Electronic Journal of Statistics
Vol. 19 (2025) 131–203
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2339

Strong identifiability and parameter learning in regression with heterogeneous response

Dat Do[1], **Linh Do**[2] **and XuanLong Nguyen**[1]

[1] *Department of Statistics, University of Michigan,*
e-mail: dodat@umich.edu; xuanlong@umich.edu

[2] *Department of Mathematics, Tulane University,*
e-mail: ldo3@tulane.edu

Abstract: Mixtures of regression are useful for regression learning with respect to an uncertain and heterogeneous response variable of interest. In addition to being a rich predictive model for the response given some covariates, the model parameters provide meaningful information about the heterogeneity in the data population, which is represented by the conditional distributions for the response given the covariates associated with a number of distinct but latent subpopulations. In this paper, we investigate conditions of strong identifiability, MLE rates of convergence for the conditional density and model parameters, and the Bayesian posterior contraction behavior arising in finite mixture of regression models, under exact-fitted and over-fitted settings and when the number of components is unknown. This theory is applicable to common choices of link functions and families of conditional distributions employed by practitioners. We provide simulation studies and data illustrations, which shed some light on the parameter learning behavior found in several popular regression mixture models reported in the literature.

MSC2020 subject classifications: Primary 62J05, 62F12, 62F15; secondary 62G05, 62G20.
Keywords and phrases: Regression mixture models, strong identifiability, parameter estimation, convergence rate, posterior contraction rate.

Received December 2023.

1. Introduction

Regression is often associated with the task of curve fitting — given data samples for pairs of random variables (X, Y), find a function $y = F(x)$ that captures the relationship between X and Y as well as possible. As the underlying population for the (X, Y) pairs becomes increasingly complex, much effort has been devoted to learning more complex models for the regression function F. In many data domains, however, due to the heterogeneity of the behavior of the response variable Y with respect to covariate X, no single function F can fit the data pairs well, no matter how complex F is. Many authors noticed this challenge and

arXiv: 2212.04091

adopted a mixture modeling framework into the regression problem, going back to [11, 35], and continuing with more recent applications, e.g., [1, 4, 28, 33, 32].

To capture the uncertain and highly heterogeneous behavior of response variable Y given covariate X, one needs more than one single regression model. Suppose that there are k different regression behaviors, one can represent the *conditional* distribution of Y given X by a mixture of k conditional density functions associated with k underlying (latent) subpopulations. One can draw from modeling tools of conditional densities such as generalized linear models or more complex components to increase model fitness for the regression task [22, 16]. Making inferences in regression mixtures can be achieved in a frequentist framework (e.g., maximum conditional likelihood estimation (MLE) [3]), or a Bayesian framework [21]. In addition to enhanced predictability for the response variable given the covariate, a key benefit of regression mixture models is that the model parameters may be used to explicate the relationship between these variables more accurately and meaningfully.

Despite the aforementioned long history of applications, a satisfactory level of understanding of several key issues concerning model parameters' identifiability and a large sample theory of regression mixture models remains far from being complete. This is perhaps due to the somewhat unusual position where regression mixture model based methods sit — like any regression problem one is interested in prediction performance, but unlike the traditional viewpoint of a single curve-fitting task one must come to terms with the multi-modality of the response variable due to the underlying data population's latent heterogeneity. Thus, one must also be interested in the quality of parameter estimates representing such heterogeneity. There is a slowly growing theoretical literature, but most existing works are limited to the questions of consistency of estimation for the mixture of generalized linear models with some specific classes of conditional densities and link functions, or simulation-based methods [15, 24, 23, 40, 12, 39]. In particular, [15] investigates the identifiability of the mixture of Gaussian regression models with linear link functions. [24] generalizes the results for the exponential families. [40] further extends the identifiability results to more general link functions, but no analysis of parameter estimation. [39] shows the consistency for density learning of this model under the Bayesian setting. On parameter estimation behavior, more recently [25] proposed a penalized MLE method for model selection for the class of identifiable mixture of regression models with linear link functions and established rates of parameter estimation. [20] investigated the parameter estimation behavior for the Gaussian mixture of regression models.

In this paper, we study parameter identifiability, parameter estimation behavior, and prediction performance arising from the finite mixture of regression models. We work with general conditional density kernels and link functions, investigate both an MLE approach and a Bayesian approach for estimation. Consider a regression mixture model in the following form:

$$f_{G_0}(y|x) = \sum_{j=1}^{k_0} p_j^0 f(y|h_1(x, \theta_{1j}^0), h_2(x, \theta_{2j}^0)), \tag{1}$$

where $x \in \mathcal{X} \subset \mathbb{R}^p$ is a vector including the explanatory variables, $y \in \mathcal{Y}$ is the response variable. The conditional density function $f_{G_0}(y|x)$ take the mixture form, where the discrete probability measure $G_0 = \sum_{j=1}^{k_0} p_j^0 \delta_{(\theta_{1j}^0, \theta_{2j}^0)}$ encapsulates all unknown parameters in the model, with $(p_j^0)_{j=1}^{k_0}$ being the mixing proportion, and $(\theta_{1j}^0)_{j=1}^{k_0}$ and $(\theta_{2j}^0)_{j=1}^{k_0}$ being parameters in a compact subspace Θ_1 of $\mathbb{R}^{d_1}$ and Θ_2 of $\mathbb{R}^{d_2}$, respectively. We call G_0 the latent mixing measure associated with the regression mixture model. The link functions $h_1 : \mathcal{X} \times \Theta_1 \to H_1$ and $h_2 : \mathcal{X} \times \Theta_2 \to H_2$ are known, where H_1, H_2 are compact subsets of $\mathbb{R}$. The family of densities $\{f(y|\mu, \phi) : \mu \in H_1, \phi \in H_2\}$ is given, where all of them are dominated by a common distribution ν on $\mathcal{Y}$ which can be either a counting or continuous measure. In many applications, the family f is a dispersion exponential family distribution with parameter $\mu = h_1(x, \theta_1)$ is modeled as the mean, and $\phi = h_2(x, \theta_2)$ is modeled as the variance of $f(y|h_1(x, \theta_1), h_2(x, \theta_2))$ so that the mixture of regression models can capture the average trends and dispersion of subpopulations in the data. We are interested assessing the quality of the conditional density estimates, as well as that of parameters $(p_j^0)_{j=1}^{k_0}$, $(\theta_{1j}^0)_{j=1}^{k_0}$, and $(\theta_{2j}^0)_{j=1}^{k_0}$ from i.i.d. samples $(x_i, y_i)_{i=1}^n$, where distribution of y_i given x_i is given in the model (1) and x_i follows some (unknown) marginal distribution $\mathbb{P}_X$ on $\mathcal{X}$.

Our parameter estimation theory inherits from and generalizes several recent developments in the finite mixture models literature. [2] initiated the theoretical investigation of parameter estimation in a univariate finite mixture model by introducing a notion of *strong* identifiability. [31] developed a theory for both finite and infinite mixture models in a multivariate setting using optimal transport distances. [17] studied convergence rates in various families of vector-matrix distributions. A central concept in these papers is the notion of strong identifiability of (unconditional) mixtures of density functions. This is a condition on a parametrized family of density function f of y, as there is no covariate x here. At a high level, it requires that the family of function f, along with their partial derivatives with respect to the parameters up to a certain order, are linearly independent. Once this condition is satisfied, one can establish a lower bound on the distance between mixture distributions in terms of the optimal transport distances between the corresponding latent mixing parameters. Such a bound is called an *inverse bound*, which plays a crucial role in deriving the rates of parameter estimates.

With regression mixture modeling, we move from the unconditional mixtures described above to conditional mixture models. Thus, there are several fundamental distinctions. First, one works with the family of conditional density functions in the form $f(y|h(x, \theta))$, which involves both the conditional density kernel f and the link function h. A strong identifiability condition for conditional distributions that we develop will inevitably involve both variables x and y. The focus of inference is on the conditional distribution of Y given covariate X, while the marginal distribution of X is assumed unknown and of little interest. Accordingly, the identifiability condition must ideally require as little information from the marginal distribution of the covariate as possible. Moreover, given that the identifiability condition holds, the inverse bound that we establish will be a lower bound on the *expected* distance of the conditional densities, where the expectation is taken with respect to the marginal distribution of the covariate. This is also crucial because we will obtain rates of conditional density estimation in terms of the mentioned expected distance and use the inverse bound to derive the rates of convergence for the corresponding mixing parameters of interest.

Another interesting feature that distinguishes conditional mixtures from unconditional mixtures is that the former tends to satisfy strong identifiability conditions more easily than the latter. This is because of the role the covariate x plays in providing more constraints that prevent the violation of the linear independence condition. For instance, it is trivial that an unconditional mixture of Bernoulli distributions is not identifiable, but it will be shown (not so easily) that the mixture of conditional distributions using the Bernoulli kernel is not only identifiable but also strongly identifiable. There are situations where there is a lack of strong identifiability, such as in the case of negative binomial regression mixtures, a model extensively employed in practice (e.g., see [33, 32]), but we shall show that such situations occur precisely only in a Lebesgue measure zero subset of the parameter space.

To summarize, there are several contributions made in this paper. First, we develop a rigorous notion of strong identifiability for general regression mixture models. We provide a characterization of such a notion in terms of simple conditions on the conditional density kernel f and link function h and show that they are satisfied by a broad range of density kernels and link functions often employed in practice. Second, we study several examples of regression mixtures when strong identifiability is violated and investigate the consequences. Third, we establish convergence rates for regression mixtures given strong identifiability, under both Bayesian estimation and MLE frameworks. We consider three different learning scenarios: when the number of mixture components k_0 is known (i.e., exact-fitted setting), when only an upper bound is known (i.e., overfitted setting), and when even such an upper bound is unknown. Finally, we conduct a series of simulation studies to support the theory and discuss the connections with empirical findings in the regression mixture literature [23, 33, 34].

In this paper we shall focus primarily on the "point-wise" convergence parameter estimation rate, i.e., assume that there is a fixed true set of parameters in the mixture model, instead of allowing them to vary and overlap arbitrarily as in a minimax setting, e.g., [14, 44, 41]. The uniform convergence rates arising in the latter asymptotic setting are known to be extremely slow as the level of over-fitting increases and generally challenging to establish. In theory, the distinction between the minimax and pointwise convergence rates within the finite mixture framework deserves careful attention; see the recent paper [41] for the sharpest and general treatment in the finite and multivariate mixture setting. In practice, the slow minimax convergence rates do not reflect the meaningful applicability of mixture modeling based methods. For instance, for many realistic data distributions, such as in the crash data analysis [32] (to be demonstrated in Section 5), the fixed true parameters assumption is arguably more applicable since the modeler knows that there are only some heterogeneous but unchanged sources that lead to car crashes. Furthermore, unlike vanilla mixtures, previous theoretical investigation for regression mixture models only mentioned identifiability and consistency [15, 24, 39]. Hence, the investigation of the point-wise parameter estimation behavior can be valuable and the technique we study may be useful for a more delicate analysis of this model class.

The rest of this paper is organized as follows. Section 2 provides preliminaries on the mixture of regression models. In Section 3, we present notions of strong identifiability and associated characterization for regression mixtures, followed by a set of inverse bounds. Building upon this strong identifiability theory, in Section 4, we establish the rates of conditional density estimation and parameter estimation. In Section 5, we carry out simulation studies and data illustrations to support our theory and discuss the empirical findings in the literature. Finally, Section 6 discusses future directions. All proofs are deferred to the Supplementary material.

Notation Given a mixing measure $G = \sum_{j=1}^{k} p_j \delta_{(\theta_{1j}, \theta_{2j})}$, the mixture of regression model with respect to G is denoted by $f_G(y|x) = \sum_{j=1}^{k} p_j f(y|h_1(x, \theta_{1j}), h_2(x, \theta_{2j}))$. The joint distribution of (x, y) is $d\mathbb{P}_G(x, y) = d\mathbb{P}_X(x) \times f_G(y|x) d\nu(y)$, where $\mathbb{P}_X$ is an unknown distribution of covariate X. $\mathbb{E}_X$ denotes the expectation w.r.t. $\mathbb{P}_X$. We write $f_j(y|x) = f(y|h_1(x, \theta_{1j}), h_2(x, \theta_{2j}))$ for short, for $j = 1, \ldots, k$, if there is no confusion. Denote $\Theta = \Theta_1 \times \Theta_2$, $H = H_1 \times H_2$. Let $\mathcal{E}_k(\Theta)$ be the space of mixing measures with exactly k atoms in Θ, and $\mathcal{O}_k(\Theta) = \cup_{\kappa=1}^{k} \mathcal{E}_\kappa(\Theta)$ the space of mixing measures with no more than k atoms in Θ. If there is no confusion, we write $\mathcal{E}_k(\Theta)$ and $\mathcal{O}_K(\Theta)$ as $\mathcal{E}_k$ and $\mathcal{O}_K$ for short. For two sequence $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$, we write $a_n \preccurlyeq b_n$ if there is a constant C such that $a_n \leq C b_n$ for all n. We also write $a_n \succcurlyeq b_n$ if $b_n \preccurlyeq a_n$, and $a_n \asymp b_n$ if we have both $a_n \preccurlyeq b_n$ and $a_n \succcurlyeq b_n$. The multiplicative constants in those inequalities will be specified in the main results for clarity. We use d_H, d_{TV}, and K for the Hellinger distance, total variation distance, and Kullback-Leibler (KL) divergence between densities, respectively.

2. Preliminaries

Regression mixture models A mixture of regression model may be applied with many different family distributions and link functions to fit a large range of data distributions. For example, when the response variable y is continuous, we can choose the family of (conditional) density to be normal $\{\mathcal{N}(y|\mu, \phi) : \mu \in \mathbb{R}, \phi \in \mathbb{R}_+\}$, and parametrize μ_j and ϕ_j via two link functions $\mu_j = h_1(x, \theta_{1j}), \phi_j = h_2(x, \theta_{2j})$, for $j = 1, \ldots, k$. These functions can be represented by polynomials or trigonometric polynomials with variable x and coefficients θ_{1j}, θ_{2j}. Alternatively, when y is a counting variable, one can use the Binomial distribution $\{\mathrm{Bin}(y|N, q) : q \in [0, 1]\}$ if y is bounded and the Poisson distribution $\{\mathrm{Poi}(y|\mu) : \mu \in \mathbb{R}_+\}$ otherwise. If one wishes to take into account the dispersion of y, Negative Binomial distribution $\{\mathrm{NB}(y|\mu, \phi) : \mu, \phi \in \mathbb{R}_+\}$, where

$\mathrm{NB}(y|\mu,\phi) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)y!}\left(\frac{\mu}{\phi+\mu}\right)^y\left(\frac{\phi}{\phi+\mu}\right)^\phi$ may be used. If the values of μ or ϕ need to be non-negative or belong to a compact set, one may apply functions such as exponential functions or the sigmoid (inverse logit) function compositing with a polynomial or trigonometric polynomial parametrized by θ_1, θ_2. The general theory to be presented will be applicable to all these models, and others.

Wasserstein distances As discussed in the Introduction, all parameters in the mixture model for the conditional distribution $f_G(y|x)$ of the response y given covariate x are encapsulated by the latent mixing measure $G = \sum_{j=1}^k p_j \delta_{(\theta_{1j},\theta_{2j})}$. In order to characterize identifiability and learning rates of parameter learning, one needs a suitable metric for the mixing measure G. Wasserstein distances have become a useful tool to quantify the convergence of latent mixing measures in mixture models [31]. Given two discrete measures $G = \sum_{j=1}^k p_j\delta_{\theta_j}$ and $G' = \sum_{j=1}^{k'} p'_j\delta_{\theta'_j}$ on a normed space Θ endowed with a norm $\|\cdot\|$, the W_r Wasserstein metric, in which $r \geq 1$, is defined as:

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

where the infimum is taken over all joint distribution on $[1,\ldots,k]\times[1,\ldots,k']$ such that $\sum_{i=1}^k q_{ij} = p'_j, \sum_{j=1}^{k'} q_{ij} = p_i$. Note that for $G_0 = \sum_{j=1}^{k_0} p_j^0\delta_{\theta_j^0} \in \mathcal{E}_{k_0}$, if $G = \sum_{j=1}^k p_j\delta_{\theta_j}$ varies on $\mathcal{O}_k$ such that $W_r(G,G_0)\to 0$ and Θ is compact, then

$$W_r^r(G,G_0) \asymp \sum_{i=1}^{k_0}\left|\sum_{\theta_j\in V_i} p_j - p_i^0\right| + \sum_{i=1}^{k_0}\sum_{\theta_j\in V_i} p_j\left\|\theta_j - \theta_i^0\right\|^r, \qquad (2)$$

where $V_i = \{\theta : \left\|\theta - \theta_i^0\right\| \leq \left\|\theta - \theta_{i'}^0\right\| \forall i' \neq i\}$ is the Voronoi cell of θ_i^0 in Θ (see, e.g., [19]). Hence, for every atom of G_0, there is a subset of atoms of G converging to it at the same rate as $W_r(G,G_0)\to 0$. Therefore, the convergence in a Wasserstein metric W_r implies the convergence of parameters in mixture models. In this paper, unless noted otherwise the space $\Theta = \Theta_1\times\Theta_2$ is chosen to be a compact subset of $\mathbb{R}^{d_1+d_2}$ and $\|\cdot\|$ is the usual ℓ^2 distance.

Mixtures of conditional densities In a regression mixture model, a focus of inference will be on the conditional density $f_G(y|x)$, while there will be as little assumption as possible on the marginal distribution of covariate X. It is clear from the representation of $f_G(y|x)$ that the identifiability and parameter learning behavior of the regression problem will repose upon suitable conditions specified by f, h and the unknown parameter G. The analysis of conditional density estimation requires us to control how large the conditional density family $\{f(y|h_1(x,\theta_1),h_2(x,\theta_2)) : \theta_1\in\Theta_1, \theta_2\in\Theta_2\}$ is. This can be accomplished by assuming Lipschitz conditions on f, h_1 and h_2. In particular, we say that f is uniformly Lipschitz if there exists $c_f > 0$ such that for all $\mu, \mu' \in H_1, \phi, \phi' \in H_2$:

$$\sup_{y \in \mathcal{Y}} |f(y|\mu, \phi) - f(y|\mu', \phi')| \le c_f(|\mu - \mu'| + |\phi - \phi'|). \tag{3}$$

The link functions h_1 and h_2 are called uniformly Lipschitz if there are $c_1, c_2 > 0$ such that for all $\theta_1, \theta_1' \in \Theta_1, \theta_2, \theta_2' \in \Theta_2$:

$$\sup_{x \in \mathcal{X}} |h_1(x, \theta_1) - h_1(x, \theta_1')| \le c_1 \|\theta_1 - \theta_1'\|, \ \sup_{x \in \mathcal{X}} |h_2(x, \theta_2) - h_2(x, \theta_2')| \le c_2 \|\theta_2 - \theta_2'\|. \tag{4}$$

In a regression problem, one is interested in prediction error guarantee in addition to assessing the quality of parameter estimates. For a standard (single component) regression model, we often model $f(y|x) = f(y|h_1(x, \theta_1), h_2(x, \theta_2))$, where $h_1(x, \theta)$ is the mean parameter, i.e., $\mathbb{E}[Y|X = x] = h_1(x, \theta)$. After estimating $\hat{\theta}_1$ from the data, the prediction error is customarily taken to be the mean square error $\mathbb{E}_X(h_1(X, \theta_1^0) - h_1(X, \hat{\theta}_1))^2$, where θ_1^0 is the true parameter. For a regression mixture, let the true latent mixing measure be $\sum_{j=1}^{k_0} p_j^0 \delta_{(\theta_{1j}^0, \theta_{2j}^0)}$ for which an estimate is denoted by $\sum_{j=1}^{k} \hat{p}_j \delta_{(\hat{\theta}_{1j}, \hat{\theta}_{2j})}$. In this setting, due to the heterogeneous nature of the response, the predicted value for y at any x may be taken by the quantity $\sum_{j=1}^{k} \hat{p}_j \delta_{h_1(x, \hat{\theta}_{1j})}$, or its mean $\sum_{j=1}^{k} \hat{p}_j h_1(x, \hat{\theta}_{1j})$. As a result, the prediction error for the mean estimate can be written as $\mathbb{E}_X W_2^2 \left(\sum_{j=1}^{k_0} p_j^0 \delta_{h_1(X, \theta_j^0)}, \sum_{j=1}^{k} \hat{p}_j \delta_{h_1(X, \hat{\theta}_{1j})} \right)$. If one is interested in describing the prediction error in terms of both the mean trend and dispersion, one can use $\mathbb{E}_X W_2^2 \left(\sum_{j=1}^{k_0} p_j^0 \delta_{(h_1(X, \theta_{1j}^0), h_2(X, \theta_{2j}^0))}, \sum_{j=1}^{k} \hat{p}_j \delta_{(h_1(X, \hat{\theta}_{1j}), h_2(X, \hat{\theta}_{2j}))} \right)$.

Key inequalities The following basic inequality controls the *expected* total variation distance between conditional densities by a Wasserstein distance between the corresponding parameters:

Lemma 1. *Assume conditions* (3) *and* (4) *hold. Then for every* $G \in \mathcal{O}_K(\Theta)$ *and* $K \ge 1$*, we have*

$$\mathbb{E}_X[d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X))] \precsim W_1(G, G_0), \tag{5}$$

where the multiplicative constant in this inequality only depends on c_f, c_1*, and* c_2*.*

The inequality established in the above lemma quantifies the impact of parameter estimation on the quality of conditional density estimation: if G is well estimated, then so is the conditional distribution represented by the conditional densities $f_G(Y|X)$. In order to quantify the identifiability and convergence of the unknown parameter G, we will need to establish inequalities of the following type:

$$\mathbb{E}_X[d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X))] \succsim W_r^r(G, G_0), \tag{6}$$

for all G in some space of latent mixing measures, and r depends on that space. Following [31, 42, 6], we refer to this as *inverse bounds*, because in our setting, they allow us to lower bound the distance between conditional probability models (f_G and f_{G_0}) by the distance between the parameters of inferential interest (G and G_0). Unlike prior works, our inverse bounds control the *expected* total variational distance under the marginal distribution of the covariate X. A simple observation is that these inverse bounds are quantitative versions of the classical identifiability condition [37] for the regression problem, because if $f_G = f_{G_0}$ for a.e. x, y, then the bound (6) entails that $G = G_0$. Moreover, the inverse bounds play an important role in establishing the convergence rate for parameter estimation. They allow us to translate convergence rates for density estimation (left-hand side of Eq. (6)) into that of parameter estimation (right-hand side of Eq. (6)). The technique to prove inverse bounds is to rely on a notion of strong identifiability to be developed for regression mixture models in the following section.

3. Strong identifiability and inverse bounds

3.1. Conditions of strong identifiability

Identifiability and strong identifiability conditions play important roles in the theoretical analysis of mixture models [37, 2, 17]. They provide a finer characterization of the non-singularity of the Fisher information for mixtures of distributions [19]. In plain words, these conditions require that the kernel density function of interest and its derivatives up to a certain order with respect to all relevant parameters be linearly independent. For the mixture of regression model (1), the kernel density function is that of the conditional probability of variable y given covariate x. The following definition is our formulation of strong identifiability for the conditional density functions:

Definition 1. *The family of conditional densities* $\{f(y|h_1(x,\theta_1), h_2(x,\theta_2)) : \theta_1 \in \Theta_1, \theta_2 \in \Theta_2\}$ *(or in short,* $f(\cdot|h_1,h_2)$*) is identifiable in order r, where* $r=1$ *(resp.,* $r=2$*) with complexity level k, if* $f(y|h_1(x,\theta_1), h_2(x,\theta_2))$ *is differentiable up to order r with respect to* (θ_1,θ_2)*, and (A1.) (resp., (A2.)) holds.*

(A1.) (First order identifiable) For any given k distinct elements $(\theta_{11},\theta_{21}),\ldots,(\theta_{1k},\theta_{2k}) \in \Theta_1\times\Theta_2$*, if there exist* $\alpha_j\in\mathbb{R}, \beta_j\in\mathbb{R}^{d_1}, \gamma_j\in\mathbb{R}^{d_2}$ *as* $j=1,\ldots,k$ *such that for almost all* x,y *(w.r.t.* $\mathbb{P}_X\times\nu$*)*

$$\sum_{j=1}^{k} \alpha_j f_j(y|x) + \beta_j^\top \frac{\partial}{\partial\theta_1} f_j(y|x) + \gamma_j^\top \frac{\partial}{\partial\theta_2} f_j(y|x) = 0,$$

then $\alpha_j=0, \beta_j=0\in\mathbb{R}^{d_1}, \gamma_j=0\in\mathbb{R}^{d_2}$ *for* $j=1,\ldots,k$*;*

(A2.) (Second order identifiable) For any given k distinct elements $(\theta_{11},\theta_{21}),\ldots,(\theta_{1k},\theta_{2k}) \in \Theta_1\times\Theta_2$ *and* $s_1,\ldots,s_k\geq 1$*, if there exist* $\alpha_j\in\mathbb{R}, \beta_j\in\mathbb{R}^{d_1}, \gamma_j\in\mathbb{R}^{d_2}$*, and* $\rho_{jt}\in\mathbb{R}^{d_1}, \nu_{jt}\in\mathbb{R}^{d_2}$ *as* $j=1,\ldots,k, t=1,\ldots,s$ *such that for almost all* x,y *(w.r.t.* $\mathbb{P}_X\times\nu$*)*

$$\sum_{j=1}^{k} \alpha_j f_j(y|x) + \beta_j^\top \frac{\partial}{\partial\theta_1} f_j(y|x) + \gamma_j^\top \frac{\partial}{\partial\theta_2} f_j(y|x) + \sum_{t=1}^{s_j}\left(\rho_{jt}^\top \frac{\partial}{\partial\theta_1^2} f_j(y|x)\rho_{jt}\right)$$

$$+\sum_{t=1}^{s_j}\left(\nu_{jt}^{\top}\frac{\partial}{\partial\theta_2^2}f_j(y|x)\nu_{jt}\right)+\sum_{t=1}^{s_j}\left(\rho_{jt}^{\top}\frac{\partial}{\partial\theta_1\partial\theta_2}f_j(y|x)\nu_{jt}\right)=0,$$

then $\alpha_j = 0, \beta_j = \rho_{jt} = 0 \in \mathbb{R}^{d_1}, \gamma_j = \nu_{jt} = 0 \in \mathbb{R}^{d_2}$ *for* $t = 1, \dots, s_j, j = 1, \dots, k$.

When we speak of strong identifiability without specifying the complexity level, it should be understood that the condition is satisfied for any complexity level $k \geq 1$. These strong identifiability conditions for conditional density functions are useful in deriving rates of convergence for the regression mixture model's parameters even when the associated Fisher information matrices are singular, e.g., when the model has redundant parameters. Indeed, when showing the convergence rate of an estimator G to the true mixing measure G_0 in the over-fitted setting, there might exist several redundant atoms of G converge to a common atom of G_0. The customary technique of applying the first-order Taylor expansion around $f_{G_0}(\cdot|X)$ may fail because the coefficients of these redundant components can be combined and canceled out. Instead, one needs to perform a Taylor expansion up to the second order around $f_{G_0}(\cdot|X)$, necessitating the second-order identifiability condition developed here. It will be shown in the sequel that the strong identifiability conditions hold for most popular mixtures of regression models. There are notable exceptions which shall be discussed separately. For instance, a mixture of binomial regression models generally satisfies strong identifiability only up to a finite complexity level.

Since our model (1) is hierarchical with two levels of parameters:

$$G_0 = \sum_{j=1}^{k_0} p_j^0 \delta_{(\theta_{1j},\theta_{2j})} \mapsto \sum_{j=1}^{k_0} p_j^0 \delta_{(h_1(x,\theta_{1j}),h_2(x,\theta_{2j}))} \mapsto \sum_{j=1}^{k_0} p_j^0 f(y|h(x,\theta_{1j}),h(x,\theta_{2j})), \tag{7}$$

it is difficult to directly verify conditions (A1.) and (A2.). We will show in the following that they can be deduced from the identifiability conditions of a family of (unconditional) distribution $\{f(y|\mu,\phi) : \mu,\phi\}$ and family of functions (h_1, h_2). Recall from [17, 31]:

Definition 2. *The family of (unconditional) distributions* $\{f(y|\mu,\phi) : (\mu,\phi) \in H\}$ *(or in short,* f*) is identifiable in order* r *with complexity level* k*, for some* $r, k \geq 0$*, if* $f(y|\mu,\phi)$ *is differentiable up to order* r *in* (μ,ϕ) *and the following holds:*

(A3.) For any given k *distinct elements* $(\mu_1,\phi_1),\dots,(\mu_k,\phi_k) \in H$*, if for each pair of* $n := (n_1,n_2)$*, where* $n_1 \geq n_2 \geq 0, n_1+n_2 \leq r$*, we have* $\alpha_n^{(j)} \in \mathbb{R}$ *such that*

$$\sum_{l=0}^{r}\sum_{n_1+n_2=l}\sum_{j=1}^{k}\alpha_n^{(j)}\frac{\partial^{n_1+n_2}f}{\partial\mu^{n_1}\partial\phi^{n_2}}(y|\mu_j,\phi_i)=0$$

for almost all y*, then* $\alpha_n^{(j)} = 0$ *for all* $1 \leq j \leq k$ *and pair* $n = (n_1, n_2)$*.*

Condition (A3.) for $r = 0$ simply ensures that the mixture of f distributions model uniquely identifies the mixture components. The strong identifiability conditions ($r \geq 1$) are required to establish the convergence rates [2, 31]. In particular, we only need to consider $r = 1$ and 2 (first and second-order identifiability) in this article. In the model (1), there is a hierarchically higher level of parameters (θ_1, θ_2) that we want to learn, and it connects to the observations through the link functions h_1, h_2 as $\mu = h_1(x, \theta_1), \phi = h_2(x, \theta_2)$. To ensure that θ_1 and θ_2 can be learned efficiently, we also need suitable conditions for h_1 and h_2.

Definition 3. *The family of functions* $\{(h_1(x, \theta_1), h_2(x, \theta_2)) : \theta_1 \in \Theta_1, \theta_2 \in \Theta_2\}$ *is called* <u>*identifiable with complexity level k*</u> *respect to* $\mathbb{P}_X$ *if the following conditions hold:*

(A4.) For every set of $k+1$ *distinct elements* $(\theta_{11}, \theta_{21}), ..., (\theta_{1(k+1)}, \theta_{2(k+1)}) \in \Theta_1 \times \Theta_2$*, there exists a subset* $A \subset \mathcal{X}$, $\mathbb{P}_X(A) > 0$ *such that* $(h_1(x, \theta_{11}), h_2(x, \theta_{21})), ..., (h_1(x, \theta_{1(k+1)}), h_2(x, \theta_{2(k+1)}))$ *are distinct for every* $x \in A$*;*

(A5.) Moreover, if there are vector $\beta_1 \in \mathbb{R}^{d_1}, \beta_2 \in \mathbb{R}^{d_2}$ *such that*

$$\beta_1^\top \frac{\partial}{\partial \theta_1} h_1(x, \theta_{1j}) = 0, \quad \beta_2^\top \frac{\partial}{\partial \theta_2} h_2(x, \theta_{2j}) = 0 \quad \forall x \in A \setminus N, \quad j = 1, \ldots, k+1,$$

where N *is a zero-measure set (i.e.,* $\mathbb{P}_X(N) = 0$*), then* $\beta_1 = 0$ *and* $\beta_2 = 0$.

Remark 1. *1. Condition (A4.) is necessary for identifying regression mixture components. Indeed, for two distinct pairs* (θ_1, θ_2) *and* (θ_1', θ_2') *in* $\Theta_1 \times \Theta_2$*, there may exists some point* $x \in \mathcal{X}$ *so that* $h_1(x, \theta_1) = h_1(x, \theta_1'), h_2(x, \theta_2) = h_2(x, \theta_2')$*. If we only observe data* (x, y) *at such* x*, it is not possible to distinguish between* (θ_1, θ_2) *and* (θ_1', θ_2')*.*

2. In linear models, condition (A5.) reads that there is no multicollinearity: If we model $h_1(x, \theta) = \theta_1 \psi_1(x) + \cdots + \theta_{d_1} \psi_{d_1}(x)$*, where* ψ_i*'s are pre-defined functions, then by substitute this into condition (A5.), we have* $\psi_1, \ldots, \psi_{d_1}$ *must be linearly independent as functions of* x*. Otherwise, the model is not identifiable with respect to parameters* θ_j*'s.*

3. (A4.) and (A5.) can be viewed as generalization (to non-linear) and population versions of condition (1b) and (2) in [13] (or condition in Theorem 2.2 in [15]).

