Convergence of latent mixing measures in finite and infinite mixture models [1]

XuanLong Nguyen
xuanlong@umich.edu
Department of Statistics
University of Michigan

Abstract

We consider Wasserstein distances for assessing the convergence of latent discrete measures, which serve as mixing distributions in hierarchical and nonparametric mixture models. We clarify the relationships between Wasserstein distances of mixing distributions and f-divergence functionals such as Hellinger and Kullback-Leibler distances on the space of mixture distributions using various identifiability conditions. The convergence in Wasserstein metrics for discrete measures has a natural interpretation of the convergence of individual atoms that provide support for the discrete measure. It is also typically stronger than the weak convergence induced by standard f-divergence metrics. We establish rates of convergence of posterior distributions for latent discrete measures in several mixture models, including finite mixtures of multivariate distributions and infinite mixtures based on the Dirichlet process.

1 Introduction

A notable feature in the development of hierarchical and Bayesian nonparametric models is the role of discrete probability measures, which serve as mixing distributions to combine relatively simple models into richer classes of statistical models [20, 22]. In recent years the mixture modeling methodology has been significantly extended, by many authors taking the mixing measure to be random and infinite dimensional via suitable priors constructed in a nested, hierarchical and nonparametric manner. This results in rich models that can fit more complex and high dimensional data (see, e.g., [10, 28, 26, 25, 23] for several examples of such models, as well as a recent book [16]).

The focus of this paper is to analyze convergence behavior of the posterior distribution of latent mixing measures as they arise in several mixture models, including finite mixtures and the infinite Dirichlet process mixtures. Let $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ denote a discrete probability measure. Atoms θ_i's are elements in space Θ, while vector of probabilities $\boldsymbol{p} = (p_1, \ldots, p_k)$ lies in a $k-1$ dimensional probability simplex. In a mixture setting, G is combined with a likelihood density $f(\cdot|\theta)$ with respect to a dominating measure μ on $\mathcal{X}$, to

[1] AMS 2000 subject classification. Primary 62F15, 62G05; secondary 62G20.

Key words and phrases: Mixture distributions, hierarchical models, Bayesian nonparametrics, Wasserstein distance, f-divergence, rates of convergence, Dirichlet processes.

yield the mixture density: $p_G(x) = \int f(x|\theta)dG(\theta) = \sum_{i=1}^k p_i f(x|\theta_i)$. In a clustering application, atoms θ_i's represent distinct behaviors in a heterogeneous data population, while mixing probabilities p_i's are the associated proportions of such behaviors. Under this interpretation, there is a need for comparing and assessing the quality of the discrete measure $\hat{G}$ estimated on the basis of available data. An important work in this direction is by Chen [5], who used the L_1 metric on the cumulative distribution functions on the real line to study convergence rates of the mixing measure G. Chen's results were subsequently extended to a Bayesian estimation setting for a univariate mixture model [17]. These works were limited to only univariate and finite mixture models, with k bounded by a known constant, while our interest is when k may be unbounded, and Θ is multi-dimensional or even an abstract space.

The analysis of consistency and convergence rates of posterior distributions for Bayesian estimation have seen much progress in the past decade. Key recent references include [1, 13, 27, 33, 14, 34]. Analysis of specific mixture models in a Bayesian setting have also been studied [12, 11, 18, 15]. All these work primarily focus on the convergence behavior of the posterior distribution of the data density p_G. On the other hand, there seems to be very few results concerning with the convergence behavior of latent mixing measures G. Notably, the analysis of convergence for mixing (smooth) densities often arised in the context of frequentist estimation for deconvolution problems, mainly within the kernel density estimation method (e.g., [4, 35, 8]).

The primary contribution of this paper is to show that the Wasserstein distances provide a natural and useful metric for the analysis of convergence for latent and discrete mixing measures in mixture models, and to establish convergence rates of posterior distributions in a number of well-known Bayesian nonparametric and mixture models. Wasserstein distances originally arised in the problem of optimal transportation [31]. They have been utilized in a number of statistical contexts (e.g., [7, 21, 2, 6]). For discrete probability measures, they can be obtained by a minimum matching (or moving) procedure between the sets of atoms that provide support for the measures under comparison, and consequentially are simple to compute. Suppose that Θ is equipped with a metric ρ. Let $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$. Then, the L_r Wasserstein distance metric on the space of discrete probability measures with support in Θ, namely, $\bar{\mathcal{G}}(\Theta)$, is:

$$d_\rho(G, G') = \left[\inf_{\boldsymbol{q}} \sum_{i,j} q_{ij} \rho^r(\theta_i, \theta'_j) \right]^{1/r},$$

where the infimum is taken over all joint probability distributions on $[1, \ldots, k] \times [1, \ldots, k']$ such that $\sum_j q_{ij} = p_i$ and $\sum_i q_{ij} = p'_j$.

As clearly seen from this definition, the Wasserstein distances inherit directly the metric of the space of atomic support Θ, suggesting that they can be useful for assessing estimation procedures for discrete measures in hierarchical models. It is worth noting that if $(G_n)_{n\geq 1}$ is a sequence of discrete probability measures with k distinct atoms and G_n tends to some discrete measure G_0 in d_ρ metric, then G_n's ordered set of atoms must converge to G_0's

atoms in ρ after some permutation of atom labels. Thus, in the clustering application illustrated above, convergence of mixing measure G may be interpreted as the convergence of distinct typical behavior θ_i's that characterize the heterogeneous data population. A hint for the relevance of the Wasserstein distances can be drawn from an observation that the L_1 distance for the CDF's of univariate random variables, as studied by Chen [5], is in fact a special case of the L_1 Wasserstein metric when $\Theta = \mathbb{R}$.

The plan for the paper is as follows. Section 2 explores the relationship between Wasserstein distances for mixing measures and well-known divergence functionals for mixture densities in a mixture model. We produce a simple lemma which gives an upper bound on f-divergences between mixture densities by certain Wasserstein distances between mixing measures. This implies that d_ρ topology can be stronger than those induced by divergences between mixture densities. Next, we consider various identifiability conditions under which convergence of mixture densities entails convergence of mixing measures in the Wasserstein metric. We present two theorems, which provide upper bounds on $d_\rho(G, G')$ in terms of divergences between p_G and $p_{G'}$. Theorem 1 is applicable to mixing measures with a bounded number of atomic support, generalizing a result from [5]. Theorem 2 is applicable to mixing measures with unbounded number of support points, but is restricted to only convolution mixture models.

Section 3 focuses on the convergence of posterior distributions of latent mixing measures in a Bayesian nonparametric setting. Here, the mixing measure G is endowed with a prior distribution Π. Assuming an n-sample $X_1, \ldots, X_n$ that is generated according to p_{G_0}, we study conditions under which the postetrior distribution of G, namely, $\Pi(\cdot|X_1, \ldots, X_n)$, contracts to the "truth" G_0 under the d_ρ metric, and provide the contraction rates. In Theorems 3 and 4 of Section 3, we establish the convergence rates for the posterior distribution for G in terms of d_ρ metric. These results are proved using the standard approach of Ghosal, Ghosh and van der Vaart [13]. Our convergence theorems have several notable features. They rely on separate conditions for the prior Π and likelihood function f, which are typically simpler to verify than conditions formulated in terms of mixture densities. The claim of convergence in Wasserstein metrics is typically stronger than the weak convergence induced by the Hellinger metric in the existing work mentioned above.

In Section 4 posterior consistency and convergence rates of latent mixing measures are derived for a number of well-known mixture models in the literature, including finite mixtures of multivariate distributions, and infinite mixtures based on Dirichlet processes. For finite mixtures with bounded number of atomic support in $\mathbb{R}^d$, the posterior convergence rate for mixing measures is the minimax optimal $n^{-1/4}$ under suitable identifiability conditions. For Dirichlet process mixtures defined on $\mathbb{R}^d$, specific rates are established under smoothness conditions of the likelihood density function f. In particular, for ordinary smooth likelihood densities with smoothness β (e.g., Laplace), the rate achieved is $(\log n/n)^\gamma$ for any $\gamma < \frac{2}{(d+2)(4+(2\beta+1)d)}$. For supersmooth likelihood densities with smoothness β (e.g., normal), the rate achieved is $(\log n)^{-1/\beta}$.

