\right) + \cdots + \beta_k \exp\left(s_k^2 x^2 + m_k x\right) = 0,$$

for all $x \in \mathbb{R}$, which is equivalent to equation (25). □

C.2 Proof of Proposition 2.5

If k_σ is the Gaussian kernel with $m > K$, then we get the conclusions as direct consequences of Example 2.3(a). If k_σ is the multivariate Student kernel, then h_0 has a tail heavier than Gaussian tail, so that we get the conclusions as consequences of Proposition 2.4(a).

C.3 Proof of Proposition 2.6

Because T has a finite and postive number of layers, it is a piecewise linear and non-linear function. So the conclusions are direct consequences of Proposition 2.4(b).

C.4 Proof of Theorem 3.1

This result can be obtained by modifying the proof of Theorem 7.4 in [26]. Recall that we defined the function class

$$\overline{\mathcal{P}}_k^{1/2}(\Theta, \epsilon) = \left\{\bar{p}_{\lambda G}^{1/2} : G \in \mathcal{O}_k(\Theta),\ h(\bar{p}_{\lambda G}, p_{\lambda^* G_*}) \le \epsilon\right\}, \tag{28}$$

where for any $G \in \mathcal{O}_K(\Theta)$, we write $\bar{p}_{\lambda G} = (p_{\lambda G} + p_{\lambda^* G_*})/2$, and measure the complexity of this class through the bracketing entropy integral

$$\mathcal{J}_B(\epsilon, \overline{\mathcal{P}}_k^{1/2}(\Theta, \epsilon), \nu) = \int_{\epsilon^2/2^{13}}^{\epsilon} \sqrt{\log N_B(u, \overline{\mathcal{P}}_k^{1/2}(\Theta, u), \nu)}du \vee \epsilon,$$

where $N_B(\epsilon, X, \eta)$ denotes the ϵ-bracketing number of a metric space (X, η) and ν is the Lebesgue measure. We denote by $P_{\lambda G}$ the distribution corresponding to the density $p_{\lambda G}$. The technique to prove this theorem is to bound the convergence rate by the increments of an empirical processes:

$$\nu_n(\lambda G) = \sqrt{n} \int_{\{p_{\lambda^* G_*}\} > 0} \frac{1}{2} \log \frac{\overline{p}_{\lambda G}}{p_{\lambda^* G_*}} d(P_n - P_{\lambda^* G_*}),$$

where $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure ($X_1, \ldots, X_n \overset{iid}{\sim} p_{\lambda^* G_*}$). We first recall Theorem 5.11 in [26] with the notations adapted from our setting:

Theorem C.2. *Let $R > 0$, $k \geq 1$, and $\mathcal{G}$ be a subset of $\mathcal{O}_k(\Theta)$, which contains G_*. Given $C_1 < \infty$, for all C sufficiently large, and for $n \in \mathbb{N}$ and $t > 0$ satisfying*

$$t \le \sqrt{n}((8R) \wedge (C_1 R^2)), \tag{29}$$

and

$$t \ge C^2(C_1 + 1)\left(R \vee \int_{t/(2^6\sqrt{n})}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \overline{\mathcal{P}}_k^{1/2}(\Theta, R), \nu\right) du\right), \tag{30}$$

we have

$$\mathbb{P}_{\lambda^* G_*}\left(\sup_{G \in \mathcal{G}, h(\overline{p}_{\lambda G}, p_{\lambda^* G_*}) \le R} |\nu_n(\lambda G)| \ge t\right) \le C \exp\left(-\frac{t^2}{C^2(C_1+1)R^2}\right). \tag{31}$$

Now we proceed to prove Theorem 3.1, the proof is divided into three parts: Bounding the tail probability of $h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*})$ by sums of empirical processes increments using chaining technique, bounding the empirical processes increments using Theorem C.2, and bounding the expectation of $h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*})$ using its tail probability.

Step 1 (Bounding the tail probability $h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*})$ by sums of empirical processes increments): Firstly, by Lemma 4.1 and 4.2 of [26], we have

$$\frac{1}{16} h^2(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) \leq h^2(\overline{p}_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) \leq \frac{1}{\sqrt{n}} \nu_n(\hat{\lambda}_n \hat{G}_n).$$

Hence, for any $\delta > \delta_n := (\log n/n)^{1/2}$, we have

$$\begin{aligned}
\mathbb{P}_{\lambda^* G_*}(h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) \geq \delta) &\leq \mathbb{P}_{\lambda^* G_*} \Bigg(\nu_n(\hat{\lambda}_n \hat{G}_n) - \sqrt{n} h^2(\overline{p}_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) \geq 0, \\
&\qquad\qquad\qquad\qquad h(\overline{p}_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) \geq \delta/4 \Bigg) \\
&\leq \mathbb{P}_{\lambda^* G_*} \left(\sup_{\lambda, G : h(\overline{p}_{\lambda G}, p_{\lambda^* G_*}) \geq \delta/4} [\nu_n(\lambda G) - \sqrt{n} h^2(\overline{p}_{\lambda G}, p_{\lambda^* G_*})] \geq 0 \right) \\
&\leq \sum_{s=0}^{S} \mathbb{P}_{\lambda^* G_*} \left(\sup_{\lambda, G : 2^s \delta/4 \leq h(\overline{p}_{\lambda G}, p_{\lambda^* G_*}) \leq 2^{s+1}\delta/4} |\nu_n(\lambda G)| \geq \sqrt{n} 2^{2s} (\delta/4)^2 \right) \\
&\leq \sum_{s=0}^{S} \mathbb{P}_{\lambda^* G_*} \left(\sup_{\lambda, G : h(\overline{p}_{\lambda G}, p_{\lambda^* G_*}) \leq 2^{s+1}\delta/4} |\nu_n(\lambda G)| \geq \sqrt{n} 2^{2s} (\delta/4)^2 \right),
\end{aligned}$$

where S is a smallest number such that $2^S \delta/4 > 1$, as $h(\overline{p}_{\lambda G}, p_{\lambda^* G_*}) \leq 1$. Now we will bound the each term above using Theorem C.2.

Step 2 (Bounding the empirical processes increments using Theorem C.2): In Theorem C.2, choose $R = 2^{s+1}\delta, C_1 = 15$ and $t = \sqrt{n} 2^{2s} (\delta/4)^2$, we can readily check that Condition (29) satisfies (because $2^{s-1}\delta/4 \leq 1$ for all $s = 0, \dots, S$). Condition (30) satisfies thanks to Assumption A3:

$$\begin{aligned}
\int_{t/(2^6\sqrt{n})}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \mathcal{P}_k^{1/2}(\Theta, R), \nu \right) du \vee 2^{s+1}\delta &= \sqrt{2} \int_{R^2/2^{13}}^{R/\sqrt{2}} H_B^{1/2}\left(u, \mathcal{P}_k^{1/2}(\Theta, R), \nu \right) du \vee 2^{s+1}\delta \\
&\leq 2 \mathcal{J}_B(R, \mathcal{P}^{1/2}(\Theta, R), \nu) \\
&\leq 2J\sqrt{n} 2^{2s+1} \delta^2 = 2^6 J t.
\end{aligned}$$

So the conclusion of Theorem C.2 gives us

$$\mathbb{P}_{\lambda^* G_*}(h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) > \delta) \leq C \sum_{s=0}^{\infty} \exp\left(\frac{2^{2s} n \delta^2}{J^2 2^{14}} \right) \leq c \exp\left(\frac{n\delta^2}{c^2} \right), \tag{32}$$

where c is a large constants that does not depend on λ^*, G_*.

Step 3 (Implying the bound on supremum of expectation): Thus, we have

$$\mathbb{E} h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) = \int_0^{\infty} \mathbb{P}(h(p_{\hat{\lambda}_n \hat{G}_n}, p_{\lambda^* G_*}) > \delta) d\delta \leq \delta_n + c \int_{\delta_n}^{\infty} \exp\left(-\frac{n\delta^2}{c^2} \right) \leq \tilde{c}\delta_n,$$

for some $\tilde{c}$ does not depend on λ^*, G_*. Hence, we finally proved that

$$\sup_{G_* \in \mathcal{O}_k(\Theta), \lambda^* \in [0,1]} \mathbb{E}_{\lambda^*, G_*} h(p_{\widehat{\lambda}_n \widehat{G}_n}, p_{\lambda^* G_*}) \leq C\sqrt{\log n/n}.$$

As a consequence, we obtain the conclusion of the theorem.

D Proof of Section 3

D.1 Proof of Proposition 3.2

We first need to denote some notations that are required for the proof. Those notations are well-known in Empirical Processes field [26]. Denote by

$$\mathcal{P}_k(\Theta) = \{p_{\lambda G} : \lambda \in [0, 1], G \in \mathcal{O}_k(\Theta)\},$$

and let $N(\epsilon, \mathcal{P}_k(\Theta), \|\cdot\|_\infty)$ be the $\epsilon-$covering number of $(\mathcal{P}_k(\Theta, \|\cdot\|_\infty)$ and $N_B(\epsilon, \mathcal{P}_k(\Theta), h)$ be the bracketing number of $\mathcal{P}_k(\Theta)$ measured by Hellinger metric h. $H_B(\epsilon, \mathcal{P}_k(\Theta), h) = \log N_B(\epsilon, \mathcal{P}_k(\Theta), h)$ is called the bracketing entropy of $\mathcal{P}_k(\Theta)$ under metric h. Let $\overline{\mathcal{P}}_k(\Theta) = \{(p_{\lambda G} + p_{\lambda^* G_*})/2 : \lambda \in [0,1], G \in \mathcal{O}_k(\Theta)\}$ and $\overline{\mathcal{P}}_k^{1/2}(\Theta) = \{p^{1/2} : p \in \overline{\mathcal{P}}_k(\Theta)\}$. We want to show that

$$\mathcal{J}_B(\epsilon, \overline{\mathcal{P}_k}^{1/2}(\Theta, \epsilon), L^2(\mu)) = \left(\int_{\epsilon^2/2^{2^{13}}}^{\epsilon} H_B^{1/2}(\delta, \overline{\mathcal{P}_k}^{1/2}(\Theta, \delta), \nu) d\delta \vee \delta\right) \lesssim \sqrt{n}\epsilon^2, \tag{33}$$

for all $n > N$ large enough and $\epsilon > (\log n/n)^{1/2}$. We proceed to show that claim (33) will be proved if

$$\log N(\epsilon, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\epsilon), \tag{34}$$

$$H_B(\epsilon, \mathcal{P}_k(\Theta), h) \lesssim \log(1/\epsilon), \tag{35}$$

and then prove claim (34) and (35).

Proof of that claim (35) **implies claim** (33) Because $\overline{\mathcal{P}}_k^{1/2}(\Theta, \delta) \subset \overline{\mathcal{P}}_k^{1/2}(\Theta)$ and from the definition of Hellinger distance,

$$H_B(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta, \delta), \mu) \leq H_B(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta), \mu) = H_B(\frac{\delta}{\sqrt{2}}, \overline{\mathcal{P}}_k(\Theta), h).$$

Now use the fact that for densities f_*, f_1, f_2, we have $h^2((f_1 + f_*)/2, (f_2 + f_*)/2) \leq h^2(f_1, f_2)/2$, oen can readily check that $H_B(\frac{\delta}{\sqrt{2}}, \overline{\mathcal{P}}_k(\Theta), h) \leq H_B(\delta, \mathcal{P}_k(\Theta), h)$. Hence, if claim (35) holds true, then

$$H_B(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta, \delta), \mu) \leq H_B(\delta, \mathcal{P}_k(\Theta), h) \lesssim \log(1/\delta),$$

which implies that

$$\mathcal{J}_B(\epsilon, \overline{\mathcal{P}}_k^{1/2}(\Theta, \delta), \mu) \lesssim \epsilon(\log(2^{13}/\epsilon^2))^{1/2} < n\epsilon^2,$$

for all $\epsilon > (\log n/n)^{1/2}$. Hence, claim (33) is proved.