Hence, the two conditions in Definition 3 are necessary for learning parameters of the mixture of regression models. The following result shows that Definition 2 and Definition 3 give sufficient conditions to deduce the strong identifiability given by Definition 1, where the chain rule plays an essential role in its proof.

Theorem 1. *For any complexity level* k*, if the family of distributions* f *is strongly identifiable in order* r *(via (A3.)) and the family of functions* h *is identifiable (via (A4.) and (A5.)), then the family of conditional density* $f(y|x)$ *is strongly identifiable in order* r*, where* $r = 1, 2$*.*

3.2. Characterization of strong identifiability

Theorem 1 provides a simple recipe for establishing the strong identifiability of the conditional densities arising in regression mixture models (1) by checking the identifiability conditions of family f and family h. In the following, we provide specific examples.

Proposition 1. *(a) The family of location normal distribution* $\{\mathcal{N}(y|\mu,\sigma^2) : \mu \in \mathbb{R}\}$ *with fixed variance* σ^2 *is identifiable in the second order, for* $\mathcal{N}(y|\mu,\sigma^2) = \exp(-(y-\mu)^2/2\sigma^2)$*. The location-scale family* $\{\mathcal{N}(y|\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ *is identifiable in the first order;*
(b) The Poisson family $\{\mathrm{Poi}(y|\lambda) : \lambda \in \mathbb{R}^+\}$ *is identifiable in the second order;*
(c) The family of Binomial distributions $\{\mathrm{Bin}(y|N,q) : q \in [0,1]\}$ *with fixed number of trials* N *is identifiable in the first order with complexity level* k *if* $2k \leq N+1$*, and is identifiable in the second order with complexity level* k *if* $3k \leq N+1$*;*
(d) The family of negative binomial distributions $\{\mathrm{NB}(y|\mu,\phi) : \mu\}$ *with fixed* $\phi \in \mathbb{R}_+$ *is identifiable in the second order.*

The identifiability conditions (A4.) and (A5.) usually hold for parametric models, as we see below. We first define a general class of functions:

Definition 4. *We say a family of functions* $\{h(x,\theta) : \theta \in \Theta\}$ *is completely identifiable if for any* $\theta \neq \theta' \in \Theta$*, we have* $h(x,\theta) \neq h(x,\theta')$ *almost surely in* $\mathbb{P}_X$*.*

Proposition 2. *If* h_1 *and* h_2 *are both completely identifiable, then the family of functions* $\{(h_1,h_2)\}$ *satisfies condition (A4.).*

Most functions used in parametric regression mixture models are completely identifiable.

Proposition 3. *Suppose that* $\mathbb{P}_X$ *has a density with respect to Lebesgue measure on* $\mathcal{X}$*, then the following families of functions are completely identifiable and satisfy condition (A5.):*
(a) Polynomial of finite dimensions $h(x,\theta) = \sum_{d_1+\ldots+d_p \leq d, d_i \geq 0} \theta_{(d_1,\ldots,d_p)} x_1^{d_1} \ldots x_p^{d_p}$*, where* $d \in \mathbb{N}_+$ *and* $\theta = (\theta_{(d_1,\ldots,d_p)} : d_i \geq 0, \sum_{i=1}^p d_i \leq d)$*;*
(b) Trigonometric polynomials in $\mathbb{R}$*:* $h(x,\theta) = a_0 + \sum_{n=1}^d b_n \sin(nx) + \sum_{n=1}^d c_n \cos(nx)$*, where* $\theta = (a_0, b_1, \ldots, b_d, c_1, \ldots, c_d)$*;*
(c) Mixtures of polynomials and trigonometric polynomials as in (a) and (b): $h(x,\theta) = \sum_{n=0}^d a_n x^n + \sum_{n=1}^d b_n \sin(nx) + \sum_{n=1}^d c_n \cos(nx)$*, where* $\theta = (a_0, \ldots, a_d, b_1, \ldots, b_d, c_1, \ldots, c_d)$*;*
(d) $h(x,\theta) = g(p(x,\theta))$*, where* g *is a diffeomorphism, i.e., a continuously differentiable bijective function, and* $p(x,\theta)$ *is completely identifiable and satisfies condition (A5.).*

Remark 2. *In a general linear model,* $h(x,\theta) = \exp(\theta^\top x) \in \mathbb{R}_+$ *or* $h(x,\theta) = \sigma(\theta^\top x) \in [0,1]$*, where* σ *is the sigmoid (inverse logit) function. Both the expo-*

nential function and sigmoid function are one-to-one, and $\theta^\top x$ is a first-order polynomial, so the above results apply.

3.3. Inverse bounds for mixture of regression models

At the heart of our convergence theory for parameter learning in regression mixture models lies a set of inverse bounds, which are given as follows.

Theorem 2. *(a) (Exact-fitted) Given $G_0 \in \mathcal{E}_{k_0}(\Theta)$ for $k_0 \in \mathbb{N}_+$. Suppose that the family of conditional densities $\{f(\cdot|h_1, h_2)\}$ is identifiable in the first order, and the family of functions (h_1, h_2) is identifiable (with the complexity level k_0). Then for all $G \in \mathcal{E}_{k_0}(\Theta)$, there holds*

$$\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X)) \succsim W_1(G, G_0), \tag{8}$$

where the constant in this inequality depends only on $G_0, h_1, h_2, f, \mathbb{P}_X$, and ν (but not on G).

(b) (Over-fitted) Given $G_0 \in \mathcal{E}_{k_0}(\Theta)$ for $k_0 \in \mathbb{N}_+$ and $k_0 \leq \bar{K}$ for some natural number $\bar{K}$. Suppose that the family of conditional densities $\{f(\cdot|h_1, h_2)\}$ is identifiable in the second order, and the family of functions (h_1, h_2) is identifiable (with the complexity level $\bar{K}$). Then for all $G \in \mathcal{O}_{\bar{K}}(\Theta)$, there holds

$$\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X)) \succsim W_2^2(G, G_0). \tag{9}$$

where the constant in this inequality depends only on $G_0, h_1, h_2, f, \mathbb{P}_X$, and ν (but not on G).

If the true number of components k_0 is known, then Theorem 2 entails that the convergence rate for parameter estimations can be as fast as the convergence rate for conditional densities under the total variation distance. However, in practice, we may not know k_0 and fit the system by a large number $\bar{K}$. In this over-fitted regime, provided that the identifiability conditions for distribution f and function h are satisfied in the second order, the convergence rate for parameter estimation may be twice as slow as that of the conditional densities.

Based on the convergence behavior of the regression mixture model's parameters, we can establish guarantees on the prediction error for the response variable. The following bounds will be useful for deducing the prediction error bounds from that of parameter estimates.

Proposition 4. *Suppose that the density f and link functions h_1, h_2 are uniformly Lipschitz, then for all $G \in \mathcal{O}_{\bar{K}}(\Theta), G = \sum_{j=1}^{\bar{K}} p_j \delta_{(\theta_{1j}, \theta_{2j})}$ and $r \geq 1$, we have*

$$W_r(G, G_0) \succsim \mathbb{E}_X W_r\left(\sum_{j=1}^{\bar{K}} p_j \delta_{(h_1(X,\theta_{1j}), h_2(X,\theta_{2j}))}, \sum_{j=1}^{k_0} p_j^0 \delta_{(h_1(X,\theta_{1j}^0), h_2(X,\theta_{2j}^0))}\right),$$

and

$$W_r(G, G_0) \succsim \mathbb{E}_X \left| \sum_{j=1}^{\bar{K}} p_j h_u(X, \theta_{uj}) - \sum_{i=1}^{k_0} p_i^0 h_u(X, \theta_{ui}^0) \right| \quad \forall u = 1, 2,$$

where the constants in those inequalities only depend on Lipschitz constants of h_1 *and* h_2*.*

3.4. Consequences of lack of strong identifiability

Strong identifiability notions characterize the favorable conditions under which efficient regression learning is possible in the mixture setting. Next, we turn our attention to the consequence of the lack of strong identifiability. Firstly, we note that the normal distributions satisfy the well-known heat equation: $\frac{\partial^2}{\partial \mu^2}\mathcal{N}(y|\mu,\sigma^2) = \frac{\partial}{\partial(\sigma^2)}\mathcal{N}(y|\mu,\sigma^2), \quad \forall \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$, so the mixture of normal regression model no longer satisfies the strong identifiability condition in the second order. This (Fisher information matrix's) singularity structure is universal (i.e., holds for all μ, σ^2). Therefore, the inverse bounds presented in Theorem 2 for the *over-fitted setting* may not hold and potentially lead to slow convergence rates for the mixture of regression with normal kernel, which can be seen in some recent work [20, 30, 29]. However, this kernel does satisfy the strong identifiability in the first order, so the parameter estimation rate in the mixture of normal regression is as fast as the density estimation rate in the *exact-fitted setting*, as established by Theorem 1 and 2.

Another interesting example arises in negative binomial regression mixture models, which have been utilized in the traffic analysis of heterogeneous environments [33, 32]. These authors observed via many empirical experiments that the quality of parameter estimates and the prediction performance may be affected by the (overlapped) sample-mean values obtained from the data. However, there was a lack of precise theoretical understanding. Our theoretical framework can be applied to shed light on the behavior of this class of regression mixture model. It starts with the observation that the mixture of negative binomial distributions does not even satisfy the first-order strongly identifiable condition. Moreover, we can identify precisely the instances where strong identifiability fails to hold and investigate the impact on the quality of parameter estimates and the prediction performance in such instances.

First, we note that the mean-dispersion negative binomial conditional density $\{\mathrm{NB}(y|\mu,\phi) : \mu \in \mathbb{R}_+, \phi \in \mathbb{R}_+\}$ satisfies the following equation:

$$\frac{\partial}{\partial \mu}\mathrm{NB}(y|\mu,\phi) = \frac{\phi}{\mu}\mathrm{NB}\left(y|\mu\frac{\phi+1}{\phi}, \phi+1\right) - \frac{\phi}{\mu}\mathrm{NB}(y|\mu,\phi), \forall y \in \mathbb{N}. \tag{10}$$

Thus, a 2-mixture of negative binomial distributions $\mathrm{NB}(y|\mu_1,\phi_1)$ and NB $(y|\mu_2,\phi_2)$ such that

$$\frac{\mu_1}{\phi_1} = \frac{\mu_2}{\phi_2} \text{ and } \phi_1 = \phi_2 + 1 \tag{11}$$

does not satisfy the strong identifiability condition in the first order. As a result, the inverse bounds of the general type (6) cannot hold, even in the exact-fitted setting. The following proposition makes this clear.

Proposition 5. *Consider a mixture of negative binomial regression model with link functions* $h(x, \theta_1) = \exp(\theta_{10} + (\bar{\theta}_1)^\top x), h(x, \theta_2) = \theta_2$ *for* $\theta_1 = [\theta_{10}, \bar{\theta}_1] \in \mathbb{R}^{p+1}$ *and* $\theta_2 \in \mathbb{R}_+$. *For any* $k_0 \geq 2$ *and* $r \geq 1$, *there exists* G_0 *and a sequence* $G_n \in \mathcal{E}_{k_0}$ *such that* $W_r(G_n, G_0) \to 0$ *and*

$$\mathbb{E} d_{TV}(f_{G_n}(\cdot|X), f_{G_0}(\cdot|X)) \preccurlyeq W_r^{2r}(G_n, G_0). \tag{12}$$

As a consequence of (12), *for any measurable estimate* $\hat{G}_n$ *of the mixing measure* G, *the following holds for any* $r \geq 1$:

$$\inf_{\hat{G}_n \in \mathcal{E}_{k_0}} \sup_{G \in \mathcal{E}_{k_0}} \mathbb{E}_{\mathbb{P}_G} W_r(\hat{G}_n, G) \succcurlyeq n^{-1/(4r)}. \tag{13}$$

Remark 3. *By plugging* $r = 1$ *into* (12), *it is clear that the inverse bound* (8) *for the exact-fitted setting no longer holds for the negative binomial regression mixture family. Part of the reason for the slow minimax rate is due to the use of* W_r. *For a general finite mixture, it is not difficult to show a minimax lower bound* $n^{-1/(2r)}$ *under* W_r, *provided that the number of mixture components is known. Here, a non-trivial minimax rate* $n^{-1/(4r)}$ *twice as slow arises due to the violation of strong identifiability for the negative binomial mixtures. The sharpest minimax rate remains unknown for this model setting.*

Nonetheless, we will show that non-identifiability occurs only in a Lebesgue measure zero subset of the parameter space.

Proposition 6. *Given* k *distinct pairs* $(\mu_1, \phi_1), \ldots, (\mu_k, \phi_k) \in \mathbb{R}_+ \times \mathbb{R}_+$ *such that there does not exist two indices* $i \neq j$ *satisfying* $\frac{\mu_i}{\phi_i} = \frac{\mu_j}{\phi_j}$ *and* $|\phi_i - \phi_j| = 1$, *then the mixture of negative binomials* $(\mathrm{NB}(\mu_i, \phi_i))_{i=1}^k$ *is strongly identifiable in the first order. If we further assume that there does not exist two indices* $i \neq j$ *satisfying* $\frac{\mu_i}{\phi_i} = \frac{\mu_j}{\phi_j}$ *and* $|\phi_i - \phi_j| \in \{1, 2\}$, *then the mixture of negative binomials* $(\mathrm{NB}(\mu_i, \phi_i))_{i=1}^k$ *is strongly identifiable in the second order.*

Finally, we note that the theory established earlier (Theorem 1 and Theorem 2) represent sufficient conditions. There still may exist non-identifiable or non-strongly identifiable families f and (h_1, h_2) that lead to strong identifiable $f(\cdot|h_1, h_2)$. As an example, the mixture of two Binomial distributions $p(y) = p_1 \mathrm{Bin}(y|1, q_1) + p_2 \mathrm{Bin}(y|1, q_2)$, with parameter (p_1, p_2, q_1, q_2) is not even identifiable: e.g., the model is the same with $(p_1, p_2, q_1, q_2) = (.5, .5, .3, .7)$ and $(p_1, p_2, q_1, q_2) = (.5, .5, .2, .8)$. However, the mixture of two logistic regression models $f_G(y|x) = p_1 \mathrm{Bin}(y|1, \sigma(\theta_1^\top x)) + p_2 \mathrm{Bin}(y|1, \sigma(\theta_2^\top x)))$ is strongly identifiable and enjoys the inverse bound as well as standard convergence rates, as shown in the following proposition.

Proposition 7. *Suppose that the link function $h(x,\theta) = \sigma(\theta x) := 1/(1+e^{\theta x})$ for $\theta, x \in \mathbb{R}$, and the density kernel $f(y) = Bin(y|1,q)$, where $q = h(x,\theta)$. Moreover, the support of $\mathbb{P}_X$ contains an open set in $\mathbb{R}$. Then, the mixture of two binomial regression components associated with the mixing measure $G = p_1\delta_{\theta_1} + p_2\delta_{\theta_2}$, where $\theta_1 + \theta_2 \neq 0$, is strongly identifiable in the first order.*

The proof of such a result involves some specific analytical properties of kernel f and function h and is difficult to generalize. This again highlights our general theory developed in this section, which is applicable to a vast range of kernels f and link functions (h_1, h_2). The pathological phenomena described will be revisited in Section 5 via simulations.

4. Statistical efficiency in learning regression mixtures

Building on the previous section, we are ready to present convergence rates of the maximum likelihood estimator, and a Bayesian posterior contraction theory for the quantities of interest.

4.1. Maximum (conditional) likelihood estimation

Given n i.i.d. observations $(x_1,y_1),\ldots,(x_n,y_n)$, where $x_j \overset{i.i.d.}{\sim} \mathbb{P}_X$ and $y_j|x_j \sim f_{G_0}(y|x), j = 1,\ldots,n$, for $G_0 = \sum_{i=1}^{k_0} p_j^0 \delta_{(\theta_{1j}^0,\theta_{2j}^0)}$. Denote the maximum likelihood estimate by

$$\widehat{G}_n := \underset{G\in\mathcal{E}_{k_0}(\Theta)}{\arg\max} \sum_{j=1}^{n} \log f_G(y_j|x_j),$$

in the exact-fitted setting, and we change the $\mathcal{E}_{k_0}(\Theta)$ in the above formula to $\mathcal{O}_K(\Theta)$, where $K \geq k_0$ in the over-fitted setting. It is implicitly assumed in this section that $\widehat{G}_n$ is measurable, otherwise a standard treatment using an outer measure of $\mathbb{P}_{G_0}$ instead of $\mathbb{P}_{G_0}$ can be invoked. To obtain the rate of convergence of $\widehat{G}_n$ to G_0, we combine the inverse bounds above with the convergence of density estimates based on the standard theory of M-estimation for regression problems [38]. For conditional density estimation, the convergence behavior of $f_{\widehat{G}_n}$ to f_{G_0} is evaluated in the sense of the *expected* Hellinger distance:

$$\begin{aligned}\overline{d}_H^2(f_G, f_{G'}) &:= \mathbb{E}_X d_H^2(f_G(\cdot|X), f_{G'}(\cdot|X)) \\ &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{Y}} (\sqrt{f_G(y|x)} - \sqrt{f_{G'}(y|x)})^2 d\nu(y) d\mathbb{P}_X(x),\end{aligned}$$

for all $G, G' \in \cup_{k=1}^{\infty}\mathcal{O}_k(\Theta)$. To this end, recall several basic notions related to the entropy numbers of a class of functions. For any $k \in \mathbb{N}$, set

$$\mathcal{F}_k(\Theta) = \left\{ f_G(y|x) : G \in \mathcal{O}_k(\Theta) \right\}, \quad \overline{\mathcal{F}}_k^{1/2}(\Theta) = \left\{ f_{(G+G_0)/2}^{1/2}(y|x) : G \in \mathcal{O}_k(\Theta) \right\},$$

and the Hellinger ball centered around f_{G_0}:

$$\overline{\mathcal{F}}_k^{1/2}(\delta) = \overline{\mathcal{F}}_k^{1/2}(\Theta, \delta) = \left\{ f^{1/2} \in \overline{\mathcal{F}}_k^{1/2}(\Theta) : \overline{d}_H(f, f_{G_0}) \leq \delta \right\}.$$

The complexity (richness) of this set is characterized by the following entropy integral:

$$\mathcal{J}(\delta) := \mathcal{J}(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta, \delta)) = \left(\int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(u, \overline{\mathcal{F}}_k^{1/2}(\delta), L_2(\mathbb{P}_X \times \nu)) du \right) \vee \delta, \tag{14}$$

where H_B is the bracketing entropy number, L_2 distance between two conditional densities f and g is defined as $\|f - g\|_{L_2} := (\int d\mathbb{P}_X \int d\nu (f(y|x) - g(y|x))^2)^{1/2}$, and the constant 2^{13} appears due to the chaining technique in bounding supremum of empirical processes [38, 10]. A useful tool for establishing the rate of convergence under expected conditional density estimation by the MLE is given by the following theorem, which is an adaptation of Theorem 7.4. in [38] or Theorem 7.2.1. in [10]).

Theorem 3. *Take $\Psi(\delta) \geq \mathcal{J}(\delta, \overline{\mathcal{F}}_k^{1/2}(\delta))$ in such a way that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then, for a universal constant c and for*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n), \tag{15}$$

we have for all $\delta \geq \delta_n$ that $\mathbb{P}_{G_0}(\overline{d}_H(f_{\widehat{G}_n}, f_{G_0}) > \delta) \leq c \exp\left(-\frac{n\delta^2}{c^2}\right)$.

Combining Theorem 3 with the inverse bounds established in Section 3, we readily arrive at the following concentration inequalities for the MLE's parameter estimates based on the bracketing entropy numbers and its entropy integral given by Eq. (14).

Theorem 4. *(a) (Exact-fitted) Suppose that k_0 is known, the entropy condition (15) holds, the family of conditional densities $f(\cdot|h_1, h_2)$ is identifiable in the first order, and (h_1, h_2) is identifiable. Then, for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$, there exist a constant C depending on G_0 and universal constant c such that for all $\delta > \delta_n$:*

$$\mathbb{P}_{G_0}\left(W_1(\widehat{G}_n, G_0) > C\delta\right) \leq c\exp(-n\delta^2/c^2).$$

(b) (Over-fitted) Suppose that k_0 is unknown but $k_0 < \bar{K}$ known, the entropy condition (15) holds, the family of conditional densities $f(\cdot|h_1, h_2)$ is identifiable in the second order, and (h_1, h_2) is identifiable. Then, for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$, there exist a constant C depending on G_0 and K and universal constant c such that for all $\delta > \delta_n$:

$$\mathbb{P}_{G_0}\left(W_2^2(\widehat{G}_n, G_0) > C\delta\right) \leq c\exp(-n\delta^2/c^2).$$

From the definition and condition of the entropy integral (Eq. (14)), we see that the key to establishing convergence rate δ_n is by controlling the bracketing entropy number $H_B(u, \overline{\mathcal{F}}_k^{1/2}, L_2)$. Generally, under some conditions on the smoothness and boundedness of the model, we can show that the bracketing entropy number of the model space is in the same order as that of the parameter space, which then yields the parametric convergence rates for finite-dimensional models. By generalizing the proof technique for bounding entropy number of Gaussian mixture models in [9], in the following, we shall present a set of mild assumptions for general classes of regression mixture models and obtain rates of convergence under these assumptions.

(B1.) (Assumptions on kernel densities) For the family of distribution $\{f(y|\mu,\phi)|\mu \in H_1, \phi \in H_2\}$, $\|f(\cdot|\mu,\phi)\|_\infty$ is uniformly bounded and uniformly light tail probability, i.e., there exist constant $\underline{c}, \overline{c}$ and constant $d_1, d_2, d_3 > 0$ such that $f(y|\mu,\phi) \leq d_1 \exp(-d_2|y|^{d_3})$, for all $y \geq \overline{c}$ or $y \leq \underline{c}$ and $\mu \in H_1, \phi \in H_2$.

(B2.) (Lipschitz assumptions) Both the kernel densities $\{f(y|\mu,\phi)|\mu \in H_1, \phi \in H_2\}$ and link functions h_1 and h_2 are uniformly Lipschitz in the sense of (3) and (4).

We note that:

Proposition 8. *The families of Normal, Poisson, Binomial, and negative binomial distribution satisfy condition (B1.).*

The predictive performance of MLE for the regression mixture model is given as follows.

Theorem 5. *Given assumptions (B1.) and (B2.), and the dominating measure ν is Lebesgue over $\mathbb{R}$ or counting measure on $\mathbb{Z}$. Then, there exists a constant C depending on Θ, f, h_1, h_2, and a universal constant c such that*

$$\mathbb{P}_{G_0}\left(\overline{d}_H(f_{\widehat{G}_n}, f_{G_0}) > C\sqrt{\frac{\log n}{n}}\right) \leq c\exp(-\log n/c^2).$$

Combining Theorem 2 and Theorem 5, we arrive at the convergence rates for the maximum (conditional) likelihood estimates for the model parameters:

Theorem 6. *(a) (Exact-fitted) Suppose that k_0 is known, the family of conditional densities $f(\cdot|h_1,h_2)$ is identifiable in the first order, and the family of functions (h_1,h_2) is identifiable. Furthermore, assume (B1.) and (B2.) hold, then for any $G_0 \in \mathcal{E}_{k_0}(\Theta)$, there exist constant C depending on G_0, Θ, f, h_1, h_2 and a universal constant c such that*

$$\mathbb{P}_{G_0}\left(W_1(\widehat{G}_n, G_0) > C(\log n/n)^{1/2}\right) \leq c\exp(-\log n/c^2).$$

(b) (Over-fitted) Suppose that k_0 is unknown and is upper bounded by a known number $\bar{K} < +\infty$, the family of conditional densities $f(\cdot|h_1,h_2)$ is identifiable in the second order, and the family of functions (h_1,h_2) is identifiable. Furthermore, assume (B1.) and (B2.) hold, there exist constant C

depending on $G_0, \bar{K}, \Theta, f, h_1, h_2$ and a universal constant c such that

$$\mathbb{P}_{G_0}\left(W_2(\widehat{G}_n, G_0) > C(\log n/n)^{1/4}\right) \leq c\exp(-\log n/c^2).$$

Remark 4. *The tail bounds in Theorem 6 can simply imply the convergence in expectation $\mathbb{E}W_1(\widehat{G}_n, G_0) \leq C(\log n/n)^{1/2}$ (in part (a)) and $\mathbb{E}W_2(\widehat{G}_n, G_0) \leq C(\log n/n)^{1/4}$ (in part (b)), where constant C depends on G_0, by using Markov's inequality. We present a proof in Appendix A.2.*

We have established that in the over-fitted setting, a standard application of the MLE yields a parameter estimation rate that is twice as slow compared to the exact-fitted setting. This rate is not optimal in the pointwise sense. However, the established convergence behavior motivate the following model selection procedure that can asymptotically choose the correct k, which in turns yields the optimal estimation rate, modulo a logarithmic factor, for the parameters.

Proposition 9. *Let $\widehat{G}_n^{(k)} \in \mathcal{O}_k$ be the MLE estimated from data with at most k atoms, for $k = 1, \dots, \overline{K}$. For a sequence $(a_n) \subset \mathbb{R}_+$, let*

$$k_n = \inf\{k : W_2(\widehat{G}_n^{(k')}, \widehat{G}_n^{(k'+1)}) \leq a_n \forall k' \geq k\}.$$

If $(\log n/n)^{1/4} \ll a_n \ll 1$, there exist a constant C depending on $\Theta, h_1, h_2, f, \overline{K}$, and a universal constant c such that for all sufficiently large n depending on G_0 and (a_n),

$$\mathbb{P}_{G_0}(k_n = k_0) \geq 1 - cn^{-1/c^2}, \tag{16}$$

and

$$\mathbb{P}_{G_0}\left(W_1(\widehat{G}_n^{(k_n)}, G_0) \geq C\left(\frac{\log n}{n}\right)^{1/2}\right) \leq 2cn^{-1/c^2}. \tag{17}$$

4.2. Bayesian posterior contraction theorems for parameter inference

Given i.i.d. pairs $(x_1, y_1), \dots, (x_n, y_n)$ such that $y_i|x_i \sim f_{G_0}(y|x)$ for some true latent mixing measure $G_0 = \sum_{j=1}^{k_0} p_i^0 \delta_{(\theta_{1j}^0, \theta_{2j}^0)}$, and $x_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X$. In the Bayesian regime, we model the data as $y_i|x_i \sim f_G(y|x)$, where $G \sim \Pi$ with Π being some prior distribution on the space mixing measures. Let $\mathcal{G}$ denote the support of the prior Π on the mixing measure G. The nature of the prior distribution Π depends on the several different settings that we will consider. In the exact-fitted setting, we assume $\bar{K} \equiv k_0 < +\infty$ is known, whereas the over-fitted setting means that the upper bound $k_0 \leq \bar{K} < +\infty$ is given, but k_0 unknown. In both cases, $\mathcal{G} = \mathcal{O}_{\bar{K}}$ so Π is in effect a prior distribution on $(\mathcal{O}_{\bar{K}}, \mathcal{B}(\mathcal{O}_{\bar{K}}))$. Later, we shall assume that neither k_0 nor an upper bound $\bar{K}$ is given; instead, a random variable K is used to represent the number of mixture components and endowed with a prior distribution.

By the Bayes' rule, the posterior distribution of the parameter G is given by

$$\Pi(G \in B|x^{[n]}, y^{[n]}) = \frac{\int_B \prod_{i=1}^n f_G(y_i|x_i) d\Pi(G)}{\int_{\mathcal{O}_{\bar{K}}} \prod_{i=1}^n f_G(y_i|x_i) d\Pi(G)}$$

for any measurable set $B \subset \mathcal{G}$. Now, we want to study the posterior contraction rate of $\Pi(\cdot|x^{[n]}, y^{[n]})$ to the true latent mixing measure G_0 as $n \to \infty$. We proceed to describe several standard assumptions on the prior often employed in practice.

The case of known upper bound $\bar{K}$

(B3.) (Prior assumption) Prior Π on space $\mathcal{O}_{\bar{K}}$ is induced by prior $\Pi_p \times \Pi_{\theta_1}^{\bar{K}} \times \Pi_{\theta_2}^{\bar{K}}$ on $\{(p_1, \ldots, p_{\bar{K}}, \theta_{11}, \ldots, \theta_{1\bar{K}}, \theta_{21}, \ldots, \theta_{2\bar{K}})|p_i \geq 0, \sum_{i=1}^{\bar{K}} p_i = 1, \theta_{1j} \in \Theta_1, \theta_{2j} \in \Theta_2\}$, where Π_p is a prior distribution of $(p_j)_{i=1}^{\bar{K}}$ on $\Delta^{\bar{K}-1}$, Π_{θ_1} is a prior distribution of θ_{1j} on Θ_1, and Π_{θ_2} is a prior distribution of θ_{2j} on Θ_2, independently for $i = 1, \ldots, \bar{K}$. We further assume that $\Pi_p, \Pi_{\theta_1}, \Pi_{\theta_2}$ have a density with respect to Lebesgue measure on $\Delta^{\bar{K}-1}, \Theta_1, \Theta_2$, respectively, which are bounded away from zero and infinity.

(B4.) There exists $\epsilon_0 > 0$ such that for all $G \in \mathcal{O}_{\bar{K}}(\Theta)$ satisfying $W_1(G, G_0) \leq \epsilon_0$, we have $\mathbb{E}_{\mathbb{P}_{G_0}}(f_{G_0}/f_G) \leq M_0$ for M_0 only depends on $\epsilon_0, G_0, \bar{K}, \Theta$.

The posterior contraction behavior for conditional densities is given as follows.

Theorem 7. *Assume that (B2.)-(B4.) hold. For any $G_0 \in \mathcal{G}$, there exists constant C depending on Θ, f, h_1, h_2 such that as $n \to \infty$,*

$$\Pi\left(G{:}\overline{d}_H(f_G(y|X), f_{G_0}(y|X)) \geq C\sqrt{\frac{\log n}{n}}\middle| x^{[n]}, y^{[n]}\right) \to 0 \text{ in } \otimes_{i=1}^n \mathbb{P}_{G_0} \text{-probability.} \quad (18)$$

Combining the above result with the inverse bounds developed in Section 3 leads to the contraction rates of the posterior distribution in the exact-fitted and over-fitted settings.

Theorem 8. *Suppose that assumptions (A1.)-(A3.) and (B2.)-(B4.) hold. Fix any $G_0 \in \mathcal{G}$.*

(a) (Exact-fitted) If $\bar{K} = k_0$, there exists some constant C_1 depending on G_0, Θ, f, h_1, h_2 such that

$$\Pi\left(G{:}W_1(G, G_0) \geq C_1 \left(\frac{\log n}{n}\right)^{1/2} \middle| x^{[n]}, y^{[n]}\right) \stackrel{n\to\infty}{\longrightarrow} 0 \text{ in } \otimes_{i=1}^n \mathbb{P}_{G_0} \text{-probability.}$$

(b) (Over-fitted) If $\bar{K} \geq k_0$, there exists some constant C_2 depending on $G_0, \bar{K}, \Theta, f, h_1, h_2$ such that

$$\Pi\left(G{:}W_2(G, G_0) \geq C_2 \left(\frac{\log n}{n}\right)^{1/4} \middle| x^{[n]}, y^{[n]}\right) \stackrel{n\to\infty}{\longrightarrow} 0 \text{ in } \otimes_{i=1}^n \mathbb{P}_{G_0} \text{-probability.}$$

The case of unknown $\bar{K}$

Finally, when the number of components K and its upper bound are unknown, there are various approaches for prior specification. Here, we adopt the widely utilized "mixture-of-finite-mixtures" prior [27]. In particular, the prior distribution Π on the space $\mathcal{G}$ of mixing measures is induced by the following specification. We show that the optimal rate of parameter estimation, up to a logarithmic factor, is achieved.

(B5.) A prior distribution Π_K on K with support in $\mathbb{N}$, i.e., $\Pi_K(K = k) > 0$ for all $k \in \mathbb{N}$.