Notations. For ease of notations, we also use f_i in place of $f(\cdot|\theta_i)$, and f'_j in place of $f(\cdot|\theta'_j)$ for likelihood density functions. Divergences (distances) studied in the paper include the total variational distance: $V(p_G, p_{G'}) := \frac{1}{2}\int |p_G(x) - p_{G'}(x)|d\mu(x)$, the Hellinger distance: $h^2(p_G, p_{G'}) := \frac{1}{2}\int(\sqrt{p_G(x)} - \sqrt{p_{G'}(x)})^2 d\mu(x)$, and the Kullback-Leibler divergence: $K(p_G, p_{G'}) = \int p_G(x)\log(p_G(x)/p_{G'}(x))d\mu(x)$. These divergences are related by $V^2/2 \leq h^2 \leq V$ and $h^2 \leq K/2$. $N(\epsilon, \Theta, \rho)$ denotes the covering number of the metric space (Θ, ρ), i.e., the minimum number of ϵ-balls needed to cover the entire space Θ. $D(\epsilon, \Theta, \rho)$ denotes the packing number of (Θ, ρ), i.e., the maximum number of points that are mutually separated by at least ϵ in distance. They are related by $N(\epsilon, \Theta, \rho) \leq D(\epsilon, \Theta, \rho) \leq N(\epsilon/2, \Theta, \rho)$. Diam$(\Theta)$ denotes the diameter of Θ.

2 Wasserstein distances for mixing measures

2.1 Definition and a basic inequality

Let (Θ, ρ) be a space equiped with a non-negative distance function $\rho : \Theta \times \Theta \to \mathbb{R}_+$, i.e. function that satisfies $\rho(\theta_1, \theta_2) = 0$ if and only if $\theta_1 = \theta_2$. If in addition, ρ is symmetric ($\rho(\theta_1, \theta_2) = \rho(\theta_2, \theta_1)$), and satisfies the triangle inequality, then it is a proper metric. A discrete probability measure G on a measure space equipped with the Borel sigma algebra takes the form $G = \sum_{i=1}^k p_i\delta_{\theta_i}$ for some $k \in \mathbb{N} \cup \{+\infty\}$, where $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ denotes the proportion vector, while $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are the associated atoms in Θ. $\boldsymbol{p}$ has to satisfy $0 \leq p_i \leq 1$ and $\sum_{i=1}^k p_k = 1$. (With a bit abuse of notations, we write $k = \infty$ when $G = \sum_{i=1}^\infty p_i\delta_{\theta_i}$ has countably infinite support points represented by the infinite sequence of atoms $\boldsymbol{\theta} = (\theta_1, \ldots)$ and the associated sequence of probability mass $\boldsymbol{p}$.) Likewise, $G' = \sum_{j=1}^{k'} p'_j\delta_{\theta'_j}$ is another discrete probability measure that has at most k' distinct atoms. Let $\mathcal{G}_k(\Theta)$ denote the space of all discrete probability measures with at most k atoms. Let $\mathcal{G}(\Theta) = \cup_{k\in\mathbb{N}_+}\mathcal{G}_k(\Theta)$, the set of all discrete measures with finite support. Finally, $\bar{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures (including those with countably infinite support).

Let $\boldsymbol{q} = (q_{ij})_{i\leq k; j\leq k'} \in [0,1]^{k\times k'}$ denote a joint probability distribution on $\mathbb{N}_+ \times \mathbb{N}_+$ that satisfies the marginal constraints: $\sum_{i=1}^k q_{ij} = p'_j$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, \ldots, k; j = 1, \ldots, k'$. We also call $\boldsymbol{q}$ a coupling of $\boldsymbol{p}$ and $\boldsymbol{p'}$. Let $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ denote the space of all such couplings. We start with the L_1 Wasserstein distance:

Definition 1. *Let ρ be a distance function on Θ. The Wasserstein distance for two discrete measures $G(\boldsymbol{p}, \boldsymbol{\theta})$ and $G'(\boldsymbol{p'}, \boldsymbol{\theta'})$ is:*

$$d_\rho(G, G') = \inf_{\boldsymbol{q}\in\mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})} \sum_{i,j} q_{ij}\rho(\theta_i, \theta'_j). \tag{1}$$

Our main focus is on the L_1 Wasserstein d_ρ, and L_2 Wasserstein distance. The latter corresponds to the square root of d_{ρ^2} in our definition, where $\rho(\theta_i, \theta'_j)$ is replaced by $\rho^2(\theta_i, \theta'_j)$. When Θ is a metric space (e.g., $\mathbb{R}^d$), and ρ is taken to be its metric, we revert

to the more standard notations, $W_1(G, G')$ for $d_\rho(G, G')$, and $W_2^2(G, G')$ for $d_{\rho^2}(G, G')$. However, d_ρ will be employed when ρ may be a general or a non-standard distance function or metric.

From here on, discrete measure $G \in \bar{\mathcal{G}}(\Theta)$ plays the role of the mixing distribution in a mixture model. Let $f(x|\theta)$ denote the density (with respect to a dominating measure μ) of a random variable X taking values in $\mathcal{X}$, given parameter $\theta \in \Theta$. For the ease of notations, we also use $f_i(x)$ for $f(x|\theta_i)$. Combining G with the likelihood function f yields a mixture distribution for X that takes the following density:

$$p_G(x) = \int f(x|\theta) dG(\theta) = \sum_{i=1}^{k} p_i f_i(x).$$

A central theme in this paper is to explore the relationship between Wasserstein distances of mixing measures G, G', e.g., $d_\rho(G, G')$, and divergences of mixture densities $p_G, p_{G'}$. Divergences that play important roles in this paper are the total variational distance, Hellinger distance, and the Kullback-Leibler distance. All these are in fact instances of a broader class of divergences known as the f-divergences (Csizar, 1966; Ali & Silvey, 1967):

Definition 2. *Let $\phi : \mathbb{R} \to \mathbb{R}$ denote a convex function. An f-divergence (or Ali-Silvey distance) between two probability densities f_i and f'_j is defined as $\rho_\phi(f_i, f'_j) = \int \phi(f'_j/f_i) f_i d\mu$. Likewise, the f-divergence between p_G and $p_{G'}$ is $\rho_\phi(p_G, p_{G'}) = \int \phi(p_{G'}/p_G) p_G d\mu$.*

f-divergences can be used as a distance function or metric on Θ. When ρ is taken to be an f-divergence, $\rho(\theta_i, \theta'_j) := \rho_\phi(f_i, f'_j)$, for a convex function ϕ, we call the corresponding Wasserstein distance a *composite* Wasserstein distance:

$$d_{\rho_\phi}(G, G') := \inf_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')} \sum_{ij} q_{ij} \rho_\phi(f_i, f'_j).$$

For $\phi(u) = \frac{1}{2}(\sqrt{u} - 1)^2$ we obtain the squared Hellinger ($\rho_h^2 \equiv h^2$), which induces the composite Wasserstein distance $d_{\rho_h^2}$. For $\phi(u) = \frac{1}{2}|u - 1|$ we obtain the variational distance ($\rho_V \equiv V$), which induces d_{ρ_V}. For $\phi(u) = -\log u$, we obtain the Kullback-Leiber divergence ($\rho_K \equiv K$), which induces d_{ρ_K}.

Lemma 1. *Let $G, G' \in \bar{\mathcal{G}}(\Theta)$ such that both $\rho_\phi(p_G, p_{G'})$ and $d_{\rho_\phi}(G, G')$ are finite for some convex function ϕ. Then, $\rho_\phi(p_G, p_{G'}) \leq d_{\rho_\phi}(G, G')$.*

This lemma highlights a simple direction in the aforementioned relationship: any f-divergence between mixture distributions p_G and $p_{G'}$ is dominated by a Wasserstein distance between mixing measures G and G'. As will be evident in the sequel, this basic inequality is also handy in enabling us to obtain upper bounds on the power of tests. It also proves useful for establishing lower bounds on small Kullback-Leibler ball probabilities in the space of mixture densities p_G in terms of small ball probabilities in the metric space (Θ, ρ). The latter quantities are typically easier to obtain estimates for than the former.