Proof of claim (34) By invoking the proof of Lemma 2.1. of [16], we have a ϵ-net $\mathcal{S}$ for $(\{p_G : G \in \mathcal{O}_k(\Theta), h\})$ with the cardinality being bounded as follows

$$|\mathcal{S}| \leq \left(\frac{2d\overline{\lambda}}{\epsilon}\right)^{d(d+1)k/2} \times \left(\frac{2a}{\epsilon}\right)^{dk} \left(\frac{5}{\epsilon}\right)^{k}.$$

Denote by $\mathcal{G}$ the set of latent mixing measures G in that net. Let $\mathcal{S}_0$ be an $\epsilon-$net in $[0,1]$ for λ, it is seen that $|\mathcal{S}_0| \leq 1/\epsilon$. Now we form a net for $\mathcal{P}_k(\Theta)$ by $\{p_{\lambda G} : \lambda \in \mathcal{S}_0, G \in \mathcal{G}\}$. Hence, for any λ, G, there exists $\tilde{\lambda} \in \mathcal{S}_0, G \in \mathcal{G}$ such that

$$|\lambda - \tilde{\lambda}| \leq \epsilon, \|p_G - p_{\tilde{G}}\|_\infty \leq \epsilon.$$

This implies

$$\begin{aligned}\left\|p_{\lambda G} - p_{\tilde{\lambda}\tilde{G}}\right\|_\infty &\leq \left\|p_{\lambda G} - p_{\tilde{\lambda} G}\right\|_\infty + \left\|p_{\tilde{\lambda} G} - p_{\tilde{\lambda}\tilde{G}}\right\|_\infty \\ &\leq |\lambda - \tilde{\lambda}|(\|h_0\|_\infty + \|p_G\|_\infty) + \tilde{\lambda}\left\|p_G - p_{\tilde{G}}\right\|_\infty \\ &\leq \epsilon\left(\|h_0\|_\infty + \frac{1}{(\sqrt{2\pi}\underline{\lambda})^{d/2}}\right) + \epsilon \\ &\lesssim \epsilon.\end{aligned}$$

Hence, we get an $\epsilon-$net for $\mathcal{P}_k(\Theta)$ with the cardinality less than or equal

$$|\mathcal{S}_0| \times |\mathcal{S}| = \frac{1}{\epsilon} \times \left(\frac{2d\overline{\lambda}}{\epsilon}\right)^{d(d+1)k/2} \times \left(\frac{2a}{\epsilon}\right)^{dk} \left(\frac{5}{\epsilon}\right)^{k}.$$

Thus,

$$\log N(\epsilon, \mathcal{P}_k(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\epsilon).$$

Proof of claim (35) Now, from the entropy number to get the bracketing number, we let $\eta \le \epsilon$ which will be chosen later. Let $f_1, \dots, f_N$ be a η-net for $\mathcal{P}_k(\Theta)$. We have

$$(x-\theta)^T\Sigma^{-1}(x-\theta) \ge \frac{\|x-\theta\|_2^2}{\bar{\lambda}} \ge \frac{\|x\|_2^2}{4\bar{\lambda}}, \quad \forall\ \|x\| \ge 2\sqrt{d}a, (\theta, \Sigma) \in \Theta, \tag{36}$$

Moreover, h_0 has an exponential tail $-\log h_0(x) \gtrsim \|x\|_2^\beta$ for some $\beta > 0$, and $\|h_0\|_\infty < C$ for some constant C. Therefore, if we let $\beta' = \min\{\beta, 2\} > 0$ and $C' = \max\left\{C, \dfrac{1}{(2\pi)^{d/2}\underline{\lambda}^d}\right\}$, then

$$H(x) = \begin{cases} C_1 \exp(-\|x\|_2^{\beta'}), & \|x\|_2 \ge B_1, \\ C', & \text{otherwise} \end{cases} \tag{37}$$

is an envelop for $\mathcal{P}_k(\Theta)$, where C_1 depends only on $\underline{\lambda}$ and h_0, B_1 depends on $a, \bar{\lambda}, h_0$. We can construct brackets $[p_i^L, p_i^U]$ as follows.

$$p_i^L(x) = \max\{f_i(x) - \eta, 0\}, p_i^U(x) = \min\{f_i(x) + \eta, H(x)\}. \tag{38}$$

Because for each $f \in \mathcal{P}_k(\Theta)$, there is f_i such that $\|f - f_i\|_\infty < \eta$, therefore $p_i^L \le f \le p_i^U$. Moreover, for any $B \ge B_1$,

$$\begin{aligned}\int_{\mathbb{R}^d} (p_i^U - p_i^L) d\mu &\le \int_{\|x\|_2 \le B} 2\eta dx + \int_{\|x\|_2 \ge B} H(x) dx \\ &\lesssim \eta B^d + B^d \exp\left(-B^{\beta'}\right),\end{aligned} \tag{39}$$

where we use spherical coordinate to have

$$\int_{\|x\| \le B} dx = \frac{\pi^{d/2}}{\Gamma(d/2+1)} B^d \lesssim B^d,$$

and

$$\begin{aligned}\int_{\|x\| \ge B} \exp\left(-\|x\|_2^{\beta'}\right) &\lesssim \int_{r \ge B} r^{d-1} \exp\left(-r^{\beta'}\right) dr \\ &= \frac{1}{\beta} \int_{B^{\beta'}}^\infty u^{d/\beta'-1} \exp(-u) du \quad (\text{change of variable } u = r^{\beta'}) \\ &\le \frac{1}{\beta'} B^{d-\beta'} \exp(-B^{\beta'}),\end{aligned}$$

in which the last step we use the inequality (with change of variable formula)

$$\int_z^\infty u^{d/\beta-1} e^{-u} du = z^{d/\beta} e^{-z} \int_0^\infty (1+s)^{d/\beta-1} e^{-zs} ds \le z^{d/\beta} e^{-z} \frac{1}{z - d/\beta + 1} < z^{d/\beta} e^{-z}, \tag{40}$$

whenever $z > d/\beta'$, and we use $z = B^{\beta'}$. Hence, in (39), if we choose $B = B_1(\log(1/\eta))^{1/\beta'}$ then

$$\int_{\mathbb{R}^d} (p_i^U - p_i^L) d\mu \lesssim \eta \left(\log\left(\frac{1}{\eta}\right)\right)^{d/\beta'}. \tag{41}$$

Therefore, there exists a positive constant c which does not depend on η such that

$$H_B(c\eta \log(1/\eta)^{d/\beta'}, \mathcal{P}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\eta).$$

Let $\epsilon = c\eta(\log(1/\eta))^{d/\beta'}$, we have $\log(1/\epsilon) \asymp \log(1/\eta)$, which combines with inequality $\|\cdot\|_1 \le h^2$ leads to

$$H_B(\epsilon, \mathcal{P}_k(\Theta), h) \le H_B(\epsilon^2, \mathcal{P}_k(\Theta), \|\cdot\|_1) \lesssim \log(1/\epsilon^2) \lesssim \log(1/\epsilon).$$

Thus, we have proved claim (35).

We put a remark here that the technique in this proof can be generalized for any family of $f(x|\theta)$ that have sub-exponential tails, i.e. $f(x|\theta) \lesssim \exp(-\|x\|^\gamma)$ for all x large enough and $\gamma > 0$. We can substitute this condition into equation (36), then proceed to continue the proof similarly.

Next, we provide proofs for inverse bounds in Section 3 of the paper. Because there are several results with the same spirit in this section, to make it easy for reader, we recall each result before proving it.

D.2 Proof of Theorem 3.3

Theorem 3.3. Assume that k_* is known, f is first order identifiable and (f, k_*) is distinguishable from h_0. Then, for any $G \in \mathcal{E}_{k_*}(\Theta)$, there exist positive constant C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following holds:

(a) When $\lambda^* = 0$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \lambda$.

(b) When $\lambda^* \in (0, 1]$, then
$$V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \underbrace{[|\lambda - \lambda^*| + (\lambda + \lambda^*) W_1(G, G_*)]}_{\overline{W}_1(\lambda G, \lambda^* G_*)}.$$

We first provide the proof of the theorem for the setting $\lambda^* \in (0, 1]$ in Section D.2.1. Then, the proof for the setting $\lambda^* = 0$ is presented in Section D.2.2.

D.2.1 Proof of setting $\lambda^* \in (0, 1]$

Recall that, we define $\overline{W}_1(\lambda G, \lambda^* G_*) := |\lambda - \lambda^*| + (\lambda + \lambda^*) W_1(G, G_*)$. Besides that, $G_* = \sum_{i=1}^{k_*} p_i^* \delta_{\theta_i^*}$. In order to obtain the proof of the theorem for the setting $\lambda^* \in (0, 1]$, it is sufficient to verify the following two claims:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{E}_{k_*}(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\overline{W}_1(\lambda G, \lambda^* G_*)} : \overline{W}_1(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0, \tag{42}$$

$$\inf_{\lambda \in [0,1], G \in \mathcal{E}_{k_*(\Theta)} : \overline{W}_1(\lambda G, \lambda^* G_*) > \epsilon'} \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\overline{W}_1(\lambda G, \lambda^* G_*)} > 0, \tag{43}$$

for any $\epsilon' > 0$.

Proof of claim (42)**:** Assume that claim (42) does not hold. It indicates that there exists a sequence of probability measures $G_n \in \mathcal{E}_{k_*}(\Theta)$ and a sequence of $\lambda_n \in [0, 1]$ such that $\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$. Therefore, we have $\lambda_n \to \lambda^*$ and $W_1(G_n, G_*) \to 0$ as $n \to \infty$. We can relabel the atoms and weights of G_n such that it admits the following form:

$$G_n = \sum_{i=1}^{k_*} p_i^n \delta_{\theta_i^n}, \tag{44}$$

where $p_i^n \to p_i^*$ and $\theta_i^n \to \theta_i^*$ for all $i \in [k_*]$. To ease the ensuing presentation, we denote $\Delta\theta_i^n := \theta_i^n - \theta_i^*$ and $\Delta p_i^n := p_i^n - p_i^*$ for $i \in [k_*]$. Then, using the coupling between G_n and G_* such that it put mass $\min\{p_i^n, p_i^*\}$ on $\delta_{(\theta_i^n, \theta_i^*)}$, we can verify that

$$W_1(G_n, G_*) \asymp \sum_{i=1}^{k_*} |\Delta p_i^n| + p_i^n \left\| \Delta\theta_i^n \right\|_2. \tag{45}$$

Our proof is divided into three steps.

Step 1 - Taylor expansion: Invoking Taylor expansion up to the first order, we find that

$$f(x|\theta_i^n) = f(x|\theta_i^*) + (\Delta\theta_i^n)^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) + R_i(x),$$

where $R_i(x)$ is Taylor remainder such that $R_i(x) = o(\|\Delta\theta_i^n\|_2)$ for $i \in [k_*]$. Given the above expressions, we obtain that

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = (\lambda^* - \lambda_n) h_0(x) + \sum_{i=1}^{k_*} \left(\lambda_n p_i^n - \lambda^* p_i^*\right) f(x|\theta_i^*) + \lambda_n p_i^n \left(\Delta\theta_i^n\right)^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) + R(x), \tag{46}$$

where $R(x) = \lambda_n \sum_{i=1}^n p_i^n R_i(x) = o\left(\lambda_n \sum_{i=1}^{k_*} p_i^n \left\|\Delta\theta_i^n\right\|_2\right)$. From the expression of $W_1(G_n, G_*)$ in (45), we have $R(x)/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$ for all x.

Step 2 - Non-vanishing coefficients: From equation (46), we can represent the ratio $(p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x))/\overline{W}_1(\lambda_n G_n, \lambda^* G_*)$ as a linear combination of elements of $h_0(x)$, $f(x|\theta_i^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_i^*)$ for $i \in [k_*]$. Assume that all of the coefficients associated with these terms go to 0 as $n \to \infty$. As the coefficient with $h_0(x)$ goes to 0, we obtain that $(\lambda^* - \lambda_n)/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$. Furthermore, the coefficients of $f(x|\theta_i^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_i^*)$ vanish to 0 are equivalent to the following limits

$$(\lambda_n p_i^n - \lambda^* p_i^*)/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0, \quad p_i^n \left\|\Delta \theta_i^n\right\|_2 /\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0.$$

As we have $(\lambda^* - \lambda_n)/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0$, the above limits lead to

$$\lambda^* (\Delta p_i^n)/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0.$$

Putting the above results together, we obtain $1 = \overline{W}_1(\lambda_n G_n, \lambda^* G_*)/\overline{W}_1(\lambda_n G_n, \lambda^* G_*) \to 0$, which is a contraction. As a consequence, not all the coefficients of $h_0(x)$, $f(x|\theta_i^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_i^*)$ go to 0 for $i \in [k_*]$.