(B6.) For each $k \in \mathbb{N}$, given the event $K = k$, the conditional prior distribution of the mixing measure $G = \sum_{j=1}^{k} p_j \delta_{(\theta_{1j},\theta_{2j})} \in \mathcal{E}_k$ is induced by the following specification: $\Pi_p \times \Pi_{\theta_1}^k \times \Pi_{\theta_2}^k$ on $\{(p_1, \dots, p_k, \theta_{11}, \dots, \theta_{1k}, \theta_{21}, \dots, \theta_{2k})|p_i \geq 0, \sum_{i=1}^{k} p_i = 1, \theta_{1j} \in \Theta_1, \theta_{2j} \in \Theta_2\}$, where Π_p is a prior distribution of $(p_j)_{i=1}^k$ on Δ^{k-1}, Π_{θ_1} is a prior distribution of θ_{1j} on Θ_1, and Π_{θ_2} is a prior distribution of θ_{2j} on Θ_2, independently for $i = 1, \dots, k$. Assume that $\Pi_p, \Pi_{\theta_1}, \Pi_{\theta_2}$ have a density with respect to Lebesgue measure on $\Delta^{k-1}, \Theta_1, \Theta_2$, respectively, which are bounded away from zero and infinity.

(B7.) For each $k \in \mathbb{N}$, there exists $\epsilon_0 > 0$ such that for all $G \in \mathcal{E}_k(\Theta)$ satisfying $W_1(G, G_0) \leq \epsilon_0$, we have $\mathbb{E}_{\mathbb{P}_{G_0}}(f_{G_0}/f_G) \leq M_0$ for M_0 only depends on $\epsilon_0, G_0, k, \Theta$.

Theorem 9. *Assume that (A1.)-(A3.), (B2.), and (B5.)-(B7.) hold. There exists a subset $\mathcal{G}_0 \subset \mathcal{G}$, where $\Pi(\mathcal{G}_0) = 1$ such that for all $k_0 \in \mathbb{N}$ and $G_0 \in \mathcal{G}_0 \cap \mathcal{E}_{k_0}$, there hold as $n \to \infty$*

(a) $\Pi(K = k_0|x^{[n]}, y^{[n]}) \to 1$ *a.s. under* $\otimes_{i=1}^n \mathbb{P}_{G_0}$;

(b) there is a constant C depending on G_0, Θ, f, h_1, h_2 such that

$$\Pi\left(G{:}W_1(G,G_0)\geq C\left(\frac{\log n}{n}\right)^{1/2}\middle|x^{[n]},y^{[n]}\right) \to 0 \textit{ in } \otimes_{i=1}^n \mathbb{P}_{G_0} \textit{ probability.}$$

5. Simulations and data illustrations

Parameter estimation in exact-fitted and over-fitted settings We first illustrate the parameter estimation rates of the mixture of regression models under general settings. Consider a mixture of normal regression model with a polynomial link function and fixed variance: $f_G(y|x) = \sum_{i=1}^{k} p_i \mathcal{N}(y|h(x, \theta_i), \sigma^2)$, where $x \in \mathbb{R}, G = \sum_{i=1}^{k} p_i \delta_{\theta_i}, \theta_i \in \mathbb{R}^3, h(x, \theta_i) = \theta_{i1} + \theta_{i2}x + \theta_{i3}x^2$. We simulate data $(x_i)_{i=1}^n$ from an uniform distribution on $[-3, 3]$ and then generated y_i given x_i from this model using $G_0 = p_1^0 \delta_{\theta_1^0} + p_2^0 \delta_{\theta_2^0}$, where $\sigma = 1, \theta_1^0 = (1, -5, 1), \theta_2^0 = (2, 5, 2), p_1^0 = p_2^0 = 0.5$. Because the variances of both components are known, and the link function is a polynomial, this model is strongly identifiable in the second order (by Theorem 1). The maximum (conditional) likelihood estimate $\hat{G}_n$ of G_0 is obtained by the Expectation-Maximization (EM) algorithm. We

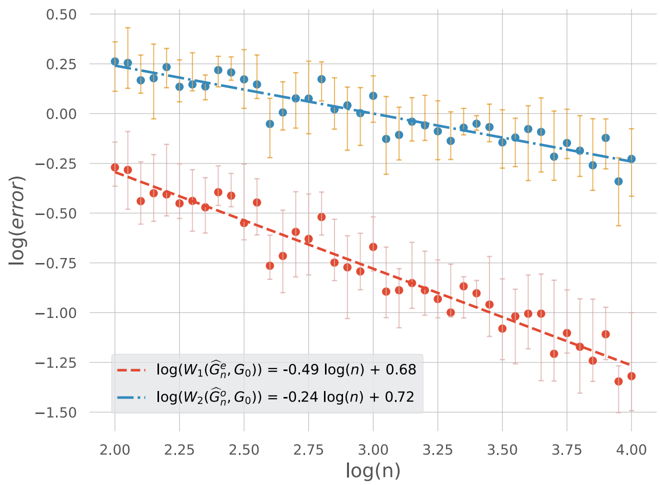
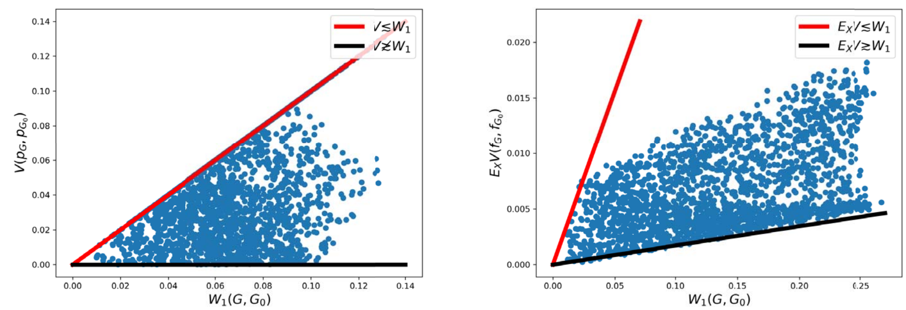


Fig 1. *Convergence rates of parameters estimation in the exact-fitted and over-fitted setting.*

considered two cases where the number of components of G_0 is known to be 2 for the exact-fitted setting and set $k = 3$ for the over-fitted setting. For each setting, we consider the logarithm of the sample sizes $\log_{10}(n)$ ranging from 2 to 4 (so that n ranges from 100 to $10,000$) and generate n samples from the true distribution. Each experiment is repeated 16 times, and we plot the average (with quartile bar) of the logarithm of the error in Figure 1 and perform a simple linear regression against $\log n$ (best to see with color). It can be seen that the convergence rate of the exact-fitted mixing measures is $n^{-1/2}$ (in Wasserstein-1 distance), and the over-fitted mixing measures is $n^{-1/4}$ (in Wasserstein-2 distance). These results are compatible with the theoretical results established in Theorem 4.

Regression mixtures vs unconditional mixtures The characterization results (Theorem 1 and propositions in Section 3.2) provide easy-to-check sufficient conditions for strong identifiability (in the sense of Def. 1) and lead to the parameter estimation rate as we see above. Part of the sufficient conditions requires that the kernel density f be strongly identifiable up to a certain order, a standard condition considered in the asymptotic theory for finite (and unconditional) mixture models [31, 18, 19]. It is noteworthy that the strong identifiability condition of a mixture of regression model given in Def. 1 is typically a weaker condition than that of a standard unconditional mixture model, because the presence of the covariate x makes the conditional mixture model more constrained. Hence, it is possible that for an unconditional mixture of kernel densities f strong identifiability and hence the inverse bound $V \succcurlyeq W_1$ may not hold, but when f is utilized in a regression mixture model instead, the strong identifiability and hence the inverse bound still holds.

We demonstrate this observation by a theoretical result given in Proposi-

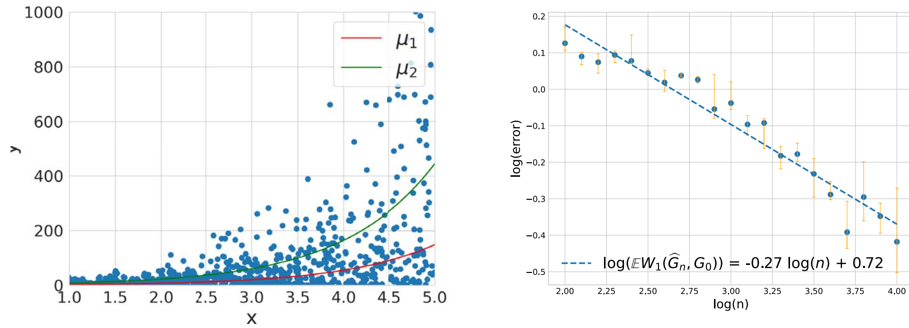


(a) Mixture of binomial distributions (b) Mixture of binomial regression models

FIG 2. *Impact of identifiability on inverse bounds: unconditional mixtures with binomial kernel are not identifiable (left) while regression mixtures with the same kernel are strongly identifiable (right).*

tion 7 for the mixture of binomial regression models. To illustrate this result by a simulation study, let $p_G(y) := \sum_{i=1}^k p_i \text{Bin}(y|N, q_i)$, for $G = \sum_{i=1}^k p_i \delta_{q_i} \in \mathcal{O}_K([0,1])$, where N is a fixed natural number. From the discussion in Section 3, we know that this model is only strongly identifiable in the first order if $2k \leq N+1$. It means the inverse bound may not hold when $2k > N+1$. Let $k=2$, $N=1$ so that $2k > N+1$, and $G_0 = 0.5\delta_{0.3} + 0.5\delta_{0.7}$, and then uniformly generate 2000 random samples of G around G_0. We compare $W_1(G, G_0)$ against $d_{TV}(p_G, p_{G_0})$ to see if the inverse bound $d_{TV}(p_G, p_{G_0}) \succcurlyeq W_1(G, G_0)$ holds or not. It can be seen in Fig. 2(a) that such an inverse bound does not hold. In contrast, for the mixture of binomial regression model under the same setting $k=2, N=1$ (a.k.a. mixture of two logistic regression): $f_G(y|x) = p_1 \text{Bin}(y|1, \sigma(\theta_1 x)) + p_2 \text{Bin}(y|1, \sigma(\theta_2 x))$ for $G = p_1\delta_{\theta_1} + p_2\delta_{\theta_2}$ and σ being the sigmoid function, the inverse bound as established by Theorem 2 still holds. We uniformly sample 2000 measure G around $G_0 = 0.5\delta_{0.5} + 0.5\delta_5$ and plot $W_1(G, G_0)$ against $\mathbb{E}_X d_{TV}(f_G(y|X), f_{G_0}(y|X))$, for $X \sim \text{Uniform}([-6, 6])$. The relationship $\mathbb{E}_X d_{TV}(f_G(y|X), f_{G_0}(y|X)) \succcurlyeq W_1(G, G_0)$ holds in this scenario (Fig. 2(b)). As a consequence, the mixture of logistic regression model still enjoys the convergence rate of $n^{-1/2}$ for parameter estimation in the exact-fitted setting.

Investigating the lack of strong identifiability In Section 3.4, we discussed the lack of strong identifiability in the mixture of negative binomial regression models. Here, we conduct an experiment to show the posterior contraction rate of parameter estimation where the true model is not strongly identifiable (cf. Eq. (11) holds). A dataset is drawn from a negative binomial regression mixture model: $f_G(y|x) = \sum_{j=1}^2 p_j \text{NB}(y|h(x, \theta_j), \phi_j)$, where the covariate, x_i, is randomly generated from a uniform distribution over the interval $[3, 6]$; the component means are $h(x_i, \theta_1) = \exp(X_i\theta_1)$, $h(x_i, \theta_2) = \exp(X_i\theta_2)$, where $X_i = (1, x_i), i = 1, ..., n$. The true latent mixing measure



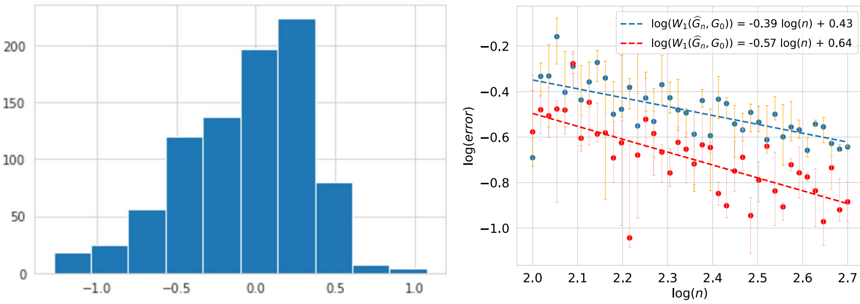
(a) Simulated data from a NB mixture (b) Posterior contraction rate

FIG 3. *Convergence rate in exact-fitted case where the model is non-strongly-identifiable.*

is $G_0 = p_1^0 \delta_{\{\theta_1^0, \phi_1^0\}} + p_2^0 \delta_{\{\theta_2^0, \phi_2^0\}}$, where the regression coefficients are $\theta_1^0 = (0,1)'$, $\theta_2^0 = (\log 3, 1)'$, the dispersion parameters are $\phi_1^0 = 0.5$, $\phi_2^0 = 1.5$ and the mixing proportions are $p_1^0 = 0.4$, $p_2^0 = 0.6$. Under this simulation, Eq. (11) is satisfied. The simulated dataset is illustrated in Figure 3(a). This model is not strongly identifiable.

For model fitting, we adopt the Bayesian approach and consider the exact-fitted case. Similar to [32], we choose the prior p to be Beta$(1,1)$, $\theta_j \sim \mathcal{N}(0, I_2)$ (normal distribution with identity covariance matrix), and $\phi_j^{-1} \sim$ Gamma $(0.01, 0.01)$ (a non-informative gamma distribution) for $j = 1, 2$. The full posterior distribution is approximated using a Gibbs MCMC algorithm with a Metropolis-Hasting step for ϕ_j due to the non-conjugate prior. Details are given in Appendix E.1. For each different sample size n, we run the experiment 8 times. For each time running, we produce 2500 MCMC samples, discard the first 500, and use the remaining samples to estimate the expected Wasserstein distances to G_0. The logarithm of estimation error averaged over the 8 runs (with quartile error bars) is reported in Fig 3(b). By performing a simple linear regression, we can see that W_1 error is of order $n^{-1/4}$, suggesting a slow rate of pointwise convergence even in the exact-fitted setting. We note that this slow convergence behavior is hinted (but not proved) by the inequality and minimax bound in Proposition 5.

Analysis of crash data To validate the applicability of the proposed theoretical result in Section 3.2 illustrated via the simulation study above, we use the dataset collected in 1995 at urban 4-legged signalized intersections in Toronto, Canada. The same data has been explored and fitted by a mixture of negative binomial regression models by [32], where they showed the good quality of the dataset as well as the best-fitted model for it. This crash data set contains 868 intersections, which have a total of $10,030$ reported crashes. In their paper, the authors explicated the heterogeneity in the dataset which came from the existence of several different sub-populations (i.e., the data collected from the



(a) Histogram of difference $\mu_1/\phi_1 - \mu_2/\phi_2$

(b) The red dashed line corresponds to the subsamples from the whole dataset while the blue dashed line to subsamples chosen near a pathological situation.

FIG 4. *The impact of Crash data being near pathological cases of non-strong identifiability.*

different business environments, contains a mix of fixed and actuated traffic signals, and so on). Accordingly, the mean functional form has been used for each component as below:

$$\mu_{j,i} = h(\boldsymbol{F}_i, \theta_j) = \theta_{j,0} F_{1i}^{\theta_{j,1}} F_{2i}^{\theta_{j,2}}, \text{ for } j = 1, 2 \text{ and } i = 1, ..., 868, \tag{19}$$

where $\mu_{j,i}$ is the jth component's estimated number of crashes for intersection i; F_{1i} counts the entering flows in vehicles/day from the major approaches at intersection i; F_{2i} the entering flows in vehicles/day from the minor approaches at intersection i; and $\theta_j = (\log(\theta_{j,0}), \theta_{j,1}, \theta_{j,2})'$ the estimated regression for component j. According to [32], the best model for describing the dataset is a two-mixture of negative binomial regression where $\phi_1 = 9.3692$, $\phi_2 = 8.2437$, $p_1 = 0.43, p_2 = 0.57, \theta_1 = (-10.9407, 0.8588, 0.5056)'$, $\theta_2 = (-9.7842, 0.3987, 0.8703)'$.

It can be seen that the two values of ϕ_1 and ϕ_2 nearly satisfy the second condition of the pathological case mentioned in Section 3.2 (i.e., $\phi_1 \approx \phi_2 + 1$). If the first condition holds (i.e., $\frac{\mu_1}{\phi_1} = \frac{\mu_2}{\phi_2}$), then we would be in a singular situation. To verify this, we calculate $\left(\dfrac{\mu_{1,i}}{\phi_1} - \dfrac{\mu_{2,i}}{\phi_2}\right)$ for all samples $(\boldsymbol{F}_i)_{i=1}^{868}$. The histogram of this difference can be seen in Fig. 4(a). By the Anderson-Darling test, we see that this distribution is significantly different from the degenerate distribution at 0, with the calculated p-value for this test being 1.28×10^{-13}. Hence, we are quite far from the pathological situations of non-strong identifiability. In theory, the method should still enjoy the $n^{-1/2}$ convergence rate if the model is well-specified and exact-fitted. We further subsample this data and calculate the error of the estimator from the subsampled dataset to that of the whole dataset. For each sample size, we replicate the experiment 8 times and report the average error (in red) and the interquartile error bar (in orange) in Fig. 4(b). We can see that the error is approximate of the order $n^{-1/2}$.

Finally, we conduct another subsampling experiment to focus on the data corresponding to $\left|\frac{\mu_{1,i}}{\phi_1} - \frac{\mu_{2,i}}{\phi_2}\right| \leq 0.3$. This data subset (with sample size 502) represents a data population that is closer to the pathological cases of non-strong identifiability than that of the previous experiment. A noteworthy observation is that the closer the data population is to a pathological situation, the slower the actual convergence to the true parameters will be. In particular, the blue dashed line in Fig. 4(b) represents the average errors in this case after an 8-time running of the experiment. This result provides an interesting demonstration of the population theory given in the previous section.

6. Conclusion

We developed a strong identifiability theory for general finite mixture of regression models and derived rates of convergence for density estimation and parameter estimation in both Bayesian and MLE frameworks. This theory was shown to be applicable to a wide range of models employed in practice. It also invites interesting new questions. First, in our models mixture weights p_j's do not vary with covarate x. It would be interesting to extend the theory to the situation of co-varying weights. Second, while our theory is applicable to the case of unknown but finite number of mixture components, it remains challenging to extend such a theory for infinite conditional mixtures motivated from Bayesian nonparametrics [31, 16]. Finally, as demonstrated with the negative binomial mixtures, although the singularity situation (i.e., strong identifiability is violated) is rare, being in the vicinity of a singular model can be a far more common scenario. We would like to investigate more precisely the impact on parameter estimates when the true model is in the vicinity of a singular model, and to provide suitable statistical methods to overcome the inefficiency of inference in such situations.

Supplementary material

In the supplementary material, we collect the proofs and additional information deferred from the main text. Section A includes the proofs for all main theoretical results. Section B includes the proofs for all remaining theoretical results. Section C and Section D provide a general theory of M-estimators convergence rates and posterior contraction rates in the regression setting, respectively. The EM algorithms for mixtures of regression are described in Section E.

Appendix A: Proofs of main results

A.1. Identifiablity conditions and inverse bounds

Proof of Theorem 1. We will divide the proof into first order and second order identifiability cases.

The case $r = 1$: For some $k \in \mathbb{N}$, suppose that there exist k distinct elements $(\theta_{11}, \theta_{21}), \dots, (\theta_{1k}, \theta_{2k}) \in \Theta_1 \times \Theta_2$, and $\alpha_j \in \mathbb{R}, \beta_j \in \mathbb{R}^{d_1}, \gamma_j \in \mathbb{R}^{d_2}$ as $j = 1, \dots, K$ such that for almost all x, y (w.r.t. $\mathbb{P}_X \times \nu$)

$$\sum_{j=1}^{k} \alpha_j f_j(y|x) + \beta_j^\top \frac{\partial}{\partial \theta_1} f_j(y|x) + \gamma_j^\top \frac{\partial}{\partial \theta_2} f_j(y|x) = 0,$$

then we want to show that $\alpha_j = 0, \beta_j = 0 \in \mathbb{R}^{d_1}, \gamma_j = 0 \in \mathbb{R}^{d_2}$ for $j = 1, \dots, k$. Indeed, by the chain rule, the left-hand side of the above expression is equal to

$$\sum_{j=1}^{k} \alpha_j f_j(y|x) + \beta_j^\top \frac{\partial h_1(x, \theta_{1j})}{\theta_{1j}} \frac{\partial}{\partial \mu} f(y|h_1(x, \theta_{1j}), h_2(x, \theta_{2j})) \\ + \gamma_j^\top \frac{\partial h_2(x, \theta_{2j})}{\partial \theta_{2j}} \frac{\partial}{\partial \phi} f(y|h_1(x, \theta_{1j}), h_2(x, \theta_{2j})).$$

Because of the identifiability of (h_1, h_2) (condition (A4.)), there exists a set A with $\mathbb{P}_X(A) > 0$ such that $(h_1(x, \theta_{1j}), h_2(x, \theta_{2j}))_{j=1}^k$ are distinct. Therefore, by the first order identifiability of f, we have

$$\beta_j^\top \frac{\partial h_1(x, \theta_1)}{\partial \theta_1} = 0, \quad \gamma_j^\top \frac{\partial h_2(x, \theta_2)}{\partial \theta_2} = 0,$$

for all $x \in A$ (possibly except a $\mathbb{P}_X$ zero-measure set). Hence, by condition (A5.), we have $\beta_j = 0, \gamma_j = 0$ for all $j = 1, \dots, k$. This concludes the first-order identifiability of the family of conditional densities $f(\cdot|h_1, h_2)$.

The case $r = 2$: For any $k, s \geq 1$, given k distinct elements $(\theta_{11}, \theta_{21}), \dots, (\theta_{1k}, \theta_{2k}) \in \Theta_1 \times \Theta_2$, if there exist $\alpha_j \in \mathbb{R}, \beta_j \in \mathbb{R}^{d_1}, \gamma_j \in \mathbb{R}^{d_2}$, and $\rho_{jt} \in \mathbb{R}^{d_1}, \nu_{jt} \in \mathbb{R}^{d_2}$ as $t = 1, \dots, s_j$ and $j = 1, \dots, k$ such that for almost all x, y (w.r.t. $\mathbb{P}_X \times \nu$)

$$\sum_{j=1}^{k} \alpha_j f_j(y|x) + \beta_j^\top \frac{\partial}{\partial \theta_1} f_j(y|x) + \gamma_j^\top \frac{\partial}{\partial \theta_2} f_j(y|x) + \sum_{t=1}^{s_j} \left(\rho_{jt}^\top \frac{\partial}{\partial \theta_1^2} f_j(y|x) \rho_{jt} \right) \\ + \sum_{t=1}^{s_j} \left(\nu_{jt}^\top \frac{\partial}{\partial \theta_2^2} f_j(y|x) \nu_{jt} \right) + \sum_{t=1}^{s_j} \left(\rho_{jt}^\top \frac{\partial}{\partial \theta_1 \partial \theta_2} f_j(y|x) \nu_{jt} \right) = 0,$$

then we want to show that $\alpha_j = 0, \beta_j = \rho_{jt} = 0 \in \mathbb{R}^{d_1}, \gamma_j = \nu_{jt} = 0 \in \mathbb{R}^{d_2}$ for $j = 1, \dots, k, t = 1, \dots, s$. Indeed, again, by the chain rule, the expression above is equivalent to

$$\sum_{j=1}^{k} \alpha_j f_j(y|x) + \left(\beta_j^\top \frac{\partial h_1(x, \theta_1)}{\partial \theta_1} + \sum_{t=1}^{s_j} \rho_{jt}^\top \frac{\partial^2 h_1(x, \theta_{1j})}{\partial \theta_1^2} \rho_{jt} \right) \frac{\partial}{\partial \mu} f_j(y|x) \\ + \sum_{t=1}^{s_j} \left(\rho_{jt}^\top \frac{\partial h_1(x, \theta_{1j})}{\partial \theta_1} \right)^2 \frac{\partial}{\partial \mu^2} f_j(y|x)$$

$$+\left(\gamma_j^\top \frac{\partial h_2(x,\theta_2)}{\partial \theta_2} + \sum_{t=1}^{s_j} \nu_{jt}^\top \frac{\partial^2 h_2(x,\theta_{2j})}{\partial \theta_2^2} \nu_{jt}\right) \frac{\partial}{\partial \phi} f_j(y|x)$$
$$+\sum_{t=1}^{s_j} \left(\nu_{jt}^\top \frac{\partial h_2(x,\theta_{2j})}{\partial \theta_2}\right)^2 \frac{\partial^2}{\partial \phi^2} f_j(y|x)$$
$$+2\sum_{t=1}^{s_j} \left(\rho_{jt}^\top \frac{\partial^2}{\partial \theta_1 \partial \theta_2} f_j(y|x) \nu_{jt}\right) \frac{\partial^2}{\partial \mu \partial \phi} f_j(y|x) = 0. \quad (20)$$

Because of the identifiability of (h_1, h_2) (condition (A4.)), there exists a set A with $\mathbb{P}_X(A) > 0$ such that $(h_1(x,\theta_{1j}), h_2(x,\theta_{2j}))_{j=1}^k$ are distinct. By the second-order identifiability of f, then we have

$$\alpha_j = 0,$$
$$\beta_j^\top \frac{\partial h_1(x,\theta_1)}{\theta_1} + \sum_{t=1}^{s_j} \rho_{jt}^\top \frac{\partial^2 h_1(x,\theta_{1j})}{\partial \theta_1^2} \rho_{jt} = 0, \quad \sum_{t=1}^{s_j} \left(\rho_{jt}^\top \frac{\partial h_1(x,\theta_{1j})}{\partial \theta_1}\right)^2 = 0,$$
$$\gamma_j^\top \frac{\partial h_2(x,\theta_2)}{\theta_2} + \sum_{t=1}^{s_j} \nu_{jt}^\top \frac{\partial^2 h_2(x,\theta_{2j})}{\partial \theta_2^2} \nu_{jt} = 0, \quad \sum_{t=1}^{s_j} \left(\nu_{jt}^\top \frac{\partial h_2(x,\theta_{2j})}{\partial \theta_2}\right)^2 = 0,$$
$$\sum_{t=1}^{s_j} \left(\rho_{jt}^\top \frac{\partial^2}{\partial \theta_1 \partial \theta_2} f_j(y|x) \nu_{jt}\right) = 0,$$

for all $x \in A$, possibly except a $\mathbb{P}_X$ zero-measure set. Hence, by condition (A5.), from the third and fifth equation above, we have $\alpha_j = 0, \rho_{jt} = 0, \nu_{jt} = 0$ for all $t = 1, \ldots, s_j, j = 1, \ldots, k$. These together with the second and fourth equation further imply that $\beta_j = 0, \gamma_j = 0$ for all $j = 1, \ldots, k$. We arrive at the second-order identifiability of family of conditional densities $f(y|x)$ as desired. □

Proof of Theorem 2. In order to prove part (a) of the theorem, we divide it into two regimes, local and global, and establish two following claims: for any $\epsilon' > 0$,

$$\inf_{G \in \mathcal{E}_{k_0}(\Theta): W_1(G,G_0) > \epsilon'} \frac{\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X))}{W_1(G,G_0)} > 0, \quad (21)$$

$$\lim_{\epsilon \to 0} \inf_{G \in \mathcal{E}_{k_0}(\Theta)} \left\{ \frac{\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X))}{W_1(G,G_0)} : \ W_1(G,G_0) \le \epsilon \right\} > 0. \quad (22)$$

Proof of claim (21): Suppose that this is not true, that is, there exists a sequence $G_\ell \in \mathcal{E}_k(\Theta)$ such that $W_1(G_\ell, G_0) > \epsilon'$ for all n and as $n \to \infty$,

$$\frac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{W_1(G_\ell, G_0)} \to 0.$$

This implies $\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X)) \to 0$. Since Δ^{k_0-1}, Θ_1 and Θ_2 are compact, there exists a subsequence of $(G_\ell)_\ell$ (which is assumed to be $(G_\ell)_\ell$ itself without loss of generality) that converges weakly to some $G' = \sum_{j=1}^{k'} p'_j \delta_{(\theta'_{1j}, \theta'_{2j})} \in$

$\mathcal{O}_{k_0}(\Theta_1\times\Theta_2)$, where $(\theta'_{11},\theta'_{21}),\ldots,(\theta'_{1k'},\theta'_{2k'})$ are distinct. Hence, $W_1(G',G_0)>\epsilon'$, and $\mathbb{E}_X d_{TV}(f_{G'}(\cdot|X),f_{G_0}(\cdot|X))=0$. Because $d_{TV}(f_{G'}(\cdot|x),f_{G_0}(\cdot|x))\geq 0$ for all x, this implies

$$f_{G'}(y|x)=f_{G_0}(y|x)\quad \text{a.s. in } x,y,$$

which means

$$\sum_{j=1}^{k'} p'_j f(y|h_1(x,\theta'_{1j}),h_2(x,\theta'_{2j}))=\sum_{j=1}^{k_0} p^0_j f(y|h_1(x,\theta^0_{2j}),h_2(x,\theta^0_{2j}))\quad \text{a.s. in } x,y.$$

By the zero-order identifiability of f, we conclude that

$$\sum_{j=1}^{k'} p'_j \delta_{(h_1(x,\theta'_{1j}),h_2(x,\theta'_{2j}))}=\sum_{j=1}^{k_0} p^0_j \delta_{(h_1(x,\theta^0_{1j}),h_2(x,\theta^0_{2j}))}\quad \text{a.s. in } x. \tag{23}$$

If family (h_1,h_2) is identifiable, then we argue that the set $\{(\theta'_{11},\theta'_{21}),\ldots,(\theta'_{1k'},\theta'_{1k'})\}$ must equal $\{(\theta^0_{11},\theta^0_{21}),\ldots,(\theta^0_{1k_0},\theta^0_{2k_0})\}$. Otherwise, we can assume (without loss of generality) that $(\theta'_{11},\theta'_{21})\notin\{(\theta^0_{11},\theta^0_{21}),\ldots,(\theta^0_{1k_0},\theta^0_{2k_0})\}$, which means there exists a set A with $\mathbb{P}_X(A)>0$ such that $(h_1(x,\theta'_{11}),\ h_2(x,\theta'_{21}))$ is distinct from any pair among $(h_1(x,\theta^0_{11}),h_2(x,\theta^0_{21}))\ldots,(h_1(x,\theta^0_{1k_0}),h_2(x,\theta^0_{2k_0}))$ for all $x\in A$. But this contradicts (23), because the left-hand side put a positive weight to the atom $(h(x,\theta'_{11}),h(x,\theta'_{21}))$, which the right-hand side does not have. Hence, we have the set $\{(\theta'_{11},\theta'_{21}),\ldots,(\theta'_{1k_0},\theta'_{1k_0})\}$ equals $\{(\theta^0_{11},\theta^0_{21}),\ldots,(\theta^0_{1k_0},\theta^0_{2k_0})\}$. Now, without loss of generality, we can assign $(\theta'_{11},\theta'_{21})=(\theta^0_{11},\theta^0_{21})$, $\ldots,(\theta'_{1k_0},\theta'_{2k_0})=(\theta^0_{1k_0},\theta^0_{2k_0})$. Because $(\theta^0_{11},\theta^0_{21}),\ldots,(\theta^0_{1k_0},\theta^0_{2k_0})$ are distinct, there exists x' such that

$$(h_1(x',\theta^0_{11}),h_2(x',\theta^0_{21})),\ldots,(h_1(x',\theta^0_{1k_0}),h_2(x',\theta^0_{2k_0}))$$

are distinct, which together with Eq. (23) entail that $p'_i=p^0_i$ for all $i\in\{1,\ldots,k_0\}$. Thus, $G'=G_0$, while $W_1(G',G_0)>\epsilon'$, a contradiction. Hence, claim (21) is proved.