Example 1. Suppose that $\Theta = \mathbb{R}^d$, ρ is the Euclidean metric, $f(x|\theta)$ is the multivariate normal density $N(\theta, I_{d\times d})$ with mean θ and identity covariance matrix, then $h^2(f_i, f_j') = 1 - \exp -\frac{1}{8}\|\theta_i - \theta_j'\|^2 \leq \frac{1}{8}\|\theta_i - \theta_j'\|^2 = \rho^2(\theta_i, \theta_j')^2/8$. So, $d_{\rho_h^2}(G, G') \leq d_{\rho^2}(G, G')/8$. The above lemma then entails that $h^2(p_G, p_{G'}) \leq d_{\rho^2}(G, G')/8 = W_2^2(G, G')/8$.

Similarly, for the Kullback-Leibler divergence, since $K(f_i, f_j') = \frac{1}{2}\|\theta_i - \theta_j'\|^2$, by Lemma 1, $K(p_G, p_{G'}) \leq d_{\rho_K}(G, G') = \frac{1}{2}d_{\rho^2}(G, G') = W_2(G, G')^2/2$.

For another example, if $f(x|\theta)$ is a Gamma density with location parameter θ, Θ is a compact subset of $\mathbb{R}$ that is bounded away from 0. Then $K(f_i, f_j') = O(|\theta_i - \theta_j|)$. This entails that $K(p_G, p_{G'}) \leq d_{\rho_K}(G, G') \leq O(W_1(G, G'))$.

2.2 Wasserstein metric identifiability in finite mixture models

Lemma 1 shows that for many choices of ρ, d_ρ yields a stronger topology on $\bar{\mathcal{G}}(\Theta)$ than the topology induced by f-divergences on the space of mixture distributions p_G. In other words, convergence of p_G may not imply convergence of G in d_ρ metric. To ensure this property, additional conditions are needed on the space of discrete measures $\bar{\mathcal{G}}(\Theta)$, along with identifiability conditions for the family of likelihood functions $\{f(\cdot|\theta), \theta \in \Theta\}$.

The classical definition of Teicher [30] specifies the family $\{f(\cdot|\theta), \theta \in \Theta\}$ to be identifiable if for any $G, G' \in \mathcal{G}(\Theta)$, $\|p_G - p_{G'}\|_\infty = 0$ implies that $G = G'$. We need a slightly stronger version, allowing for the inclusion for discrete measures with infinite support:

Definition 3. *The family $\{f(\cdot|\theta), \theta \in \Theta\}$ is finitely identifiable if for any $G \in \mathcal{G}_\Theta$ and $G' \in \bar{\mathcal{G}}_\Theta$, $|p_G(x) - p_{G'}(x)| = 0$ for almost all $x \in \mathcal{X}$ implies that $G = G'$.*

To obtain convergence rates, we also need the notion of strong identifiability of [5], herein adapted to a multivariate setting.

Definition 4. *Assume that $\Theta \subseteq \mathbb{R}^d$ and ρ is the Euclidean metric. The family $\{f(\cdot|\theta), \theta \in \Theta\}$ is strongly identifiable if $f(x|\theta)$ is twice differentiable in θ and for any finite k and k different $\theta_1, \ldots, \theta_k$, the equality*

$$\operatorname*{ess\,sup}_{x\in\mathcal{X}} \left| \sum_{i=1}^{k} \alpha_i f(x|\theta_i) + \beta_i^T Df(x|\theta_i) + \gamma_i^T D^2 f(x|\theta_i)\gamma_i \right| = 0 \tag{2}$$

implies that $\alpha_i = 0$, $\beta_i = \gamma_i = \mathbf{0} \in \mathbb{R}^d$ for $i = 1, \ldots, k$. Here, for each x, $Df(x|\theta_i)$ and $D^2 f(x|\theta_i)$ denote the gradient and the Hessian at θ_i of function $f(x|\cdot)$, respectively.

Finite identifiability is satisfied for the family of Gaussian distributions for both mean and variance parameters [29], see also Theorem 1 of [18]. Chen identified a broad class of families, including the Gaussian family, for which the strong identifiability condition holds [5].

Define $\psi(G, G') = \sup_x |p_G(x) - p_{G'}(x)|/W_2^2(G, G')$ if $G \neq G'$ and ∞ otherwise. Also define $\psi_1(G, G') = V(p_G, p_{G'})/W_2^2(G, G')$ if $G \neq G'$ and ∞ otherwise. The notion of strong identifiability is useful via the following key result, which generalizes Chen's result to Θ of arbitrary dimensions.

Theorem 1. (Strong identifiability). *Suppose that Θ is compact subset of $\mathbb{R}^d$, the family $\{f(\cdot|\theta), \theta \in \Theta\}$ is strongly identifiable, and for all $x \in \mathcal{X}$, the Hessian matrix $D^2 f(x|\theta)$ satisfies a uniform Lipschitz condition*

$$|\gamma^T (D^2 f(x|\theta_1) - D^2 f(x|\theta_2))\gamma| \leq C\|\theta_1 - \theta_2\|^\delta \|\gamma\|^2, \tag{3}$$

for all x, θ_1, θ_2 and some fixed C and $\delta > 0$. Then, for fixed $G_0 \in \mathcal{G}_k(\Theta)$, where $k < \infty$:

$$\lim_{\epsilon \to 0} \inf_{G, G' \in \mathcal{G}_k(\Theta)} \left\{ \psi(G, G') : W_2(G_0, G) \vee W_2(G_0, G') \leq \epsilon \right\} > 0. \tag{4}$$

The assertion also holds with ψ being replaced by ψ_1.

Remark. Suppose that G_0 has exactly k distinct support points in Θ (i.e., $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ where $p_i > 0$ for all $i = 1, \ldots, k$). Then, an examination of the proof reveals that the requirement that Θ be compact is not needed. Indeed, if there is a sequence of $G_n \in \mathcal{G}_k(\Theta)$ such that $d_\rho(G_0, G_n) \to 0$, then it is simple to show that there is a subsequence of G_n that also has k distinct atoms, which converge in ρ metric to the set of k atoms of G_0 (up to some permutation of the labels). The proof of the theorem proceeds as before.

For the rest of this paper, by strong identifiability we always mean conditions specified in Theorem 1 so that Eq. (4) can be deduced. This practically means that the conditions specified by Eq. (2) and Eq. (3) be given, while the compactness of Θ may sometimes be required.

2.3 Wasserstein metric identifiability in infinite mixture models

Next, we state a counterpart of Theorem 1 for $G, G' \in \bar{\mathcal{G}}(\Theta)$, i.e., mixing measures with potentially unbounded number of support points. We restrict our attention to convolution mixture models on $\mathbb{R}^d$. That is, the likelihood density function $f(x|\theta)$, with respect to Lebesgue, takes the form $f(x - \theta)$ for some multivariate density function f on $\mathbb{R}^d$. Thus, $p_G(x) = G * f(x) = \sum_{i=1}^k p_i f(x - \theta_i)$ and $p_{G'}(x) = G' * f(x) = \sum_{j=1}^{k'} p'_j f(x - \theta'_j)$.

As before, we need a compactness condition for Θ. Additional key assumptions concern with the smoothness of density function f. This is characterized in terms of the tail behavior of the Fourier transform $\tilde{f}$ of f: $\tilde{f}(\omega) = \int_{\mathbb{R}^d} e^{-i\langle \omega, x \rangle} f(x) dx$. We consider both ordinary smooth densities (e.g., Laplace and Gamma), and supersmooth densities (e.g., normal). The following result does not require that G, G' be discrete.

Theorem 2. *Suppose that G, G' are probability measures that place full support on compact set $\Theta \subset \mathbb{R}^d$. f is a density function on $\mathbb{R}^d$ that is symmetric (around 0), i.e., $\int_A f dx = \int_{-A} f dx$ for any Borel set $A \subset \mathbb{R}^d$. Moreover, assume that $\tilde{f}(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$.*

(1) **Ordinary smooth likelihood.** *Suppose that $|\tilde{f}(\omega) \prod_{j=1}^d |\omega_j|^\beta| \geq d_0$ as $\omega_j \to \infty$ $(j = 1, \ldots, d)$ for some positive constants d_0 and β. Then for any $m <$*

$4/(4+(2\beta+1)d)$, *there is some constant* $C(d,\beta,m)$ *dependent only on* d,β *and* m *such that*

$$W_2^2(G,G') \le C(d,\beta,m)V(p_G,p_{G'})^m,$$

as $V(p_G,p_{G'}) \to 0$.