Step 3: Show the contradiction using the distinguishability condition and Fatou's lemma: Denote m_n as the maximum of the absolute values of the coefficients of $h_0(x)$, $f(x|\theta_i^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_i^*)$ as $i \in [k_*]$. Since not all of these coefficients vanish to 0, we have $m_n \not\to 0$ as $n \to \infty$. Therefore, $d_n = 1/m_n \not\to \infty$ as $n \to \infty$. Given the previous results, there exist $\alpha_0, \alpha_1, \ldots, \alpha_{k_*}$ and $\beta_1, \ldots, \beta_{k_*}$ such that not all of them are 0 and the following limit holds:

$$d_n \cdot \frac{p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x)}{\overline{W}_1(\lambda_n G_n, \lambda^* G_*)} \to \alpha_0 h_0(x) + \sum_{i=1}^{k_*} \alpha_i f(x|\theta_i^*) + \beta_i^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*).$$

By means of Fatou's lemma, we have

$$\begin{aligned} 0 = \lim_{n\to\infty} d_n \cdot \frac{V(p_{\lambda_n G_n}, p_{\lambda^* G_*})}{\overline{W}_1(\lambda_n G_n, \lambda^* G_*)} &\geq \int \liminf_{n\to\infty} d_n \cdot \frac{p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x)}{\overline{W}_1(\lambda_n G_n, \lambda^* G_*)} dx, \\ &= \int \left(\alpha_0 h_0(x) + \sum_{i=1}^{k_*} \alpha_i f(x|\theta_i^*) + \beta_i^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) \right) dx. \quad (47) \end{aligned}$$

The above equation indicates that

$$\alpha_0 h_0(x) + \sum_{i=1}^{k_*} \alpha_i f(x|\theta_i^*) + \beta_i^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) = 0,$$

for almost surely x. Since (f, k_*) is distinguishable from h_0 and f is first order identifiable, the above equation suggests that $\alpha_0 = \alpha_1 = \ldots = \alpha_{k_*} = 0$ and $\beta_1 = \ldots = \beta_{k_*} = \mathbf{0}$, which is a contradiction.

As a consequence, we achieve the conclusion of claim (42).

Proof of claim (43) Similar to the proof of claim (42), we also prove claim (43) by contradiction. Assume that claim (43) does not hold. It implies that we can find sequences $\lambda_n' \in [0,1]$ and $G_n' \in \mathcal{E}_{k_*}(\Theta)$ such that $\overline{W}_1(\lambda_n' G_n', \lambda^* G_*) > \epsilon'$ and $V(p_{\lambda_n' G_n'}, p_{\lambda^* G_*})/\overline{W}_1(\lambda_n' G_n', \lambda^* G_*) \to 0$ as $n \to \infty$. Since $[0,1]$ and Θ are bounded sets, there exist $\lambda' \in [0,1]$ and $G' \in \mathcal{E}_{k_*}(\Theta)$ such that $\lambda_n' \to \lambda'$ and $W_1(G_n', G') \to 0$ as $n \to \infty$. Since $\overline{W}_1(\lambda_n' G_n', \lambda^* G_*) > \epsilon'$ for all n, the previous limits indicate that $\overline{W}_1(\lambda' G', \lambda^* G_*) \geq \epsilon'$.

On the other hand, since $V(p_{\lambda_n' G_n'}, p_{\lambda^* G_*})/\overline{W}_1(\lambda_n' G_n', \lambda^* G_*) \to 0$, we have $V(p_{\lambda_n' G_n'}, p_{\lambda^* G_*}) \to 0$ as $n \to \infty$. An application of Fatou's lemma leads to

$$0 = \lim_{n\to\infty} V(p_{\lambda_n' G_n'}, p_{\lambda^* G_*}) \geq \frac{1}{2} \int \liminf_{n\to\infty} \left| p_{\lambda_n' G_n'}(x) - p_{\lambda^* G_*}(x) \right| dx = V(p_{\lambda' G', \lambda^* G_*}).$$

Due to the identifiability of model (1), the above equation leads to $(\lambda', G') \equiv (\lambda^*, G_*)$, which is a contradiction to the condition that $\overline{W}_1(\lambda' G', \lambda^* G_*) \geq \epsilon'$. As a consequence, we achieve the conclusion of claim (43).

D.2.2 Proof of setting $\lambda^* = 0$

We want to show that

$$\inf_{G \in \mathcal{E}_{k_*}(\Theta)} \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\lambda} > 0 \tag{48}$$

Proof of claim (48)**:** Assume that claim (48) does not hold. We can find two sequences $\bar{\lambda}_n \in [0,1]$ and $\bar{G}_n \in \mathcal{E}_{k_*}(\Theta)$ such that $V(p_{\bar{\lambda}_n \bar{G}_n}, p_{\lambda^* G_*})/\bar{\lambda}_n \to 0$ as $n \to \infty$. We denote $\bar{G}_n = \sum_{i=1}^{k_*} \bar{p}_i^n \delta_{\bar{\theta}_i^n}$. Since Θ is a bounded set, there exists $\bar{G} = \sum_{i=1}^{k_*} \bar{p}_i \delta_{\bar{\theta}_i} \in \mathcal{E}_{k_*}(\Theta)$ such that $W_1(\bar{G}_n, \bar{G}) \to 0$ as $n \to \infty$. Invoking Fatou's lemma, we obtain that

$$\begin{aligned} 0 = \lim_{n \to \infty} \frac{V(p_{\bar{\lambda}_n \bar{G}_n}, p_{\lambda^* G_*})}{\lambda_n} &\geq \frac{1}{2} \int \liminf_{n \to \infty} \left| \sum_{i=1}^{k_*} \bar{p}_i^n f(x|\bar{\theta}_i^n) - h_0(x) \right| dx \\ &= V\left(\sum_{i=1}^{k_*} \bar{p}_i f(.|\bar{\theta}_i), h_0(.) \right). \end{aligned}$$

The above equation shows that $\sum_{i=1}^{k_*} \bar{p}_i f(x|\bar{\theta}_i) = h_0(x)$ for almost surely x, which is a contradiction to the hypothesis that (f, k_*) is distinguishable from h_0. Hence, we reach the conclusion of claim (48).

D.3 Proof of Theorem 3.4

Theorem 3.4. Assume that k_* is unknown and strictly upper bounded by a given K. Besides that, f is second order identifiable and (f, K) is distinguishable from h_0. Then, for any $G \in \mathcal{O}_K(\Theta)$, there exist positive constant C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following holds:

(a) When $\lambda^* = 0$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \lambda$.

(b) When $\lambda^* \in (0, 1]$, then

$$V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \underbrace{\left[|\lambda - \lambda^*| + (\lambda + \lambda^*) W_2^2(G, G_*) \right]}_{\overline{W}_2(\lambda G, \lambda^* G_*)}.$$

The proof argument for the setting $\lambda^* = 0$ is similar to that in Section D.2.2; therefore, it is omitted. We focus only on the proof of the setting $\lambda^* \in (0, 1]$.

Similar to the proof of Theorem 3.3, in order to reach the conclusion of Theorem 3.4 for the setting $\lambda^* \in (0, 1]$, it is sufficient to demonstrate the following claims:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{O}_K(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\overline{W}_2(\lambda G, \lambda^* G_*)} : \overline{W}_2(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0, \tag{49}$$

$$\inf_{\lambda \in [0,1], G \in \mathcal{O}_{K(\Theta)}: \overline{W}_2(\lambda G, \lambda^* G_*) > \epsilon'} \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\overline{W}_2(\lambda G, \lambda^* G_*)} > 0,$$

for any $\epsilon' > 0$. Since the proof of the second claim is similar to that of claim (43) in Section D.2; therefore, it is omitted.

Proof of claim (49)**:** Similar to the proof of claim (42), we use proof by contradiction for claim (49). Assume that claim (49) does not hold. Given that assumption, we can find sequences $G_n \in \mathcal{O}_K(\Theta)$ and $\lambda_n \in [0,1]$ such that $\overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/\overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$. As $W_2(G_n, G_*) \to 0$ as $n \to \infty$, using the similar argument as that in Section 3.2 in Ho et al. [18], we can find a subsequence of G_n (without loss of generality, we replace that subsequence by the whole sequence of G_n with $k' \in [k_*, K]$ supports such that

$$G_n = \sum_{i=1}^{k_* + \bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \delta_{\theta_{ij}^n}, \tag{50}$$

where $\sum_{j=1}^{s_i} p_{ij}^n \to p_i^*$ and $\theta_{ij}^n \to \theta_i^*$ for all $i \in [k_* + \bar{l}]$. Here, $p_i^* = 0$ for $k_* + 1 \leq i \leq k_* + \bar{l}$. In addition, $s_1, \ldots, s_{k_*+\bar{l}} \geq 1$ are such that $\sum_{i=1}^{k_*+\bar{l}} s_i = k'$. To ease the ensuing presentation, we denote $\Delta\theta_{ij}^n := \theta_{ij}^n - \theta_i^*$ and $\Delta p_{i.}^n := \sum_{j=1}^{s_i} p_{ij}^n - p_i^*$ for $i \in [k_* + \bar{l}]$. Then, based on Lemma 3.1 in Ho et al. [18], we have

$$W_2^2(G_n, G_*) \asymp \sum_{i=1}^{k_*+\bar{l}} |\Delta p_{i.}^n| + \sum_{i=1}^{k_*+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2. \tag{51}$$

We divide our proof of claim (49) into three steps.

Step 1 - Taylor expansion: An application of Taylor expansion up to the second order leads to

$$f(x|\theta_{ij}^n) = f(x|\theta_i^*) + (\Delta\theta_{ij})^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) + (\Delta\theta_{ij})^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)(\Delta\theta_{ij}) + R_{ij}(x),$$

where $R_{ij}(x)$ is Taylor remainder such that $R_{ij}(x) = o(\|\Delta\theta_{ij}\|_2^2)$ for all $i \in [k_* + \bar{l}]$ and $j \in [s_i]$. Collecting the above equations, we obtain that

$$\begin{aligned} p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) &= (\lambda^* - \lambda_n) h_0(x) + \sum_{i=1}^{k_*+\bar{l}} \left(\sum_{j=1}^{s_i} \lambda_n p_{ij}^n - \lambda^* p_i^* \right) f(x|\theta_i^*) \\ &+ \lambda_n \left(\sum_{j=1}^{s_i} p_{ij}^n \Delta\theta_{ij}^n \right)^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) + \lambda_n \left(\sum_{j=1}^{s_i} p_{ij}^n \left(\Delta\theta_{ij}^n\right)^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)(\Delta\theta_{ij}^n) \right) + R(x), \end{aligned} \tag{52}$$

where $R(x) = \lambda_n \sum_{i=1}^{k_*+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n R_{ij}(x) = o\left(\lambda_n \sum_{i=1}^{k_*+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2\right)$. Given the expression of $W_2^2(G_n, G_*)$ in equation (77), we can verify that $R(x)/\overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$.

Step 2 - Non-vanishing coefficients: Given the expression in equation (52), we can view $(p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x))/\overline{W}_2(\lambda_n G_n, \lambda^* G_*)$ as a linear combination of elements of the forms $h_0(x), f(x|\theta_i^*), \frac{\partial f}{\partial \theta}(x|\theta_i^*)$, and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)$ for all $i \in [k_* + \bar{l}]$. Assume that their coefficients go to 0 as n tends to infinity. As the coefficient of $h_0(x)$ goes to 0, we have $(\lambda_n - \lambda^*)/\overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0$.

Similarly, by learning the coefficients of $f(x|\theta_i^*)$ and $\left[\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)\right]_{jj}$ for $j \in [d]$, we obtain the following limits:

$$\left(\sum_{j=1}^{s_i} \lambda_n p_{ij}^n - \lambda^* p_i^* \right) / \overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0, \quad \lambda_n \left(\sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2 \right) / \overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0.$$

Collecting the above limits, we find that

$$\frac{\lambda^* \Delta p_{i.}^n}{\overline{W}_2(\lambda_n G_n, \lambda^* G_*)} = \frac{(\lambda^* - \lambda_n)\left(\sum_{j=1}^{s_i} p_{ij}^n\right) + \left(\sum_{j=1}^{s_i} \lambda_n p_{ij}^n - \lambda^* p_i^*\right)}{\overline{W}_2(\lambda_n G_n, \lambda^* G_*)} \to 0.$$

Putting the above results together, we achieve that $1 = \overline{W}_2(\lambda_n G_n, \lambda^* G_*)/\overline{W}_2(\lambda_n G_n, \lambda^* G_*) \to 0$, which is a contraction. Therefore, not all the coefficients associated with $h_0(x), f(x|\theta_i^*), \frac{\partial f}{\partial \theta}(x|\theta_i^*)$, and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)$ for $i \in [k_* + \bar{l}]$ go to 0 as n tends to infinity.