Proof of claim (22): Suppose this does not hold. Then there exists a sequence $G_\ell\in\mathcal{E}_{k_0}(\Theta)$ such that

$$W_1(G_\ell,G_0)\to 0,\quad \frac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X),f_{G_0}(\cdot|X))}{W_1(G_\ell,G_0)}\to 0. \tag{24}$$

We can relabel the atoms and weights of G_ℓ such that it admits the following form:

$$G_\ell=\sum_{j=1}^{k_0} p^\ell_j \delta_{(\theta^\ell_{1j},\theta^\ell_{2j})}, \tag{25}$$

where $p^\ell_j\to p^0_j$, $\theta^\ell_j\to\theta^0_j$ and $\theta^\ell_{2j}\to\theta^0_{2j}$ for all $i\in[k_0]$. To ease the ensuing presentation, we denote $\Delta p^\ell_j:=p^\ell_j-p^0_j$, $\Delta\theta^\ell_{1i}:=\theta^\ell_{1j}-\theta^0_{1j}$ and $\Delta\theta^\ell_{2j}:=\theta^\ell_{2j}-\theta^0_{2j}$

for $i \in [k_0]$. Then, using the coupling between G_ℓ and G_0 such that it put mass $\min\{p_i^\ell, p_i^0\}$ on $\delta_{((\theta_{1j}^\ell,\theta_{2j}^\ell),(\theta_{1j}^0,\theta_{2j}^0))}$, we can verify that

$$W_1(G_\ell, G_0) \preccurlyeq \sum_{i=1}^{k_0} \left|\Delta p_j^\ell\right| + p_j^\ell(\left\|\Delta\theta_{1j}^\ell\right\| + \left\|\Delta\theta_{2j}^\ell\right\|) =: D_1(G_\ell, G_0). \tag{26}$$

The remainder of the proof is composed of three steps.

Step 1 (Taylor expansion) To ease the notation, we write for short

$$f_j^0(y|x) = f(y|h_1(x, \theta_{1j}^0), h_2(x, \theta_{2j}^0)),$$

and

$$f_j^\ell(y|x) = f(y|h_1(x, \theta_{1j}^\ell), h_2(x, \theta_{2j}^\ell)),$$

for all $j = 1, \ldots, k_0$. Because $f(y|h_1(x, \theta_1), h_2(x, \theta_2))$ is differentiable with respect to θ for all x, y, by applying Taylor expansion up to the first order and the chain rule, we find that for all $j = 1, \ldots, k_0$,

$$f_j^\ell(y|x) - f_j^0(y|x) = (\Delta\theta_{1j}^\ell)^\top \frac{\partial}{\partial\theta_1} f_j^0(y|x) + (\Delta\theta_{2j}^\ell)^\top \frac{\partial}{\partial\theta_2} f_j^0(y|x) + R_j(x, y),$$

where $R_j(x)$ is Taylor remainder such that $R_j(x, y) = o(\left\|\Delta\theta_{1j}^\ell\right\| + \left\|\Delta\theta_{2j}^\ell\right\|)$ for $i \in [k_0]$. Combine the above expression for $j = 1, \ldots, k_0$, we have

$$\begin{aligned} f_{G_\ell}(y|x) - f_{G_0}(y|x) = \sum_{j=1}^{k_0} \left(\Delta p_j^\ell\right) f_j^0(y|x) + p_j^\ell(\Delta\theta_{1j}^\ell)^\top \frac{\partial}{\partial\theta_1} f_j^0(y|x) \\ + p_j^\ell(\Delta\theta_{2j}^\ell)^\top \frac{\partial}{\partial\theta_2} f_j^0(y|x) + R(x, y), \end{aligned} \tag{27}$$

where $R(x, y) = \sum_{i=1}^{\ell} p_j^\ell R_i(x, y) = o\left(\sum_{i=1}^{k_0} p_j^\ell \left(\left\|\Delta\theta_{1j}^\ell\right\| + \left\|\Delta\theta_{2j}^\ell\right\|\right)\right)$. From Eq. (26), we have $R(x, y)/D_1(G_\ell, G_0) \to 0$ as $\ell \to \infty$ for all x, y.

Step 2 (Extracting non-vanishing coefficients) From Eq. (24) and (26), we have that

$$\frac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{D_1(G_\ell, G_0)} \to 0 \quad (\ell \to \infty). \tag{28}$$

We can write

$$\begin{aligned} \frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_1(G_\ell, G_0)} = \sum_{j=1}^{k_0} \alpha_j^\ell f_j^0(y|x) + (\beta_j^\ell)^\top \frac{\partial}{\partial\theta_1} f_j^0(y|x) + (\gamma_i^\ell)^\top \frac{\partial}{\partial\theta_2} f_j^0(y|x) \\ + \frac{R(x, y)}{D_1(G_\ell, G_0)}, \end{aligned} \tag{29}$$

where $\alpha_i^\ell = \dfrac{(\Delta p_j^\ell)}{D_1(G_\ell, G_0)} \in \mathbb{R}$, $\beta_i^\ell = \dfrac{p_j^\ell(\Delta\theta_{1j}^\ell)}{D_1(G_\ell, G_0)} \in \mathbb{R}^{d_1}$ and $\gamma_i^\ell = \dfrac{p_j^\ell(\Delta\theta_{2j}^\ell)}{D_1(G_\ell, G_0)} \in \mathbb{R}^{d_2}$. From the definition of $D_1(G_\ell, G_0)$, we have

$$\sum_{i=1}^{k_0} |\alpha_i^\ell| + \sum_{i=1}^{k_0} \|\beta_i^\ell\| + \sum_{i=1}^{k_0} \|\gamma_i^\ell\| = 1,$$

therefore (α_i^ℓ) is in $[-1,1]$, (β_i^ℓ) is in $[-1,1]^{d_1}$ and (γ_i^ℓ) is in $[-1,1]^{d_2}$, by the compactness of those sets, there exist subsequences of (α_i^ℓ), β_i^ℓ and γ_i^ℓ (without loss of generality, we assume it is the whole sequence itself) such that $\alpha_i^\ell \to \alpha_i \in [-1,1], \beta_i^\ell \to \beta_i \in [-1,1]^d$ as $\ell \to \infty$ for all $i = 1, \ldots, k_0$. As $\sum_{i=1}^{k_0} |\alpha_i| + \sum_{i=1}^{k_0} \|\beta_i\| + \sum_{i=1}^{k_0} \|\gamma_i\| = 1$, at least one of them is not zero.

Step 3 (Deriving contradiction via Fatou's lemma) By Fatou's lemma, we have

$$\begin{aligned}
&\liminf_{\ell\to\infty} \frac{2\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{D_1(G_\ell, G_0)} \\
&= \liminf_{\ell\to\infty} \int_{\mathcal{X}} d\mathbb{P}_X(x) \int_{\mathcal{Y}} d\nu(y) \left| \frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_1(G_\ell, G_0)} \right| \\
&\geq \int_{\mathcal{X}} d\mathbb{P}_X(x) \int_{\mathcal{Y}} d\nu(y) \left(\liminf_{n\to\infty} \left| \frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_1(G_\ell, G_0)} \right| \right) \\
&\geq \int_{\mathcal{X}} d\mathbb{P}_X(x) \int_{\mathcal{Y}} d\nu(y) \left| \liminf_{\ell\to\infty} \frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_1(G_\ell, G_0)} \right| \\
&= \int_{\mathcal{X}} d\mathbb{P}_X(x) \int_{\mathcal{Y}} d\nu(y) \left| \sum_{j=1}^{k_0} \alpha_j f_j^0(y|x) + (\beta_j)^\top \frac{\partial}{\partial\theta_1} f_j^0(y|x) + (\gamma_i)^\top \frac{\partial}{\partial\theta_2} f_j^0(y|x) \right|.
\end{aligned}$$

Since $\lim_{n\to\infty} \dfrac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{D_1(G_\ell, G_0)} = 0$, we have

$$\sum_{j=1}^{k_0} \alpha_j f_j^0(y|x) + (\beta_j)^\top \frac{\partial}{\partial\theta_1} f_j^0(y|x) + (\gamma_i)^\top \frac{\partial}{\partial\theta_2} f_j^0(y|x) = 0, \text{ a.s. in } x, y, \quad (30)$$

where at least one of $\alpha_i, \beta_i, \gamma_i$ are not 0. But by the identifiability of family of conditional densities f (condition (A1.)), we have $\alpha_1 = \cdots = \alpha_{k_0} = 0$, $\beta_1 = \cdots = \beta_{k_0} = 0$ and $\gamma_1 = \cdots = \gamma_{k_0} = 0$. Hence, we arrive at a contradiction and conclude claim (22).

For part (b) of the theorem, in a similar spirit we can achieve the conclusion by proving the following claims:

$$\inf_{G\in\mathcal{O}_K(\Theta): W_2(G,G_0)>\epsilon'} \frac{\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X))}{W_2^2(G, G_0)} > 0, \quad (31)$$

for any $\epsilon' > 0$, and

$$\lim_{\epsilon \to 0} \inf_{G \in \mathcal{O}_K(\Theta)} \left\{ \frac{\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X))}{W_2^2(G, G_0)} : \ W_2(G, G_0) \le \epsilon \right\} > 0. \tag{32}$$

The proof of claim (31) is similar to that of claim (21) and is omitted. Now we proceed to prove claim (32). Suppose this does not hold. So there exists a sequence $G_\ell \in \mathcal{O}_K(\Theta)$ such that

$$W_2(G_\ell, G_0) \to 0, \quad \frac{\mathbb{E}_X d_{TV}(p_{G_\ell}(\cdot|X), p_{G_0}(\cdot|X))}{W_2^2(G_\ell, G_0)} \to 0. \tag{33}$$

We can assume that all G_ℓ have the same number of atoms (by extracting a subsequence if needed) and relabel the atoms and weights of G_ℓ such that it admits the following form:

$$G_\ell = \sum_{j=1}^{k_0+r} \sum_{t=1}^{s_j} p_{jt}^\ell \delta_{(\theta_{1jt}^\ell, \theta_{2jt}^\ell)}, \tag{34}$$

where $\sum_{t=1}^{s_j} p_{jt}^\ell \to p_j^0$, $\theta_{1jt}^\ell \to \theta_{1j}^0$ and $\theta_{2jt}^\ell \to \theta_{2j}^0$ for all $j \in [k_0 + r]$, $p_j^0 = 0$ for all $j > k_0$, and $(\theta_{11}^0, \theta_{21}^0), \ldots, (\theta_{1(k_0+r)}^0, \theta_{2(k_0+r)}^0)$ are distinct. For all j, t, we denote $\Delta\theta_{1jt}^\ell := \theta_{1jt}^\ell - \theta_{1j}^0$, $\Delta\theta_{2jt}^\ell := \theta_{2jt}^\ell - \theta_{2j}^0$, and $\Delta p_j^\ell := \sum_{t=1}^{s_j} p_{jt}^\ell - p_j^0$. We have

$$W_2^2(G_\ell, G_0) \preccurlyeq \sum_{j=1}^{k_0+r} \left(\left|\Delta p_j^\ell\right| + \sum_{t=1}^{s_j} p_{jt}^\ell \left(\left\|\Delta\theta_{1jt}^\ell\right\|^2 + \left\|\Delta\theta_{2jt}^\ell\right\|^2 \right) \right) =: D_2(G_\ell, G_0) \tag{35}$$

As in part (a) the remainder of the proof is divided into three steps.

Step 1 (Taylor expansion) To ease the notation, we write for short

$$f_j^0(y|x) = f(y|h_1(x, \theta_{1j}^0), h_2(x, \theta_{2j}^0)), \quad f_{jt}^\ell(y|x) = f(y|h_1(x, \theta_{1jt}^\ell), h_2(x, \theta_{2jt}^\ell)),$$

for all $t = 1, \ldots, s_j, j = 1, \ldots, k_0 + r$. Because $f(y|h_1(x, \theta_1), h_2(x, \theta_2))$ is differentiable up to the second order with respect to θ_1, θ_2 for all x, y, by applying Taylor expansion up to the second order and the chain rule, we find that

$$\begin{aligned} f_{jt}^\ell(y|x) - f_j^0(y|x) &= (\Delta\theta_{1jt}^\ell)^\top \frac{\partial}{\partial\theta_1} f_j^0(y|x) + (\Delta\theta_{2jt}^\ell)^\top \frac{\partial}{\partial\theta_2} f_j^0(y|x) \\ &+ \frac{1}{2} (\Delta\theta_{1jt}^\ell)^\top \frac{\partial^2}{\partial\theta_1^2} f_j^0(y|x) (\Delta\theta_{1jt}^\ell) + \frac{1}{2} (\Delta\theta_{2jt}^\ell)^\top \frac{\partial^2}{\partial\theta_2^2} f_j^0(y|x) (\Delta\theta_{2jt}^\ell) \\ &+ (\Delta\theta_{1jt}^\ell)^\top \frac{\partial^2}{\partial\theta_1 \partial\theta_2} f_j^0(y|x) (\Delta\theta_{2jt}^\ell) + R_i(x, y) \end{aligned}$$

where $R_i(x,y)$ is Taylor remainder such that $R_{ij}(x,y) = o\left(\left\|\Delta\theta^{\ell}_{1jt}\right\|^2\right)$ for $i \in [k_0+r]$. Therefore,

$$\begin{aligned}
f_{G_\ell}(y|x) - f_{G_0}(y|x) &= \sum_{j=1}^{k_0+r} \left(\Delta p^{\ell}_j\right) f^0_j(y|x) + \sum_{j=1}^{k_0+r}\sum_{t=1}^{s_j} p^{\ell}_{jt}(f^{\ell}_{jt}(y|x) - f^0_j(y|x)) \\
&= \sum_{j=1}^{k_0+r} \left(\Delta p^{\ell}_j\right) f^0_j(y|x) + \sum_{j=1}^{k_0+r} \left(\sum_{t=1}^{s_j} p^{\ell}_{ij}(\Delta\theta^{\ell}_{1jt})^\top\right) \frac{\partial}{\partial\theta_1} f^0_j(y|x) \\
&+ \sum_{j=1}^{k_0+r} \left(\sum_{t=1}^{s_j} p^{\ell}_{ij}(\Delta\theta^{\ell}_{2jt})^\top\right) \frac{\partial}{\partial\theta_2} f^0_j(y|x) \\
&+ \sum_{j=1}^{k_0+r} \Big(\sum_{t=1}^{s_j} \frac{p^{\ell}_{ij}}{2}(\Delta\theta^{\ell}_{1jt})^\top \frac{\partial^2}{\partial\theta_1^2} f^0_j(y|x)(\Delta\theta^{\ell}_{1jt})\Big) \\
&+ \sum_{j=1}^{k_0+r} \Big(\sum_{t=1}^{s_j} \frac{p^{\ell}_{ij}}{2}(\Delta\theta^{\ell}_{2jt})^\top \frac{\partial^2}{\partial\theta_2^2} f^0_j(y|x)(\Delta\theta^{\ell}_{2jt})\Big) \\
&+ \sum_{j=1}^{k_0+r} \Big(\sum_{t=1}^{s_j} \frac{p^{\ell}_{ij}}{2}(\Delta\theta^{\ell}_{1jt})^\top \frac{\partial^2}{\partial\theta_1\partial\theta_2} f^0_j(y|x)(\Delta\theta^{\ell}_{2jt})\Big) + R(x,y),
\end{aligned} \tag{36}$$

where $R(x,y) = \sum_{j,t} p^{\ell}_{jt} w_{jt}(x,y) = o\left(\sum_{j,t} p^{\ell}_{jt}\left(\left\|\Delta\theta^{\ell}_{1jt}\right\|^2 + \left\|\Delta\theta^{\ell}_{2jt}\right\|^2\right)\right)$. From the expression in Eq. (35), we have $R(x,y)/D_2(G_\ell, G_0) \to 0$ as $\ell \to \infty$ for all x, y.

Step 2 (Extracting non-vanishing coefficients) From Eq. (33) and (35), we have that

$$\frac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{D_2(G_\ell, G_0)} \to 0 \quad (\ell \to \infty). \tag{37}$$

We can write

$$\begin{aligned}
\frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_2(G_\ell, G_0)} &= \sum_{i=1}^{k_0+r} a^{\ell}_j f^0_j(y|x) + \sum_{j=1}^{k_0+r} b^{n}_j \frac{\partial}{\partial\theta_1} f^0_j(y|x) + \sum_{j=1}^{k_0+r} c^{n}_j \frac{\partial}{\partial\theta_2} f^0_j(y|x) \\
&+ \sum_{j=1}^{k_0+r} \Big(\sum_{t=1}^{s_j}(w^{\ell}_{jt})^\top \frac{\partial^2}{\partial\theta_1^2} f^0_j(y|x)(w^{\ell}_{jt})\Big) + \sum_{j=1}^{k_0+r} \Big(\sum_{t=1}^{s_j}(v^{\ell}_{jt})^\top \frac{\partial^2}{\partial\theta_2^2} f^0_j(y|x)(v^{\ell}_{jt})\Big) \\
&+ 2\sum_{j=1}^{k_0+r} \Big(\sum_{t=1}^{s_j}(w^{\ell}_{jt})^\top \frac{\partial^2}{\partial\theta_1\partial\theta_2} f^0_j(y|x)(v^{\ell}_{jt})\Big) + R(x,y),
\end{aligned} \tag{38}$$

where

$$a^{\ell}_j = \frac{(\Delta p^{\ell}_j)}{D_2(G_\ell, G_0)},\ b^{\ell}_j = \frac{\sum_{t=1}^{s_j} p^{\ell}_{jt}(\Delta\theta^{\ell}_{1jt})}{D_2(G_\ell, G_0)},\ c^{\ell}_j = \frac{\sum_{t=1}^{s_j} p^{\ell}_{jt}(\Delta\theta^{\ell}_{2jt})}{D_2(G_\ell, G_0)},$$

and

$$w_{jt}^{\ell} = \frac{\sqrt{p_{jt}^{\ell}}(\Delta\theta_{1jt}^{\ell})}{\sqrt{2D_2(G_\ell, G_0)}},\ v_{jt}^{\ell} = \frac{\sqrt{p_{jt}^{\ell}}(\Delta\theta_{2jt}^{\ell})}{\sqrt{2D_2(G_\ell, G_0)}},$$

for all $j \in [k_0 + l]$. From the definition of $D_2(G_\ell, G_0)$, we have

$$\sum_{j=1}^{k_0+r} |a_j^n| + 2\sum_{j=1}^{k_0+r}\sum_{t=1}^{s_i} \left\|w_{jt}^{\ell}\right\|^2 + 2\sum_{j=1}^{k_0+r}\sum_{t=1}^{s_i}\left\|v_{jt}^{\ell}\right\|^2 = 1,$$

so that $M_\ell := \max_{j,t}\{|a_j^\ell|, \|b_j^n\|, \|c_j^n\|, \|w_{jt}^\ell\|^2, \|v_{jt}^\ell\|^2\}$ is always bounded below by $\dfrac{1}{5\bar{K}}$ for all n, and does not converge to 0. Denote

$$\alpha_j^\ell = a_j^\ell/M_\ell,\ \beta_j^\ell = b_j^\ell/M_\ell,\ \gamma_j^\ell = c_j^\ell/M_\ell,\ \rho_{jt}^\ell = w_{jt}^\ell/\sqrt{M_\ell},\ \nu_{jt}^\ell = v_{jt}^\ell/\sqrt{M_\ell}.$$

for all $t = 1, \dots, s_j, j = 1, \dots, k_0 + r$. By compactness and subsequence argument, we can have that $\alpha_j^n \to \alpha_j \in [-1, 1], \beta_j^n \to \beta_j \in [-1, 1]^{d_1}$ and $\gamma_{j,t}^\ell \to \gamma_j \in [-1, 1]^{d_2}$, and $\rho_{jt}^\ell \to \rho_{jt} \in [-1, 1]^{d_1}, \nu_{jt}^\ell \to \nu_{jt} \in \mathbb{R}^{d_2}$ as $\ell \to \infty$ for all t, j, and at least one of those limits is not zero.

Step 3 (Deriving contradiction via Fatou's lemma) By Fatou's lemma, we have

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{2}{M_\ell}\frac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{D_2(G_\ell, G_0)}\\
&= \liminf_{n\to\infty} \frac{2}{M_\ell}\int_{\mathcal{X}} d\mathbb{P}_X(x)\int_{\mathcal{Y}} d\nu(y)\left|\frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_2(G_\ell, G_0)}\right|\\
&\geq \int_{\mathcal{X}} d\mathbb{P}_X(x)\int_{\mathcal{Y}} d\nu(y)\left(\liminf_{n\to\infty}\frac{1}{M_\ell}\left|\frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_2(G_\ell, G_0)}\right|\right)\\
&\geq \int_{\mathcal{X}} d\mathbb{P}_X(x)\int_{\mathcal{Y}} d\nu(y)\left|\liminf_{n\to\infty}\frac{1}{M_\ell}\frac{f_{G_\ell}(y|x) - f_{G_0}(y|x)}{D_2(G_\ell, G_0)}\right|\\
&= \int_{\mathcal{X}} d\mathbb{P}_X(x)\int_{\mathcal{Y}} d\nu(y)\Bigg|\sum_{j=1}^{k_0+l}\alpha_j f_j^0(y|x) + (\beta_j)^\top\frac{\partial}{\partial\theta_1}f_j^0(y|x) + (\gamma_j)^\top\frac{\partial}{\partial\theta_2}f_j^0(y|x)\\
&+ \sum_{t=1}^{s_j}(\rho_{jt})^\top\frac{\partial}{\partial\theta_1^2}f_j^0(y|x)(\rho_{jt}) + \sum_{t=1}^{s_j}(\nu_{jt})^\top\frac{\partial}{\partial\theta_2^2}f_j^0(y|x)(\nu_{jt})\\
&+ 2\sum_{t=1}^{s_j}(\rho_{jt})^\top\frac{\partial}{\partial\theta_1\partial\theta_2}f_j^0(y|x)(\nu_{jt})\Bigg|.
\end{aligned}$$

Since $\lim_{n\to\infty}\dfrac{1}{M_\ell}\dfrac{\mathbb{E}_X d_{TV}(f_{G_\ell}(\cdot|X), f_{G_0}(\cdot|X))}{D_2(G_\ell, G_0)} = 0$, the integrand in the right hand side of the above display is 0 for almost all x, y. By the second order identifiability of $f(y|x)$, all the coefficients are 0, which contradicts with the fact derived in the end of Step 2. We arrive at the conclusion of claim (32). □

Proof of Proposition 2. We want to prove that for h_1 and h_2 being completely identifiable, then for any $k \geq 1$ and distinct pairs $(\theta_{11}, \theta_{21}), \ldots, (\theta_{1k}, \theta_{2k})$ we have

$$(h_1(x, \theta_{11}), h_2(x, \theta_{21})), \ldots, (h_1(x, \theta_{1k}), h_2(x, \theta_{2k}))$$

are distinct almost surely. For any $i \neq j$, because $(\theta_{1i}, \theta_{2i}) \neq (\theta_{1j}, \theta_{2j})$, we have either $\theta_{1i} \neq \theta_{1j}$ or $\theta_{2i} \neq \theta_{2j}$. By the complete identifiability of h_1 and h_2, we have either

$$\mathbb{P}_X(\{x : h_1(x, \theta_{1i}) = h_1(x, \theta_{1j})\}) = 0,$$

or

$$\mathbb{P}_X(\{x : h_2(x, \theta_{2i}) = h_2(x, \theta_{2j})\}) = 0.$$

Hence,

$$\begin{aligned}
&\mathbb{P}_X(\{x : (h_1(x, \theta_{1i}), h_2(x, \theta_{2i})) = (h_1(x, \theta_{1j}), h_2(x, \theta_{2j}))\}) \\
&= \mathbb{P}_X(\{x : h_1(x, \theta_{1i}) = h_1(x, \theta_{1j}), h_2(x, \theta_{2i}) = h_2(x, \theta_{2j})\}) \\
&\leq \min\{\mathbb{P}_X(\{x : h_1(x, \theta_{1i}) = h_1(x, \theta_{1j})\}), \mathbb{P}_X(\{x : h_2(x, \theta_{2i}) = h_2(x, \theta_{2j})\})\} \\
&= 0.
\end{aligned}$$

Now consider the set

$$A = \cup_{1 \leq i < j \leq k} \{x : (h_1(x, \theta_{1i}), h_2(x, \theta_{2i})) = (h_1(x, \theta_{1j}), h_2(x, \theta_{2j}))\},$$

we have

$$\mathbb{P}_X(A) \leq \sum_{1 \leq i < j \leq k} \mathbb{P}_X(\{x : (h_1(x, \theta_{1i}), h_2(x, \theta_{2i})) = (h_1(x, \theta_{1j}), h_2(x, \theta_{2j}))\}) = 0.$$

Therefore, $(h_1(x, \theta_{11}), h_2(x, \theta_{21})), \ldots, (h_1(x, \theta_{1k}), h_2(x, \theta_{2k}))$ are distinct on A^c, where $\mathbb{P}_X(A^c) = 1$. □

Proof of Proposition 4. (a) This comes directly from the fact that if h_1 and h_2 are Lipschitz. For any $G_0 = \sum_{j=1}^{k_0} p_i^0 \delta_{(\theta_{1j}^0, \theta_{2j}^0)} \in \mathcal{E}_{k_0}(\Theta), G = \sum_{i=1}^{\bar{K}} p_i \delta_{(\theta_{1i}, \theta_{2i})} \in \mathcal{E}_{\bar{K}}(\Theta)$, we have

$$|h_1(x, \theta_{1j}^0) - h_1(x, \theta_{1i})| \leq c_1 \left\| \theta_{1j}^0 - \theta_{1i} \right\|,$$

and

$$|h_2(x, \theta_{2j}^0) - h_2(x, \theta_{2i})| \leq c_2 \left\| \theta_{2j}^0 - \theta_{2i} \right\|,$$

for any $j = 1, \ldots, k_0, i = 1, \ldots, \bar{K}$ and all $x \in \mathcal{X}$, where c_1 and c_2 are constants which only depend on h_1 and h_2. Hence, for any coupling $(q_{ij})_{i,j=1}^{\bar{K}, k_0}$ of $(p_i)_{i=1}^{\bar{K}}$ and $(p_j^0)_{j=1}^{k_0}$, we have

$$\begin{aligned}
&\sum_{i,j} q_{ij} (\left\| \theta_{1i} - \theta_{1j}^0 \right\| + \left\| \theta_{2i} - \theta_{2j}^0 \right\|) \\
&\geq \overline{C}_1 \sum_{i,j} q_{ij} (|h(x, \theta_{1i}) - h(x, \theta_{1j}^0)| + |h(x, \theta_{1i}) - h(x, \theta_{2j}^0)|)
\end{aligned}$$

for $\overline{C}_1 = 1/\max\{c_1, c_2\}$ and for all $x \in \mathcal{X}$. Taking infimum with respect to the LHS, we have

$$W_r(G, G_0) \geq \overline{C}_1 W_r \left(\sum_{j=1}^{k_0} p_j^0 \delta_{(h_1(x,\theta_{1j}^0), h_2(x,\theta_{2j}^0))}, \sum_{j=1}^{K} p_j \delta_{(h_1(x,\theta_{1j}), h_2(x,\theta_{2j}))} \right),$$

Taking the expectation with respect to $\mathbb{P}_X$ we obtain

$$W_r(G, G_0) \geq \overline{C}_1 \mathbb{E}_X W_r \left(\sum_{j=1}^{k_0} p_j^0 \delta_{(h_1(X,\theta_{1j}^0), h_2(X,\theta_{2j}^0))}, \sum_{j=1}^{K} p_j \delta_{(h_1(X,\theta_{1j}), h_2(X,\theta_{2j}))} \right).$$

(b) For any coupling $(q_{ij})_{i,j=1}^{K,k_0}$ of $(p_i)_{i=1}^{K}$ and $(p_j^0)_{j=1}^{k_0}$ we have

$$\begin{aligned} \mathbb{E}_X \left| \sum_{i=1}^{K} p_i h_1(X, \theta_{1i}) - \sum_{j=1}^{k_0} p_j^0 h_1(X, \theta_{1j}^0) \right| &\leq \sum_{i,j=1}^{K,k_0} q_{ij} |h_1(X, \theta_{1i}) - h_1(X, \theta_{1j}^0)| \\ &\leq \sum_{i,j=1}^{K,k_0} q_{ij} c_1 \left\| \theta_{1i} - \theta_{1j}^0 \right\|. \end{aligned}$$

Taking infimum of all feasible $(q_{ij})_{i,j}$, this implies

$$\mathbb{E}_X \left| \sum_{i=1}^{K} p_i h_1(X, \theta_{1i}) - \sum_{j=1}^{k_0} p_j^0 h_1(X, \theta_{1j}^0) \right| \leq c_1 W_1(G, G_0).$$

Doing similarly for h_2, we have the conclusion. □

A.2. Convergence rates for conditional density estimation and parameter estimation

Firstly, we combine the inverse bounds (Theorem 2) with the convergence theory for density estimation to derive convergence rates for parameter estimation that arise in regression mixture models as presented in Theorem 4.

Proof of Theorem 4. Recall that with the assumptions in this theorem, we have

$$\mathbb{E}_X d_{TV}(f_G, f_{G_0}) \geq C_1 W_1(G, G_0) \quad \forall G \in \mathcal{E}_{k_0}(\Theta),$$

and for any $K > k_0$,

$$\mathbb{E}_X d_{TV}(f_G, f_{G_0}) \geq C_2 W_2^2(G, G_0) \quad \forall G \in \mathcal{O}_K(\Theta),$$

for $C_1, C_2 > 0$ only depend on Θ, G_0, f, h_1, h_2, and K. Besides,

$$\sqrt{2} d_H(f_G, f_{G_0}) \geq \mathbb{E}_X d_{TV}(f_G, f_{G_0}) \quad \forall G, G_0.$$

Combining these inequalities with the concentration inequality given in Theorem 3 to have

$$\mathbb{P}_{G_0}(W_1(\widehat{G}_n, G_0) > C\delta) \leq \mathbb{P}_{G_0}(\overline{d}_H(f_G, f_{G_0}) > \sqrt{2}CC_1\delta) \leq c\exp(-n\delta^2/c^2),$$

for the exact-fitted setting, since $\hat{G}_n \in \mathcal{E}_{k_0}(\Theta)$. In the over-fitted setting, a similar argument yields

$$\mathbb{P}_{G_0}(W_2^2(\widehat{G}_n, G_0) > C\delta) \leq \mathbb{P}_{G_0}(\overline{d}_H(f_G, f_{G_0}) > \sqrt{2}CC_2\delta) \leq c\exp(-n\delta^2/c^2).$$

□

Next, we proceed to prove Theorem 5 which is concerned with the convergence rates of conditional density estimation.

Proof of Theorem 5. The proof is a generalization of proof of Theorem 3.1 in [9]. First we prove that if for any fixed k and for all $\epsilon \in (0, 1/2)$, these claims hold

$$\log N(\epsilon, \mathcal{F}_k(\Theta), \|\cdot\|_\infty) \preccurlyeq \log(1/\epsilon) \tag{39}$$

$$H_B(\epsilon, \mathcal{F}_k(\Theta), \overline{d}_H) \preccurlyeq \log(1/\epsilon), \tag{40}$$

then by applying Theorem 3, we can arrive at our conclusion. Indeed, since

$$\left\| \left(\frac{f + f_0}{2}\right)^{1/2} - \left(\frac{g + f_0}{2}\right)^{1/2} \right\|_2 \leq d_H(f, g)$$

for all densities f, g, f_0, we have

$$H_B(u, \overline{\mathcal{F}}_k^{1/2}(\Theta, u), \|\cdot\|_2) \leq H_B(u, \mathcal{P}_k(\Omega), \overline{d}_H),$$

for all $u > 0$. Thus, we can bound the bracketing entropy integral as follows

$$\begin{aligned}
\mathcal{J}(\delta, \overline{\mathcal{F}}_k^{1/2}(\Theta, \delta)) &\leq \left(\int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(u, \mathcal{F}_k(\Omega), \overline{d}_H) du\right) \vee \delta \\
&\preccurlyeq \left(\int_{\delta^2/2^{13}}^{\delta} \log(1/u) du\right) \vee \delta \\
&\leq \left(\int_{\delta^2/2^{13}}^{\delta} \log(2^{13}/\delta^2) du\right) \vee \delta \\
&\leq \delta \log(2^{13}/\delta^2) \vee \delta \\
&\preccurlyeq \delta \log(1/\delta),
\end{aligned}$$

where we use the fact that $\log(1/u)$ is a decreasing function. Hence, if we choose $\Psi(\delta) = \delta\log(1/\delta)$, then $\Psi(\delta) \succcurlyeq \mathcal{J}(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta, \delta))$, $\Psi(\delta)/\delta^2 = \log(1/\delta)(1/\delta)$ is a non-increasing function, and for $\delta_n = O((\log n/n)^{1/2})$, we have

$$\sqrt{n}\delta_n^2 \succcurlyeq \log n/\sqrt{n} \succcurlyeq \Psi(\delta_n).$$

Therefore, the result of Theorem 3 says that there exist constant C and c such that

$$\mathbb{P}\left(\overline{d}_H(f_{\widehat{G}_n}, f_{G_0}) > C\sqrt{\frac{\log n}{n}}\right) \preccurlyeq \exp(-c\log n),$$

which is the conclusion. It remains to verify (39) and (40).