(2) **Supersmooth likelihood.** *Suppose that* $|\tilde{f}(\omega)\prod_{j=1}^d \exp(|\omega_j|^\beta/\gamma)| \ge d_0$ *as* $\omega_j \to \infty$ $(j=1,\ldots,d)$ *for some positive constants* β,γ,d_0. *Then there is some constant* $C(d,\beta)$ *dependent only on* d *and* β *such that*

$$W_2^2(G,G') \le C(d,\beta)(-\log V(p_G,p_{G'}))^{-2/\beta},$$

as $V(p_G,p_{G'}) \to 0$.

Example 2. For the standard normal density on $\mathbb{R}^d$, $\tilde{f}(\omega)=\prod_{j=1}^d \exp -\omega_i^2/2$, we obtain that $d_\rho^2(G,G') \lesssim (-\log V(p_G,p_{G'}))^{-1}$ as $d_\rho(G,G') \to 0$ (so that $V(p_G,p_{G'}) \to 0$, by Lemma 1). For a Laplace density on $\mathbb{R}$, e.g., $\tilde{f}(\omega)=\frac{1}{1+\omega^2}$, then $d_\rho^2(G,G') \lesssim V(p_G,p_{G'})^m$ for any $m<4/9$, as $d_\rho(G,G') \to 0$.

3 Convergence of posterior distributions of mixing measures

We turn to a study of convergence of discrete mixing measures in a Bayesian setting. Let $X_1,\ldots,X_n$ be an iid sample according to the mixture density $p_G(x)=\int f(x|\theta)dG(\theta)$, where f is known, while $G=G_0$ for some unknown mixing measure in $\mathcal{G}_k(\Theta)$. The true number support points for G may be unknown. In the Bayesian estimation framework, G is endowed with a prior distribution Π on a suitable measure space of discrete probability measures in $\bar{\mathcal{G}}(\Theta)$. The posterior distribution of G is given by, for any measurable set B:

$$\Pi(B|X_1,\ldots,X_n)=\int_B \prod_{i=1}^n p_G(X_i)\Pi(G)/\int \prod_{i=1}^n p_G(X_i)\Pi(G).$$

We shall study conditions under which the posterior distribution is consistent, i.e., it concentrates on arbitrarily small W_2 neighborhoods of G_0, and establish the rates of the convergence. We follow the general framework of Ghosal, Ghosh and van der Vaart [13], who analyzed convergence behavior of posterior distributions in terms of f-divergences such as Hellinger and variational distances on the mixture densities of the data. In the following we formulate two convergence theorems for mixture model setting (which can be viewed as counterparts of Theorem 2.1 and 2.4 of [13]). A notable feature of our theorems is that conditions (e.g., entropy and prior concentration) are stated directly in terms of the Wasserstein metric, as opposed to f-divergences on the mixture densities. They may be typically separated into independent conditions for the prior for G and the likelihood family and are simpler to verify for mixture models. In addition, the convergence of the posterior distribution of mixing measures is established in terms of L_2 Wasserstein distance metric.

The following notion plays a central role in our general results.

Definition 5. *Fix $k < \infty$ and $G_0 \in \mathcal{G}_k(\Theta)$. Let $\mathcal{G}$ be a subset of $\bar{\mathcal{G}}(\Theta)$. Define the Hellinger information of W_2 metric for subset $\mathcal{G}$ as a real-valued function on the real line $C_k(\mathcal{G}, \cdot) : \mathbb{R} \to \mathbb{R}$:*

$$C_k(\mathcal{G}, r) = \inf_{G \in \mathcal{G}: W_2(G_0, G) \geq r/2} h^2(p_{G_0}, p_G). \tag{5}$$

It is obvious that C_k is a non-negative and non-decreasing function. The following characterizations of C_k are simple consequences of Theorems 1 and 2:

Proposition 1. *(a) Suppose that $\mathcal{G}$ and $\mathcal{G}_k(\Theta)$ are both compact in the Wasserstein topology, and the family of likelihood functions is finitely identifiable. Then, $C_k(\mathcal{G}, r) > 0$ for any $r > 0$.*

(b) Suppose that $\Theta \subset \mathbb{R}^d$ is compact, and the family of likelihood functions is strongly identifiable as specified in Theorem 1. Then, for each *k there is a constant $c(k, G_0) > 0$ such that $C_k(\mathcal{G}_k(\Theta), r) \geq c(k, G_0) r^4$ for all $r > 0$.*

(c) Suppose that $\Theta \subset \mathbb{R}^d$ is compact, and the family of likelihood functions is ordinary smooth with parameter β, as specified in Theorem 2. Then, for any $\epsilon > 0$, there is some constant $c(d, \beta)$ such that $C_k(\bar{\mathcal{G}}(\Theta), r) \geq c(d, \beta) r^{4+(2\beta+1)d+\epsilon}$ for any $r > 0$, and any $k > 0$. For supersmooth likelihood family, we have $C_k(\bar{\mathcal{G}}(\Theta), r) \geq \exp[-c(d, \beta) r^{-\beta}]$ for any $r > 0$ and any $k > 0$.

A main ingredient in the analysis of convergence of posterior distributions is through proving the existence of tests for subsets of parameters of interest. A test φ_n is a measurable indicator function of the iid sample $X_1, \ldots, X_n$. For a fixed pair of discrete measures $(G_0, G_1) \in (\mathcal{G}_k(\Theta) \times \mathcal{G})$, where $\mathcal{G}$ is a given subset of $\bar{\mathcal{G}}(\Theta)$, consider tests for discriminating G_0 against a closed Wasserstein ball centered at G_1. Write

$$B_W(G_1, r) = \{G \in \bar{\mathcal{G}}(\Theta) : W_2(G_1, G) \leq r\}.$$

The following lemma highlights the role of the Hellinger information:

Lemma 2. *Suppose that (Θ, ρ) is a metric space. Fix $k < \infty$ and $G_0 \in \mathcal{G}_k(\Theta)$. $\mathcal{G}$ is a subset of mixing measures such that $\mathcal{G}_k(\Theta) \subseteq \mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$. Given $G_1 \in \mathcal{G}$, let $r = W_2(G_0, G_1)$. Suppose that either one of the following two sets of conditions holds:*

(I) $\mathcal{G}$ is a convex set, in which case, set $M(\mathcal{G}, G_1, r) = 1$.

(II) $\mathcal{G}$ is non-convex, while Θ is a totally bounded and bounded set. In addition, for some constants $C_1 > 0, \alpha \geq 1$, $h(f_i, f'_j) \leq C_1 \rho^\alpha(\theta_i, \theta'_j)$ for any likelihood functions f_i, f'_j in the family. In this case, define

$$M(\mathcal{G}, G_1, r) = D\left(\frac{C_k(\mathcal{G}, r)^{1/4}}{\mathit{Diam}(\Theta)^{\alpha-1}\sqrt{2C_1}}, \mathcal{G} \cap B_W(G_1, r/2), W_2\right). \tag{6}$$

Then, there exist tests $\{\varphi_n\}$ that have the following properties:

$$P_{G_0}\varphi_n \leq M(\mathcal{G}, G_1, r)\exp[-nC_k(\mathcal{G}, r)/8], \tag{7}$$

$$\sup_{G\in\mathcal{G}\cap B_W(G_1,r/2)} P_G(1-\varphi_n) \leq \exp[-nC_k(\mathcal{G}, r)/8]. \tag{8}$$

Here, P_G denotes the expectation under the mixture distribution given by density p_G.