Step 3: Show the contradiction using the distinguishability condition and Fatou's lemma: Similar to Step 3 in Section D.2.1, by denoting $d_n = 1/m_n$ where m_n is the maximum values of the absolute values of the coefficients of $h_0(x), f(x|\theta_i^*), \frac{\partial f}{\partial \theta}(x|\theta_i^*)$, and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)$, we have

$$d_n \cdot \frac{p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x)}{\overline{W}_1(\lambda_n G_n, \lambda^* G_*)} \to \alpha_0 h_0(x) + \sum_{i=1}^{k_*+\bar{l}} \alpha_i f(x|\theta_i^*) + \beta_i^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) + \gamma_i^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)\gamma_i,$$

where $\alpha_i, \beta_i, \gamma_i$ are some coefficients such that not all of them are 0. However, the Fatou's lemma suggests that the RHS of the above equation is 0 for almost surely x. Since (f, K) is distinguishable from h_0, it shows that $\alpha_i = 0$, $\beta_i = \mathbf{0} \in \mathbb{R}^d$, and $\gamma_i = \mathbf{0} \in \mathbb{R}^{d\times d}$ for all $i \in [k_* + \bar{l}]$— a contradiction. As a consequence, we obtain the conclusion of claim (49).

D.4 Proof of Theorem 3.5

Theorem 3.5. Assume that k_* is unknown and strictly upper bounded by a given K. Besides that, f is location-scale Gaussian distribution and (f, K) with fixed variance is distinguishable in any order from h_0. Then, for any $G \in \mathcal{O}_K(\Theta)$, there exist positive constant C_1 and C_2 depending only on $\lambda^*, G_*, h_0, \Theta$ such that the following holds:

(a) When $\lambda^* = 0$, then $V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \lambda$.

(b) When $\lambda^* \in (0, 1]$, then

$$V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*).$$

The proof argument for the setting $\lambda^* = 0$ is similar to that in Section D.2.2; therefore, it is omitted. We focus only on the proof of the setting $\lambda^* \in (0, 1]$.

Denote by $\overline{r}_1 = \overline{r}(K - k_*)$. Similar to the proof of Theorem 3.3, in order to reach the conclusion of Theorem 3.5 for the setting $\lambda^* \in (0, 1]$, it is sufficient to demonstrate the following claims:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{O}_K(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\overline{W}_{\overline{r}_1}(\lambda G, \lambda^* G_*)} : \overline{W}_{\overline{r}_1}(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0, \tag{53}$$

$$\inf_{\lambda \in [0,1], G \in \mathcal{O}_{K(\Theta)} : \overline{W}_{\overline{r}_1}(\lambda G, \lambda^* G_*) > \epsilon'} \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{\overline{W}_{\overline{r}_1}(\lambda G, \lambda^* G_*)} > 0,$$

for any $\epsilon' > 0$. Since the proof of the second claim is similar to that of claim (43) in Section D.2; therefore, it is omitted. We now proceed to prove claim (53). Suppose that it is not correct, that is, there exist sequences λ_n and $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_K(\Theta)$ such that $\overline{W}_{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*) \to 0$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*}) / \overline{W}_{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*) \to 0$. For the ease of presentation, we consider the one dimension Gaussian case where $(\mu, \Sigma) = (\theta, v)$, the higher dimension cases are treated similar.

We can use the subsequence argument to have $\lambda^* \geq \lambda_n$ for all n and G_n can be assumed to have a fixed number of atoms k' (less than or equals K) and have a representation as in (54), that is,

$$G_n = \sum_{i=1}^{k_* + \bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{ij}^n, v_{ij}^n)}, \tag{54}$$

where $\sum_{j=1}^{s_i} p_{ij}^n \to p_i^*$ and $\theta_{ij}^n \to \theta_i^*, v_{ij}^n \to v_i^*$ for all $i \in [k_* + \bar{l}]$. Here, $p_i^* = 0$ for $k_* + 1 \leq i \leq k_* + \bar{l}$. In addition, $s_1, \ldots, s_{k_* + \bar{l}} \geq 1$ are such that $\sum_{i=1}^{k_* + \bar{l}} s_i = k'$.

Step 1 - Taylor expansion: Using Taylor expansion of f around $\{(\theta_i^*, v_i^*)\}_{i=1}^{k_*}$ to the $\overline{r}_1-$th order we have

$$\begin{aligned} p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) &= (\lambda^* - \lambda_n) h_0(x) + \lambda_n \Big(\sum_{i=1}^{k_* + \underline{l}} \sum_{j=1}^{s_i} p_{ij}^n f(x | \theta_{ij}^n, v_{ij}^n) \Big) - \sum_{i=1}^{k_*} p_i^* f(x | \theta_i^*, v_i^*) \\ &= (\lambda^* - \lambda_n) h_0(x) + \sum_{i=1}^{k_* + \underline{l}} \sum_{j=1}^{s_i} \lambda_n p_{ij}^n \sum_{|\boldsymbol{\alpha}|=1}^{\overline{r}_1} (\Delta \theta_{ij}^n)^{\alpha_1} (\Delta v_{ij}^n)^{\alpha_2} \frac{1}{\boldsymbol{\alpha}!} \frac{\partial^{|\boldsymbol{\alpha}|} f(\theta_i^*, v_i^*)}{\partial^{\alpha_1} \theta \partial^{\alpha_2} v} \\ &+ \sum_{i=1}^{k_* + \underline{l}} (\Delta p_{i\cdot}^n) f(x | \theta_i^*, v_i^*) + R(x), \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2$, $\boldsymbol{\alpha}! = \alpha_1! \alpha_2!$, $\Delta \overline{p}_{i\cdot}^n = \lambda_n \sum_j p_{ij}^n - p_i^*$, $\Delta \theta_{ij}^n = \theta_{ij}^n - \theta_i^*$, $\Delta v_{ij}^n = v_{ij}^n - v_i^*$ and $R(x) = o(\sum_{i=1}^{k_* + \underline{l}} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta \theta_{ij}^n|^{\overline{r}_1} + |\Delta v_{ij}^n|^{\overline{r}_1}))$. Now we can use the character equation $\frac{\partial^2 f}{\partial \theta^2} = 2\frac{\partial f}{\partial v}$ to rewrite the formula above as

$$(\lambda^* - \lambda_n)h_0(x) + \sum_{\alpha=1}^{2\overline{r}_1} \sum_{i=1}^{k_*+\underline{l}} \left(\sum_{j=1}^{s_i} \lambda_n p_{ij}^n \sum_{n_1,n_2} \frac{(\Delta\theta_{ij}^n)^{n_1}(\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \right) \frac{\partial^\alpha f(\theta_i^*, v_i^*)}{\partial \theta^\alpha}$$
$$+ \sum_{i=1}^{k_*+\underline{l}} (\Delta p_{i\cdot}^n) f(x|\theta_i^*, v_i^*) + R(x), \tag{55}$$

where we sum over n_1, n_2 such that $n_1 + 2n_2 = \alpha, n_1 + n_2 \leq \overline{r}_1$.

Step 2 - Non-vanishing coefficients: Assume that all coefficients in the formula above vanish when dividing by $W_{\overline{r}_1}^{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*)$ when $n \to \infty$. Because

$$W_{\overline{r}_1}^{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*) \asymp |\lambda_n - \lambda^*| + (\lambda_n + \lambda^*) \left(\sum_{i=1}^{k_*+\bar{l}} |\Delta p_{i.}^n| + \sum_{i=1}^{k_*+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n (\left\|\Delta\theta_{ij}^n\right\|_2^{\overline{r}_1} + \left\|\Delta v_{ij}^n\right\|_2^{\overline{r}_1}) \right) := D_{\overline{r}_1}(G_n, G_*), \tag{56}$$

we have

$$\frac{\lambda^* - \lambda_n}{D_{\overline{r}_1}(G_n, G_*)} \to 0, \quad \frac{\Delta p_{i\cdot}^n}{D_{\overline{r}_1}(G_n, G_*)} \to 0. \tag{57}$$

These limits together imply

$$\frac{(\lambda^* + \lambda_n)\Delta p_{i\cdot}^n}{D_{\overline{r}_1}(G_n, G_*)} \to 0, \quad \forall i = 1, \ldots, k_* + \bar{l}.$$

From the definition of $D_{\overline{r}_1}$, it can be deduced that there exists at least an index i^* such that

$$\sum_{j=1}^{s_{i*}} \frac{(\lambda_n + \lambda^*) p_{i^* j}^n ((\theta_{ij}^n)^{\overline{r}_1} + (v_{ij}^n)^{\overline{r}_1})}{D_{\overline{r}_1}(G_n, G_*)} \not\to 0.$$

Without loss of generality, assign $i^* = 1$. But as we assume all the coefficients in equation (55) go to 0 for all α and i, we have

$$\frac{\sum_{j=1}^{s_1} \lambda_n p_{1j}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \leq \overline{r}_1}} \frac{(\theta_{1j}^n)^{n_1}(v_{1j}^n)^{n_2}}{2^{n_2} n_1! n_2!}}{D_{\overline{r}_1}(G_n, G_*)} \to 0,$$

for all $\alpha = 1, \ldots, 2\overline{r}_1$. From two expressions above combining with equation (57), we have for all $\alpha = 1, \ldots, 2\overline{r}_1$,

$$F_\alpha := \frac{\sum_{j=1}^{s_1} p_{1j}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \leq \overline{r}_1}} \frac{(\Delta\theta_{1j}^n)^{n_1}(\Delta v_{1j}^n)^{n_2}}{2^{n_2} n_1! n_2!}}{\sum_{j=1}^{s_1} p_{1j}^n ((\Delta\theta_{ij}^n)^{\overline{r}_1} + (\Delta v_{ij}^n)^{\overline{r}_1})} \to 0. \tag{58}$$

If $s_1 = 1$ then substituting $\alpha = 1$ and $\alpha = 2\overline{r}_1$ gives

$$\frac{|\Delta\theta_{11}^n|^{\overline{r}_1}}{|\Delta\theta_{11}^n|^{\overline{r}_1} + |\Delta v_{11}^n|^{\overline{r}_1}}, \frac{|\Delta v_{11}^n|^{\overline{r}_1}}{|\Delta\theta_{11}^n|^{\overline{r}_1} + |\Delta v_{11}^n|^{\overline{r}_1}} \to 0,$$

which is impossible as they are sum up to 1 for all n. Hence $s_1 \geq 2$. Now we proceed to show the contradiction using the system of equations (6). Denote by $\overline{p}_n = \max_{1 \leq j \leq s_1}\{p_{1j}^n\}, \overline{M}_n = \max_{1 \leq j \leq s_1}\{|\Delta\theta_{1j}^n|, |\Delta v_{1j}^n|^{1/2}\}$. By the subsequence argument in compact sets, without loss of generality, we can denote $c_j^2 := \lim_{n\to\infty} p_{1j}^n / \overline{p}_n$, $a_j = \lim \Delta\theta_{1j}^n / \overline{M}_n$, and $b_j = \lim \Delta v_{1j}^n / \overline{M}_n$ for all $j = 1, \ldots, k_* + \bar{l}$. Because of the definition of $\mathcal{O}_{K,c_0}$, we have $p_{ij}^n \geq c_0$ for all j, which implies all c_j are different from 0 and at least one of them is 1. Similarly, in $(a_j, b_j)_j$, there is at least one of them equals to 1 or -1. Dividing both numerators and denominators of equation (58) by $\overline{p}_n \overline{M}_n^\alpha$, we have

$$\sum_{j=1}^{s_1} \sum_{n_1+2n_2=\alpha} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0,$$

for all $\alpha = 1, \dots, \overline{r}_1$. Hence, we get the contradiction, where we use the fact that $s_1 \leq K - k_* + 1$ (as $s_i \geq 1$ for all $i \geq 2$) and $\overline{r}_1 = \overline{r}(K - k_*)$ is the smallest number such that equation (6), where $k = K - k_*$, has the trivial solution only. Hence, when dividing by $W_{\overline{r}_1}^{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*)$, not all coefficients of equation (55) vanish as $n \to \infty$.