Proof of claim (39) Since Θ_1 and Θ_2 are compact, for all $\epsilon > 0$, there exists a ϵ-net B_1 of $(\Theta_1, \|\cdot\|)$ and B_2 of $(\Theta_2, \|\cdot\|)$ with the cardinality $|B_1| \leq \left(\frac{\text{diam}(\Theta_1)}{\epsilon}\right)^{d_1}$ and $|B_2| \leq \left(\frac{\text{diam}(\Theta_2)}{\epsilon}\right)^{d_2}$. We also know that there exists a ϵ-net A for $(\Delta^{k-1}, \|\cdot\|_\infty)$ such that $|A| \leq \left(\frac{5}{\epsilon}\right)^k$ ([9]). We consider the following subset of $\mathcal{F}_k(\Theta)$

$$C = \{p_G : G = \sum_{i=1}^k p_i\delta_{(\theta_{1j},\theta_{2j})}, (p_i)_{i=1}^k \in A, \theta_{1j} \in B_1, \theta_{2j} \in B_2 \forall\, i\}.$$

We can see that

$$|C| = |A| \times |B_1|^k \times |B_2|^k \leq \left(\frac{5}{\epsilon}\right)^k \left(\frac{\text{diam}(\Theta_1)}{\epsilon}\right)^{d_1 k} \left(\frac{\text{diam}(\Theta_2)}{\epsilon}\right)^{d_2 k}.$$

For any $G = \sum_{i=1}^k p_i\delta_{\theta_{1j}} \in \mathcal{O}_k(\Theta)$, there exist $(\tilde{p}_i)_{i=1}^k \in A$ and $\tilde{\theta}_i \in B$ such that $|p_i - \tilde{p}_i| \leq \epsilon$ and $\left\|\theta_{1j} - \tilde{\theta}_i\right\| \leq \epsilon$ for all i. Let $\tilde{G} = \sum_{i=1}^k \tilde{p}_i\delta_{\tilde{\theta}_i}$ and $G' = \sum_{i=1}^k \tilde{p}_i\delta_{\theta_{1j}}$, by triangle inequality, we have

$$\begin{aligned}
&\|f_G(y|x) - f_{\tilde{G}}(y|x)\|_\infty \leq \|f_G(y|x) - f_{G'}(y|x)\|_\infty + \|f_{G'}(y|x) - f_{\tilde{G}}(y|x)\|_\infty \\
&\leq \sum_{j=1}^k |p_j - \tilde{p}_j|\, \|f(y|h_1(x,\theta_{1j}), h_2(x,\theta_{2j}))\|_\infty \\
&+ \sum_{j=1}^k \tilde{p}_j \left\|(f(y|h_1(x,\theta_{1j}), h_2(x,\theta_{2j})) - f(y|h_1(x,\tilde{\theta}_{1j}), h_2(x,\tilde{\theta}_{2j}))\right\|_\infty \\
&\preccurlyeq \sum_{j=1}^k |p_j - \tilde{p}_j|\, \|f(y|h_1(x,\theta_{1j}), h_2(x,\theta_{2j}))\|_\infty \\
&+ \sum_{i=1}^k \tilde{p}_j \left(\left\|\theta_{1j} - \tilde{\theta}_{1j}\right\| + \left\|\theta_{2j} - \tilde{\theta}_{2j}\right\|\right) \\
&\preccurlyeq \epsilon,
\end{aligned}$$

where we apply the assumptions that $\|f(y|\mu,\phi)\|_\infty$ is bounded uniformly in $(\mu,\phi) \in H$, and the uniform Lipschitz of f and h_1, h_2. Hence C forms a ϵ-net of $\mathcal{F}_k(\Theta)$. This implies that

$$N(\epsilon, \mathcal{F}_k(\Theta), \|\cdot\|_\infty) \preccurlyeq \left(\frac{5}{\epsilon}\right)^k \left(\frac{\text{diam}(\Theta_1)}{\epsilon}\right)^{d_1 k} \left(\frac{\text{diam}(\Theta_2)}{\epsilon}\right)^{d_2 k}.$$

Taking the logarithm of both sides, we arrive at the conclusion of claim (39).

Proof of claim (40) We first construct an ϵ-bracketing for $\mathcal{F}_k(\Theta)$ under ℓ_1 norm. Let η be a small number that we can choose later, and $f_1, \ldots, f_M$ is a η-net for $\mathcal{F}_k(\Theta)$ under $\|\cdot\|_\infty$, for $M \preccurlyeq \log(1/\epsilon)$. Denote by C_1 an upper bound for $\|f(y|\mu,\phi)\|_\infty$ for all $(\mu,\phi) \in H$. From our assumptions, we can construct an envelope for $\mathcal{F}_k(\Theta)$ as follows

$$H(x,y) = \begin{cases} d_1 \exp(-d_2|y|^{d_3}), & \forall y > \overline{c} \text{ or } y < \underline{c} \\ C_1, & \forall y \in [\underline{c}, \overline{c}], \end{cases}$$

where we can assume that $\overline{c} > 0$ and $\underline{c} < 0$. Then, we can create the brackets $[f_i^L(x,y), f_i^U(x,y)]_{i=1}^M$ by

$$f_i^L(x,y) := \max\{f_i(y|x) - \eta, 0\}, \quad f_i^U(x,y) := \max\{f_i(y|x) + \eta, H(x,y)\}.$$

Because for all $f \in \mathcal{F}_k(\Theta)$, we have a f_i such that $\|f - f_i\|_\infty \le \eta$, it can be seen that $f(y|x) \in [f_i^L(x,y), f_i^U(x,y)]$ for all x, y. Therefore, $\mathcal{F}_k(\Theta) \subset \cup_{i=1}^M [f_i^L, f_i^U]$. Besides, for any $\overline{C} > \overline{c}$ and $\underline{C} < \underline{c}$, we have

$$\begin{aligned}
&\int (f_i^U - f_i^L) d\mathbb{P}(x) d\nu(y) \le \int_x \int_{y=\underline{C}}^{y=\overline{C}} (f_i^U - f_i^L) d\mathbb{P}(x) d\nu(y) \\
&+ \int_x \int_{\{y<\underline{C}\}\cup\{y>\overline{C}\}} (f_i^U - f_i^L) d\mathbb{P}(x) d\nu(y) \\
&\le \eta(\overline{C} - \underline{C}) + \int_{\{y<\underline{C}\}\cup\{y>\overline{C}\}} d_1 \exp(-d_2|y|^{d_3}) d\nu(y) \\
&\le \eta(\overline{C} - \underline{C}) + \int_{\{u<d_2|\underline{C}|^{d_3}\}\cup\{u>d_2\overline{C}^{d_3}\}} d_1 \exp(-|u|)|u|^{1/d_3-1} d\nu(y) \\
&\preccurlyeq \eta(\overline{C} - \underline{C}) + \overline{C} \exp(-d_2\overline{C}^{d_3}) - \underline{C} \exp(d_2\underline{C}^{d_3}),
\end{aligned} \tag{41}$$

where we use the change of variable formula $u = d_2|y|^{d_3}$, and the fact that when ν is the Lebesgue measure:

$$\begin{aligned}
\int_{u \ge z} \exp(-u)\,(u)^{1/d_3-1}\,du &= z^{1/d_3} e^{-z} \int_0^\infty (1+s)^{1/d_3-1} e^{-zs} \\
&\le z^{1/d_3} e^{-z} \frac{1}{z - 1/d_3 + 1} < z^{1/d_3} e^{-z},
\end{aligned}$$

for all $z \ge 0$. Notice that if f is a probability mass function (i.e., ν is discrete), we can change the integral to sum, and the result still holds because

$$\sum_{y=\overline{C}+1}^{\infty} \exp(-d_2|y|^{d_3}) \le \int_{y=\overline{C}}^{\infty} \exp(-d_2|y|^{d_3}) dy.$$

Now, let $\overline{C} = \overline{c}(\log(1/\eta))^{1/d_3}, \underline{C} = \underline{c}(\log(1/\eta))^{1/d_3}$, we have

$$\left\|f_i^U - f_i^L\right\|_1 \preccurlyeq \eta^{d_4}\left(\log\left(\frac{1}{\eta}\right)\right)^{1/d_3},$$

where $d_4 = \max\{1, d_2\overline{c}^{d_3}, d_2|\underline{c}|^{d_3}\}$. Hence, there exists a positive constant c which does not depend on η such that

$$H_B(c\eta^{d_4}(\log(1/\eta))^{1/d_3}, \mathcal{F}_k(\Theta), \|\cdot\|_1) \preccurlyeq \log(1/\eta).$$

Let $\epsilon = c\eta^{d_4}(\log(1/\eta))^{1/d_3}$, we have $\log(1/\epsilon) \asymp \log(1/\eta)$. Combining with inequality $\|\cdot\|_1 \leq h^2$ yields

$$H_B(\epsilon, \mathcal{F}_k(\Theta), h) \leq H_B(\epsilon^2, \mathcal{F}_k(\Theta), \|\cdot\|_1) \preccurlyeq \log(1/\epsilon^2) \preccurlyeq \log(1/\epsilon).$$

Thus, we have proved claim (40). □

Finally, we obtain upper bounds on the tail probability for some popular family of distributions in order to verify that they satisfy all conditions of Theorem 3.

Proof of Proposition 8. Since the parameter space Λ is compact, we can assume it is a subset of some $[\underline{\lambda}, \overline{\lambda}]$, where $\overline{\lambda} > 0$ and $\underline{\lambda} < 0$. If the family of distribution is discrete, then obviously its probability mass function is bounded uniformly by 1.

(a) For the family of normal distribution $\{f(y|\mu, \sigma^2) : \mu \in [\overline{\lambda}, \underline{\lambda}], \sigma^2\}$, we have that

$$f(y|\mu, \sigma^2) \leq \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-y/8\sigma^2), \tag{42}$$

for all $y > 2\overline{\lambda}$ or $y < 2\underline{\lambda}$.

(b) For the family of Binomial distribution $\mathrm{Bin}(N, q)$, we can see that it is discrete and domain of q is bounded. Therefore the conclusion is immediate.

(c) For the family of Poisson distribution $f(y|\lambda)$, we have that $f(y|\lambda) = 0 \forall\, y < 0$ and

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \leq \exp(-y),$$

for all $y \geq 2(\overline{\lambda}e)^2$ due to the inequality $y! \geq \left(\frac{y}{2}\right)^{y/2}$.

(d) For the family of negative binomial distribution $f(y|\mu, \phi)$, we also have $f(y|\mu, \phi) = 0 \forall\, y < 0$, and

$$f(y|\mu, \phi) \preccurlyeq y^{[\phi]+1}\left(\frac{\mu}{\mu+\theta}\right)^y \leq y^{[\phi]+1}\left(\frac{\overline{\mu}}{\overline{\mu}+\theta}\right)^y \leq \left(\frac{\overline{\mu}}{\overline{\mu}+\theta}\right)^{y/2}.$$

for all y large enough compared to $\overline{\mu}$ and ϕ. □

Proof of Theorem 6. Similar to the proof of Theorem 4, with $\delta_n = \sqrt{\log(n)/n}$ (and using Theorem 5 instead of Theorem 3). □

Proof of Remark 4. From Theorem 6, there exist constants C depending on G_0, Θ, f, h_1, h_2 and an universal constant c such that

$$\mathbb{P}_{G_0}\left(W_1(\widehat{G}_n, G_0) > \delta\right) \le c\exp(-n\delta^2/c^2),$$

for all $\delta > C(\log n/n)^{1/2} =: \delta_n$. Therefore,

$$\begin{aligned}
\mathbb{E}_{G_0} W_1(\widehat{G}_n, G_0) = & \mathbb{E}_{G_0} W_1(\widehat{G}_n, G_0) 1_{[W_1(\widehat{G}_n, G_0) < \delta_n]} \\
& + \mathbb{E}_{G_0} W_1(\widehat{G}_n, G_0) 1_{[W_1(\widehat{G}_n, G_0) > \delta_n]} \\
\le & \delta_n + \int_{\delta_n}^{\infty} \mathbb{P}_{G_0}\left(W_1(\widehat{G}_n, G_0) > \delta\right) d\delta \\
\le & \delta_n + \int_{\delta_n}^{\infty} c\exp(-n\delta^2/c^2) d\delta \\
\le & \delta_n + \int_{\delta_n}^{\infty} c\exp(-C(n\log n)^{1/2}\delta/c^2) d\delta \\
= & \delta_n + c\frac{c^2}{(n\log n)^{1/2}C}\exp(-C(n\log n)^{1/2}\delta/c^2)\Big|_{\infty}^{\delta_n} \\
\le & C'\left(\frac{\log n}{n}\right)^{1/2},
\end{aligned}$$

where C' is a constant depending on Θ, f, h_1, h_2, G_0. Similar inequalities yield the expectation bound for $W_2(\widehat{G}_n, G_0)$ in the over-fitted setting. □

Proof of Proposition 9. From Theorem 6, we have that with probability at least $1 - cn^{-1/c^2}$,

$$\begin{aligned}
W_2(\widehat{G}_n^{(k_0)}, \widehat{G}_n^{(k_0-1)}) &\ge W_2(G_0, \widehat{G}_n^{(k_0-1)}) - W_2(G_0, \widehat{G}_n^{(k_0)}) \\
&\ge W_2(G_0, G_0^{\perp}) - C_1\left(\frac{\log n}{n}\right)^{1/4} \\
&\gg a_n,
\end{aligned}$$

where C_1 is a constant depending on G_0, Θ, h_1, h_2, f, and

$$G_0^{\perp} \in \arg\inf_{G \in \mathcal{O}_{k-1}} W_2(G_0, G).$$

Note that since $\mathcal{O}_{k_0-1}$ is compact and W_2 is a metric and continuous for both arguments, this infimum problem achieves at least a minimizer (see also [5] for the closed form solution of $G_0^{\perp}$). In the second inequality, we also use the fact that $W_2^2(G_0, \widehat{G}_n^{(k_0)}) \le W_1(G_0, \widehat{G}_n^{(k_0)})$ as $W_1(G_0, \widehat{G}_n^{(k_0)}) \to 0$. Since $G_0 \in \mathcal{E}_{k_0}(\Theta)$, $W_2(G_0, G_0^{\perp})$ is a positive constant, which justifies the last inequality above.

Besides, for all $k \in [k_0, \overline{K} - 1]$, with probability at least $1 - cn^{-1/c^2}$, we have

$$W_2(\widehat{G}_n^{(k)}, \widehat{G}_n^{(k+1)}) \le W_2(\widehat{G}_n^{(k)}, G_0) + W_2(G_0, \widehat{G}_n^{(k+1)}) \le C_2\left(\frac{\log n}{n}\right)^{1/4} \ll a_n.$$

Hence, for sufficiently large n, the selection rule of k_n gives $\mathbb{P}(k_n = k_0) \geq 1 - cn^{-1/c^2}$. By combining it with Theorem 7(a), we arrive at the conclusion (17). □

A.3. Posterior contraction theorems

Proof of Theorem 7. It suffices to verify conditions (i) and (ii) of Theorem 11 in Appendix D in order to arrive at the conclusion, with $\mathcal{F} = \mathcal{F}_n = \{f_G : G \in \mathcal{O}_K\}$ and $\epsilon_n = (\log n/n)^{1/2}$.

Checking condition (i): We need to show that the prior distribution puts enough mass around the true (conditional) density function f_{G_0}, i.e., to obtain a lower bound for $\Pi(B_2(f_{G_0}, \epsilon_n))$. First, consider the ball $\{G \in \mathcal{O}_K(\Theta) : W_1(G, G_0) \leq C\epsilon_n^2\}$ for a constant C to be chosen later. By Lemma 1, we have $\mathbb{E}_X d_H^2(f_{G_0}, f_G) \leq C_1 C\epsilon_n^2$, where C_1 depends on Θ. Because $CC_1\epsilon_n^2 \leq \epsilon_0$ for all sufficiently large n, we have $\mathbb{E}_{\mathbb{P}_X \times f_{G_0}}(f_{G_0}/f_G) \leq M$. By Theorem 5 in [43], we have

$$\mathbb{E}_{\mathbb{P}_X} K(f_{G_0}, f_G) \preccurlyeq \epsilon_n^2 \log(\sqrt{M}/\sqrt{CC_1}\epsilon_n)$$
$$\mathbb{E}_{\mathbb{P}_X} K_2(f_{G_0}, f_G) \preccurlyeq \epsilon_n^2 \log(\sqrt{M}/\sqrt{CC_1}\epsilon_n)^2.$$

Hence, for $\overline{M} = \log(\sqrt{M}/\sqrt{CC_1})$, we have

$$\Pi(B_2(f_{G_0}, \overline{M}\epsilon_n)) \geq \Pi(W_1(G, G_0) \leq C\epsilon_n).$$

However, for all $G = \sum_{i=1}^{k_0} p_i \delta_{(\theta_{1i}, \theta_{2i})}$ such that $\|\theta_{1i} - \theta_{1i}^0\| \leq \epsilon_n/(4k_0), \|\theta_{2i} - \theta_{2i}^0\| \leq \epsilon_n/(4k_0), |p_i - p_i^0| \leq \epsilon_n/(4k_0 \mathrm{diam}(\Theta_1) \times \mathrm{diam}(\Theta_2))$, we have

$$\begin{aligned} W_1(G_0, G) &\leq \sum_{i=1}^{k_0} (p_i^0 \wedge p_i)(\|\theta_{1i} - \theta_{1i}^0\| + \|\theta_{2i} - \theta_{2i}^0\|) + |p_i - p_i^0|(\mathrm{diam}(\Theta_1)\mathrm{diam}(\Theta_2)) \\ &\leq \epsilon_n. \end{aligned}$$

Due to assumption (B1.), the prior measure of this set is asymptotically greater than ϵ_n Hence

$$\Pi(W_1(G, G_0) \leq C\epsilon_n) \succcurlyeq \epsilon_n \succcurlyeq e^{-cn\epsilon_n^2},$$

as $\epsilon_n = (\log n/n)^{1/2}$.

Checking condition (ii): We need to provide an upper bound for the entropy number $\log N(\mathcal{F}, \overline{d}_H, \epsilon_n)$. By Lemma 1,

$$\overline{d}_H^2(f_G, f_{G_0}) \leq \mathbb{E}_X d_{TV}(f_G, f_{G_0}) \preccurlyeq W_1(G, G_0)$$

We use the same strategy as in the proof of Theorem 5. Since Θ_1 and Θ_2 are compact, for all $\epsilon > 0$, there exists an ϵ-net B_1 of $(\Theta_1, \|\cdot\|)$ and B_2 of $(\Theta_2, \|\cdot\|)$ with the cardinality $|B_1| \leq \left(\frac{\text{diam}(\Theta_1)}{\epsilon}\right)^{d_1}$ and $|B_2| \leq \left(\frac{\text{diam}(\Theta_2)}{\epsilon}\right)^{d_2}$. Moreover, there exists an ϵ-net A for $(\Delta^{k-1}, \|\cdot\|_\infty)$ such that $|A| \leq \left(\frac{5}{\epsilon}\right)^k$. We consider the following subset of $\mathcal{F}$

$$C = \{G : G = \sum_{i=1}^{k} p_j \delta_{(\theta_{1j},\theta_{2j})}, (p_j)_{j=1}^k \in A, \theta_{1j} \in B_1, \theta_{2j} \in B_2 \forall\, j\}.$$

Note that

$$|C| = |A| \times |B_1|^k \times |B_2|^k \leq \left(\frac{5}{\epsilon}\right)^k \left(\frac{\text{diam}(\Theta_1)}{\epsilon}\right)^{d_1 k} \left(\frac{\text{diam}(\Theta_2)}{\epsilon}\right)^{d_2 k}.$$

For any $G = \sum_{i=1}^k p_i \delta_{(\theta_{1j},\theta_{2j})} \in \mathcal{O}_k(\Theta)$, there exist $(\tilde{p}_j)_{j=1}^k \in A$ and $\tilde{\theta}_j \in B$ such that $|p_j - \tilde{p}_j| \leq \epsilon_n$ and $\|\theta_{1j} - \tilde{\theta}_{2j}\| \leq \epsilon_n$, $\|\theta_{2j} - \tilde{\theta}_{2j}\| \leq \epsilon_n$ for all j. Let $\tilde{G} = \sum_{j=1}^k \tilde{p}_i \delta_{(\tilde{\theta}_{1j},\tilde{\theta}_{2j})}$ and $G' = \sum_{i=1}^k \tilde{p}_i \delta_{(\theta_{1j},\theta_{2j})}$, by the triangle inequality, we have

$$\begin{aligned} W_1(G,\tilde{G}) &\leq W_1(G,G') + W_1(G',\tilde{G}) \\ &\leq \sum_{j=1}^{k} |p_j - \tilde{p}_j| 2(\text{diam}(\Theta_1) + \text{diam}(\Theta_2)) \\ &\quad + \sum_{j=1}^{k} \tilde{p}_j (\|\theta_{1j} - \tilde{\theta}_{1j}\| + \|\theta_{2j} - \tilde{\theta}_{2j}\|) \preccurlyeq \epsilon_n, \end{aligned}$$

This implies that the covering number

$$N(\epsilon_n, \mathcal{F}, \overline{d}_H) \preccurlyeq \left(\frac{5}{\epsilon_n}\right)^k \left(\frac{\text{diam}(\Theta_1)}{\epsilon_n}\right)^{d_1 k} \left(\frac{\text{diam}(\Theta_2)}{\epsilon_n}\right)^{d_2 k}.$$

Taking logarithm of both sides, we obtain $\log N(\epsilon_n, \mathcal{F}, \overline{d}_H) \preccurlyeq \log(1/\epsilon_n) \leq n\epsilon_n^2$. Now, we are ready to apply Theorem 11 to conclude the proof. □

Proof of Theorem 8. The proof of this theorem is similar to that of Theorem 4. It is a direct consequence of Theorem 7, where we proved that the posterior contraction rate of $\overline{d}_H(f_G, f_{G_0})$ is $(\log n/n)^{1/2}$, and the inverse bounds (Theorem 2), where we showed that $\overline{d}_H(f_G, f_{G_0}) \succcurlyeq W_1(G, G_0)$ in the exact-fitted case and $\overline{d}_H(f_G, f_{G_0}) \succcurlyeq W_2^2(G, G_0)$ in the over-fitted case. □

Now we are to establish the consistency of the number of parameters and the posterior contraction rate of the latent mixing measure in a Bayesian estimation setting, where the regression mixture model is endowed with a "mixture of finite mixture" prior. The proof makes a crucial usage of Doob's consistency theorem ([8] Theorem 6.9, or [26] Theorem 2.2).

Proof of Theorem 9. For each latent mixing measure G, we write $k(G)$ as its number of (distinct) support points. Recall that we have a prior Π on $\mathcal{G} = \cup_{k=1}^{\infty}\mathcal{E}_k$, which is a subset of the complete and separable Wasserstein space endowed with metric W_1. By assumption, G (and hence $k(G)$) is identifiable. By Doob's consistency theorem [7] (or [8] Theorem 6.9), there exists $\mathcal{G}_0 \subset \overline{\mathcal{G}}$ such that $\Pi(\mathcal{G}_0) = 1$ and for any $G_0 \in \mathcal{G}_0 \cap \mathcal{E}_{k_0}$, i.e. those $G_0 \in \mathcal{G}_0$ that that have k_0 supporting atoms, we have

$$\mathbb{P}(k(G) = k_0|x^{[n]}, y^{[n]}) = \mathbb{E}[1(k(G) = k_0)|x^{[n]}, y^{[n]}] \to 1(k(G_0) = k_0) = 1,$$

almost surely in $\otimes_{i=1}^{\infty}\mathbb{P}_{G_0}$. For the mixture of finite mixtures prior, K represents the (random) number of components. Moreover, by assumption, given $K = k$, the prior distributions on $p = (p_j)_{j=1}^k$ and $(\theta_j)_{j=1}^k$ are absolutely continuous, and set $G = \sum_{j=1}^k p_j\delta_{\theta_j}$. Thus, under the induced prior Π on the mixing measure, we have $k(G) = K$ for Π-almost all G. This entails that there exists $\mathcal{G}_0' \subset \overline{\mathcal{G}}$ such that $\Pi(\mathcal{G}_0') = 1$ and for any $G_0 \in \mathcal{G}_0'$ we have

$$\mathbb{P}(k(G) = K|x^{[n]}, y^{[n]}) = 1 \quad \forall n \geq 1 \quad \text{a.s } \otimes_{i=1}^{\infty} \mathbb{P}_{G_0}.$$

Now, for any $G_0 \in \mathcal{G}_0 \cap \mathcal{G}_0'$, by the calculus of probabilities

$$\begin{aligned}
\mathbb{P}(K = k_0|x^{[n]}, y^{[n]}) &\geq \mathbb{P}(K = k_0, k(G) = k_0|x^{[n]}, y^{[n]}) \\
&= \mathbb{P}(K = k(G), k(G) = k_0|x^{[n]}, y^{[n]}) \\
&\geq \mathbb{P}(k(G) = k_0|x^{[n]}, y^{[n]}) - \mathbb{P}(k(G) \neq K|x^{[n]}, y^{[n]}) \\
&= \mathbb{P}(k(G) = k_0|x^{[n]}, y^{[n]}).
\end{aligned}$$

Thus, $\mathbb{P}(K = k_0|x^{[n]}, y^{[n]}) \to 1$ a.s $\otimes_{i=1}^{\infty} \mathbb{P}_{G_0}$, provided that $G_0 \in \mathcal{G}_0 \cap \mathcal{G}_0' \cap \mathcal{E}_{k_0}$. Then, with $\epsilon_n = \sqrt{\log n/n}$, we can bound:

$$\begin{aligned}
\Pi(G : W_1(G, G_0) \succcurlyeq \epsilon_n|x^{[n]}, y^{[n]}) &= \sum_{k=1}^{\infty} \Pi(G \in \mathcal{E}_k(\Theta) : W_1(G, G_0) \succcurlyeq \epsilon_n|x^{[n]}, y^{[n]}) \\
&\leq \Pi(K \neq k_0|x^{[n]}, y^{[n]}) \\
&\quad + \Pi(G \in \mathcal{E}_{k_0}(\Theta) : W_1(G, G_0) \succcurlyeq \epsilon_n|x^{[n]}, y^{[n]}).
\end{aligned}$$

The first term goes to 0 $\mathbb{P}_{G_0}$-almost surely, thanks to the argument above. For the second term, we apply the first part of Theorem 8 to conclude that it tends to 0 in $\mathbb{P}_{G_0}$-probability. □

Appendix B: Proofs of remaining main results

B.1. Basic inequalities

Proof of Lemma 1. Let $G = \sum_{i=1}^K p_i\delta_{(\theta_{1i},\theta_{2i})}$ and recall that $G_0 = \sum_{j=1}^{k_0} p_i^0$ $\delta_{(\theta_{1j}^0,\theta_{2j}^0)}$. To ease the presentation, denote $f_i(y|x) = f(y|h_1(x,\theta_{1i}), h_2(x,\theta_{2i}))$

and $f_j^0(y|x) = f(y|h_1(x,\theta_{1j}^0), h_2(x,\theta_{2j}^0))$ for $i = 1,\dots,K, j = 1,\dots,k_0$. We have

$$\begin{aligned}
\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X)) &= \int_{\mathcal{X}} d\mathbb{P}_X \int_{\mathcal{Y}} d\nu(y) \left| \sum_{i=1}^{K} p_i f_i(y|x) - \sum_{j=1}^{k_0} p_j^0 f_j^0(y|x) \right| \\
&= \int_{\mathcal{X}} d\mathbb{P}_X \int_{\mathcal{Y}} d\nu(y) \left| \sum_{i,j=1}^{K,k_0} q_{ij}(f_i(y|x) - f_j^0(y|x)) \right| \\
&\leq \sum_{i,j=1}^{K,k_0} q_{ij} \int_{\mathcal{X}} d\mathbb{P}_X \int_{\mathcal{Y}} d\nu(y) \left| f_i(y|x) - f_j^0(y|x) \right|,
\end{aligned}$$

for any coupling $(q_{ij})_{i,j=1}^{K,k_0}$ of $(p_i)_{i=1}^{K}$ and $(p_j^0)_{j=1}^{k_0}$. But because of the uniform Lipschitz assumption of f and h_1, h_2, we have

$$|f_i(y|x) - f_j^0(y|x)| \leq c(|h_1(x,\theta_{1i}) - h_1(x,\theta_{1j}^0)| + |h_2(x,\theta_{2i}) - h_2(x,\theta_{2j}^0)|),$$

and then

$$|f_i(y|x) - f_j^0(y|x)| \leq cc_1 \left\|\theta_{1i} - \theta_{1j}^0\right\| + cc_2 \left\|\theta_{2i} - \theta_{2j}^0\right\| \quad \forall x, y.$$

Therefore,

$$\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X)) \leq c \max\{c_1, c_2\} \sum_{i,j=1}^{K,k_0} q_{ij}(\left\|\theta_{1i} - \theta_{1j}^0\right\| + \left\|\theta_{2i} - \theta_{2j}^0\right\|),$$

for all x, y. Taking infimum of all feasible $(q_{ij})_{i,j}$ to obtain

$$\mathbb{E}_X d_{TV}(f_G(\cdot|X), f_{G_0}(\cdot|X)) \preccurlyeq W_1(G, G_0). \qquad \square$$

Remark 5. *By inspecting the proof above, we see that the results still hold if we change the uniform Lipschitz condition of h_1 and h_2 to the integrability of the Lipschitz constants, i.e. there exist $c_1(x), c_2(x)$ for all $x \in \mathcal{X}$ such that*

$$h_1(x,\theta_1) - h_1(x,\theta_1') \leq c_1(x) \left\|\theta_1 - \theta_1'\right\|, \quad h_2(x,\theta_2) - h_2(x,\theta_2') \leq c_2(x) \left\|\theta_2 - \theta_2'\right\|,$$

for all $\theta_1, \theta_2, \theta_1', \theta_2'$, and $\mathbb{E}_X c_1(X) < \infty, \mathbb{E}_X c_2(X) < \infty$. This condition is weaker than the uniformly Lipschitz condition in x.

B.2. Identifiability results

Proof of Proposition 1. (a), (b): Can be found in [2, 17].