Remark. The set of conditions (II) is needed when $\mathcal{G}$ is not convex, an example of which is $\mathcal{G} = \mathcal{G}_k(\Theta)$, the space of measures with at most k support points in Θ. It is interesting to note that the loss in test power due to the lack of convexity is captured by the local entropy term $\log M(\mathcal{G}, G_1, r)$. This quantity is defined in terms of the packing by small W_2 balls whose radii are specified by the Hellinger information. We note briefly that this quantity can be easily bounded in some cases: for instance, if $\Theta \subset \mathbb{R}^d$ and $\mathcal{G} = \mathcal{G}_k(\Theta)$, then by Proposition 1, under strong identifiability, $C_k(\mathcal{G}, r) \geq c(k, G_0)r^4$. In addition, under the assumption that for all $G \in \mathcal{G}$, both the masses for all k support points of G, and the pairwise distances of such support points, are uniformly bounded away from 0, it can be shown that $M(\mathcal{G}, G_1, r)$ is bounded by a constant that depends on only G_0, Θ and k (Further details are elaborated in the proof of Theorem 5.)

Next, the existence of test can be shown for discriminating G_0 against the complement of a closed Wasserstein ball:

Lemma 3. *Assume that conditions of Lemma 2 hold and define $M(\mathcal{G}, G_1, r)$ as in Lemma 2. Suppose that for some non-increasing function $D(\epsilon)$, some $\epsilon_n \geq 0$ and every $\epsilon > \epsilon_n$,*

$$\sup_{G_1\in\mathcal{G}} M(\mathcal{G}, G_1, \epsilon) \times D(\epsilon/2, \mathcal{G} \cap B_W(G_0, 2\epsilon) \setminus B_W(G_0, \epsilon), W_2) \leq D(\epsilon). \tag{9}$$

Then, for every $\epsilon > \epsilon_n$, for any $t_0 \in \mathbb{N}$, there exist tests φ_n (depending on $\epsilon > 0$) such that

$$P_{G_0}\varphi_n \leq D(\epsilon) \sum_{t=t_0}^{\lceil \mathit{Diam}(\Theta)/\epsilon\rceil} \exp[-nC_k(\mathcal{G}, t\epsilon)/8] \tag{10}$$

$$\sup_{G\in\mathcal{G}:W_2(G_0,G)>t_0\epsilon} P_G(1-\varphi_n) \leq \exp[-nC_k(\mathcal{G}, t_0\epsilon)/8]. \tag{11}$$

Remark. It is interesting to observe that function $D(\epsilon)$ is used to control the packing number of thin layers of Wasserstein balls (a similar quantity that also arises via the peeling argument in [13] (Theorem 7.1)), in addition to the packing number M of small Wasserstein balls in terms of smaller Wasserstein balls whose radii are specified in terms of the Hellinger information function. As in the previous lemma, the latter packing number appears intrinsic to the analysis of convergence for mixing measures.

The preceeding two lemmas provide the core argument for establishing the following general posterior contraction theorems for latent mixing measures in a mixture model. The following two theorems have three types of conditions. The first is concerned with the size

of support of Π, often quantified in terms of its entropy number. Estimates of the entropy number defined in terms of Wasserstein metrics for several measure classes of interest are given in Lemma 4. The second condition is on the Kullback-Leibler support of Π, which is related to both space of discrete measures $\bar{\mathcal{G}}(\Theta)$ and the family of likelihood functions $f(x|\theta)$. The Kullback-Leibler neighborhood is defined as:

$$B_K(\epsilon) = \left\{ G \in \bar{\mathcal{G}}(\Theta) : -P_{G_0}\left(\log \frac{p_G}{p_{G_0}}\right) \leq \epsilon^2, P_{G_0}\left(\log \frac{p_G}{p_{G_0}}\right)^2 \leq \epsilon^2 \right\}. \tag{12}$$

The third type of conditions is on the Hellinger information of W_2 metric, function $C_k(\mathcal{G}, r)$, a characterization of which is given above.

Theorem 3. *Let $G_0 \in \mathcal{G}_k(\Theta) \subseteq \bar{\mathcal{G}}(\Theta)$ for some $k < \infty$, and the family of likelihood functions is finitely identifiable. Suppose that for a sequence $(\epsilon_n)_{n\geq 1}$ that tends to a constant (or 0) such that $n\epsilon_n^2 \to \infty$, and a constant $C > 0$, and* convex *sets $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$, we have:*

$$\log D(\epsilon_n, \mathcal{G}_n, W_2) \leq n\epsilon_n^2, \tag{13}$$

$$\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) \leq \exp[-n\epsilon_n^2(C+4)], \tag{14}$$

$$\Pi(B_K(\epsilon_n)) \geq \exp(-n\epsilon_n^2 C). \tag{15}$$

Moreover, suppose M_n is a sequence such that

$$C_k(\mathcal{G}_n, M_n\epsilon_n) \geq 8\epsilon_n^2(C+4), \tag{16}$$

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/8] \to 0. \tag{17}$$

Then, $\Pi(G : W_2(G_0, G) \geq M_n\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_{G_0}-probability.

The following theorem uses a weaker condition on the covering number, but it contains an additional condition on the likelihood functions which may be useful for handling the case of non-convex sieves $\mathcal{G}_n$:

Theorem 4. *Let $G_0 \in \mathcal{G}_k(\Theta) \subseteq \bar{\mathcal{G}}(\Theta)$ for some $k < \infty$. Assume the following:*

(a) The family of likelihood functions is finitely identifiable, and satisfies $h(f_i, f_j') \leq C_1\rho^\alpha(\theta_i, \theta_j')$ for any likelihood functions f_i, f_j' in the family, for some constants $C_1 > 0, \alpha \geq 1$.

(b) There is a sequence of sets $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$ for which $M(\mathcal{G}_n, G_1, \epsilon)$ is defined by Eq. (6).

(c) There is a sequence $\epsilon_n \to 0$ such that $n\epsilon_n^2$ is bounded away from 0 or tending to

infinity, and a sequence M_n such that

$$\log D(\epsilon/2, \mathcal{G}_n \cap B_W(G_0, 2\epsilon) \setminus B_W(G_0, \epsilon), W_2) + \sup_{G_1 \in \mathcal{G}_n} \log M(\mathcal{G}_n, G_1, \epsilon) \leq n\epsilon_n^2 \ \forall \epsilon \geq \epsilon_n, \tag{18}$$

$$\frac{\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n)}{\Pi(B_K(\epsilon_n))} = o(\exp(-2n\epsilon_n^2)), \tag{19}$$

$$\frac{\Pi(B_W(G_0, 2j\epsilon_n) \setminus B_W(G_0, j\epsilon_n))}{\Pi(B_K(\epsilon_n))} \leq \exp[nC_k(\mathcal{G}_n, j\epsilon_n)/16] \ \forall j \geq M_n, \tag{20}$$

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/16] \to 0. \tag{21}$$

Then, we have that $\Pi(G : W_2(G_0, G) \geq M_n\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_{G_0}-probability.

Before moving to specific examples, we state a simple lemma which provides estimates of the entropy under d_ρ metric for a number of classes of discrete measures of interest. Because d_ρ inherits directly the ρ metric in Θ, the entropy for classes in $(\bar{\mathcal{G}}(\Theta), d_\rho)$ can typically be bounded in terms of the covering number for subsets of (Θ, ρ). Since $d_\rho \equiv W_1$ is weaker than the L_2 Wasserstein metric W_2, the following bounds are immediately applicable if d_ρ is replaced by W_2.

Lemma 4. *(a)* $\log N(2\epsilon, \mathcal{G}_k(\Theta), d_\rho) \leq k(\log N(\epsilon, \Theta, \rho) + \log(e + e\mathit{Diam}(\Theta)/\epsilon))$.

(b) $\log N(2\epsilon, \bar{\mathcal{G}}(\Theta), d_\rho) \leq N(\epsilon, \Theta, \rho) \log(e + e\mathit{Diam}(\Theta)/\epsilon)$.

(c) Let $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$. Assume that $M = \max_{i=1}^k 1/p_i^* < \infty$ and $m = \min_{i,j \leq k} \rho(\theta_i^*, \theta_j^*) > 0$. Then,

$$\log N(\epsilon/2, \{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq k(\sup_{\Theta'} \log N(\epsilon/4, \Theta', \rho) + \log(32k\mathit{Diam}(\Theta)/m)),$$

where the supremum in the right side is taken over all bounded subsets $\Theta' \subseteq \Theta$ such that $\mathit{Diam}(\Theta') \leq 4M\epsilon$.