Step 3: Show the contradiction using the distinguishability condition and Fatou's lemma: Denote by

$$E_{i,\alpha} = \sum_{j=1}^{s_i} \lambda_n p_{ij}^n \sum_{n_1, n_2} \frac{(\Delta\theta_{ij}^n)^{n_1} (\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \Big/ W_{\overline{r}_1}^{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*), \quad \forall i, \alpha \geq 1.$$

$$E_{i,0} = \Delta p_{i\cdot}^n \Big/ W_{\overline{r}_1}^{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*), \forall i \geq 1, E_{0,0} = (\lambda^* - \lambda_n) \Big/ W_{\overline{r}_1}^{\overline{r}_1}(\lambda_n G_n, \lambda^* G_*).$$

We have proved that not all $E_{i,\alpha}$ go to 0. Let $d_n = \max_{0 \leq \alpha \leq 2\overline{r}_1, 0 \leq i \leq k'} |E_{i,\alpha}|$. Because $E_{i,\alpha}/d_n \in [-1, 1]$ for all n, by the subsequence argument if needed, we have $E_{i,\alpha}/m_n \to \beta_{i,\alpha}$ as $n \to \infty$, where at least one of the limits are different from 0. But Fatou's argument implies that

$$\beta_{0,0} h_0(x) + \sum_{i=1}^{k_*} \sum_{\alpha=0}^{2\overline{r}_1} \beta_{i,\alpha} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_i^*, v_i^*) = 0,$$

which contradicts our assumption. Hence, claim (53) is proved.

D.5 Proof Theorem A.1

Theorem A.1. Assume that h_0 takes the form (7) and $\lambda^* = 0$. Then, there exist positive constants C_1 and C_2 depending only on h_0, Θ such that the following holds:

(a) (exact-fitted) If f is first order identifiable, then for any $G \in \mathcal{E}_{k_0}(\Theta)$

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_1 \lambda W_1(G, G_0),$$

(b) (over-fitted) If f is second order identifiable, then for any $G \in \mathcal{O}_K(\Theta)$ that $K > k_0$

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \lambda W_2^2(G, G_0),$$

(c) (over-fitted and weakly identifiable) If f is location-scale Gaussian distribution and we further assume that $G_* \in \mathcal{E}_{k_*, c_0}(\Theta)$, then for any $G \in \mathcal{O}_{K, c_0}(\Theta)$ that $K > k_0$, there exists C_3 depends on h_0, Θ_0, c_0 such that

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_3 \lambda W_{\overline{r}(K-k_*)}^{\overline{r}(K-k_*)}(G, G_0)$$

(a) We can write

$$\frac{V(p_0, p_{\lambda G})}{\lambda W_1(G, G_0)} = \int \frac{|\sum_{i=1}^{k_0} p_i^0 f(x|\theta_i^0) - \sum_{i=1}^{k_0} p_i f(x|\theta_i)|}{W_1(G, G_0)} dx$$
$$= \frac{V(p_0, p_G)}{W_1(G, G_0)},$$

because this is the exact-fitted and first-order identifiable, we can apply Theorem 3.1. in Ho et al. [17]

(b) Similar to the last part, we can write

$$\frac{V(p_0, p_{\lambda G})}{\lambda W_2^2(G, G_0)} = \int \frac{|\sum_{i=1}^{k_0} p_i^0 f(x|\theta_i^0) - \sum_{i=1}^{K} p_i f(x|\theta_i)|}{W_2^2(G, G_0)} dx$$
$$= \frac{V(p_0, p_G)}{W_2^2(G, G_0)},$$

as this is the over-fitted and second-order identifiable, we can apply Theorem 3.2. in Ho et al. [17].

(c) Similar to last two cases, we can write

$$\frac{V(p_0, p_{\lambda G})}{\lambda W_{\overline{r}(K-k_*)}^{\overline{r}(K-k_*)}(G, G_0)} = \frac{V(p_0, p_G)}{W_{\overline{r}(K-k_*)}^{\overline{r}(K-k_*)}(G, G_0)},$$

and apply Proposition 2.2. in [16].

D.6 Proof of Theorem 3.6

Theorem 3.6. Assume that h_0 takes the form (7). Besides that, $K \geq k_0$ and f is location-scale Gaussian distribution. Then, for any $\lambda \in [0,1]$ and $G \in \mathcal{O}_{K,c_0}(\Theta)$ for some $c_0 > 0$, there exist positive constants C_1, C_2, C_3, C_4 depending only on $\lambda^*, G_*, G_0, \Theta$ (C_3 and C_4 also depends on δ) such that the following holds:

(a) When $K \leq k_* + k_0 - \bar{k} - 1$, then

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_1 \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*).$$

(b) When $K \geq k_* + k_0 - \bar{k}$, then

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_2 \biggl(1_{\{\lambda \leq \lambda^*\}} \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*) \\ + 1_{\{\lambda > \lambda^*\}} W^{\overline{r}(K-k_*)}_{\overline{r}(K-k_*)}(G, \overline{G}_*(\lambda)) \biggr)$$

(c) For $\delta > 0$, when $K = k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_3 1_{\{\lambda > \lambda^* + \delta\}} W_1(G, \overline{G}_*(\lambda)),$$

and when $K > k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*,G_*}, p_{\lambda,G}) \geq C_4 1_{\{\lambda > \lambda^* + \delta\}} \\ \times W^{\overline{r}(K-k_0-k_*+\bar{k})}_{\overline{r}(K-k_0-k_*+\bar{k})}(G, \overline{G}_*(\lambda)).$$

To facilitate the proof argument, we denote $\mathcal{T} := k_* + k_0 - \bar{k}$. In addition, we assume without loss of generality that $\theta_i^* = \theta_i^0$ for $i \in [\bar{k}]$. Moreover, we introduce the following shorthand:

$$D(\lambda G, \lambda^* G_*) = \begin{cases} \overline{W}_2(\lambda G, \lambda^* G_*), & \text{when } K \leq \mathcal{T} - 1 \\ 1_{\{\lambda \leq \lambda^*\}} \overline{W}_2(\lambda G, \lambda^* G_*) + 1_{\{\lambda > \lambda^*\}}(\lambda + \lambda^*) W_2^2(G, \overline{G}_*(\lambda)), & \text{when } K \geq \mathcal{T} \end{cases}.$$

Similar to the previous proofs, in order to obtain the conclusion of the theorem, we need to prove the following claims:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{O}_K(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{D(\lambda G, \lambda^* G_*)} : D(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0. \tag{59}$$

Proof of claim (59)**:** Assume that the above claim is not true. It indicates that we can find sequences $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_K(\Theta)$ and $\lambda_n \in [0,1]$ such that $D(\lambda_n G_n, \lambda^* G_*)$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*)$ go to 0 as n approaches to infinity. Given the assumption that $\theta_i^* = \theta_i^0$ for $i \in [\bar{k}]$, we obtain that

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = (\lambda^* - \lambda_n) \sum_{i=\bar{k}+1}^{k_0} p_i^0 f(x|\theta_i^0) + \lambda_n \left(\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n) \right) - \sum_{i=1}^{k_*} \bar{p}_i^* f(x|\theta_i^*), \tag{60}$$

where $\bar{p}_i^* = \lambda^* p_i^* + (\lambda_n - \lambda^*) p_i^0$ when $1 \leq i \leq \bar{k}$ and $\bar{p}_i^* = \lambda^* p_i^*$ otherwise. Now, we prove the contradiction of our assumption under two separate settings of λ_n.

Case 1: $\lambda^* \geq \lambda_n$ for infinitely many n. Without loss of generality, we assume that $\lambda^* \geq \lambda_n$ for all $n \geq 1$. Under this case, $D(\lambda_n G_n, \lambda^* G_*) = \overline{W}_2(\lambda_n G_n, \lambda^* G_*)$. As $D(\lambda_n G_n, \lambda^* G_*) \to 0$, we have $\lambda_n \to \lambda^*$ and $W_2(G_n, G_*) \to 0$ as $n \to \infty$. Therefore, we can rewrite G_n like equation (54).

In light of equation (60) and the assumption $\lambda^* \geq \lambda_n$, by means of Taylor expansion up to the second order around $\theta_1^*, \ldots, \theta_{k_*}^*$ as that in the proof of Theorem D.3, we can view $(p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x))/D(\lambda_n G_n, \lambda^* G_*)$ as a linear combination of elements of the forms $f(x|\theta_i^0)$, $f(x|\theta_j^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_j^*)$, and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^*)$ for $\bar{k}+1 \leq i \leq k_0$ and $j \in [k_*]$.

It is sufficient to argue that not all the coefficients of these elements go 0 as the remaining Fatou's argument is similar to Step 3 of the proof of Theorem D.3. Indeed, assume that all of these coefficients go to 0 as n tends to infinity. Since $\bar{k} < k_0$, we always have at least one index $I \in [\bar{k}+1, k_0]$. Studying the coefficient of $f(x|\theta_I^0)$ proves that $(\lambda^* - \lambda_n)/D(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$. From here, with similar argument as in Step 2 of claim (49), we can show that $1 = D(\lambda_n G_n, \lambda^* G_*)/D(\lambda_n G_n, \lambda^* G_*) \to 0$, which is a contradiction. Therefore, we obtain the conclusion of claim (59).

Case 2: $\lambda^* < \lambda_n$ for infinitely many n. Without loss of generality, we assume that $\lambda^* < \lambda_n$ for all $n \geq 1$. Under this case, we can rewrite equation (60) as follows:

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \lambda_n \Bigg(\underbrace{\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n)}_{:=f(x;G_n)} - \underbrace{\Bigg[\bigg(1 - \frac{\lambda^*}{\lambda_n}\bigg) \sum_{i=\bar{k}+1}^{k_0} p_i^0 f(x|\theta_i^0) + \sum_{i=1}^{k_*} \frac{\bar{p}_i^*}{\lambda_n} f(x|\theta_i^*) \Bigg]}_{:=f\left(x;\overline{G}_*(\lambda_n)\right)} \Bigg),$$

where $\overline{G}_*(\lambda_n) := \bigg(1 - \frac{\lambda^*}{\lambda_n}\bigg) G_0 + \frac{\lambda^*}{\lambda_n} G_*$. Under Case 2, $\bar{p}_i^* > \lambda^* p_i^* > 0$ for $i \in [k_*]$. Therefore, we can treat $f(x; G_n)$ and $f\left(x; \overline{G}_*(\lambda_n)\right)$ respectively as mixtures with k_n and $k_0 + k_* - \bar{k}$ elements.

Without loss of generality, we assume $k_n = K$ for all n, namely, the setting where G_n have full K supports. We consider three separate settings of K.

Case 2.1: $K \leq k_* + k_0 - \bar{k} - 1$. Under this case, G_n has fewer supports than $\overline{G}_*(\lambda_n)$. Hence, there always exists one element in the set $\{\theta_i^0 : \bar{k}+1 \leq i \leq k_0\} \cup \{\theta_j^* : 1 \leq j \leq k_*\}$ such that no supports of G_n converge to. We first show that this element cannot belong to the set $\{\theta_j^* : 1 \leq j \leq k_*\}$. Assume by contrary that this element is in that set. Without loss of generality, we assume this element is θ_1^*. Since $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*) \to 0$, we have $f(x; G_n) - f(x; \overline{G}_*(\lambda_n)) \to 0$ for almost surely x. Since θ_i^n do not converge to θ_1^*, the identifiability of f and the previous limit imply that $\bar{p}_1^*/\lambda_n$ goes to 0 as $n \to \infty$, which is a contradiction as $\bar{p}_1^*/\lambda_n > \lambda^* p_1^*$.

Therefore, there exists an element in the set $\{\theta_i^0 : \bar{k}+1 \leq i \leq k_0\}$ such that no elements of G_n converge to. We assume without loss of generality that this element is θ_1^0. In addition, all the elements in the set $\{\theta_j^* : 1 \leq j \leq k_*\}$ have at least one support of G_n converge to. By performing Taylor expansion up to the second order around the limit points of the supports of G_n, we can view $(p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x))/D(\lambda_n G_n, \lambda^* G_*)$ as a linear combination of elements of the forms $f(x|\theta_i^0)$, $f(x|\theta_j^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_i^0)$, $\frac{\partial f}{\partial \theta}(x|\theta_j^*)$, $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0)$, and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^*)$ for some but not all $\bar{k}+1 \leq i \leq k_0$ and for all $j \in [k_*]$. Assume that all of the coefficients associated with these elements go to 0 as n goes to infinity. Since no support of G_n converges to θ_1^0, the previous assumptions mean that $(\lambda_n - \lambda^*)/D(\lambda_n G_n, \lambda^* G_*) \to 0$. Given that result, we have

$$0 = \lim_{n \to \infty} \frac{V(p_{\lambda_n G_n}, p_{\lambda^* G_*})}{D(\lambda_n G_n, \lambda^* G_*)} = \lim_{n \to \infty} \frac{\lambda_n V(f(.; G_n), f(.; G_*))}{(\lambda_n + \lambda^*) W_2^2(G_n, G_*)},$$

which is a contradiction as $V(f(.; G_n), f(.; G_*))/W_2^2(G_n, G_*) \not\to 0$ based on the result of Theorem 3.2 in [17]. Hence, not all the coefficients with $f(x|\theta_i^0)$, $f(x|\theta_j^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_i^0)$, $\frac{\partial f}{\partial \theta}(x|\theta_j^*)$, $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0)$, and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^*)$ go to 0 as $n \to \infty$. From here, invoking the Fatou's argument and the identifiability of f, we conclude the claim (59) under Case 2.1.