(c) First, we will establish the first order identifiability condition when $2K \leq N+1$. Suppose that $q_1, q_2, \dots q_K \in [0,1]$ are distinct and there exist $\alpha_1, \dots, \alpha_K$, $\beta_1, \dots, \beta_K$ such that

$$\alpha_1 \mathrm{Bin}(y|q_1) + \cdots + \alpha_K \mathrm{Bin}(y|q_K) + \beta_1 \frac{\partial}{\partial q}\mathrm{Bin}(y|q_1) + \cdots + \beta_K \frac{\partial}{\partial q}\mathrm{Bin}(y|q_K) = 0, \tag{43}$$

for all $y = 0, 1, \dots, N$. Direct calculation gives

$$\sum_{i=1}^{K} q_i^y (1-q_i)^{N-y}\alpha_i + \sum_{i=1}^{K} \frac{\partial}{\partial q} q_i^y (1-q_i)^{N-y}\beta_i = 0, \quad \forall y = 0, \dots, N. \tag{44}$$

Because this is a system of linear equations of $(\alpha_i, \beta_i)_{i=1}^K$, it suffices to show that the following $(N+1) \times 2K$ matrix has independent columns

$$\begin{pmatrix} (1-q_1)^N & \cdots & (1-q_K)^N & \frac{\partial}{\partial q}(1-q_1)^N & \cdots & \frac{\partial}{\partial q}(1-q_K)^N \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q_1^N & \cdots & q_K^N & \frac{\partial}{\partial q} q_1^N & \cdots & \frac{\partial}{\partial q} q_K^N \end{pmatrix}.$$

Multiplying this matrix with the following upper triangular matrix

$$\begin{pmatrix} 1 & \binom{N}{1} & \binom{N}{2} & \cdots & \binom{N}{N} \\ 0 & 1 & \binom{N-1}{1} & \cdots & \binom{N-1}{N-1} \\ 0 & 0 & 1 & \cdots & \binom{N-2}{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

we only need to prove the following $(N+1) \times 2K$ matrix

$$\begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ q_1 & \cdots & q_K & 1 & \cdots & 1 \\ q_1^2 & \cdots & q_2^2 & 2q_1 & \cdots & 2q_2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q_1^N & \cdots & q_K^N & Nq_1^{N-1} & \cdots & Nq_K^{N-1} \end{pmatrix}$$

has independent columns. Because $2K \leq N+1$, it suffices to prove that $\det(D_1) \neq 0$, for

$$D_1 = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ q_1 & \cdots & q_K & 1 & \cdots & 1 \\ q_1^2 & \cdots & q_2^2 & 2q_1 & \cdots & 2q_2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q_1^{2K-1} & \cdots & q_K^{2K-1} & (2K-1)q_1^{2K-2} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}. \tag{45}$$

In the following, we will prove that $\det(D_1) = \prod_{1 \leq i < j \leq K}(q_i - q_j)^4$, so that it is different from 0 if $q_1, \dots, q_N$ are distinct as in our assumption. We borrow an idea in calculating the determinant of the Vandermonde matrix. Note that $\det(D_1)$ is a polynomial of $q_1, q_2, \ldots, q_K$, with the degree of each q_i no more than $4K-4$. Let us treat $q_1 = x$ as a variable, while $q_2, \ldots, q_K$ as constants, and prove that the polynomial $f(x)$ being equal to

$$\det\begin{pmatrix} 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ x & q_2 & \cdots & q_K & 1 & 1 & \cdots & 1 \\ x^2 & q_2^2 & \cdots & q_K^2 & 2x & 2q_2 & \cdots & 2q_K \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x^{2K-1} & q_2^{2K-1} & \cdots & q_K^{2K-1} & (2K-1)x^{2K-2} & (2K-1)q_2^{2K-2} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

is a polynomial having degree $4(K-1)$ of x and can be factorized as $\prod_{i=2}^{K}(x-q_i)^4$. It suffices for us to show $f(x), f'(x), f''(x), f^{(3)}(x)$ all attains q_2 as solutions, and similar for other q_i's. It can be seen that $f_1(q_2)$ is a determinant of a matrix with identical first two columns, therefore $f_1(q_2) = 0$. For the derivative of f_1, we use the derivative rule for product $(fg)' = f'g + g'f$ to have that $f_1'(x)$ equals

$$\det\begin{pmatrix} 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 1 & \cdots & q_K & 1 & 1 & \cdots & 1 \\ 2x & \cdots & q_K^2 & 2x & 2q_2 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ (2K-1)x^{2K-2} & \cdots & q_K^{2K-1} & (2K-1)x^{2K-2} & (2K-1)q_2^{2K-2} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$
$$+ \det\begin{pmatrix} 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ x & \cdots & q_K & 0 & 1 & \cdots & 1 \\ x^2 & \cdots & q_K^2 & 2 & 2q_2 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x^{2K-1} & \cdots & q_K^{2K-1} & (2K-1)(2K-2)x^{2K-3} & (2K-1)q_2^{2K-2} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}.$$

As the first matrix has identical first and $(K+1)$-th columns, its determinant equals 0. Hence, $f_1'(x)$ equals

$$\det\begin{pmatrix} 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ x & \cdots & q_K & 0 & 1 & \cdots & 1 \\ x^2 & \cdots & q_K^2 & 2 & 2q_2 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x^{2K-1} & \cdots & q_K^{2K-1} & (2K-1)(2K-2)x^{2K-3} & (2K-1)q_2^{2K-2} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

Now consider $f_1'(q_2)$. It is the determinant of a matrix that has identical first two columns, so $f_1'(q_2) = 0$. Continuing to apply the derivative rule for products of functions, $f_1''(x)$ equals

$$\det\begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 1 & \cdots & q_K & 0 & \cdots & 1 \\ 2x & \cdots & q_K^2 & 2 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ (2K-1)x^{2K-2} & \cdots & q_K^{2K-1} & \prod_{i=1}^{2}(2K-i)x^{2K-3} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

$$+\det\begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ x & \cdots & q_K & 0 & \cdots & 1 \\ x^2 & \cdots & q_K^2 & 0 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x^{2K-1} & \cdots & q_K^{2K-1} & \prod_{i=1}^{3}(2K-i)x^{2K-4} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

Substitute $x = q_2$ in the formula above, the first matrix has identical first and $(K+2)$-th column, and the second matrix has identical first two columns. Hence, $f_1''(q_2) = 0$. Continue applying derivative one more time, $f_1^{(3)}(x)$ equals

$$\det\begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & q_K & 0 & \cdots & 1 \\ 2 & \cdots & q_K^2 & 2 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \prod_{i=1}^{2}(2K-i)x^{2K-3} & \cdots & q_K^{2K-1} & \prod_{i=1}^{2}(2K-i)x^{2K-3} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

$$+\det\begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 1 & \cdots & q_K & 0 & \cdots & 1 \\ 2x & \cdots & q_K^2 & 2 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ (2K-1)x^{2K-2} & \cdots & q_K^{2K-1} & \prod_{i=1}^{3}(2K-i)x^{2K-4} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

$$+\det\begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 1 & \cdots & q_K & 0 & \cdots & 1 \\ 2x & \cdots & q_K^2 & 0 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ (2K-1)x^{2K-2} & \cdots & q_K^{2K-1} & \prod_{i=1}^{3}(2K-i)x^{2K-4} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}$$

$$+\det\begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ x & \cdots & q_K & 0 & \cdots & 1 \\ x^2 & \cdots & q_K^2 & 0 & \cdots & 2q_K \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x^{2K-1} & \cdots & q_K^{2K-1} & \prod_{i=1}^{4}(2K-i)x^{2K-5} & \cdots & (2K-1)q_K^{2K-2} \end{pmatrix}.$$

The first matrix has two identical columns, so its determinant is 0. Meanwhile, when we substitute $x = q_2$ into the other three matrices, each also has identical columns. Hence, $f^{(3)}(q_2) = 0$. We obtain that $f_1(x) \propto \prod_{i=2}^{K}(x-q_i)^4$. By treating $q_2, \ldots, q_K$ as variables respectively and applying the same argument, we have

$$\det(D_1) = \prod_{1\le i<j\neq K}(q_i - q_j)^4 \neq 0,$$

whenever $q_1, q_2, \dots, q_K$ are distinct.

The proof for establishing the second order identifiability when $3K \le N+1$ is similar, where the determinant of the derived $3K \times 3K$ matrix is $\prod_{1\le i<j\ne K}(q_i - q_j)^6 \ne 0$.

(d) For the family of negative binomial distributions, the density is given as

$$\mathrm{NB}(y|\mu,\phi) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)y!}\left(\frac{\mu}{\phi+\mu}\right)^y\left(\frac{\phi}{\phi+\mu}\right)^\phi.$$

Suppose that $\mu_1, ..., \mu_K$ are distinct, and there exist $\alpha_1, ..., \alpha_K, \beta_1, ..., \beta_K, \gamma_1, ..., \gamma_K$ such that for every $y \in \mathbb{N}$

$$\sum_{i=1}^K \alpha_i \mathrm{NB}(y|\mu_i,\phi) + \sum_{i=1}^K \beta_i \frac{\partial}{\partial\mu}\mathrm{NB}(y|\mu_i,\phi) + \sum_{i=1}^K \gamma_i \frac{\partial^2}{\partial\mu^2}\mathrm{NB}(y|\mu_i,\phi) = 0. \tag{46}$$

We will show that $\alpha_1 = \dots = \alpha_K = \beta_1 = \dots = \beta_K = \gamma_1 = \dots \gamma_K = 0$. Indeed, Eq. (46) is simplified as below

$$\sum_{i=1}^K \alpha_i \left(\frac{\mu_i}{\phi+\mu_i}\right)^y \left(\frac{\phi}{\phi+\mu_i}\right)^\phi + \sum_{i=1}^K \beta_i \left(\frac{\mu_i}{\phi+\mu_i}\right)^{y-1}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1}\frac{y-\mu_i}{\phi+\mu_i}$$
$$+\sum_{i=1}^K \gamma_i \left(\frac{\mu_i}{\phi+\mu_i}\right)^{y-2}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1}\frac{1}{(\phi+\mu_i)^3}[\phi(y-\mu_i)^2 - y(2\mu_i+\phi)+\mu_i^2]$$
$$= 0, \tag{47}$$

for all $y \in \mathbb{N}$. Without loss of generality, assume μ_1 is the largest value in the set of $\{\mu_1, ..., \mu_K\}$. This implies that $\dfrac{\mu_1}{\phi+\mu_1}$ is also the largest value in the set of $\left\{\dfrac{\mu_1}{\phi+\mu_1}, ..., \dfrac{\mu_K}{\phi+\mu_K}\right\}$. Dividing both sides of Eq. (47) by $\left(\dfrac{\mu_1}{\phi+\mu_1}\right)^{y-2}[\phi(y-\mu_1)^2 - y(2\mu_1+\phi)+\mu_1^2] = \left(\dfrac{\mu_1}{\phi+\mu_1}\right)^{y-2} A_1(y)$, we obtain

$$\sum_{i=1}^K \alpha_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y-2}\left(\frac{\mu_i}{\phi+\mu_i}\right)^2\left(\frac{\phi}{\phi+\mu_i}\right)^\phi \frac{1}{A_1(y)}$$
$$+\sum_{i=1}^K \beta_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y-2}\left(\frac{\mu_i}{\phi+\mu_i}\right)\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1}\frac{y-\mu_i}{(\phi+\mu_i)A_1(y)}$$
$$+\sum_{i=2}^K \gamma_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y-2}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1}\frac{A_i(y)}{(\phi+\mu_i)^3 A_1(y)}$$
$$+\gamma_1\left(\frac{\phi}{\phi+\mu_1}\right)^{\phi+1}\frac{1}{(\phi+\mu_1)^3} = 0, \forall y \in \mathbb{N}. \tag{48}$$

Let $y \to \infty$ in (48), we get $\gamma_1 = 0$. After dropping γ_1 in (47), the remaining terms of the equation is divided by $\left(\frac{\mu_1}{\phi+\mu_1}\right)^{y-1}(y-\mu_1)$, we have

$$\begin{aligned}
&\sum_{i=1}^{K} \alpha_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y-1} \left(\frac{\mu_i}{\phi+\mu_i}\right) \left(\frac{\phi}{\phi+\mu_i}\right)^{\phi} \frac{1}{y-\mu_1} \\
&+ \sum_{i=2}^{K} \beta_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y-1} \left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1} \left(\frac{y-\mu_i}{y-\mu_1}\right) \frac{1}{\phi+\mu_i} \\
&+ \sum_{i=2}^{K} \gamma_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y-2} \left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1} \frac{1}{(\phi+\mu_i)^3} \frac{A_i(y)}{y-\mu_1} \\
&+ \beta_1 \left(\frac{\phi}{\phi+\mu_1}\right)^{\phi+1} \frac{1}{\phi+\mu_1} = 0, \forall y \in \mathbb{N}.
\end{aligned} \tag{49}$$

Taking the limit $y \to \infty$ both sides of Eq. (49), we get $\beta_1 = 0$. Continuing this procedure, we set $\beta_1 = 0$ and $\gamma_1 = 0$ in Eq. (46), then divide $\left(\frac{\mu_1}{\phi+\mu_1}\right)^{y}$ on both sides of the remaining equation. The final result leads to the following:

$$\begin{aligned}
&\sum_{i=2}^{K} \alpha_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y} \left(\frac{\phi}{\phi+\mu_i}\right)^{\phi} \\
&+ \sum_{i=2}^{K} \beta_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y} \left(\frac{\mu_i}{\phi+\mu_i}\right)^{-1} \left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1} \frac{y-\mu_i}{\phi+\mu_i} \\
&+ \sum_{i=2}^{K} \gamma_i \left(\frac{\mu_i(\phi+\mu_1)}{\mu_1(\phi+\mu_i)}\right)^{y} \left(\frac{\mu_i}{\phi+\mu_i}\right)^{-2} \left(\frac{\phi}{\phi+\mu_i}\right)^{\phi+1} \frac{A_i(y)}{(\phi+\mu_i)^3} \\
&+ \alpha_1 \left(\frac{\phi}{\phi+\mu_1}\right)^{\phi} = 0, \forall y \in \mathbb{N}.
\end{aligned} \tag{50}$$

It is clear to see that $\alpha_1 = 0$ when y approaches ∞ in Eq. (50). We have shown that $\alpha_1, \beta_1, \gamma_1 = 0$. Inductively, we obtain that $\alpha_i, \beta_i, \gamma_i = 0$ for $i = 2, ..., K$. □

Proof of Proposition 3. Since $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, it is sufficient to prove the result in this proposition with respect to the Lebesgue measure.

(a) We will prove this part by applying an inductive argument with respect to p (dimension of covariate x). Suppose $p = 1$. For $\theta \neq \theta'$, the equation $h(x, \theta) = h(x, \theta')$ is a non-trivial polynomial equation, it only has a finite number of solutions. Thus the set of solution has Lebesgue measure zero, so we have $h(x, \theta) \neq h(x, \theta')$ a.s.

Assume now that the proposition is valid up to the parameter space dimension $p-1$. Now we prove that it is correct for p. Using a similar argument as above, it suffices to show that the set of solutions for any non-trivial polynomial has zero measure. Indeed, consider any such polynomial of degree d of variable $x \in \mathbb{R}^p$, we can write $x = (X_{p-1}, x_p)$, where $X_{p-1} \in \mathbb{R}^{p-1}$ and $x_p \in \mathbb{R}$. The polynomial then can be written as:

$$\sum_{j=0}^{d} p_j(X_{p-1}) x_p^i = 0, \tag{51}$$

where $(p_j(X_{p-1}))_{j=0}^d$ are polynomial of $X_{p-1} \in \mathbb{R}^{p-1}$, at least one of which is non-trivial.

Now, partition the set Z of the solutions for this polynomial into two measurable sets $Z = A \cup B$, where

$$A = \{(X_{p-1}, x_p) : p_j(X_{p-1}) = 0 \ \forall j = 1, ..., d\}$$
$$B = \{(X_{p-1}, x_p) : \text{at least one } p_j(X_{p-1}) \neq 0, \text{ and } x_p \text{ satisfies (51).}\}$$

The Lebesgue measure of set A is 0 using the induction hypothesis. While for any $x = (X_{p-1}, x_p)$ in set B, for each such X_{p-1}, there exist only a finite number of $x_p \in \mathbb{R}$ to satisfy (51), which has zero Lebesgue measure in $\mathbb{R}$. Therefore, we can use Fubini's theorem to deduce that the measure of B is also zero. Thus, Z has measure zero. We have established that $h(x, \theta)$ is completely identifiable for any polynomial $h(x, \theta)$.

Turning to the verification of Assumption (A5.), since $\frac{\partial}{\partial \theta} h(x, \theta)$ is again a non-trivial polynomial of x, (A5.) is also satisfied using the same argument above.

(b) Similar to part (a), we only need to prove that a non-trivial (not all coefficients are 0) trigonometric polynomial of x:

$$a_0 + \sum_{n=1}^{d} b_n \cos(nx) + \sum_{n=1}^{d} \sin(nx) = 0 \tag{52}$$

has a countable number of solutions. Write $\cos(nx) = \frac{1}{2}(e^{inx} + e^{-inx}), \sin(nx) = \frac{1}{2i}(e^{inx} - e^{-inx})$, where i is the imaginary unit, we can rewrite a non-trivial trigonometric polynomial above as

$$a_0 + \sum_{n=1}^{d} \tilde{b}_n e^{inx} + \sum_{n=1}^{d} \tilde{c}_n e^{-inx} = 0, \tag{53}$$

where $\tilde{b}_n$ and $\tilde{c}_n$ are computed from b_n, c_n, and the tuple $(a_0, \tilde{b}_n, \tilde{c}_n)$ is non-trivial. Set $y = e^{ix}$, this becomes a polynomial in $y \in \mathbb{C}$, which has a finite number of solutions, by the fundamental theorem of algebra. Combining this with the fact that $e^{ix} = y$ only has a countable solution in x, we arrive at the conclusion. The condition spelled out in Assumption (A5.) also holds because the derivative of a trigonometric polynomial is still of the same form.

(c) Similar to above, we express a non-trivial mixture of polynomials and trigonometric polynomials in the form:

$$\sum_{n=0}^{d} a_n x^n + \sum_{n=1}^{d} \tilde{b}_n e^{inx} + \sum_{n=1}^{d} \tilde{c}_n e^{-inx} = 0, \tag{54}$$

which is a holomorphic function in $\mathbb{C}$. This function is known to have an isolated set of solutions, which has zero measure [36]. Thus, these functions are completely identifiable. Conditions in Assumption (A5.) also follow because the derivative of a function of this type is still of the same form.

(d) Given $h(x,\theta) = g(p(x,\theta))$, where g is diffeomorphic.

To verify the complete identifiability condition, it can be seen that for $\theta \neq \theta'$: $h(x,\theta) = h(x,\theta') \Leftrightarrow p(x,\theta) = p(x,\theta')$, so that the complete identifiability of h can be deduced from what of p.

To verify Assumption (A5.), note that $\beta^\top \dfrac{\partial h(x,\theta)}{\partial \theta} = \beta^\top g'(p(x,\theta)) \dfrac{\partial p(x,\theta)}{\partial \theta}$. Since $g'(p(x,\theta)) \neq 0$ (as g is a diffeomorphism), the two equations below are equivalent.

$$\beta^\top \frac{\partial h(x,\theta)}{\partial \theta} = 0 \Longleftrightarrow \beta^\top \frac{\partial p(x,\theta)}{\partial \theta} = 0.$$

Hence, h satisfies assumption (A5.) if p does. □

Now we turn into the proof of Proposition 7, which illustrates the discussion in Section 5 by showing that a mixture of binomial regression model may be strongly identifiable even though the (unconditional) mixture of binomial distributions is not identifiable in even in the classical sense.

Proof of Proposition 7. Consider $\theta_1 \neq \pm\theta_2$. Suppose that for some $a_1, a_2, b_1, b_2 \in \mathbb{R}$ we have

$$\begin{aligned} &a_1 \mathrm{Bin}(y|1, h(x,\theta_1)) + a_2 \mathrm{Bin}(y|1, h(x,\theta_2)) \\ &+ b_1 \frac{\partial}{\partial \theta} \mathrm{Bin}(y|1, h(x,\theta_1)) + b_2 \frac{\partial}{\partial \theta} \mathrm{Bin}(y|1, h(x,\theta_2)) = 0, \end{aligned}$$

for all $y = 0, 1,$ and $x \in \mathrm{supp}(\mathbb{P}_X)$. Then we will show that $a_1 = a_2 = b_1 = b_2 = 0$. Denote $\sigma_i(x) = h(x,\theta_i)$, we have $\mathrm{Bin}(1|1, h(x,\theta_i)) = \sigma_i = 1 - \mathrm{Bin}(0|1, h(x,\theta_i))$. Besides, $\dfrac{\partial}{\partial \theta} \mathrm{Bin}(1|1, h(x,\theta_i)) = x\sigma_i(x)(1 - \sigma_i(x)) = -\dfrac{\partial}{\partial \theta} \mathrm{Bin}(0|1, h(x,\theta_i))$, so that

$$a_1 + a_2 = 0, \tag{55}$$

and

$$a_1\sigma_1(x) + a_2\sigma_2(x) + b_1 x\sigma_1(x)(1-\sigma_1(x)) + b_2 x\sigma_2(x)(1-\sigma_2(x)) = 0,\ \forall\, x \in \mathrm{supp}(\mathbb{P}_X). \tag{56}$$

Because Eq. (56) satisfies for all x in an open set, and it is an analytic function of x, it satisfies for all $x \in \mathbb{R}$ (identity theorem) [36]. Without the loss of generality, we assume $\theta_1 < \theta_2$. If $0 \le \theta_1 < \theta_2$, then by dividing both sides of Eq. (56) by $\sigma_1(x)x$, one obtains

$$\frac{a_1}{x}+a_2\frac{1+\exp(\theta_1 x)}{1+\exp(\theta_2 x)}\frac{1}{x}+b_1(1-\sigma_1(x))+b_2(1-\sigma_2(x))\frac{1+\exp(\theta_1 x)}{1+\exp(\theta_2 x)}=0, \quad \forall\, x\in\mathbb{R}.$$

Let $x \to \infty$, we have $b_1(1-\sigma_1(x)) \to b_1/2$ or b_1 (depending on whether $\theta_1 = 0$ or $\theta_1 > 0$) and all other terms go to 0. Hence $b_1 = 0$. Next, dividing both sides of Eq. (56) by $\sigma_1(x)$, one obtains

$$a_1+a_2\frac{1+\exp(\theta_1 x)}{1+\exp(\theta_2 x)}+b_2(1-\sigma_2(x))\frac{(1+\exp(\theta_1 x))x}{1+\exp(\theta_2 x)}=0, \quad \forall\, x\in\mathbb{R}.$$

Let $x \to \infty$, we have $a_1 = 0$. Therefore,

$$a_2+b_2\frac{\exp(\theta_2 x)}{1+\exp(\theta_2 x)}x=0 \quad \forall\, x\in\mathbb{R},$$

which implies $a_2 = b_2 = 0$. In the other case where $\theta_1 < 0 < \theta_2$, we let $x \to \infty$ in Eq. (56) and notice that $\theta_1(x) \to 1, \theta_2(x) \to 0$ then $a_1 = 0$. Similarly let $x \to -\infty$, we have $a_2 = 0$. Then Eq. (56) becomes

$$b_1 h(x,\theta_1)(1-h(x,\theta_1))+b_2 h(x,\theta_2)(1-h(x,\theta_2))=0.$$

But notice that $h(x,\theta_1)(1-h(x,\theta_1)) = h(x,-\theta_1)(1-h(x,-\theta_1))$, so by letting $\theta_1' = -\theta_1$, we are back to the case $\theta_1', \theta_2 > 0$. Similar to the case $\theta_1, \theta_2 < 0$, we can transform $\theta_1 \mapsto -\theta_1, \theta_2 \mapsto \theta_2'$ to go back to the first case $\theta_1, \theta_2 > 0$ (because Eq. (56) satisfies for all $x \in \mathbb{R}$). Hence, in all cases we have $a_1 = a_2 = b_1 = b_2 = 0$. Hence, strong identifiability in the first order is established. □

Remark 6. *1. The fact that mixture of Binomial distributions is not identifiable in general can be seen from a simple example:* $0.5Bin(y|1,q_1) + 0.5Bin(y|1,q_2) = 0.5Bin(y|1,q_1+\epsilon)+0.5Bin(y|1,q_2-\epsilon)$ *for all valid* $\epsilon > 0$. *That is also the reason why one cannot include the intercept parameter in the definition of* h *in the proposition above.*

2. The proof technique of this proposition is to perform analytic continuation so that the identifiability equation satisfies for all $x \in \mathbb{R}$ *then we can examine the limits* $x \to \pm\infty$. *Extending this proof technique mixture of more components (more than 2) is generally more challenging because several components can have the same limit as* $x \to \pm\infty$. *We once more highlight the usefulness of Theorem 1 and Theorem 2 for providing the guarantee for a large class of identifiable mixture densities.*

B.3. Minimax bound for mean-dispersion negative binomial regression mixtures

Proof of Proposition 5. **Step 1.** We will prove that for any $k_0 \ge 2$ there exist $G_0 \in \mathcal{E}_{k_0}(\Theta)$ and a sequence $G_n \in \mathcal{E}_{k_0}(\Theta)$ such that:

$$W_r(G_n,G_0) \to 0, \quad \sup_x d_H(f_{G_n}(\cdot|x), f_{G_0}(\cdot|x)) = O(W_r^{2r}(G_n,G_0)). \qquad (57)$$

Then, the first claim in the proposition will be proved since

$$\sup_x d_H(f_{G_n}(\cdot|x), f_{G_0}(\cdot|x)) \geq \mathbb{E}_X d_{TV}(f_{G_n}(\cdot|X), f_{G_0}(\cdot|X)).$$

To achieve (57), intuitively, we want to pick G_0 to satisfy the pathological case described in equation (10). In particular, choose $G_0 = \sum_{j=1}^{k_0} p_j^0 \delta_{(\beta_j^0, \phi_j^0)}$ where $\beta_j^0 = (\beta_{jt}^0)_{t=0}^p \in \mathbb{R}^{p+1}$ such that $\phi_2^0 = \phi_1^0 + 1, \beta_{20}^0 = \beta_{10}^0 + \log\left(\dfrac{\phi_2^0}{\phi_1^0}\right), \beta_{2i}^0 = \beta_{1i}^0$ for all $i = 1, \ldots, p$. Let $\mu_1^0 \equiv \mu_1^0(x) = \exp((\beta_1^0)^\top x), \mu_2^0 \equiv \mu_2^0(x) = \exp((\beta_2^0)^\top x)$, we have $\dfrac{\mu_1^0}{\phi_1^0} = \dfrac{\mu_2^0}{\phi_2^0}$. A combination of chain rule with equation (10) yields:

$$\begin{aligned}\frac{\partial}{\partial \beta_{10}^0} \text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0) &= \frac{d\mu_1^0}{d\beta_{10}^0} \frac{\partial}{\partial \mu_1^0} \text{NB}(y|\mu_1^0, \phi_1^0)\Big|_{\mu_1^0 = \exp((\beta_1^0)^\top x)} \\ &= \mu_1^0 \left(\frac{\phi_1^0}{\mu_1^0} \text{NB}(y|\exp((\beta_2^0)^\top x), \phi_2^0) - \frac{\phi_1^0}{\mu_1^0} \text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0) \right) \\ &= \phi_1^0 \text{NB}(y|\exp((\beta_2^0)^\top x), \phi_2^0) - \phi_1^0 \text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0).\end{aligned} \tag{58}$$

for all $x = [1, \bar{x}] \in \mathbb{R}^{p+1}, y \in \mathbb{R}$. Now, choose a sequence $G_n = \sum_{j=1}^{k_0} p_j^n \delta_{(\beta_{1j}^n, \phi_j^0)}$ such that

$$p_1^n = p_1^0 + \frac{p_1^0 \phi_1^0}{n},\ p_2^n = p_2^0 - \frac{p_1^0 \phi_1^0}{n}, \quad p_j^n = p_j^0 \quad \forall j \geq 3$$

and $\beta_{10}^n = \beta_{10}^0 + \dfrac{1}{n}$, $\beta_{ji}^n = \beta_{ji}^0$ for all $(j,i) \neq (1,0)$; and $\phi_j^n = \phi_j^0$ for all j. It can be checked that

$$W_r^r(G_n, G_0) \asymp \frac{1}{n} + (p_1^0 - \phi_1^0/n)(\beta_{10}^n - \beta_{10}^0)^r \asymp \frac{1}{n} =: \epsilon_n.$$

Meanwhile, using Taylor's expansion up to second order with integral remainder, we have

$$\begin{aligned}f_{G_n}(y|x) - f_{G_0}(y|x) &= (p_1^n - p_1^0)\text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0) \\ &+ (p_2^n - p_2^0)\text{NB}(y|\exp((\beta_2^0)^\top x), \phi_2^0) \\ &+ p_1^0(\text{NB}(y|\exp((\beta_1^n)^\top x, \phi_1^0)) - \text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0))) \\ &= \frac{1}{n} p_1^0 \phi_1^0 \text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0) - \frac{1}{n} p_1^0 \phi_1^0 \text{NB}(y|\exp((\beta_2^0)^\top x), \phi_2^0) \\ &+ p_1^0(\beta_{10}^n - \beta_{10}^0) \frac{\partial}{\partial(\beta_{10}^0)} \text{NB}(y|\exp((\beta_1^0)^\top x), \phi_1^0) \\ &+ p_1^0 \frac{(\beta_{10}^n - \beta_{10}^0)^2}{2} \int_0^1 dt(1-t) \frac{\partial^2}{(\partial \beta_{10}^0)^2} \text{NB}(y|\mu_1^0 \exp(t\epsilon_n), \phi_1^0) \\ &= p_1^0 \frac{\epsilon_n^2}{2} \int_0^1 dt(1-t) \frac{\partial^2}{(\partial \beta_{10}^0)^2} \text{NB}(y|\mu_1^0 \exp(t\epsilon_n), \phi_1^0),\end{aligned}$$

where the zero and first-order terms are canceled out due to the equation (58). Therefore,

$$\begin{aligned}
d_H^2(f_{G_n}(\cdot|x), f_{G_0}(\cdot|x)) &= \sum_{y=0}^{\infty}(f_{G_n}^{1/2}(y|x) - f_{G_0}^{1/2}(y|x))^2 \\
&= \sum_{y=0}^{\infty}\frac{(f_{G_n}(y|x) - f_{G_0}(y|x))^2}{(f_{G_n}^{1/2}(y|x) + f_{G_0}^{1/2}(y|x))^2} \\
&\leq \sum_{y=0}^{\infty}\frac{(f_{G_n}(y|x) - f_{G_0}(y|x))^2}{(f_{G_0}^{1/2}(y|x))^2} \\
&\leq \sum_{y=0}^{\infty}\frac{(f_{G_n}(y|x) - f_{G_0}(y|x))^2}{p_1^0\mathrm{NB}(y|\mu_1^0, \phi_1^0)} \\
&\leq p_1^0\frac{\epsilon_n^4}{2}\sum_{y=0}^{\infty}\int_0^1 dt\frac{\left((1-t)\frac{\partial^2}{(\partial\beta_{10}^0)^2}\mathrm{NB}(y|\mu_1^0\exp(t\epsilon_n), \phi_1^0)\right)^2}{\mathrm{NB}(y|\mu_1^0, \phi_1^0)} \\
&= p_1^0\frac{\epsilon_n^4}{2}\int_0^1 dt(1-t)^2\sum_{y=0}^{\infty}\frac{\left(\frac{\partial^2}{(\partial\beta_{10}^0)^2}\mathrm{NB}(y|\mu_1^0\exp(t\epsilon_n), \phi_1^0)\right)^2}{\mathrm{NB}(y|\mu_1^0, \phi_1^0)} \\
&\precsim \epsilon_n^4,
\end{aligned}$$

uniformly in x, where the first two inequalities are due to the fact that NB is non-negative, the third inequality is because of Holder's inequality, the equality after that is because of Fubini theorem, and the last comparison is an application of Lemma 2 with ϵ chosen to be $t\epsilon_n$. Hence,

$$\sup_x d_H(f_{G_n}(\cdot|x), f_{G_0}(\cdot|x)) = O(W_r^{2r}(G_n, G_0)),$$

as $G_n \xrightarrow{W_r} G_0$.