4 Examples

In this section we derive posterior contraction rates for two classes of mixture models, e.g, finite mixtures of multivariate distributions, and infinite mixtures based on the Dirichlet process.

4.1 Finite mixtures of multivariate distributions

Let Θ be a subset of $\mathbb{R}^d$, ρ be the Euclidean metric, and Π is a prior distribution for discrete measures in $\mathcal{G}_k(\Theta)$, where $k < \infty$ is known. Suppose that the "truth" $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$. To obtain the convergence rate of the posterior distribution of G, we need:

Assumptions A.

(A1) Θ is compact and the family of likelihood functions $f(\cdot|\theta)$ is strongly identifiable.

(A2) For some positive constant C_1, $K(f_i, f_j') \leq C_1\|\theta_i - \theta_j'\|^2$ for any $\theta_i, \theta_j' \in \Theta$. For any $G \in \text{supp}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 < C_2 K(p_{G_0}, p_G)$ for some constant $C_2 > 0$.

(A3) Under prior Π, for small $\delta > 0$, $c_3\delta^k \leq \Pi(|p_i - p_i^*| \leq \delta, i = 1\ldots, k) \leq C_3\delta^k$ and $c_3\delta^{kd} \leq \Pi(\|\theta_i - \theta_i^*\| \leq \delta, i = 1\ldots, k) \leq C_3\delta^{kd}$ for some constants $c_3, C_3 > 0$.

(A4) Under prior Π, all p_i are bounded away from 0, and all pairwise distances $\|\theta_i - \theta_j\|$ are bounded away from 0.

Remarks. Assumptions (A1) and (A2) hold for the family of Gaussian densities with mean parameter θ. Assumption (A3) holds when the prior distribution on the relevant parameters behave like a uniform distribution, up to a multiplicative constant.

Theorem 5. *Under Assumptions (A1–A4), the contraction rate in the L_2 Wasserstein distance metric of the posterior distribution of G is $n^{-1/4}$.*

Proof. Let $G = \sum_{i=1}^k p_i\delta_{\theta_i}$. Combining Lemma 1 with Assumption (A2), if $\|\theta_i - \theta_i^*\| \leq \epsilon$ and $|p_i - p_i^*| \leq \epsilon^2/(k\text{Diam}(\Theta)^2)$ for $i = 1, \ldots, k$, then $K(p_{G_0}, p_G) \leq d_{\rho_K}(G_0, G) \leq C_1 \sum_{1\leq i,j\leq k} q_{ij}\|\theta_i^* - \theta_j|^2$, for any $\boldsymbol{q} \in \mathcal{Q}$. Thus, $K(p_{G_0}, p_G) \leq C_1 W_2^2(G_0, G) \leq C_1 \sum_{i=1}^k (p_i^* \wedge p_i)\|\theta_i^* - \theta_i\|^2 + C_1 \sum_{i=1}^k |p_i - p_i^*|\text{Diam}(\Theta)^2 \leq 2C_1\epsilon^2$. Hence, under prior Π,

$$\begin{aligned}&\Pi(G : K(p_{G_0}, p_G) \leq \epsilon^2)\\ &\qquad\geq \Pi(G : \|\theta_i - \theta_i^*\| \leq \epsilon, |p_i - p_i^*| \leq \epsilon^2/(k\text{Diam}(\Theta)^2), i = 1, \ldots, k).\end{aligned}$$

In view of Assumptions (A2) and (A3), we have $\Pi(B_K(\epsilon)) \gtrsim \epsilon^{k(d+2)}$. Conversely, for sufficiently small ϵ, if $W_2(G_0, G) \leq \epsilon$ then by reordering the index of the atoms, we must have $\|\theta_i - \theta_i^*\| = O(\epsilon)$ and $|p_i - p_i^*| = O(\epsilon^2)$ for all $i = 1, \ldots, k$ (see the argument in the proof of Lemma 4(c)). This entails that under the prior Π,

$$\begin{aligned}\Pi(G : W_2^2(G_0, G) &\leq \epsilon^2) \leq \Pi(G : \|\theta_i - \theta_i^*\| \leq O(\epsilon), |p_i - p_i^*| \leq O(\epsilon^2), i = 1, \ldots, k)\\ &\lesssim \epsilon^{k(d+2)}.\end{aligned}$$

Let $\epsilon_n = n^{-1/2}$. We proceed by verifying conditions of Theorem 4. Let $\mathcal{G}_n := \mathcal{G}_k(\Theta)$. Then $\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) = 0$, so Eq. (19) trivially holds.

Next, we show that both $D(\epsilon/2, S, W_2)$, where $S = \{G \in \mathcal{G}_n : W_2(G_0, G) \leq 2\epsilon\}$, and $M(\mathcal{G}_n, G_1, \epsilon)$ are bounded above by a constant, so that (18) is satisfied. Indeed, for any $\epsilon > 0$, $\log D(\epsilon/2, S, W_2) \leq \log N(\epsilon/4, S, W_2) \leq N(\epsilon/4, S, W_1)$. By Lemma 4 (c), $N(\epsilon/4, S, W_1)$ is bounded in terms of $\sup_{\Theta'} \log N(\epsilon/8, \Theta', \rho)$, which is bounded above by a constant when Θ''s are subsets of Θ whose diameter is bounded by a multiple of ϵ.

A similar argument holds for $M(\mathcal{G}_n, G_1, \epsilon)$: Due to strong identifiability and assumption (A2), $C_k(G_n, G_1, \epsilon) \geq c\epsilon^4$ for some constant $c > 0$. By assumption (A4), and an application of Lemma 4 (c), for some constant $c_1 > 0$, $M(\mathcal{G}_n, G_1, \epsilon) \leq N(c_1\epsilon, \mathcal{G}_k(\Theta) \cap B_W(G_1, \epsilon/2), W_2)$ is bounded by a constant. Thus, Eq. (18) holds.

By Proposition 1(b) and Assumption (A4), we have

$$C_k(\mathcal{G}_n, j\epsilon_n) = \inf_{W_2(G_0,G)\geq j\epsilon/2} h^2(p_{G_0}, p_G) \geq c(j\epsilon_n)^4$$

for some constant $c > 0$. To ensure condition (21), note that (constants c change after each bounding step):

$$\begin{aligned}&\exp(2n\epsilon_n^2) \sum_{j\geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/16] \lesssim \exp(2n\epsilon_n^2) \sum_{j\geq M_n} \exp[-nc(j\epsilon_n)^4] \\ &\lesssim \exp(2n\epsilon_n^2 - ncM_n^4\epsilon_n^4).\end{aligned}$$

This upper bound goes to zero if $ncM_n^4\epsilon_n^4 \geq 4n\epsilon_n^2$, which is satisfied by taking M_n to be a large multiple of $\epsilon_n^{-1/2}$. Thus we need $M_n\epsilon_n \asymp \epsilon_n^{1/2} = n^{-1/4}$.

Under the assumptions specified above, $\Pi(G : j\epsilon_n < W_2(G, G_0) \leq 2j\epsilon_n)/\Pi(B_K(\epsilon_n)) = O(1)$. On the other hand, for $j \geq M_n$, we have $\exp[nC_k(\mathcal{G}_n, j\epsilon_n)/16] \geq \exp[nc(j\epsilon_n)^4/16]$ which is bounded below by arbitrarily large constant by choosing M_n to be a large multiple of $\epsilon_n^{-1/2}$, thereby ensuring (20).

Thus, by Theorem 4, rate of contraction for the posterior distribution of G under W_2 distance metric is $n^{-1/4}$, which is also the minimax rate $n^{-1/4}$ as proved in the univariate case by [5]. □

4.2 Dirichlet process mixtures

Given the "true" discrete measure $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$, where Θ is a metric space but $k \leq \infty$ is unknown. To estimate G_0, the prior distribution Π on discrete measure $G \in \bar{\mathcal{G}}(\Theta)$ is taken to be a Dirichlet process DP(ν, P_0) that centers at P_0 with concentration parameter $\nu > 0$ [9]. Here, parameter P_0 is a probability measure on Θ. For any $m \geq 1$, the following lemma provides a lower bound of small ball probabilities of metric space $(\bar{\mathcal{G}}(\Theta), d_{\rho^m}^{1/m})$ in terms of small ball P_0-probabilities of metric space (Θ, ρ).