Case 2.2: $K \geq k_* + k_0 - \bar{k}$. We see that the number of support points of $\bar{G}_*(\lambda_n)$ decreases to k_* if $\lambda_n \to \lambda^*$ as $n \to \infty$ or keeps being $k_* + k_0 - \bar{k}$ for any subsequence of λ_n does not converge to λ^*. In both cases, we are in the over-fitted setting as $K \geq k_* + k_0 - \bar{k}$. If $\lambda_n \to \lambda^*$, our assumption $W_2(G_n, \bar{G}_*(\lambda_n)) \to 0$ indicates that we can write G_n as in equation (54) so that the atoms of G_n converge to θ_i^* for $i \in [k_*]$ or 0. The proof of claim (59) goes through similar to what of Theorem 3.4 (or Theorem 3.2. in Ho et al. [17]).

If $\lambda_n \not\to \lambda^*$ as $n \to \infty$ then $\bar{G}_*(\lambda_n)$ has $k_0 + k_* - \bar{k}$ in any of its limits. Hence this is over-fitted setting when $K \geq k_* + k_0 - \bar{k}$ and we can proceed similar to above to have claim (59).

Case 2.3: $K = k_* + k_0 - \bar{k}$ and $\lambda_n > \lambda^* + \delta > \lambda^*$ for all n. In this case, $\lambda_n \not\to \lambda^*$, so that $\bar{G}_*(\lambda_n)$ has $k_0 + k_* - \bar{k}$ in any of its limits. Hence, this is an exact-fitted setting and we can apply Theorem 3.1. in Ho et al. [17]. As a consequence, claim (59) is shown under Case 2.3.

D.7 Proof of Theorem A.3

Theorem A.3. Assume that h_0 takes the form (7). Besides that, $K \geq k_0$ and f is location-scale Gaussian distribution. Then, for any $\lambda \in [0,1]$ and $G \in \mathcal{O}_{K,c_0}(\Theta)$ for some $c_0 > 0$, there exist positive constants C_1, C_2, C_3, C_4 depending only on $\lambda^*, G_*, G_0, \Theta$ (C_3 and C_4 also depends on δ) such that the following holds:

(a) When $K \leq k_* + k_0 - \bar{k} - 1$, then

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_1 \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*).$$

(b) When $K \geq k_* + k_0 - \bar{k}$, then

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \bigg(1_{\{\lambda \leq \lambda^*\}} \overline{W}_{\overline{r}(K-k_*)}(\lambda G, \lambda^* G_*) \\ + 1_{\{\lambda > \lambda^*\}} W_{\overline{r}(K-k_*)}^{\overline{r}(K-k_*)}(G, \overline{G}_*(\lambda)) \bigg)$$

(c) For $\delta > 0$, when $K = k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_3 1_{\{\lambda > \lambda^* + \delta\}} W_1(G, \overline{G}_*(\lambda)),$$

and when $K > k_* + k_0 - \bar{k}$, we have

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_4 1_{\{\lambda > \lambda^* + \delta\}} \\ \times W_{\overline{r}(K-k_0-k_*+\bar{k})}^{\overline{r}(K-k_0\ -k_*+\bar{k})}(G, \overline{G}_*(\lambda)).$$

We still denote $\mathcal{T} = k_* + k_0 - \bar{k}$ and follow the path of Theorem 3.6 to prove by contradiction. We denote by $\overline{r}_1 = \overline{r}(K - k_*)$, $\overline{r}_2 = \overline{r}(K - k_0 - k_* + \bar{k})$, and

$$D(\lambda G, \lambda^* G_*) = \begin{cases} \overline{W}_{\overline{r}_1}(\lambda G, \lambda^* G_*), \text{ when } K \leq \mathcal{T} - 1 \\ 1_{\{\lambda \leq \lambda^*\}} \overline{W}_{\overline{r}_1}(\lambda G, \lambda^* G_*) + 1_{\{\lambda > \lambda^*\}} (\lambda + \lambda^*) W_{\overline{r}_2}^{\overline{r}_2}(G, \overline{G}_*(\lambda)), \text{ when } K \geq \mathcal{T} \end{cases}.$$

We need to show the following claim:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{O}_K(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{D(\lambda G, \lambda^* G_*)} : D(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0. \tag{61}$$

There exists sequences λ_n and $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_K(\Theta)$ such that $D(\lambda_n G_n, \lambda^* G_*) \to 0$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*) \to 0$, where D is the lower bound in the theorem statement. For the ease of presentation, we consider the one dimension Gaussian case where $(\mu, \Sigma) = (\theta, v)$, the higher dimension cases are treated similar.

Case 1: $\lambda^* \geq \lambda_n$ for infinitely many n. We can use the subsequence argument to have $\lambda^* \geq \lambda_n$ for all n and G_n can be assumed to have a fixed number of atoms (less than or equals K) and have a representation as in (54). In this case,

$$D(\lambda_n G_n, \lambda^* G_*) = |\lambda_n - \lambda^*| + (\lambda_n + \lambda^*) \overline{W}_{\overline{r}_1}^{\overline{r}_1}(G_n, G_*) \to 0, \quad \frac{V(p_{\lambda^* G_*}, p_{\lambda_n G_n})}{D(\lambda_n G_n, \lambda^* G_*)} \to 0. \tag{62}$$

Using Taylor expansion of f around $\{(\theta_i^*, v_i^*)\}_{i=1}^{k_*}$ to the $\overline{r}_1-$th order we have

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = (\lambda^* - \lambda_n) \sum_{i=\bar{k}+1}^{k_0} p_i^0 f(x|\theta_i^0, v_i^0) + \lambda_n (\sum_{i=1}^{k_*+\underline{l}} \sum_{j=1}^{s_i} p_{ij}^n f(x|\theta_{ij}^n, v_{ij}^n)) - \sum_{i=1}^{k_*} \overline{p}_i^* f(x|\theta_i^*, v_i^*)$$

$$= (\lambda^* - \lambda_n) \sum_{i=\bar{k}+1}^{k_0} p_i^0 f(x|\theta_i^0, v_i^0) + \sum_{i=1}^{k_*+\underline{l}} \sum_{j=1}^{s_i} \lambda_n p_{ij}^n \sum_{|\boldsymbol{\alpha}|=1}^{\overline{r}_1} (\Delta \theta_{ij}^n)^{\alpha_1} (\Delta v_{ij}^n)^{\alpha_2} \frac{1}{\boldsymbol{\alpha}!} \frac{\partial^{|\boldsymbol{\alpha}| f(\theta_i^*, v_i^*)}}{\partial^{\alpha_1} \theta \partial^{\alpha_2} v}$$

$$+ \sum_{i=1}^{k_*+\underline{l}} (\Delta \overline{p}_{i\cdot}^n) f(x|\theta_i^*, v_i^*) + R(x),$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2$, $\boldsymbol{\alpha}! = \alpha_1!\alpha_2!$, $\Delta \overline{p}_{i\cdot}^n = \lambda_n \sum_j p_{ij}^n - \overline{p}_i^*$, $\Delta\theta_{ij}^n = \theta_{ij}^n - \theta_i^*$, $\Delta v_{ij}^n = v_{ij}^n - v_i^*$ and $R(x) = O(\sum_{i=1}^{k_*+\underline{l}} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta\theta_{ij}^n|^{\overline{r}_1} + |\Delta v_{ij}^n|^{\overline{r}_1}))$. Now we can use the character equation $\dfrac{\partial^2 f}{\partial \theta^2} = 2\dfrac{\partial f}{\partial v}$ to rewrite the formula above as

$$(\lambda^* - \lambda_n) \sum_{i=\bar{k}+1}^{k_0} p_i^0 f(x|\theta_i^0, v_i^0) + \sum_{\alpha=1}^{2\overline{r}_1} \sum_{i=1}^{k_*+\underline{l}} \left(\sum_{j=1}^{s_i} \lambda_n p_{ij}^n \sum_{n_1,n_2} \frac{(\Delta\theta_{ij}^n)^{n_1} (\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \right) \frac{\partial^\alpha f(\theta_i^*, v_i^*)}{\partial \theta^\alpha} + \sum_{i=1}^{k_*+\underline{l}} (\Delta \overline{p}_{i\cdot}^n) f(x|\theta_i^*, v_i^*) + R(x), \quad (63)$$

where we sum over n_1, n_2 such that $n_1 + 2n_2 = \alpha, n_1 + n_2 \leq \overline{r}_1$. Now we turn into proving the non-vanishing coefficients. Assume that all coefficients in the formula above vanish when dividing by $D(\lambda_n G_n, \lambda^* G_*)$ when $n \to \infty$. Because

$$D(\lambda_n G_n, \lambda^* G_*) \asymp |\lambda_n - \lambda^*| + (\lambda_n + \lambda^*) \left(\sum_{i=1}^{k_*+\bar{l}} |\Delta p_{i.}^n| + \sum_{i=1}^{k_*+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n (\left\|\Delta\theta_{ij}^n\right\|_2^{\overline{r}_1} + \left\|\Delta v_{ij}^n\right\|_2^{\overline{r}_1}) \right) := D_{\overline{r}_1}(G_n, G_*), \quad (64)$$

we have

$$\frac{\lambda^* - \lambda_n}{D_{\overline{r}_1}(G_n, G_*)} \to 0, \quad \frac{\Delta \overline{p}_{i\cdot}^n}{D_{\overline{r}_1}(G_n, G_*)} \to 0. \quad (65)$$

These limits together imply

$$\frac{(\lambda^* + \lambda_n) \Delta \overline{p}_{i\cdot}^n}{D_{\overline{r}_1}(G_n, G_*)} \to 0, \quad \forall i = 1, \ldots, k_* + \bar{l}. \quad (66)$$

From the definition of $D_{\overline{r}_1}$, it can be deduced that there exists at least an index i^* such that

$$\sum_{j=1}^{s_{i*}} \frac{(\lambda_n + \lambda^*) p_{i^* j}^n ((\theta_{ij}^n)^{\overline{r}_1} + (v_{ij}^n)^{\overline{r}_1})}{D_{\overline{r}_1}(G_n, G_*)} \not\to 0. \quad (67)$$

Without loss of generality, assign $i^* = 1$. But as we assume all the coefficients in equation (63) go to 0 for all α and i, we have

$$\frac{\sum_{j=1}^{s_1} \lambda_n p_{1j}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \leq \overline{r}_1}} \dfrac{(\theta_{1j}^n)^{n_1} (v_{1j}^n)^{n_2}}{2^{n_2} n_1! n_2!}}{D_{\overline{r}_1}(G_n, G_*)} \to 0, \quad (68)$$

for all $\alpha = 1, \ldots, 2\overline{r}_1$. From two expressions above combining with equation (65), we have for all $\alpha = 1, \ldots, 2\overline{r}_1$,

$$F_\alpha := \frac{\sum_{j=1}^{s_1} p_{1j}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \leq \overline{r}_1}} \dfrac{(\Delta\theta_{1j}^n)^{n_1} (\Delta v_{1j}^n)^{n_2}}{2^{n_2} n_1! n_2!}}{\sum_{j=1}^{s_1} p_{1j}^n ((\Delta\theta_{ij}^n)^{\overline{r}_1} + (\Delta v_{ij}^n)^{\overline{r}_1})} \to 0. \quad (69)$$

If $s_1 = 1$ then substituting $\alpha = 1$ and $\alpha = 2\overline{r}_1$ gives

$$\frac{|\Delta\theta_{11}^n|^{\overline{r}_1}}{|\Delta\theta_{11}^n|^{\overline{r}_1} + |\Delta v_{11}^n|^{\overline{r}_1}}, \frac{|\Delta v_{11}^n|^{\overline{r}_1}}{|\Delta\theta_{11}^n|^{\overline{r}_1} + |\Delta v_{11}^n|^{\overline{r}_1}} \to 0,$$

which is impossible as they are sum up to 1 for all n. Hence $s_1 \geq 2$. Now we proceed to show the contradiction using the system of equations (6). Denote by $\overline{p}_n = \max_{1 \leq j \leq s_1} \{p_{1j}^n\}, \overline{M}_n = \max_{1 \leq j \leq s_1} \{|\Delta\theta_{1j}^n|, |\Delta v_{1j}^n|^{1/2}\}$. By the subsequence argument in compact sets, without loss of generality, we can denote $c_j^2 := \lim_{n\to\infty} p_{1j}^n / \overline{p}_n$, $a_j = \lim \Delta\theta_{1j}^n / \overline{M}_n$, and $b_j = \lim \Delta v_{1j}^n / \overline{M}_n$ for all $j = 1, \ldots, k_* + \bar{l}$. Because of the definition of $\mathcal{O}_{K,c_0}$, we have $p_j \geq c_0$ for all j, which implies all c_j are different from 0 and at least one of them is 1. Similarly, in $(a_j, b_j)_j$, there is at least one of them equals to 1 or -1. Dividing both numerators and denominators of equation (69) by $\overline{p}_n \overline{M}_n^{\alpha}$, we have

$$\sum_{j=1}^{s_1} \sum_{n_1+2n_2=\alpha} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0,$$

for all $\alpha = 1, \ldots, \overline{r}_1$. Hence, we get the contradiction, where we use the fact that $s_1 \leq K - k_* + 1$ (as $s_i \geq 1$ for all $i \geq 2$) and $\overline{r}_1 = \overline{r}(K - k_*)$ is the smallest number such that equation (6), where $k = K - k_*$, has the trivial solution only. After that, we can argue as in the Step 9 of Proposition 2.2. in [16] to get the contradiction to the assumption proposed in the beginning, where we use the fact that Gaussian family is identifiable up to any order with respect to the location parameters.