Step 2 After having the limit above, the rest of this proof follows a standard proof technique for minimax lower bound (e.g., see Theorem 4.4. in [17]). Indeed, for any sufficient small $\epsilon > 0$, there exist $G_0, G_0' \in \mathcal{E}_{k_0}$ such that $W_r(G_0, G_0') = 2\epsilon$ and $\sup_x d_H(f_{G_0}, f_{G_0'}) \leq C\epsilon^{2r}$. Applying Lemma 1 in [45], we have the following inequality for any sequence of estimator $\hat{G}_n$ in $\mathcal{E}_{k_0}$:

$$\sup_{G\in\{G_0, G_0'\}}\mathbb{E}_{\mathbb{P}_G}W_r(\hat{G}_n, G) \geq \epsilon(1 - \mathbb{E}_X^n d_{TV}(f_{G_0}^n, f_{G_0'}^n)).$$

where $\mathbb{E}_X^n := \mathbb{E}_{\mathbb{P}_X^n}$ and $f_G^n := \prod_{i=1}^n f_G(y_i|X_i)$. Besides, we have

$$\begin{aligned}
d_{TV}(f_{G_0}^n, f_{G_0'}^n) &\leq d_H(f_{G_0}^n, f_{G_0'}^n) \\
&= \sqrt{1 - (1 - d_H^2(f_{G_0}, f_{G_0'}))^n} \\
&\leq \sqrt{1 - (1 - C^2\epsilon^{4r})^n}
\end{aligned}$$

Selecting $\epsilon = (1/(C^2 n))^{4r}$, we have $(1 - C^2\epsilon^{4r})^n \to e^{-1}$ so that

$$\sup_{G \in \{G_0, G_0'\}} \mathbb{E}_{\mathbb{P}_G} W_r(\hat{G}_n, G) \succcurlyeq \epsilon \asymp 1/n^{4r}.$$

Hence,

$$\inf_{\hat{G}_n \in \mathcal{E}_{k_0}} \sup_{G_0 \in \mathcal{E}_{k_0}} \mathbb{E}_{\mathbb{P}_G} W_r(\hat{G}_n, G) \succcurlyeq 1/n^{4r}.$$ □

Remark 7. *The construction of G_n in the proof above, combined with the lack of identifiability of the negative binomial kernels, allows us to cancel the zero and first-order term in the Taylor expansion of $f_{G_n} - f_{G_0}$, leading to the asymptotic bound $d_H(f_{G_n}, f_{G_0}) \preccurlyeq \epsilon_n^2 = O(W_r^{2r}(G_n, G_0))$. Therefore, we obtain the minimax rate $n^{-1/(4r)}$, which is as twice as slow compared to the usual rate $n^{-1/(2r)}$ in parametric models (under W_r).*

Lemma 2. *Let $\mu = \exp(\beta^\top x)$ for $\beta = (\beta_j)_{j=0}^p$ and $x = (1, \overline{x}) \in \mathbb{R}^{p+1}$, both range in compact subspaces of $\mathbb{R}^{p+1}$ and $\phi > 0$. There exists $\epsilon_0 > 0$ such that*

$$\sup_x \sup_{\epsilon \in [0, \epsilon_0]} \sum_{y=0}^{\infty} \frac{\left(\frac{\partial^2}{(\partial \beta_0)^2} \mathrm{NB}(y|e^\epsilon \mu, \phi)\right)^2}{\mathrm{NB}(y|\mu, \phi)} < \infty.$$

Proof. Because both x and β range in compact sets, we have $\beta^\top x$ is bounded away from $\pm\infty$. Therefore, for sufficiently small ϵ_0, we have $q_\epsilon := \frac{\exp(\beta^\top x + \epsilon)}{\exp(\beta^\top x + \epsilon) + \phi}$ is bounded away from 0 and 1 for all $\epsilon \in [0, \epsilon_0]$ and x, β. Denote $c = \inf_{x,\beta,\epsilon} q_\epsilon > 0$ and $C = \sup_{x,\beta,\epsilon} q_\epsilon < 1$. Direct calculation gives:

$$\frac{\partial}{\partial \beta_0} \mathrm{NB}(y|e^\epsilon \mu, \phi) = [q_\epsilon(\phi + y) - y] \mathrm{NB}(y|e^\epsilon \mu, \phi),$$

$$\frac{\partial^2}{(\partial \beta_0)^2} \mathrm{NB}(y|e^\epsilon \mu, \phi) = \underbrace{[q_\epsilon(1 - q_\epsilon)(\phi + y) + (q_\epsilon(\phi + y) - y)^2]}_{P_\epsilon(y)} \mathrm{NB}(y|e^\epsilon \mu, \phi)$$

where $P_\epsilon(y)$ is a polynomial of the fourth order of y, and

$$\frac{\mathrm{NB}(y|e^\epsilon \mu, \phi)}{\mathrm{NB}(y|\mu, \phi)} = \left(\frac{q_\epsilon}{q_0}\right)^y \left(\frac{1 - q_\epsilon}{1 - q_0}\right)^\phi.$$

Hence,

$$\begin{aligned} \sum_{y=0}^{\infty} \frac{\left(\frac{\partial^2}{(\partial \beta_0)^2} \mathrm{NB}(y|e^\epsilon \mu, \phi)\right)^2}{\mathrm{NB}(y|\mu, \phi)} &= \sum_y (P_\epsilon(y))^2 \left(\frac{q_\epsilon}{q_0}\right)^{2y} \left(\frac{1 - q_\epsilon}{1 - q_0}\right)^{2\phi} \mathrm{NB}(y|\mu, \phi) \\ &= \left(\frac{1 - q_\epsilon}{1 - q_0}\right)^{2\phi} \mathbb{E}_{Y \sim \mathrm{NB}(\mu, \phi)} (P_\epsilon(Y))^2 \left(\frac{q_\epsilon}{q_0}\right)^{2Y}. \end{aligned}$$

The first term is easily bounded by the comment on the range of q_ϵ in the beginning. To uniformly bound the expectation in the expression above, we will bound the expectation of $(P_\epsilon(Y))^4$ and $(q_\epsilon/q_0)^{4Y}$ separately, and then an application of Cauchy-Schwarz inequality yields the result. Because $P_\epsilon(Y)$ is a polynomial of Y with bounded coefficient, we have $\mathbb{E}(P_\epsilon(Y))^4 < \infty$ uniformly in x. For the second term, recall that the moment-generating function of $\mathbb{E}e^{Yt}$ exists and equals $(\frac{q_0}{1-(1-q_0)e^t})^\phi$ for all $t < \log(1/(1-q_0))$. Given an arbitrary $\delta > 0$, we can choose ϵ_0 sufficient small so that $4\log(q_\epsilon/q_0) < 1+\delta < \log(1/(1-q_0))$ uniformly in x. So that $\mathbb{E}(q_\epsilon/q_0)^{4Y} \le \mathbb{E}e^{(1+\delta)Y} = (\frac{q_0}{1-(1-q_0)e^{1+\delta}})^\phi$, which is also uniformly bounded in x. These claims together conclude the lemma. □

B.4. Strong identifiability for negative binomial regression mixtures

Theorem 5 and its proof indicate that the family of negative binomial distributions does not enjoy first order identifiability in general. However, we shall show that the set of parameter values where first order identitifiability fails to hold has Lebesgue measure zero. In particular, the following holds.

Proposition 10. *Given k distinct pairs $(\mu_1,\phi_1),\dots,(\mu_k,\phi_k) \in \mathbb{R}_+ \times \mathbb{R}_+$ such that there does not exist two indices $i \neq j$ satisfying*

$$\begin{cases} \dfrac{\mu_i}{\phi_i} = \dfrac{\mu_j}{\phi_j} \\ \phi_i = \phi_j + 1, \end{cases}$$

then the mixture of negative binomials $(\mathrm{NB}(\mu_i,\phi_i))_{i=1}^k$ is strongly identifiable in the first order.

Proof. We need to prove that if there exist $(a_i,b_i,c_i)_{i=1}^k$ such that

$$\sum_{i=1}^k a_i \mathrm{NB}(y|\mu_i,\phi_i) + b_i \frac{\partial}{\partial\mu}\mathrm{NB}(y|\mu_i,\phi_i) + c_i \frac{\partial}{\partial\phi}\mathrm{NB}(y|\mu_i,\phi_i) = 0, \tag{59}$$

for all $y \in \mathbb{N}$, then $a_i = b_i = c_i = 0 \; \forall\, y = 1,\dots,k$. To simplify the presentation, we will write the negative binomial in terms of probability-dispersion parameters, i.e., set $q = \mu/(\mu+\phi)$ (and $q_i = \mu_i/(\mu_i+\phi_i)$ for all i), then the negative binomial mass function becomes

$$f(y|q,\phi) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)y!} q^y (1-q)^\phi, \quad \forall y \in \mathbb{N}.$$

Under this presentation, we have

$$\frac{\partial}{\partial\mu}\mathrm{NB}(y|\mu,\phi) = \frac{\partial q}{\partial\mu} \times \frac{\partial f(y|q,\phi)}{\partial q} = \frac{\phi}{(\mu+\phi)^2}\frac{\partial f(y|q,\phi)}{\partial q},$$

and

$$\frac{\partial}{\partial\phi}\mathrm{NB}(y|\mu,\phi) = \frac{\partial q}{\partial\phi} \times \frac{\partial f(y|q,\phi)}{\partial q} + \frac{\partial f(y|q,\phi)}{\partial\phi} = -\frac{\mu}{(\mu+\phi)^2}\frac{\partial f(y|q,\phi)}{\partial q} + \frac{\partial f(y|q,\phi)}{\partial\phi},$$

therefore, we can write Eq. (59) as

$$\sum_{i=1}^{k} \alpha_i \mathrm{NB}(y|q_i,\phi_i) + \beta_i \frac{\partial}{\partial q}\mathrm{NB}(y|q_i,\phi_i) + \gamma_i \frac{\partial}{\partial \phi}\mathrm{NB}(y|q_i,\phi_i) = 0, \tag{60}$$

where $\alpha_i = a_i, \beta_i = \dfrac{\phi_i}{(\mu_i+\phi_i)^2} b_i - \dfrac{\mu_i}{(\mu_i+\phi_i)^2} c_i$, and $\gamma_i = c_i$. If we can prove that $\alpha_i = \beta_i = \gamma_i = 0$, it immediately follows that $a_i = b_i = c_i = 0$ for all $i = 1, \ldots, k$, and we get the identifiability result. We recall that pairs $(q_1,\phi_1), \ldots, (q_k,\phi_k)$ are distinct (implied from the assumption $(\mu_1,\phi_1), \ldots, (\mu_k,\phi_k)$ are distinct) and there does not exist indices $i \neq j$ such that

$$\begin{cases} q_i = q_j \\ \phi_i = \phi_j + 1. \end{cases}$$

We can simplify Eq. (60) as

$$\sum_{i=1}^{k} \left[\alpha_i + \beta_i \left(\frac{y}{q_i} - \frac{\phi_i}{1-q_i}\right) + \gamma_i \left(f_y(\phi_i) + \log(1-q_i)\right)\right] P_y(\theta_i) q_i^y (1-q_i)^{\phi_i} = 0, \tag{61}$$

where

$$P_y(\phi) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)} = \phi(\phi+1)\ldots(\phi+y-1), \quad f_y(\phi) = \sum_{i=0}^{y-1} \frac{1}{\phi+i}.$$

The function $P_y(\phi)$ has the following properties:

1. By Stirling's formula, $P_y(\phi) \asymp \dfrac{1}{\Gamma(\phi)} \sqrt{2\pi(\phi+y-1)} \left(\dfrac{\phi+y-1}{e}\right)^{\phi+y-1}$ as $y \to \infty$;
2. $P_y(\phi)$ is an increasing polynomial of ϕ, and $\dfrac{P_y(\phi)}{P_y(\phi')\log(\phi')} \to \infty$ as $y \to \infty$ if $\phi > \phi'$;
3. For all $\phi, \phi' \in \mathbb{R}_+$, $0 \le q < q' \le 1$, and polynomial $p(y)$ we have $p(y)\dfrac{P_y(\phi)}{P_y(\phi')}\left(\dfrac{q}{q'}\right)^y \to 0$ as $y \to \infty$;
4. $f_y(\phi) \asymp \log(y)$ as $y \to \infty$.

The second and third properties are consequences of the first one. Now, consider the subset of $(q_{1i})_{i=1}^{k_1}$ of $(q_i)_{i=1}^{k}$ which consists of all maximal elements, i.e. $q_{11} = q_{12} = \ldots q_{1k_1} = \max_{1 \le i \le k} q_i$. Dividing both sides of Eq. (61) by q_{11}^y and letting $y \to \infty$, from the third property above, we obtain:

$$\sum_{i=1}^{k_1} \left[\left(\alpha_{1i} - \beta_{1i}\frac{\phi_{1i}}{1-q_{11}} + \gamma_{1i}\log(1-q_{11})\right) + \gamma_{1i} f_y(\phi_{1i}) + \beta_{1i}\frac{y}{q_{11}}\right] \times P_y(\phi_{1i})(1-q_{11})^{\phi_{1i}} \to 0 \tag{62}$$

as $y \to \infty$. Without loss of generality, assume $\phi_{11} > \phi_{12} > \cdots > \phi_{1k_1} > 0$. Because $\phi_{11} \neq \phi_{12} + 1$, consider two cases: $\phi_{11} > \phi_{12} + 1$ and $\phi_{11} < \phi_{12} + 1$. For the first case, we notice that

$$\frac{P_y(\phi_{11})}{yP_y(\phi_{1i})} = \frac{P_y(\phi_{11})}{P_y(\phi_{1i}+1)} \frac{\phi_{1i}+y}{y} \to \infty,$$

as $y \to \infty$. Hence, by dividing both sides of Eq. (62) by $P_y(\phi_{11})$, we have

$$\left(\alpha_{11} - \beta_{11}\frac{\phi_{1i}}{1-q_{11}} + \gamma_{11}\log(1-q_{11})\right) + \gamma_{11} f_y(\phi_{11}) + \beta_{11}\frac{y}{q_{11}} \to 0 \quad (y \to \infty), \tag{63}$$

which implies that $\beta_{11} = 0$, followed by $\gamma_{11} = 0$ and $\alpha_{11} = 0$. For the second case where $\phi_{11} < \phi_{12} + 1$. We have $P_y(\phi_{12})/P_y(\phi_{11}) \to 0$ and $P_y(\phi_{11})/(yP_y(\phi_{12})) \to 0$. By dividing both sides of Eq. (62) by $yP_y(\phi_{11})$ and let $y \to \infty$, we have $\beta_{11} = 0$. Dividing both sides of Eq. (62) by $yP_y(\phi_{11})$, we have $\beta_{12} = 0$. Finally, dividing both sides of Eq. (62) by $P_y(\phi_{11})$ and letting $y \to \infty$, we also obtain that the limit (63) holds. Hence, in all cases we obtain $\alpha_{11} = \beta_{11} = \gamma_{11} = 0$. Continuing this argument, we have $\alpha_{1i} = \beta_{1i} = \gamma_{1i} = 0$ for all $i = 1, \ldots, k_1$, then $\alpha_i = \beta_i = \gamma_i = 0$ for all $i = 1, \ldots, k$. □

Proposition 11. *Given k distinct pairs $(\mu_1, \phi_1), \ldots, (\mu_k, \phi_k) \in \mathbb{R} \times \mathbb{R}_+$ such that there does not exist two indices $i \neq j$ satisfying*

$$\begin{cases} \dfrac{\mu_i}{\phi_i} = \dfrac{\mu_j}{\phi_j} \\ |\phi_i - \phi_j| \in \{1, 2\}, \end{cases}$$

then the mixture of negative binomials $(\mathrm{NB}(\mu_i, \phi_i))_{i=1}^k$ is strongly identifiable in the second order.

Proof. Using the same transformation $q_i = \mu_i/(\mu_i + \phi_i)$ as in the proof of Proposition 10, we only need to prove that if there exist $(a_i, b_i, c_i, d_i, e_i, f_i)_{i=1}^k$ such that for all $y \in \mathbb{N}$:

$$\sum_{i=1}^{k} a_i \mathrm{NB}(y|q_i, \phi_i) + b_i \frac{\partial}{\partial q}\mathrm{NB}(y|q_i, \phi_i) + c_i \frac{\partial}{\partial \phi}\mathrm{NB}(y|q_i, \phi_i)$$
$$d_i \frac{\partial^2}{\partial q^2}\mathrm{NB}(y|q_i, \phi_i) + e_i \frac{\partial^2}{\partial q \partial \phi}\mathrm{NB}(y|q_i, \phi_i) + f_i \frac{\partial^2}{\partial \phi^2}\mathrm{NB}(y|q_i, \phi_i) = 0, \tag{64}$$

then $a_i = b_i = c_i = d_i = e_i = f_i = 0$ for all $i = 1, \ldots, k$. We recall that pairs $(q_1, \phi_1), \ldots, (q_k, \phi_k)$ are distinct (implied from the assumption $(\mu_1, \phi_1), \ldots, (\mu_k, \phi_k)$ are distinct) and there does not exist indices $i \neq j$ such that

$$\begin{cases} q_i = q_j \\ |\phi_i - \phi_j| \in \{1, 2\}. \end{cases}$$

Taking all the derivatives and rewrite Eq. (64) as

$$\sum_{i=1}^{k}\Bigg[\frac{d_i}{q_i^2}y^2+\left(2d_i\frac{\phi_i}{q_i(1-q_i)}-\frac{d_i}{q_i^2}+\frac{b_i}{q_i}\right)y+\frac{e_i}{q_i}f_y(\phi_i)y+f_if_y^2(\phi_i)$$
$$+\left(2f_i\log(1-q_i)+c_i+e_i\frac{\phi_i}{1-q_i}\right)f_y(\theta_i)+C_i(y)\Bigg]P_y(\phi_i)q_i^y(1-q_i)^{\phi_i}=0, \tag{65}$$

where

$$P_y(\phi)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)}=\phi(\phi+1)\dots(\phi+y-1),\quad f_y(\phi)=\sum_{i=0}^{y-1}\frac{1}{\phi+i},$$

and

$$C_i(y)=a_i+b_i\frac{\phi_i}{1-q_i}+c_i\log(1-q_i)+d_i\frac{\phi_i^2+\phi_i}{(1-q_i)^2}$$
$$+e_i\left(\frac{\phi_i}{1-q_i}\log(q_i)+\frac{1}{1-q_i}\right)+f_i(\log(1-q_i)^2+f_y'(\phi_i)).$$

Recall some facts as follows:

1. By Stirling's formula, $P_y(\phi)\asymp\frac{1}{\Gamma(\phi)}\sqrt{2\pi(\phi+y-1)}\left(\frac{\phi+y-1}{e}\right)^{\phi+y-1}$ as $y\to\infty$;
2. $P_y(\phi)$ is an increasing polynomial of ϕ, and $\frac{P_y(\phi)}{P_y(\phi')\log(\phi')}\to\infty$ as $y\to\infty$ if $\phi>\phi'$;
3. For all $\phi,\phi'\in\mathbb{R}_+$, $0\le q<q'\le 1$, and polynomial $p(y)$ we have $p(y)\frac{P_y(\phi)}{P_y(\phi')}\left(\frac{q}{q'}\right)^y\to 0$ as $y\to\infty$;
4. $f_y(\phi)\asymp\log(y)$ as $y\to\infty$;
5. $f_y'(\phi_i)=-\sum_{j=0}^{y}\frac{1}{(\phi_i+j)^2}\in[-\pi^2/6-1/\phi_i^2,0]$ for all $\phi_i>0,y\in\mathbb{N}$.

Now, consider the subset of $(q_{1i})_{i=1}^{k_1}$ of $(q_i)_{i=1}^{k}$ which consists of all maximal elements, i.e. $q_{11}=q_{12}=\dots q_{1k_1}=\max_{1\le i\le k}q_i$, then by dividing both sides of Eq. (65) by q_{11}^y and let $y\to\infty$, from the third property above, we obtain:

$$\sum_{i=1}^{k_1}\Bigg[\frac{d_{1i}}{q_1^2}y^2+\left(2d_{1i}\frac{\phi_{1i}}{q_1(1-q_1)}-\frac{d_{1i}}{q_1^2}+\frac{b_{1i}}{q_1}\right)y+\frac{e_{1i}}{q_1}f_y(\phi_{1i})y+f_{1i}f_y^2(\phi_{1i})$$
$$+\left(2f_{1i}\log(1-q_1)+c_{1i}+e_{1i}\frac{\phi_{1i}}{1-q_1}\right)f_y(\phi_{1i})+C_i(y)\Bigg]P_y(\phi_{1i})(1-q_1)^{\phi_{1i}}=0, \tag{66}$$

as $y \to \infty$. Without loss of generality, assume $\phi_{11} > \phi_{12} > \cdots > \phi_{1k_1} > 0$. Because $|\phi_{11} - \phi_{12}| \neq 1, 2$, there are three cases:

$$\begin{cases} \phi_{11} > \phi_{12} + 2, \\ \phi_{12} + 2 > \phi_{11} > \phi_{12} + 1, \\ \phi_{12} + 1 > \phi_{11} > \phi_{12}. \end{cases}$$

For the first case, note that

$$\frac{P_y(\phi_{11})}{y^2 P_y(\phi_{1i})} = \frac{P_y(\phi_{11})}{P_y(\phi_{1i}+2)} \frac{(\phi_{1i}+y+1)(\phi_{1i}+y)}{y^2} \to \infty,$$

as $y \to \infty$. Hence, by dividing both sides of Eq. (66) by $P_y(\phi_{11})$, we have

$$\frac{d_{11}}{q_1^2} y^2 + \left(2d_{11}\frac{\phi_{1i}}{q_1(1-q_1)} - \frac{d_{11}}{q_1^2} + \frac{b_{11}}{q_1}\right) y + \frac{e_{11}}{q_1} f_y(\phi_{11}) y + f_{11} f_y^2(\phi_{11}) \\ + \left(2f_{11}\log(1-q_1) + c_{11} + e_{11}\frac{\phi_{11}}{1-q_1}\right) f_y(\phi_{11}) + C_{11}(y) \to 0 \quad (y \to \infty), \quad (67)$$

which, by considering the order of y, implies that $d_{11} = b_{11} = e_{11} = f_{11} = c_{11} = a_{11} = 0$, respectively. For the second case, where $\phi_{12} + 1\phi_{11} < \phi_{12} + 2$, we have $P_y(\phi_{11})/(y^2 P_y(\phi_{12})) \to 0$ and $P_y(\phi_{12})/(y P_y(\phi_{11})) \to 0$. By dividing both sides of Eq. (66) by $y^2 P_y(\phi_{11})$ and let $y \to \infty$, we have $b_{11} = 0$. Then dividing two sides by $y f_y(\phi_{11}) P_y(\phi_{11})$ and $y P_y(\phi_{11})$, we have $d_{11} = e_{11} = 0$. Continuing in the same way with $y^2 P_y(\phi_{11})$, we have $b_{12} = 0$. Now dividing both sides of Eq. (66) by $P_y(\phi_{11})$ and letting $y \to \infty$, we obtain that the limit (63) holds, which once again entails that $e_{11} = f_{11} = c_{11} = a_{11} = 0$. In the final case, dividing both sides of Eq. (66) by $y^2 P_y(\phi_{11}), y^2 P_y(\phi_{12}), y f_y(\phi_{11}) P_y(\phi_{11}), y P_y(\phi_{11}), y f_y(\phi_{12}) P_y(\phi_{12})$, and $y P_y(\phi_{12})$, respectively, we arrive at the same conclusion. Hence, in all cases, we have $d_{11} = b_{11} = e_{11} = f_{11} = c_{11} = a_{11} = 0$. Applying repreatedly this argument, we have $a_{1i} = b_{1i} = c_{1i} = d_{1i} = e_{1i} = f_{1i} = 0$ for all $i = 1, \ldots, k_1$, then $a_i = b_i = c_i = d_i = e_i = f_{1i} = 0$ for all $i = 1, \ldots, k$. □

Implication in negative binomial regression mixtures From the argument above, we can see that the family of binomial regression mixture model is strongly identifiable in the first order if we adjust the assumption (A4.) as follows:

(A4'.) For every set of $k+1$ distinct elements $(\theta_{11}, \theta_{21}), ..., (\theta_{1(k+1)}, \theta_{2(k+1)}) \in \Theta_1 \times \Theta_2$, there exists a subset $A \subset \mathcal{X}$, $\mathbb{P}_X(A) > 0$ such that

$$(h_1(x, \theta_{11}), h_2(x, \theta_{21})), ..., \left(h_1(x, \theta_{1(k+1)}), h_2(x, \theta_{2(k+1)})\right)$$

are distinct and $|h_2(x, \theta_{2i}) - h_2(x, \theta_{2j})| \neq 1\, (\forall\, i, j)$ for every $x \in A$.

Similarly, the second-order strong identifiability condition is satisfied if we adjust assumption (A4.) as:

(A4".) For every set of $k+1$ distinct elements $(\theta_{11}, \theta_{21}), ..., (\theta_{1(k+1)}, \theta_{2(k+1)}) \in \Theta_1 \times \Theta_2$, there exists a subset $A \subset \mathcal{X}$, $\mathbb{P}_X(A) > 0$ such that

$$(h_1(x, \theta_{11}), h_2(x, \theta_{21})), ..., \left(h_1(x, \theta_{1(k+1)}), h_2(x, \theta_{2(k+1)})\right)$$

are distinct and $|h_2(x, \theta_{2i}) - h_2(x, \theta_{2j})| \notin \{1, 2\}$ $(\forall\, i, j)$ for every $x \in A$.

Appendix C: Convergence rates for conditional densities via MLE

We present in this section a proof of Theorem 3, which provides general convergence rates of conditional densities estimation. The proof technique follows a general framework of M-estimation theory [38, 10], with a suitable adaptation for handling conditional density functions. Assume that we have n i.i.d. observations $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \overset{i.i.d.}{\sim} \mathbb{P}_X$ and $y_i|x_i \sim f_0(y|x), i = 1, \ldots, n$, for $f_0 \in \mathcal{F}$ being some family of conditional densities of y given x (commonly dominated by ν). Assume that there exists

$$\widehat{f}_n \in \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(y_i|x_i),$$

Set

$$\overline{\mathcal{F}} = \{((f + f_0)/2) : f \in \mathcal{F}\}, \quad \overline{\mathcal{F}}^{1/2} = \{\overline{f}^{1/2} : \overline{f} \in \overline{\mathcal{F}}\},$$

and denote the (expected) Hellinger ball centered around f_0 by

$$\overline{\mathcal{F}}^{1/2}(\delta) = \{\overline{f}^{1/2} \in \overline{\mathcal{F}}^{1/2}(\Theta) : \overline{d}_H(\overline{f}, f_0) \le \delta\}.$$

The size of this set is characterized by the bracket entropy integral

$$\mathcal{J}(\delta) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(u, \overline{\mathcal{F}}^{1/2}(\delta), L_2(\mathbb{P}_X \times \nu)) du \vee \delta, \tag{68}$$

where $H_B(u, \mathcal{F}, L_2(\mathbb{P}_X \times \nu)) = \log N_B(u, \mathcal{F}, L_2(\mathbb{P}_X \times \nu))$, and $N_B(u, \mathcal{F}, L_2(\mathbb{P}_X \times \nu))$ is the minimal number of pairs $(f_j^L, f_j^U)_j$ such that for every $f \in \mathcal{F}$, there exists j to have $f_j^L \le f \le f_j^U$ and $\left\|f_j^U - f_j^L\right\|_{L^2} \le \epsilon$. Define $d\mathbb{P}_0(x, y) = d\mathbb{P}_X(x) \times f_0(y|x) d\nu(y)$ be the true joint distribution of (x, y). Denote by $\mathbb{P}_n$ the empirical distribution of (x, y), i.e., $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ and $g_f = \frac{1}{2} \log \frac{f + f_0}{2 f_0} 1(f_0 > 0)$. For a probability distribution $\mathbb{P}$ and a function g, sometimes we write $\mathbb{P}g$ for $\int g d\mathbb{P}$. We start with a basic inequality that links the quality of the conditional density estimate to the associated empirical process:

Lemma 3. *With the notations defined as above, we have*

$$\frac{1}{2} \overline{d}_H^2 \left(\frac{\widehat{f}_n + f_0}{2}, f_0 \right) \le (\mathbb{P}_n - \mathbb{P}_0) g_{\widehat{f}_n}. \tag{69}$$

Proof. Due to the concavity of logarithm,

$$\frac{1}{2}\log\frac{f+f_0}{2f_0}1_{(f_0>0)} \geq \frac{1}{2}\log\frac{f}{f_0}1_{(f_0>0)}.$$

Combining the above with the fact that $\widehat{f}_n$ is the maximum conditional likelihood estimate to obtain

$$\begin{aligned} 0 &\leq \int_{f_0>0}\frac{1}{4}\log\frac{\widehat{f}_n}{f_0}\mathbb{P}_n \leq \int_{f_0>0}\frac{1}{2}\log\frac{\widehat{f}_n+f_0}{2f_0}d\mathbb{P}_n \\ &= \int_{f_0>0} g_{\widehat{f}_n} d(\mathbb{P}_n-\mathbb{P}_0) + \int_{f_0>0}\frac{1}{2}\log\frac{\widehat{f}_n+f_0}{2f_0}d\mathbb{P}_0. \end{aligned}$$

Equivalently,

$$\int_{f_0>0}\frac{1}{2}\log\frac{2f_0}{\widehat{f}_n+f_0}d\mathbb{P}_0 \leq \int_{f_0>0} g_{\widehat{f}_n} d(\mathbb{P}_n-\mathbb{P}_0).$$

By the inequality $d_H^2(\frac{1}{2}(\widehat{f}_n(\cdot|x)+f_0(\cdot|x))), f_0(\cdot|x)) \leq K(f_0(\cdot|x)\|\frac{1}{2}(\widehat{f}_n(\cdot|x)+f_0(\cdot|x))$ for almost all x, we can take the expectation with respect to $\mathbb{P}_X$ to arrive at

$$\frac{1}{2}\bar{d}_H^2\left(\frac{\widehat{f}_n+f_0}{2}, f_0\right) \leq \int_{f_0>0}\frac{1}{2}\log\frac{2f_0}{\widehat{f}_n+f_0}d\mathbb{P}_0 \leq \int g_{\widehat{f}_n} d(\mathbb{P}_n-\mathbb{P}_0). \qquad \square$$

For each $f\in\mathcal{F}$, define the squared “Bernstein norm”:

$$\rho^2(f) := 2\mathbb{P}_0(e^{|f(X)|}-|f(X)|-1). \tag{70}$$

Let $H_{B1}(\epsilon,\mathcal{F},\mathbb{P}_X\times\nu)$ be the bracketing number with respect to Bernstein norm of $\mathcal{F}$ (cf. Definition 3.5.20. in [10]). We shall make use of the concentration behavior of empirical processes associated with the class $\mathcal{F}$ by the following theorem, which is essentially Theorem 3.5.21. in [10] adapted to our setting.

Theorem 10. *Let $\mathcal{F}$ be a class of measurable functions such that $\rho(f)\leq R$ for all $f\in\mathcal{F}$. Given $C_1<\infty$, for all C sufficiently large and C_0 satisfying*

$$C_0^2 \geq C^2(C+1), \tag{71}$$

and for $n\in\mathbb{N}$ and $t>0$ satisfying

$$C_0\left(R\vee\int_{t/(2^6\sqrt{n})}^{R}\sqrt{H_{B1}(\epsilon,\mathcal{F},\mathbb{P}_X\times\nu)}\right) \leq t \leq \sqrt{n}((8R)\wedge(C_1R^2/K)), \tag{72}$$

we have

$$\mathbb{P}_0(\sqrt{n}\sup_{f\in\mathcal{F}}|(\mathbb{P}_n-\mathbb{P}_0)f|\geq t) \leq C\exp\left(-\frac{t^2}{C^2(C_1+1)R^2}\right). \tag{73}$$

Now we are ready to prove Theorem 3.