Lemma 5. *Let $G \sim DP(\nu, P_0)$, where P_0 is a non-atomic base probability measure on a compact set Θ. For a small $\epsilon > 0$, let $D = D(\epsilon, \Theta, \rho)$ denote the packing number of Θ under ρ metric. Then, under the Dirichlet process distribution,*

$$\Pi(G : d_{\rho^m}(G_0, G) \leq (2^m + 1)\epsilon^m) \geq \Gamma(\nu)[\epsilon^m(2D)^{-1}\mathit{Diam}(\Theta)^{-m}]^{D-1}\nu^D \prod_{i=1}^D P_0(S_i).$$

Here, $(S_1, \ldots, S_D)$ denotes the D disjoint $\epsilon/2$-balls that form a maximal packing of Θ. $\Gamma(\cdot)$ is the gamma function.

Proof. Since every point in Θ is of distance at most ϵ to one of the centers of $S_1, \ldots, S_D$, there is a D-partition $(S'_1, \ldots, S'_D)$ of Θ, such that $S_i \subseteq S'_i$, and $\text{Diam}(S'_i) \leq 2\epsilon$ for each $i = 1, \ldots, D$. Let $m_i = G(S'_i)$, $\mu_i = P_0(S'_i)$, and $\hat{p}_i = G_0(S'_i)$. From the definition of Dirichlet processes, $\boldsymbol{m} = (m_1, \ldots, m_D) \sim \text{Dir}(\nu\mu_1, \ldots, \nu\mu_D)$. To obtain an upper bound for $d_{\rho^m}(G_0, G)$, consider a coupling between G_0 and G, by associating $m_i \wedge \hat{p}_i$ probability mass of supporting atoms for G_0 contained in subset S'_i with the same probability mass of supporting atoms for G contained in the same subset, for each $i = 1, \ldots, D$. The remaining mass (of probability $\|\boldsymbol{m} - \hat{\boldsymbol{p}}\|$) for both measures are coupled in an arbitrary way. The expectation under this coupling of the ρ^m distance provides one such upper bound, i.e.:

$$d_{\rho^m}(G_0, G) \leq (2\epsilon)^m + \|\boldsymbol{m} - \hat{\boldsymbol{p}}\|_1 [\text{Diam}(\Theta)]^m.$$

Due to the non-atomicity of P_0, for ϵ sufficiently small, $\nu\mu_i \leq 1$ for all $i = 1, \ldots, D$. Let $\delta = \epsilon/\text{Diam}(\Theta)$. Then, under Π,

$$\begin{aligned}
&\Pr(d_{\rho^m}(G_0, G) \leq (2^m + 1)\epsilon^m) \geq \Pr(\|\boldsymbol{m} - \hat{\boldsymbol{p}}\|_1 \leq \delta^m) \\
&\geq \Pr(|m_i - \hat{p}_i| \leq \delta^m/2D, i = 1, \ldots, D-1) \\
&= \frac{\Gamma(\nu)}{\prod_{i=1}^D \Gamma(\nu\mu_i)} \int_{\Delta_{D-1} \cap |m_i - \hat{p}_i| \leq \delta^m/2D} \prod_{i=1}^{D-1} m_i^{\nu\mu_i - 1} (1 - \sum_{i=1}^{D-1} m_i)^{\nu\mu_D - 1} dm_i \\
&\geq \frac{\Gamma(\nu)}{\prod_{i=1}^D \Gamma(\nu\mu_i)} \prod_{i=1}^{D-1} \int_{\max(\hat{p}_i - \delta^m/2D, 0)}^{\min(\hat{p}_i + \delta^m/2D, 1)} m_i^{\nu\mu_i - 1} dm_i \\
&\geq \Gamma(\nu)(\delta^m/2D)^{D-1} \prod_{i=1}^{D} (\nu\mu_i).
\end{aligned}$$

The second inequality in the previous display is due to the fact that $\|\boldsymbol{m} - \hat{\boldsymbol{p}}\|_1 \leq 2\sum_{i=1}^{D-1} |m_i - \hat{p}_i|$. The third inequality is due to $(1 - \sum_{i=1}^{D-1} m_i)^{\nu\mu_D - 1} = m_D^{\nu\mu_D - 1} \geq 1$, since $\nu\mu_D \leq 1$ and $0 < m_D < 1$ almost surely. The last inequality is due to the fact that $\Gamma(\alpha) \leq 1/\alpha$ for $0 < \alpha \leq 1$. This gives the desired claim. □

Assumptions B.

(B1) The non-atomic base measure P_0 places full support on a compact set $\Theta \subset \mathbb{R}^d$. Moreover, P_0 has a Lebesgue density that is bounded away from zero.

(B2) For some constants $C_1, m_1 > 0$, $K(f_i, f'_j) \leq C_1 \rho^{m_1}(\theta_i, \theta'_j)$ for any $\theta_i, \theta'_j \in \Theta$.

For any $G \in \text{supp}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 \leq C_2 K(p_{G_0}, p_G)^{m_2}$ for some constants $C_2, m_2 > 0$.

Theorem 6. *Given Assumptions (B1) and (B2) and the smoothness conditions for the likelihood family as specified in Theorem 2, there is a sequence $\beta_n \searrow 0$ such that $\Pi(W_2(G_0, G) \geq \beta_n | X_1, \ldots, X_n) \to 0$ in P_{G_0} probability. Specifically,*

(1) For ordinary smooth likelihood functions, take $\beta_n \asymp (\log n/n)^{\frac{2}{(d+2)(4+(2\beta+1)d)+\delta}}$*, for any small* $\delta > 0$*.*

(2) For supersmooth likelihood functions, take $\beta_n \asymp (\log n)^{-1/\beta}$*.*

Proof. The proof consists of two main steps. First, we shall prove that under Assumptions (B1–B2), conditions specified by Eqs. (13) (14) (15) in Theorem 3 are satisfied by taking $\mathcal{G}_n = \bar{\mathcal{G}}(\Theta)$, which is a convex set, and ϵ_n to be a large multiple of $(\log n/n)^{1/(d+2)}$. The second step involves constructing a sequence of M_n and $\beta_n = M_n\epsilon_n$ for which Theorem 3 can be applied.

Step 1: By Lemma 1 and (B2), $K(p_{G_0}, p_G) \leq d_{\rho_K}(G_0, G) \leq C_1 d_{\rho^{m_1}}(G_0, G)$. Also, $\int p_{G_0}[\log(p_{G_0}/p_G)]^2 \lesssim C_2 d_{\rho^{m_1}}(G_0, G)^{m_2}$. Assume without loss of generality that $m_1 \leq m_2$ (the other direction is handled similarly). We obtain that $\Pi(G \in B_K(\epsilon_n)) \geq \Pi(G : d_{\rho^{m_1}}(G_0, G) \leq C_3\epsilon_n^{2\vee 2/m_2})$ for some constant $C_3 > 0$.

From (B1), there is a universal constant $c_3 > 0$ such that for any ϵ and any D-partition $(S_1, \ldots, S_D)$ specified in Lemma 5, there holds:

$$\log \prod_{i=1}^{D} P_0(S_i) \geq c_3 D \log(1/D).$$

Moreover, the packing number satisfies $D \asymp [\text{Diam}(\Theta)/\epsilon_n]^d$. Combining these facts with Lemma 5 we have: $\log \Pi(G \in B_K(\epsilon_n)) \gtrsim (D-1)\log(\epsilon_n/\text{Diam}(\Theta)) + (2D-1)\log(1/D) + D\log\nu$, where the approximation constant is dependent on m_1, m_2. It is simple to check that condition (15) holds, $\log \Pi(G \in B_K(\epsilon_n)) \geq -Cn\epsilon_n^2$, by the given rate of ϵ_n, for any constant $C > 0$.