Case 2: $\lambda^* \leq \lambda_n$ for all n. We rewrite

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \lambda_n \Bigg(\underbrace{\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n)}_{:=f(x;G_n)} - \underbrace{\Bigg[\Big(1 - \frac{\lambda^*}{\lambda_n}\Big) \sum_{i=\overline{k}+1}^{k_0} p_i^0 f(x|\theta_i^0) + \sum_{i=1}^{k_*} \frac{\overline{p}_i^*}{\lambda_n} f(x|\theta_i^*) \Bigg]}_{:=f\left(x;\overline{G}_*(\lambda_n)\right)} \Bigg), \tag{70}$$

Cases 2.1. $K \leq \mathcal{T} - 1$, argue similarly to Case 2.1. of the proof of Theorem 3.6, we have $\dfrac{\lambda_n - \lambda^*}{D(\lambda_n G_n, \lambda^* G_*)} \to 0$ as $n \to \infty$. Now we arrive at the equation (65) of Case 1. Follow the argument above, we can prove claim (61).

Case 2.2. $K \geq \mathcal{T}$, we can see equation (70) as an over-fitted mixture of location-scale Gaussian setting where the number of over-fitted atoms is at most $K - k_*$. Hence we can argue similar to Case 1 or the Proposition 2.2. in [16] to obtain the conclusion.

Cases 2.3. $K = \mathcal{T}$ and $\lambda_n > \lambda^* + \delta$ for all n. From the presentation as in equation (70), we can see that $1 - \dfrac{\lambda^*}{\lambda_n}$ does not vanish in any of it limits. Therefore $\overline{G}_*(\lambda_n)$ has $k_* + k_0 - \overline{k} = \mathcal{T}$ number of components in its limits. Because this is an exact-fitted setting, we can apply Theorem 3.1. in Ho et al. [17] to get the result of claim (61)

Cases 2.4. $K > \mathcal{T}$ and $\lambda_n > \lambda^* + \delta$ for all n, we can also see that $\overline{G}_*(\lambda_n)$ has $k_* + k_0 - \overline{k} = \mathcal{T}$ number of components in its limits. We can apply Proposition 2.2. in Ho et al. [16] to get the result of claim (61).

D.8 Proof of Theorem A.4

Theorem A.4. Assume that h_0 takes the form (7) and $\overline{k} = k_0$. Besides that, f is second order identifiable. Then, for any $\lambda \in [0, 1]$ and $G \in \mathcal{O}_K(\Theta)$ that $K \geq k_*$, there exist positive constants C_1 and C_2 depending only on $\lambda^*, G_*, G_0, \Theta$ such that the following holds:

(a) If $\mathcal{I}(\lambda)$ is not ratio-independent, then

$$V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \Big[1_{\{\lambda \in \mathcal{B}^c\}} + 1_{\{\lambda \in \mathcal{B}\}} W_2^2(G, \bar{G}_*(\lambda)) \Big]. \tag{71}$$

(b) If $\mathcal{I}(\lambda)$ is ratio-independent, then

$$\begin{aligned} V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_2 \Bigg[1_{\{\lambda \in \mathcal{B}^c\}} \Bigg(\sum_{i \in \mathcal{I}(\lambda)} \Big[(\lambda^* - \lambda) p_i^0 - \lambda^* p_i^* \Big] \\ + \mathcal{S}(\mathcal{I}(\lambda)) W_2^2(G, \widetilde{G}_*(\lambda)) \Bigg) \\ + 1_{\{\lambda \in \mathcal{B}\}} W_2^2(G, \bar{G}_*(\lambda)) \Bigg]. \end{aligned} \tag{72}$$

To ease the ensuing presentation, we denote $D(\lambda G, \lambda^* G_*) = 1_{\{\lambda \in \mathcal{B}^c\}} \biggl(\sum_{i \in \mathcal{I}(\lambda)} \biggl[(\lambda^* - \lambda) p_i^0 - \lambda^* p_i^* \biggr] + \mathcal{S}(\mathcal{I}(\lambda)) W_2^2(G, \widetilde{G}_*(\lambda)) \biggr) + 1_{\{\lambda \in \mathcal{B}\}} W_2^2(G, \bar{G}_*(\lambda))$ when $\mathcal{I}(\lambda)$ is ratio-independent or $D(\lambda G, \lambda^* G_*) = 1_{\{\lambda \in \mathcal{B}^c\}} + 1_{\{\lambda \in \mathcal{B}\}} W_2^2(G, \bar{G}_*(\lambda))$ when $\mathcal{I}(\lambda)$ is not ratio-independent.

In order to prove the theorem, it is sufficient to verify the following inequality:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{E}_{k_*}(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{D(\lambda G, \lambda^* G_*)} : D(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0. \tag{73}$$

Proof of claim (73)**:** Assume that the above claim is not true. It implies that there exist sequences $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_K(\Theta)$ and $\lambda_n \in [0,1]$ such that $D(\lambda_n G_n, \lambda^* G_*)$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*)$ go to 0 as n approaches to infinity. Since $\bar{k} = k_0$ and G_* admits the form (10), we find that

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \lambda_n \left(\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n) \right) - \sum_{i=1}^{k_*} \bar{p}_i^* f(x|\theta_i^*), \tag{74}$$

where $\bar{p}_i^* = \lambda^* p_i^* + (\lambda_n - \lambda^*) p_i^0$ when $1 \leq i \leq k_0$ and $\bar{p}_i^* = \lambda^* p_i^*$ otherwise. In addition, $\theta_i^* = \theta_i^0$ for $i \in [k_0]$. From this presentation, we see that there must exists a constant C depending on λ^*, G_*, G_0 such that $\liminf \lambda_n > C$. Indeed, suppose it is not the case, then by the subsequence argument, we can assume that $\lambda_n \to 0$. Besides, $V(\lambda_n G_n, \lambda^* G_*) \to 0$, we have $\bar{p}_i^* \to 0$ for all $i \in [k_*]$. These conditions lead to $p_i^* = 0$ for all $i > k_0$ and $p_i^0 = p_i^*$ for all $i \in [k_0]$, which mean that $G_* = G_0$ (a contradiction to our assumption). Hence, limits of (λ_n) is bounded below. We have two settings with λ_n.

Case 1: $\lambda_n \in \mathcal{B}$ for infinitely many n. Without loss of generality, we assume that $\lambda_n \in \mathcal{B}$ for all $n \geq 1$. If $k_* = k_0$ then we see that $\bar{p}_i^*$ can not vanish simultaneously when $n \to \infty$ for all i, otherwise we have $G^* = G_0$, which contradicts to the assumption in this section. Otherwise, $k_* > k_0$, and $\bar{p}_i^*$ does not vanish for all $i > k_0$. Therefore, every limit of $\sum_{i=1}^{k_*} \bar{p}_i^* f(x|\theta_i^*)$ has a number of atoms ranging from $\max\{1, k_* - k_0\}$ to k_*, which is less than or equal to K. So that this is an over-fitted scenario. In addition, $D(\lambda_n G_n, \lambda^* G_*) = W_2^2(G_n, \bar{G}_*(\lambda_n))$. We can further rewrite equation (74) as:

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \lambda_n (f(x; G_n) - f(x; \bar{G}_*(\lambda_n))).$$

From Theorem 3.2 in Ho et al. [17], we have $V(f(.; G_n), f(.; \overline{G}_*(\lambda_n))/W_2^2(G_n, \bar{G}_*(\lambda_n)) \not\to 0$ as $n \to \infty$. Putting the above results together, we obtain that $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*) \not\to 0$, which is a contradiction. Hence, we reach the conclusion of claim (74).

Case 2: $\lambda_n \notin \mathcal{B}$ for infinitely many n. Without loss of generality, we assume that $\lambda_n \notin \mathcal{B}$ for all $n \geq 1$. Under this setting, $\mathcal{I}(\lambda_n) \neq \emptyset$. In addition, for any $i \in \mathcal{I}(\lambda_n)$, $\bar{p}_i^* < 0$. Given these conditions, we can rewrite equation (74) as follows:

$$\begin{aligned} p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \sum_{i \in \mathcal{I}(\lambda_n)} (-\bar{p}_i^*) f(x|\theta_i^0) + \Biggl[\lambda_n \left(\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n) \right) - \sum_{i \in \mathcal{I}(\lambda_n)^c} \bar{p}_i^* f(x|\theta_i^0) \\ - \textstyle\sum_{i=k_0+1}^{k_*} \bar{p}_i^* f(x|\theta_i^*) \Biggr] \end{aligned} \tag{75}$$

We have two separate settings with $\mathcal{I}(\lambda_n)$.

Case 2.1: $\mathcal{I}(\lambda_n)$ is not ratio-independent. Under this case, $D(\lambda_n G_n, \lambda^* G_*) = 1$. Since $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*) \to 0$, we have $V(p_{\lambda_n G_n}, p_{\lambda^* G_*}) \to 0$. It indicates that $p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) \to 0$ almost surely x. Since $-\bar{p}_i^* > 0$ for all $i \in \mathcal{I}(\lambda_n)$, the previous limit demonstrates that $\bar{p}_i^* \to 0$ for all $i \in \mathcal{I}(\lambda_n)$, which leads to $p_i^*/p_i^0 = p_j^*/p_j^0$ for all $i, j \in \mathcal{I}(\lambda_n)$. It contradicts the assumption that $\mathcal{I}(\lambda_n)$ is not ratio-independent. Hence, we achieve the conclusion of claim (74) under Case 2.1.