Proof of Theorem 3. We have

$$
\begin{aligned}
&\mathbb{P}_0(\overline{d}_H(\widehat{f}_n, f_0) \geq \delta) \\
&\leq \mathbb{P}\left(\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(g_{\widehat{f}_n}) - \sqrt{n}\overline{d}_H^2\left(\frac{\widehat{f}_n + f_0}{2}, f_0\right) \geq 0, \overline{d}_H^2\left(\frac{\widehat{f}_n + f_0}{2}, f_0\right) \geq \delta^2/C\right) \\
&\leq \mathbb{P}_0\left(\sup_{f: \overline{d}_H^2(\overline{f}, f_0) \geq \delta^2/C} [\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(g_f) - \sqrt{n}\overline{d}_H^2(\overline{f}, f_0)] \geq 0\right) \\
&\leq \sum_{s=0}^{S} \mathbb{P}_0\left(\sup_{f: 2^s\delta^2/C \leq \overline{d}_H^2(\overline{f}, f_0) \leq 2^{s+1}\delta^2/C} |\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(g_f)| \geq \sqrt{n}2^s\delta^2/C\right) \\
&\leq \sum_{s=0}^{S} \mathbb{P}_0\left(\sup_{f: \overline{d}_H^2(\overline{f}, f_0) \leq 2^{s+1}\delta^2/C} |\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(g_f)| \geq \sqrt{n}2^s\delta^2/C\right),
\end{aligned}
$$

where S is a smallest number such that $2^S\delta^2/C > 1$, as $\overline{d}_H(\overline{f}, f_0) \leq 1$. Now we will bound the tail probability of the empirical process

$$
\mathbb{P}_0\left(\sup_{f: \overline{d}_H^2(\overline{f}, f_0) \leq 2^{s+1}\delta^2/C} |\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(g_f)| \geq \sqrt{n}2^s\delta^2/C\right)
$$

by using Theorem 10. Indeed, since $p(x) = \dfrac{(e^{|x|} - |x| - 1)}{(e^x - 1)^2}$ is a decreasing function, and $g_f \geq -(\log 2)/2$ for all f,

$$
\exp(|g_f|) - |g_f| - 1 \leq p(-(\log 2)/2)(\exp(g_f) - 1)^2 \leq \left(\sqrt{\frac{f_0 + f}{2f_0}} - 1\right)^2.
$$

Taking expectation with respect to $\mathbb{P}_0$ both sides to obtain

$$
\rho^2(g_f) \leq 2\overline{d}_H^2(\overline{f}, f_0) \leq 2^{s+2}\delta/C.
$$

Applying Theorem 10 with $R = 2^{s/2+1}\delta/C^{1/2}, t = \sqrt{n}2^s\delta^2/C$, we obtain

$$
\mathbb{P}_0\left(\sup_{f: \overline{d}_H^2(\overline{f}, f_0) \leq 2^{s+1}\delta^2/C} |\sqrt{n}(\mathbb{P}_n - \mathbb{P}_0)(g_f)| \geq \sqrt{n}2^s\delta^2/C\right) \leq C' \exp\left(-\frac{2^{2s}n\delta^2}{C'}\right).
$$

Hence,

$$
\mathbb{P}_0(\overline{d}_H(\widehat{f}_n, f_0) \geq \delta) \leq \sum_{s=0}^{S} C' \exp\left(-\frac{2^{2s}n\delta^2}{C'}\right) \leq c\exp(-n\delta^2/c). \qquad \square
$$

Appendix D: Convergence rates of conditional densities via Bayesian estimation

We present in this section a general theorem for the Bayesian posterior contraction behavior of conditional density functions that arise in the regression problem. The proof technique follows a general approach of Bayesian estimation theory [8], with a suitable adaptation for handling conditional density functions. Let us recall the setup. Given i.i.d. pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ in $\mathcal{X} \times \mathcal{Y}$ from the true generating model

$$y_i|x_i \sim f_0(y|x),$$
$$x_i \sim \mathbb{P}_X.$$

Here, $\mathbb{P}_X$ is some unknown distribution of covariate X, while f_0 is assumed to belong to a family of conditional probability functions $\{f(y|x) : f \in \mathcal{F}\}$, which are absolutely continuous with respect to a common dominating σ-finite measure ν. To make inference of f_0 from the data using the Bayesian approach, we assume

$$y_i|x_i, f \sim f(y|x),$$
$$f \sim \Pi,$$

for some prior distribution Π on the space of conditional probability functions $\mathcal{F}$. The posterior distribution of f is given by, for any measurable subset $B \subset \mathcal{F}$,

$$\Pi(f \in B|(x_i, y_i)_{i=1}^n) = \frac{\int_B \prod_{i=1}^n f(y_i|x_i) d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n f(y_i|x_i) d\Pi(f)}.$$

As in the MLE analysis, the posterior contraction behavior of f will be assessed by the expected (squared) Hellinger distance $\overline{d}_H(f, f_0) = (\mathbb{E}_X d_H^2(f(y|X), f_0(y|X)))^{1/2}$. That is, we will find a sequence $(\epsilon_n) \to 0$ such that

$$\Pi(\overline{d}_H(f(y|X), f_0(y|X)) \geq M_n \epsilon_n | x_1, \ldots, x_n, y_1, \ldots, y_n) \to 0, \tag{74}$$

in $\otimes_{i=1}^n \mathbb{P}_0$-probability, as $n \to \infty$. Here, M_n is an arbitrary diverging sequence.

Recall the following basic fact (cf. [8] Lemma D.2., or [10] Chapter 7).

Lemma 4. *Given arbitrary probability density p and q, there exist probability densities $\overline{p}$ and $\overline{q}$ such that for any probability density r (all are commonly dominated by ν)*

$$\mathbb{E}_{y\sim r}\sqrt{\frac{\overline{q}(y)}{\overline{p}(y)}} \leq 1 - \frac{1}{6} d_H^2(p, q) + d_H^2(p, r), \quad \mathbb{E}_{y\sim r}\sqrt{\frac{\overline{p}(y)}{\overline{q}(y)}} \leq 1 - \frac{1}{6} d_H^2(p, q) + d_H^2(q, r)$$

From now, for every conditional density $f(y|x)$, denote by $\mathbb{P}_f$ the joint distribution of x, y, i.e., $d\mathbb{P}_f(x, y) = f(y|x) d\nu(y) \times d\mathbb{P}_X(x)$. Using Lemma 4 one arrives at the following result on the existence of tests for conditional density functions:

Lemma 5. *For any two conditional density functions f_0, f_1 such that $\overline{d}_H^2(f_0, f_1) = \epsilon^2$, there exists a test Ψ_n based on $x_1, \ldots, x_n, y_1, \ldots, y_n$ such that*

$$\mathbb{P}_{f_0}^n \Psi_n \le e^{-n\epsilon^2/6}, \qquad \sup_{f \in B(f_1, \epsilon/4)} \mathbb{P}_f^n (1 - \Psi_n) \le e^{-n\epsilon^2/12}, \tag{75}$$

where $B(f, \epsilon) := \{g \in \mathcal{F} : \overline{d}_H(f, g) \le \epsilon\}$ for all $\epsilon \ge 0$ and $f \in \mathcal{F}$.

Proof. For any $x \in \mathcal{X}$, consider probability density functions $f_0(\cdot|x)$ and $f_1(\cdot|x)$. By Lemma 4, there exist density functions $\overline{f}_0(\cdot; x)$ and $\overline{f}_1(\cdot; x)$ such that for all probability density functions $f(\cdot|x)$

$$\mathbb{E}_{y \sim f_0(\cdot|x)} \sqrt{\frac{\overline{f}_1(y|x)}{\overline{f}_0(y|x)}} \le 1 - \frac{1}{6} d_H^2(f_0(\cdot|x), f_1(\cdot|x)),$$

and

$$\mathbb{E}_{y \sim f(\cdot|x)} \sqrt{\frac{\overline{f}_0(y|x)}{\overline{f}_1(y|x)}} \le 1 - \frac{1}{6} d_H^2(f_0(\cdot|x), f_1(\cdot|x)) + d_H^2(f(\cdot|x), f_1(\cdot|x)).$$

Define test function $\Psi_n(x_1, \ldots, x_n, y_1, \ldots, y_n) = 1\left(\prod_{i=1}^n \frac{\overline{f}_1(y_i; x_i)}{\overline{f}_0(y_i; x_i)} \ge 1\right)$. To verify (75), by Markov's inequality,

$$\begin{aligned}
\mathbb{P}_{f_0}^n \Psi_n &\le \mathbb{P}_{f_0}^n \prod_{i=1}^n \sqrt{\frac{\overline{f}_1(y_i|x_i)}{\overline{f}_0(y_i|x_i)}} \\
&= \prod_{i=1}^n \mathbb{E}_X \mathbb{E}_{Y \sim f_0(\cdot|X)} \sqrt{\frac{\overline{f}_1(Y|X)}{\overline{f}_0(Y|X)}} \\
&\le \left(1 - \frac{1}{6} \mathbb{E}_X d_H^2(f_0(\cdot|X), f_1(\cdot|X))\right)^n \\
&= (1 - \epsilon^2/6)^n \\
&\le e^{-n\epsilon^2/6},
\end{aligned}$$

and for every $f \in B(f_1, \epsilon/4)$,

$$\begin{aligned}
\mathbb{P}_f^n \Psi_n &\le \mathbb{P}_f^n \prod_{i=1}^n \sqrt{\frac{\overline{f}_0(y_i|x_i)}{\overline{f}_1(y_i|x_i)}} \\
&= \prod_{i=1}^n \mathbb{E}_X \mathbb{E}_{Y \sim f_0(\cdot|X)} \sqrt{\frac{\overline{f}_0(Y|X)}{\overline{f}_1(Y|X)}} \\
&\le \left(1 - \frac{1}{6} \mathbb{E}_X d_H^2(f_0(\cdot|X), f_1(\cdot|X)) + \mathbb{E}_X d_H^2(f(\cdot|X), f_1(\cdot|X))\right)^n \\
&\le \left(1 - \frac{1}{12} \epsilon^2\right)^n \\
&\le e^{-n\epsilon^2/12}.
\end{aligned}$$

□

The above lemma shows the existence of tests to distinguish between f_0 and a small ball around any $f_1 \neq f_0$. Next, we establish the existence of tests for f_0 against all $f \in \mathcal{F}$ being a bounded distance away from f_0. Recall that $N(\mathcal{F}, d, \epsilon)$ denotes the covering number of $\mathcal{F}$ by d-balls with radius ϵ.

Lemma 6. *For every natural number M large enough, there exists a test Ψ_n such that*

$$\mathbb{P}^n_{f_0}\Psi_n \leq N(\mathcal{F}, d_H, \epsilon)\frac{e^{-nM^2\epsilon^2/12}}{1-e^{-nM^2\epsilon^2/12}}, \qquad \sup_{f\in\mathcal{F}:\overline{d}_H(f,f_0)>M\epsilon} \mathbb{P}^n_f(1-\Psi_n) \leq e^{-nM^2\epsilon^2/12}. \tag{76}$$

Proof. For every $j \in \mathbb{N}$ such that $j \geq M$, consider a minimal covering of the set $\mathcal{F}_j := \{f \in \mathcal{F} : j\epsilon < \overline{d}_H(f, f_0) < 2j\epsilon\}$ by balls $(F_{j,l})_l$ of radius $j\epsilon/4$. Because $\mathcal{F}_j \subset \mathcal{F}$ and $j\epsilon/4 \geq \epsilon$, the number of such balls is no more than $N(\mathcal{F}, d_H, \epsilon)$. Moreover, by Lemma 5 for each $(F_{j,l})_l$ there exists a test $\phi_{j,l}$ satisfying

$$\mathbb{P}^n_{f_0}\phi_{j,l} \leq e^{-nj^2\epsilon^2/6}, \qquad \sup_{f\in F_{j,l}} \mathbb{P}^n_f(1-\phi_{j,l}) \leq e^{-nj^2\epsilon^2/12}. \tag{77}$$

Let $\Psi_n := \max_{j\geq M;l} \phi_{j,l}$. Then

$$\mathbb{P}^n_{f_0}\Psi_n \leq N(\mathcal{F}, d_H, \epsilon)\sum_{j\geq M} e^{-nj^2\epsilon^2/6} \leq N(\mathcal{F}, d_H, \epsilon)\frac{e^{-nM^2\epsilon^2/12}}{1-e^{-nM^2\epsilon^2/12}},$$

and

$$\sup_{f\in\mathcal{F}:\overline{d}_H(f,f_0)>M\epsilon} \mathbb{P}^n_f(1-\Psi_n) \leq \sup_{j,l}\sup_{f\in F_{j,l}} \mathbb{P}^n_f(1-\phi_{j,l}) \leq e^{-nM^2\epsilon^2/12}. \qquad \square$$

Now for every $\epsilon > 0$, define a ball with radius ϵ around f_0 as

$$\begin{aligned} B_2(f_0, \epsilon) := \{&f \in \mathcal{F} : \mathbb{P}_{f_0}\log(f_0(Y|X)/f(Y|X)) \\ &\leq \epsilon^2, \mathbb{P}_{f_0}(\log(f_0(Y|X)/f(Y|X)))^2 \leq \epsilon^2\}. \end{aligned} \tag{78}$$

The following theorem establishes the posterior contraction convergence rate for conditional density functions under the expected squared Hellinger distance.

Theorem 11. *Assume that there exist sequences $\overline{\epsilon}_n, \epsilon_n$, such that $\overline{\epsilon}_n \leq \epsilon_n$, and $\sqrt{n}\overline{\epsilon}_n \to \infty$, a sequence of measurable set $\mathcal{F}_n \subset \mathcal{F}$, and a constant C such that*

(i) $\Pi(B_2(f_0, \overline{\epsilon}_n)) \geq e^{-Cn\overline{\epsilon}_n^2}$;
(ii) $\log N(\epsilon_n, \mathcal{F}_n, \overline{d}_H) \leq n\epsilon_n^2$;
(iii) $\Pi(\mathcal{F}_n^c) \leq e^{-(C+4)n\overline{\epsilon}_n^2}$.

Then, for every sequence $M_n \to \infty$, there holds

$$\Pi(f : \overline{d}_H(f, f_0) > M_n\epsilon_n | x_1, \ldots, x_n, y_1, \ldots, y_n) \to 0 \tag{79}$$

in $\mathbb{P}^n_{f_0}$-probability, as $n \to \infty$.

Proof. Write $x^{[n]}, y^{[n]} = \{x_1, \dots, x_n, y_1, \dots, y_n\}$ for short. By Lemma 6, there exists a test Ψ_n such that

$$\mathbb{P}_0^n \Psi_n \le e^{n\epsilon_n^2} \frac{e^{-nM_n^2\epsilon_n^2/12}}{1 - e^{-nM_n^2\epsilon_n^2/12}}, \qquad \sup_{f \in \mathcal{F}_n : \overline{d}_H(f,f_0) \ge M_n\epsilon_n} \mathbb{P}_f^n (1 - \Psi_n) \le e^{-nM_n^2\epsilon_n^2/12}.$$

As $M_n \to \infty$ and $n\epsilon_n^2 \to \infty$, both probabilities above go to 0. By Bayes' rule,

$$\Pi(f : \overline{d}_H(f, f_0) > M_n\epsilon_n | x^{[n]}, y^{[n]}) = \frac{\int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \prod_{i=1}^n (f/f_0)(y_i|x_i) d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n (f/f_0)(y_i|x_i) d\Pi(f)}. \tag{80}$$

Let $B_n := B_2(f_0, \epsilon_n)$ and $A_n = \{(x^{[n]}, y^{[n]}) : \int_{B_n} \prod_{i=1}^n (f/f_0)(y_i|x_i) d\Pi(f) \ge e^{-(C+2)n\overline{\epsilon}_n^2}\}$. Because a probability is always less than or equal to 1, we have

$$\begin{aligned}
&\Pi(f : \overline{d}_H(f, f_0) > M_n\epsilon_n | x^{[n]}, y^{[n]}) \\
&\le 1_{(A_n)^c} + 1_{A_n} \frac{\int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \prod_{i=1}^n (f/f_0) d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n (f/f_0)(y_i|x_i) d\Pi(f)} \\
&\le \Psi_n + 1_{(A_n)^c} + 1_{A_n} \frac{\int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \prod_{i=1}^n (f/f_0) d\Pi(f)(1 - \Psi_n)}{\int_{\mathcal{F}} \prod_{i=1}^n (f/f_0)(y_i|x_i) d\Pi(f)} \\
&\le \Psi_n + 1_{(A_n)^c} + e^{(C+2)n\overline{\epsilon}_n^2} \int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \prod_{i=1}^n (f/f_0) d\Pi(f)(1 - \Psi_n).
\end{aligned}$$

By the construction of the test Ψ_n, we have $\mathbb{P}_0^n \Psi_n \to 0$. Besides, assumption (i) and Lemma 7 imply that

$$\begin{aligned}
\mathbb{P}_0^n(A_n^c) &= \mathbb{P}_0^n \left(\int_{B_n} \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i) d\Pi(f) \le e^{-(C+2)n\overline{\epsilon}_n^2} \right) \\
&\le \mathbb{P}_0^n \left(\int_{B_n} \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i) d\Pi(f) \le e^{-2n\overline{\epsilon}_n^2} \Pi(B_n) \right) \\
&\le \frac{1}{n\overline{\epsilon}_n^2} \to 0.
\end{aligned}$$

For the last term, by Fubini's theorem,

$$\begin{aligned}
&\mathbb{P}_0^n e^{(C+2)n\overline{\epsilon}_n^2} \int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \prod_{i=1}^n (f/f_0) d\Pi(f)(1 - \Psi_n) \\
&= e^{(C+2)n\overline{\epsilon}_n^2} \int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \mathbb{P}_0^n \prod_{i=1}^n (f/f_0)(1 - \Psi_n) d\Pi(f) \\
&= e^{(C+2)n\overline{\epsilon}_n^2} \int_{\overline{d}_H(f,f_0) > M_n\epsilon_n} \mathbb{P}_f^n (1 - \Psi_n) d\Pi(f) \\
&\le e^{(C+2)n\overline{\epsilon}_n^2} \Bigg(\int_{f \in \mathcal{F}_n : \overline{d}_H(f,f_0) > M_n\epsilon_n} \mathbb{P}_f^n (1 - \Psi_n) d\Pi(f)
\end{aligned}$$

$$+ \int_{f\in\mathcal{F}_n^c} \mathbb{P}_f^n(1-\Psi_n)d\Pi(f)\Big)$$
$$\leq e^{(C+2)n\bar{\epsilon}_n^2}(e^{-nM_n^2\epsilon_n^2/12} + \Pi(\mathcal{F}_n^c)),$$

which tends to 0, thanks to the construction of the test and assumption (iii). □

The above proof made use of the following lemma, which is taken from [8] (and adapted for conditional densities). We include its proof for completeness.

Lemma 7. *For every $\epsilon > 0$, let $B = B_2(f_0, \epsilon)$. For all $c > 0$, we have*

$$\mathbb{P}_0^n\left(\int_B \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)d\Pi(f) \leq \exp(-(c+1)n\epsilon^2)\Pi(B)\right) \leq \frac{1}{c^2 n\epsilon^2}. \tag{81}$$

Proof. By dividing two sides of the inequality inside $\mathbb{P}_0^n$ by $\Pi(B)$, we can (without loss of generality) assume that $\Pi(B) = 1$. By Jensen's inequality

$$\log \int \prod_{i=1}^n (f/f_0)(y_i|x_i)d\Pi(f) \geq \sum_{i=1}^n \int \log(f/f_0)(y_i|x_i)d\Pi(f).$$

Hence, for $\mathbb{P}_n$ being the empirical distribution, we have

$$\mathbb{P}_0^n\left(\int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)d\Pi(f) \leq \exp(-(c+1)n\epsilon^2)\right)$$
$$\leq \mathbb{P}_0^n\left(\sum_{i=1}^n \int \log(f/f_0)(y_i|x_i)d\Pi(f) \leq -(c+1)n\epsilon^2\right)$$
$$\leq \mathbb{P}_0^n\Big(\sqrt{n}\int\int \log(f/f_0)d\Pi(f)d(\mathbb{P}_n - \mathbb{P}_0)$$
$$\leq -\sqrt{n}(1+c)\epsilon^2 - \sqrt{n}\int\int \log(f/f_0)d\Pi(f)d\mathbb{P}_0\Big)$$

By Fubini's theorem and the definition of $B = B_2(f_0, \epsilon)$,

$$-\sqrt{n}\int\int \log(f/f_0)d\Pi(f)d\mathbb{P}_0 = \sqrt{n}\int \mathbb{P}_0 \log(f_0/f)d\Pi(f) \leq \sqrt{n}\epsilon^2. \tag{82}$$

Therefore,

$$\mathbb{P}_0^n\left(\int \prod_{i=1}^n \frac{f}{f_0}(y_i|x_i)d\Pi(f) \leq \exp(-(c+1)n\epsilon^2)\right)$$
$$\leq \mathbb{P}_0^n\left(\sqrt{n}\int\int \log(f/f_0)d\Pi(f)d(\mathbb{P}_n - \mathbb{P}_0) \leq \sqrt{n}c\epsilon^2\right)$$
$$\overset{(*)}{\leq} \frac{\mathrm{Var}_{\mathbb{P}_0}(\int \log(f/f_0)d\Pi(f))}{c^2 n\epsilon^4}$$

$$\leq \frac{\mathbb{P}_0(\int \log(f/f_0)d\Pi(f))^2}{c^2 n\epsilon^4}$$
$$\overset{(**)}{\leq} \frac{\mathbb{P}_0 \int (\log(f/f_0))^2 d\Pi(f)}{c^2 n\epsilon^4}$$
$$\leq \frac{1}{c^2 n\epsilon^2},$$

where we apply Chebyshev's inequality in $(*)$ and Jensen's inequality in $(**)$. Hence, inequality (81) is proved. □

Appendix E: Computational details

E.1. Bayesian approach

Here we describe in details the derivation of Markov Chain Monte Carlo (MCMC) algorithm that we use in Section 5. In particular, given a mixture of k-negative binomial regression model:

$$f_G(y|x) = \sum_{j=1}^{k} p_j \mathrm{NB}(y|h(x,\theta_j), \phi_j).$$

As mentioned in Section 5, given the data, $\{x_i, y_i\}_{i=1}^n$, we chose the prior distributions of $p = (p_1, p_2, ..., p_k)$, θ_j and $\eta_j = \phi_j^{-1}$, for $j = 1, ..., k$ as the following

$$p \sim \mathrm{Dir}(1, 1, ..., 1)$$
$$\theta_j \sim \mathrm{MVN}(0, I) \text{ (multivariate normal distribution), for } j{=}1, ..., k,$$
$$\eta_j \sim \mathrm{Gamma}(0.01, 0.01) \text{ (a non-informative gamma distribution), for } j{=}1, ..., k.$$

The full conditional posterior distributions of the model parameters are given below.

$$P(Z_i = j|y, x) = \frac{p_j \mathrm{NB}\left(y_i|h(x_i, \theta_j), \eta_j\right)}{\sum_{m=1}^{2} p_m \mathrm{NB}\left(y_i|h(x_i, \theta_m), \eta_m\right)}, \tag{83}$$

$$p|(y, x, \mathbf{Z}, ...) \sim \mathrm{Dir}(1{+}n_1, .., , 1{+}n_k), \text{ where } n_j{=}\#\{i : Z_i{=}j\}; \text{ for } j{=}1, ..., k, \tag{84}$$

$$f(\theta_j|y, x, \mathbf{Z}, ...) \propto \left[\prod_{i:Z_i=j} \mathrm{NB}\left(y_i|h(x_i, \theta_j), \eta_j\right)\right] \exp\left(-\frac{1}{2}\left\|\theta_j\right\|^2\right), \tag{85}$$

$$g(\eta_j|y, x, \mathbf{Z}, ...) \propto \left[\prod_{i:Z_i=j} \mathrm{NB}\left(y_i|h(x_i, \theta_j), \eta_j\right)\right] \eta_j^{0.01-1} \exp\left(-0.01\eta_j\right), \tag{86}$$

Algorithm 1 Gibbs sampling algorithm

Input: The prior distributions of p, θ_j and η_j, for $j = 1, ..., k$;
The number of iterations ($T_{\max}$) and burn-in steps
Output: A Markov Chains $\{\mathbf{\Phi}_t\}_{t\geq 0}$ attaining posterior distribution of $(p|(y, x, \mathbf{Z}, ...))$, $\eta_1|(y, x, \mathbf{Z}, ...)$,..., $\eta_k|(y, x, \mathbf{Z}, ...)$, $\theta_j|(y, x, \mathbf{Z}, ...)$ (for $j = 1, ..., k$) as the stationary distribution.

$t = 0$. Draw $\mathbf{\Phi}_0 = (p^{(0)}, \eta_1^{(0)}, ..., \eta_k^{(0)}, \theta_1^{(0)}, ..., \theta_k^{(0)})$ randomly.
for $t = 1, 2, \ldots, T_{\max}$ **do**
  Generate $\mathbf{Z}^{(t)} \sim \mathbf{Z}|y, x, p^{(t-1)}, \eta_1^{(t-1)}, ..., \eta_k^{(t-1)}, \theta_1^{(t-1)}, ..., \theta_k^{(t-1)}$
  Generate $p^{(t)} \sim p|y, x, \mathbf{Z}^{(t)}, \theta_1^{(t-1)}, ..., \theta_k^{(t-1)}, \eta_1^{(t-1)}, ..., \eta_k^{(t-1)}$
  Generate $\eta_j^{(t)} \sim \eta_j|y, x, \mathbf{Z}^{(t)}, p^{(t)}, \theta_1^{(t-1)}, ..., \theta_k^{(t-1)}$ ▷ Using Metropolis-Hasting algorithm
  Generate $\theta_j^{(t)} \sim \theta_j|y, x, \mathbf{Z}^{(t)}, p^{(t)}, \eta_1^{(t)}, ..., \eta_k^{(t)}$ ▷ Using Metropolis-Hasting algorithm
  Set $\mathbf{\Phi}_t = (p^{(t)}, \eta_1^{(t)}, ..., \eta_k^{(t)}, \theta_1^{(t)}, ..., \theta_k^{(t)})$
end for

where

$$\text{NB}\,(y_i|h(x_i, \theta_j), \eta_j) = \frac{\Gamma(y_i + 1/\eta_j)}{\Gamma(y_i + 1)\Gamma(1/\eta_j)} \left(\frac{\exp(x_i'\theta_j)}{\exp(x_i'\theta_j + 1/\eta_j)}\right)^{y_i} \times \left(\frac{1/\eta_j}{x_i'\theta_j + 1/\eta_j}\right)^{1/\eta_j}.$$

The full posterior distribution is sampled by using Gibbs sampling algorithm (Algorithm 1). Since the posterior distributions of θ_j and η_j $(j = 1, ..., k)$ are known up to a normalizing constant, the Metropolis-Hasting (MH) algorithm has been used to sample the distribution. When it comes to θ_j, a multivariate normal distribution is used as a proposal density. In particular, for each $j = 1, ..., k$, a candidate $\theta_j^* \sim \text{MVN}(\theta_j^{(t-1)}, \Sigma')$ is accepted with probability

$$\min\left\{1, \frac{f(\theta_j^*|...)}{f(\theta_j^{(t-1)}|...)}\right\}.$$

In terms of η_j, the proposal density is from a Gamma distribution. Specifically, for each $j = 1, ..., k$, a candidate $\eta_j^* \sim \text{Gamma}(2, 2/\eta_j^{(t-1)})$ is accepted with probability

$$\min\left\{1, \frac{g(\eta_j^*|...)p\left(\eta_j^{(t-1)}|\eta_j^*\right)}{g(\eta_j^{(t-1)}|...)p\left(\eta_j^*|\eta_j^{(t-1)}\right)}\right\},$$

where $f(\theta_j|y, x, \mathbf{Z}^{(t)}, p(t), \eta^{(t)})$ and $g(\eta_j|y, x, \mathbf{Z}^{(t)}, p(t), \theta^{(t)})$ are as in Eq. (85), (86), respectively, and $p\left(\eta_j^{(t-1)}|\eta_j^*\right)$ is the gamma density $\text{Gamma}(2, 2/\eta_j^*)$.

For each different sample size n, we run the experiment 16 times. For each time of running, we produced 2500 MCMC samples and discarded the first 500 as a "burn-in" set. From among the remaining 2000, we computed the mean of a vector containing 2000 Wasserstein distances (W_1) between the MCMC results and the true mixing distribution.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

This work is supported in part by the NSF Grant DMS-2015361 and a research gift from Wells Fargo.

References

[1] Bermúdez, L., Karlis, D. and Morillo, I. (2020). Modelling unobserved heterogeneity in claim counts using finite mixture models. *Risks* **8** 10.

[2] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics* 221–233. MR1331665

[3] DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification* **5** 249–282. MR0971156

[4] Ding, C. S. (2006). Using regression mixture analysis in educational research. *Practical Assessment, Research, and Evaluation* **11** 11.

[5] Do, D., Do, L., McKinley, S. A., Terhorst, J. and Nguyen, X. (2024). Dendrogram of mixing measures: Learning latent hierarchy and model selection for finite mixture models. *arXiv preprint arXiv:2403.01684*.

[6] Do, D., Ho, N. and Nguyen, X. (2022). Beyond Black Box Densities: Parameter Learning for the Deviated Components. *arXiv preprint arXiv:2202.02651*.

[7] Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications* 23–27. MR0033460

[8] Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference* **44**. Cambridge University Press. MR3587782

[9] Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29** 1233–1263. MR1873329

[10] Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press. MR3588285

[11] Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics* **1** 3-15. `https://doi.org/10.1016/0304-4076(73)90002-X`.

[12] Grün, B. and Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of classification* **25** 225–247. MR2476114

[13] Grün, B. and Leisch, F. (2008). Finite mixtures of generalized linear regression models. In *Recent advances in linear models and related areas* 205–230. Springer. MR2523852

[14] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics* **46** 2844–2870. MR3851757

[15] Hennig and Christian (2000). Identifiablity of Models for Clusterwise Linear Regression. *Journal of classification* **17**. MR1814811

[16] Hjort, N., Holmes, C., Mueller, P. and Walker, S. (2010). *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press. MR2722987

[17] Ho, N. and Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics* **10** 271-307. MR3466183

[18] Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics* **44** 2726-2755. MR3576559

[19] Ho, N. and Nguyen, X. (2019). Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM Journal on Mathematics of Data Science* **1** 730–758. MR4016133

[20] Ho, N., Yang, C.-Y. and Jordan, M. I. (2019). Convergence rates for gaussian mixtures of experts. *arXiv preprint arXiv:1907.04377*. MR4577762

[21] Hurn, M., Justel, A. and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of computational and graphical statistics* **12** 55–79. MR1977206

[22] Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation* **3** 79–87.

[23] Jaki, T., Kim, M., Lamont, A., George, M., Chang, C., Feaster, D. and Van Horn, M. L. (2019). The effects of sample size on the estimation of regression mixture models. *Educational and Psychological Measurement* **79** 358–384.

[24] Jiang, W. and Tanner, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Networks* **12** 1253–1258.

[25] Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association* **102** 1025–1038. MR2411662

[26] Miller, J. W. (2018). A detailed treatment of Doob's theorem. *arXiv preprint arXiv:1801.03122*.

[27] Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**. MR3803469

[28] Mufudza, C. and Erol, H. (2016). Poisson mixture regression models for heart disease prediction. *Computational and Mathematical Methods in Medicine* **2016**.

[29] Nguyen, H., Akbarian, P., Yan, F. and Ho, N. (2024). Statistical

Perspective of Top-K Sparse Softmax Gating Mixture of Experts. In *The Twelfth International Conference on Learning Representations*.

[30] Nguyen, H., Nguyen, T. and Ho, N. (2024). Demystifying softmax gating function in Gaussian mixture of experts. *Advances in Neural Information Processing Systems* **36**.

[31] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **41** 370–400. MR3059422

[32] Park, B.-J. and Lord, D. (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention* **41** 683-691. https://doi.org/10.1016/j.aap.2009.03.007. MR2771520

[33] Park, B.-J., Lord, D. and Hart, J. D. (2010). Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. *Accident Analysis and Prevention* **42** 741–749.

[34] Sarstedt, M. and Schwaiger, M. (2008). Model selection in mixture regression analysis–A Monte Carlo simulation study. In *Data analysis, machine learning and applications* 61–68. Springer.

[35] Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing* **22** 367–373. MR0620064

[36] Stein, E. M. and Shakarchi, R. (2010). *Complex analysis* **2**. Princeton University Press. MR1976398

[37] Teicher, H. (1963). Identifiability of finite mixtures. *The annals of Mathematical statistics* 1265–1269. MR0155376

[38] van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

[39] Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing* **12** 315–330. MR1951705

[40] Wang, S., Yao, W. and Huang, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statistics and Probability Letters* **93** 41–45. MR3244553

[41] Wei, Y., Mukherjee, S. and Nguyen, X. (2023). Minimum Φ-distance estimators for finite mixing measures. *arXiv preprint arXiv:2304.10052*.

[42] Wei, Y. and Nguyen, X. (2022). Convergence of de Finetti's mixing measure in latent structure models for observed exchangeable sequences. *The Annals of Statistics* **50** 1859–1889. MR4474475

[43] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* 339–362. MR1332570

[44] Wu, Y. and Yang, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics* **48** 1981 – 2007. MR4134783

[45] Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam* 423–435. Springer. MR1462963