Case 2.2: $\mathcal{I}(\lambda_n)$ is ratio-independent. Under this case, $D(\lambda_n G_n, \lambda^* G_*) = \sum_{i \in \mathcal{I}(\lambda_n)} \Bigg[(\lambda^* - \lambda_n)p_i^0 - \lambda^* p_i^*\Bigg] + \mathcal{S}(\mathcal{I}(\lambda_n)) W_2^2(G_n, \widetilde{G}_*(\lambda_n)) \to 0$ and $V(p_{\lambda_n G_n, \lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*) \to 0$, which imply $V(p_{\lambda_n G_n, \lambda^* G_*}) \to 0$. We first prove that $\mathcal{S}(\mathcal{I}(\lambda_n)) \not\to 0$. Indeed, suppose it is not the case, then $p_i^* = 0$ for all $i > k_0$ and $(\lambda^* p_i^* + (\lambda_n - \lambda^*)p_i^0) \to 0$ for all $i \in \mathcal{I}(\lambda_n)$. From equation (75) and the fact that $V(p_{\lambda_n G_n, \lambda^* G_*}) \to 0$, we also see that $\overline{p}_i^* \to 0$ for all $i \in \mathcal{I}(\lambda_n)$ and $\lambda_n \to 0$. But that means

$$\lambda_n \to 0, \lambda^* p_i^* + (\lambda_n - \lambda^*)p_i^0 \to 0, \quad \forall i \in [k_0].$$

Those limits together imply that $\lambda^*(p_i^0 - p_i^*) = 0$ for all $i \in [k_0]$, which is contradictory with our assumption that $G^* \neq G_0$. Hence $\mathcal{S}(\mathcal{I}(\lambda_n)) \not\to 0$. As $D(\lambda_n G_n, \lambda^* G_*) \to 0$, we have $W_2^2(G_n, \widetilde{G}_*(\lambda_n)) \to 0$ as $n \to \infty$. It implies that we can rewrite G_n as follows:

$$G_n = \sum_{i \in \mathcal{I}(\lambda_n)^c \cup \{k_0+1, \ldots, k_* + \bar{l}\}} \sum_{j=1}^{s_i} p_{ij}^n \delta_{\theta_{ij}^n}, \tag{76}$$

where $\sum_{j=1}^{s_i} p_{ij}^n \to \overline{p}_i^*/\mathcal{S}(\mathcal{I}(\lambda_n))$ and $\theta_{ij}^n \to \theta_i^*$ for all $i \in \mathcal{J} := \mathcal{I}(\lambda_n)^c \cup \{k_0 + 1, \ldots, k_* + \bar{l}\}$. Here, $\overline{p}_i^* = 0$ for $k_* + 1 \leq i \leq k_* + \bar{l}$. In addition, $\sum_{i \in \mathcal{J}} s_i = k'$ for some k' such that $k_* - k_0 + |\mathcal{I}(\lambda_n)^c| \leq k' \leq k_*$. To faciliate the proof argument, we denote $\Delta\theta_{ij}^n := \theta_{ij}^n - \theta_i^*$ and $\Delta p_{i.}^n := \sum_{j=1}^{s_i} p_{ij}^n - \overline{p}_i^*/\mathcal{S}(\mathcal{I}(\lambda_n))$ for $i \in \mathcal{J}$. The result of Lemma 3.1 in Ho et al. [18] leads to

$$W_2^2(G_n, \tilde{G}_*(\lambda_n)) \asymp \sum_{i \in \mathcal{J}} |\Delta p_{i.}^n| + \sum_{i \in \mathcal{J}} \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2. \tag{77}$$

Invoking Taylor's expansion up to the second order, we have

$$\begin{aligned} p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) &= \sum_{i \in \mathcal{I}(\lambda_n)} (-\overline{p}_i^*) f(x|\theta_i^0) + \sum_{i \in \mathcal{J}} (\lambda_n \sum_{j=1}^{s_i} p_{ij}^n - \overline{p}_i^*) f(x|\theta_i^*) \\ &\quad + \lambda_n \left(\sum_{j=1}^{s_i} p_{ij}^n \Delta\theta_{ij}^n\right)^\top \frac{\partial f}{\partial \theta}(x|\theta_i^*) + \lambda_n \left(\sum_{j=1}^{s_i} p_{ij}^n \left(\Delta\theta_{ij}^n\right)^\top \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)(\Delta\theta_{ij}^n)\right) + R(x), \end{aligned} \tag{78}$$

where $R(x)$ is Taylor remainder such that $R(x) = o\left(\lambda_n \sum_{i \in \mathcal{J}} \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2\right)$. Therefore, we have $R(x)/D(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$.

The expression in equation (78) indicates that we can view $(p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x))/D(\lambda_n G_n, \lambda^* G_*)$ as a linear combination of elements of the forms $f(x|\theta_i^0)$, $f(x|\theta_j^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_j^*)$, $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^*)$ for $i \in \mathcal{I}(\lambda_n)$ and $j \in \mathcal{J}$. Assume that the coefficients of these terms go to 0 as n approaches infinity. By studying the coefficients of $f(x|\theta_i^0)$ when $i \in \mathcal{I}(\lambda_n)$, we find that

$$(\sum_{i \in \mathcal{I}(\lambda_n)} (-\overline{p}_i^*))/D(\lambda_n G_n, \lambda^* G_*) \to 0.$$

Given the above result, as the coefficients of $f(x|\theta_i^*)$ and $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^*)$ go to 0 when $i \in \mathcal{J}$, we obtain

$$\frac{\mathcal{S}(\mathcal{I}(\lambda_n)) \sum_{j=1}^{s_i} p_{ij}^n - \overline{p}_i^*}{D(\lambda_n G_n, \lambda^* G_*)} = \frac{[\lambda_n - (\sum_{l \in \mathcal{I}(\lambda_n)} \overline{p}_l^*))] \sum_{j=1}^{s_i} p_{ij}^n - \overline{p}_i^*}{D(\lambda_n G_n, \lambda^* G_*)} \to 0,$$

$$\frac{\mathcal{S}(\mathcal{I}(\lambda_n)) \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2}{D(\lambda_n G_n, \lambda^* G_*)} = \frac{[\lambda_n - (\sum_{l \in \mathcal{I}(\lambda_n)} \overline{p}_l^*))] \sum_{j=1}^{s_i} p_{ij}^n \left\|\Delta\theta_{ij}^n\right\|_2^2}{D(\lambda_n G_n, \lambda^* G_*)} \to 0$$

Putting the above results together, given the expression in equation (77), we obtain $1 = D(\lambda_n G_n, \lambda^* G_*)/D(\lambda_n G_n, \lambda^* G_*) \to 0$ as $n \to \infty$, which is a contradiction. Therefore, not all the coefficients of $f(x|\theta_i^0)$, $f(x|\theta_j^*)$, $\frac{\partial f}{\partial \theta}(x|\theta_j^*)$, $\frac{\partial^2 f}{\partial \theta^2}(x|\theta_j^*)$ when $i \in \mathcal{I}(\lambda_n)$ and $j \in \mathcal{J}$. From here, we utilize the Fatou's argument from the previous proofs to obtain the conclusion of claim (73) under Case 2.2.

D.9 Proof of Theorem A.5

Theorem A.5. Assume that h_0 takes the form (7) and $\bar{k} = k_0$. Besides that, f is location-scale Gaussian distribution. Then, for $\tilde{k} := \max\{k_* - k_0, 1\}$, and for any $\lambda \in [0,1]$ and $G \in \mathcal{O}_{K,c_0}(\Theta)$ for some $K \geq k_*$ and $c_0 > 0$, there exist positive constants C_1 and C_2 depending only on $\lambda^*, G_*, G_0, \Theta$ such that on $\lambda^*, G_*, G_0, \Theta$ such that

(a) If $\mathcal{I}(\lambda)$ is not ratio-independent, then

$$V(p_{\lambda^* G_*}, p_{\lambda G}) \geq C_1 \Bigg[1_{\{\lambda \in \mathcal{B}^c\}} \\ + 1_{\{\lambda \in \mathcal{B}\}} W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \bar{G}_*(\lambda)) \Bigg]. \tag{79}$$

(b) If $\mathcal{I}(\lambda)$ is ratio-independent, then

$$V(p_{\lambda^*, G_*}, p_{\lambda, G}) \geq C_2 \Bigg[1_{\{\lambda \in \mathcal{B}^c\}} \Bigg(\sum_{i \in \mathcal{I}(\lambda)} \Big[(\lambda^* - \lambda) p_i^0 - \lambda^* p_i^* \Big] \\ + \mathcal{S}(\mathcal{I}(\lambda)) W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \widetilde{G}_*(\lambda)) \Bigg) \\ + 1_{\{\lambda \in \mathcal{B}\}} W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \bar{G}_*(\lambda)) \Bigg]. \tag{80}$$

The proof of Theorem A.5 is similar to what of Theorem A.4 and with the technical details borrowed from Theorem A.3. Therefore we only highlight the main differences. Denote by $D(\lambda G, \lambda^* G_*) = 1_{\{\lambda \in \mathcal{B}^c\}} \Big(\sum_{i \in \mathcal{I}(\lambda)} \Big[(\lambda^* - \lambda) p_i^0 - \lambda^* p_i^* \Big] + \mathcal{S}(\mathcal{I}(\lambda)) W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \widetilde{G}_*(\lambda)) \Big) + 1_{\{\lambda \in \mathcal{B}\}} W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \bar{G}_*(\lambda))$ when $\mathcal{I}(\lambda)$ is ratio-independent or $D(\lambda G, \lambda^* G_*) = 1_{\{\lambda \in \mathcal{B}^c\}} + 1_{\{\lambda \in \mathcal{B}\}} W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G, \bar{G}_*(\lambda))$ when $\mathcal{I}(\lambda)$ is not ratio-independent.

In order to prove the theorem, it is sufficient to verify the following inequality:

$$\lim_{\epsilon \to 0} \inf_{\lambda \in [0,1], G \in \mathcal{E}_{k_*}(\Theta)} \left\{ \frac{V(p_{\lambda G}, p_{\lambda^* G_*})}{D(\lambda G, \lambda^* G_*)} : \ D(\lambda G, \lambda^* G_*) \leq \epsilon \right\} > 0. \tag{81}$$

Proof of claim (81)**:** Assume that the above claim is not true. It implies that there exist sequences $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_K(\Theta)$ and $\lambda_n \in [0,1]$ such that $D(\lambda_n G_n, \lambda^* G_*)$ and $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*)$ go to 0 as n approaches to infinity. Since $\bar{k} = k_0$ and G_* admits the form (10), we find that

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \lambda_n \left(\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n) \right) - \sum_{i=1}^{k_*} \bar{p}_i^* f(x|\theta_i^*), \tag{82}$$

where $\bar{p}_i^* = \lambda^* p_i^* + (\lambda_n - \lambda^*) p_i^0$ when $1 \leq i \leq k_0$ and $\bar{p}_i^* = \lambda^* p_i^*$ otherwise. In addition, $\theta_i^* = \theta_i^0$ for $i \in [k_0]$. One could argue as in Theorem A.4 to get (λ_n) being bounded below.

Case 1: $\lambda_n \in \mathcal{B}$ for infinitely many n. Without loss of generality, we assume that $\lambda_n \in \mathcal{B}$ for all $n \geq 1$. Under this case, every limit of $\sum_{i=1}^{k_*} \bar{p}_i^* f(x|\theta_i^*)$ has a number of atoms ranging from $\tilde{k}$ to k_*, which is less than or equal to K. So that this is an over-fitted scenario where the number of over-fitted atoms is at most $K - \tilde{k}$. In addition, $D(\lambda_n G_n, \lambda^* G_*) = W_{\overline{r}(K-\tilde{k})}^{\overline{r}(K-\tilde{k})}(G_n, \bar{G}_*(\lambda_n))$. We can further rewrite equation (82) as:

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \lambda_n (f(x; G_n) - f(x; \bar{G}_*(\lambda_n))).$$

Now we can argue similarly to the proof Theorem A.3 or Proposition 2.2. in [16] to get $V(p_{\lambda_n G_n}, p_{\lambda^* G_*})/D(\lambda_n G_n, \lambda^* G_*) \not\to 0$, which combines with the fact that $\lambda_n \not\to 0$ gives us a contradiction. Hence, we reach the conclusion of claim (81)

Case 2: $\lambda_n \notin \mathcal{B}$ for infinitely many n. Without loss of generality, we assume that $\lambda_n \notin \mathcal{B}$ for all $n \geq 1$. Under this setting, $\mathcal{I}(\lambda_n) \neq \emptyset$. In addition, for any $i \in \mathcal{I}(\lambda_n)$, $\bar{p}_i^* < 0$. Given these conditions, we can rewrite equation (74) as follows:

$$p_{\lambda_n G_n}(x) - p_{\lambda^* G_*}(x) = \sum_{i \in \mathcal{I}(\lambda_n)} (-\bar{p}_i^*) f(x|\theta_i^0) + \Bigg[\lambda_n \left(\sum_{i=1}^{k_n} p_i^n f(x|\theta_i^n)\right) - \sum_{i \in \mathcal{I}(\lambda_n)^c} \bar{p}_i^* f(x|\theta_i^0) \\ - \textstyle\sum_{i=k_0+1}^{k_*} \bar{p}_i^* f(x|\theta_i^*)\Bigg] \tag{83}$$

We have two separate settings with $\mathcal{I}(\lambda_n)$.

Case 2.1: $\mathcal{I}(\lambda_n)$ is not ratio-independent. This is the same as Case 2.1. of Theorem A.4. With a similar argument, we can show that $\mathcal{I}(\lambda_n)$ must be ratio-independent, which is a contradiction. Hence, we get claim (81) under this case.

Case 2.2: $\mathcal{I}(\lambda_n)$ is ratio-independent. We can see that the second term of equation (83) is in an over-fitted setting with the number of extra components being at most $K - \tilde{k}$. Arguing similar to Case 2.2. of Theorem A.3 gives us the conclusion of claim (81).