Since $\mathcal{G}_n = \bar{\mathcal{G}}(\Theta)$, (14) trivially holds. Turning to condition (13), by Lemma 4(b), we have $\log(2\epsilon_n, \bar{\mathcal{G}}(\Theta), W_2) \leq \log N(2\epsilon_n, \bar{\mathcal{G}}(\Theta), d_\rho) \leq N(\epsilon_n, \Theta, \rho)\log(e + e\text{Diam}(\Theta)/\epsilon_n) \leq (\text{Diam}(\Theta)/\epsilon_n)^d \log(e + e\text{Diam}(\Theta)/\epsilon_n) \leq n\epsilon_n^2$ by the specified rate of ϵ_n.

Step 2: For any $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$, let $R_k(\mathcal{G}, r)$ be the inverse of the Hellinger information function of d_ρ metric. Specifically, for any $t \geq 0$,

$$R_k(\mathcal{G}, t) = \inf\{r \geq 0 | C_k(\mathcal{G}, r) \geq t\}.$$

Note that $R_k(\mathcal{G}, 0) = 0$. $R_k(\mathcal{G}, \cdot)$ is non-decreasing because $C_k(\mathcal{G}, \cdot)$ is.

Let $(\epsilon_n)_{n\geq 1}$ be the sequence determined in the previous step of the proof. Let $M_n = R_k(\bar{\mathcal{G}}(\Theta), 8\epsilon_n^2(C+4))/\epsilon_n$, and $\beta_n = M_n\epsilon_n = R_k(\bar{\mathcal{G}}(\Theta), 8\epsilon_n^2(C+4))$. Condition (16) holds by definition of R_k, i.e., $C_k(\mathcal{G}(\Theta), M_n\epsilon_n) \geq 8\epsilon_n^2(C+4)$. To verify (17), note that the running sum with respect to j cannot have more than $\text{Diam}(\Theta)/\epsilon_n$ terms, and due to the monotonicity of C_k, we have

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/8]$$
$$\leq \text{Diam}(\Theta)/\epsilon_n \exp(2n\epsilon_n^2 - nC_k(\mathcal{G}_n, M_n\epsilon_n)/8) \to 0.$$

Hence, Theorem 3 can be applied to conclude that $\Pi(d_\rho(G_0, G) \geq \beta_n | X_1, \ldots, X_n) \to 0$ in P_{G_0} probability. Under ordinary smoothness condition (as specified in Theorem 2), $R_k(\bar{\mathcal{G}}(\Theta), t) = t^{\frac{1}{4+(2\beta+1)d+\delta}}$, where δ is an arbitrarily positive constant. So,

$$\beta_n \asymp \epsilon_n^{\frac{2}{4+(2\beta+1)d+\delta}} = (\log n/n)^{\frac{2}{(d+2)(4+(2\beta+1)d+\delta)}}.$$

On the other hand, under supersmoothness condition, $R_k(\bar{\mathcal{G}}(\Theta), t) = (1/\log(1/t))^{1/\beta}$. So, $\beta_n \asymp (\log(1/\epsilon_n))^{-1/\beta} \asymp (\log n)^{-1/\beta}$. □

5 Proofs of main results

5.1 Identifiability results

Proof of Theorem 1.

Proof. Suppose that Eq. (4) is not true, then there will be sequences of G_n and G'_n tending to G_0 in W_2 metric, and that $\psi(G_n, G'_n) \to 0$. We write $G_n = \sum_{i=1}^{\infty} p_{n,i}\delta_{\theta_{n,i}}$, where $p_{n,i} = 0$ for indices i greater than k_n, the number of atoms of G_n. Similar notation is applied to G'_n. Since both G_n and G'_n have finite number of atoms, there is $\boldsymbol{q}^{(n)} \in \mathcal{Q}(\boldsymbol{p}_n, \boldsymbol{p'}_n)$ so that $W_2^2(G_n, G'_n) = \sum_{ij} q_{ij}^{(n)} \|\theta_{n,i} - \theta'_{n,j}\|^2$.

Let $\mathcal{O}_n = \{(i,j) : \|\theta_{n,i} - \theta'_{n,j}\| \leq W_2(G_n, G'_n)\}$. This set exists because there are pairs of atoms $\theta_{n,i}, \theta'_{n,j}$ such that $\|\theta_{n,i} - \theta'_{n,j}\|$ is bounded away from zero in the limit. Since $\boldsymbol{q}^{(n)} \in \mathcal{Q}(\boldsymbol{p}_n, \boldsymbol{p'}_n)$, we can express

$$\begin{aligned}
\psi(G_n, G'_n) &= \sup_x \left| \sum_{i=1}^{k_n} p_{n,i} f(x|\theta_{n,i}) - \sum_{j=1}^{k'_n} p'_{n,j} f(x|\theta'_{n,j}) \right| / W_2^2(G_n, G'_n) \\
&= \sup_x \left| \sum_{ij} q_{ij}^{(n)} (f(x|\theta_{n,i}) - f(x|\theta'_{n,j})) \right| / W_2^2(G_n, G'_n),
\end{aligned}$$

and, by Taylor's expansion,

$$\begin{aligned}
\psi(G_n, G'_n) &= \sup_x \Bigg| \sum_{(i,j)\notin\mathcal{O}_n} q_{ij}^{(n)} (f(x|\theta'_{n,j}) - f(x|\theta_{n,i})) + \\
&\quad \sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)} (\theta'_{n,j} - \theta_{n,i})^T Df(x|\theta_{n,i}) \\
&\quad \sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)} (\theta'_{n,j} - \theta_{n,i})^T D^2 f(x|\theta_{n,i})(\theta'_{n,j} - \theta_{n,i}) + \\
&\quad R_n(x) \Bigg| / W_2^2(G_n, G'_n) \\
&=: \sup_x |A_n(x) + B_n(x) + C_n(x) + R_n(x)| / D_n,
\end{aligned}$$

where

$$R_n(x) = O\Big(\sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}\|\theta_{n,i} - \theta'_{n,j}\|^{2+\delta}\Big) = o\Big(\sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}\|\theta_{n,i} - \theta'_{n,j}\|^{2}\Big)$$

due to Eq. (3) and the definition of $\mathcal{O}_n$. So $R_n(x)/W_2^2(G_n, G'_n) \to 0$.

The quantities $A_n(x), B_n(x)$ and $C_n(x)$ are linear combinations of elements of $f(x|\theta)$, $Df(x|\theta)$ and $D^2f(x|\theta)$ for different θ's, respectively. Since Θ is compact, subsequences of G_n and G'_n can be chosen so that each of their support points converges to a fixed atom θ_l^*, for $l = 1, \ldots, k^* \leq k$. After being rescaled, the limits of $A_n(x)/D_n, B_n(x)/D_n$ and $C_n(x)/D_n$ are still linear combinations with constant coefficients not depending on x.

We shall now argue that not all such coefficients vanish to zero. Suppose this is not the case. It follows that for the coefficients of $C_n(x)/D_n$ we have:

$$\sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}\|\theta'_{n,j} - \theta_{n,i}\|^2/W_2^2(G_n, G'_n) \to 0.$$

This implies that $\sum_{(i,j)\notin\mathcal{O}_n} q_{ij}^{(n)}\|\theta'_{n,j} - \theta_{n,i}\|^2/W_2^2(G_n, G'_n) \to 1$. Since Θ is bounded, there exists a pair $(i, j) \notin \mathcal{O}_n$ such that $q_{ij}^n/W_2^2(G_n, G'_n)$ does not vanish to zero. But then, one of the coefficients of $A_n(x)/D_n$ does not vanish to zero, which contradicts the hypothesis.

Next, we observe that some of the coefficients of $A_n(x)/D_n, B_n(x)/D_n$ and $C_n(x)/D_n$ may tend to infinity. For each n, let d_n be the inverse of the maximum coefficient of $A_n(x)/D_n$, $B_n(x)/D_n$, and $C_n(x)/D_n$. From the conclusion in the preceding paragraph, $|d_n|$ is uniformly bounded from above by a constant for all n. Moreover $d_nA_n(x)/D_n$ converges to $\sum_{j=1}^{k^*}\alpha_j f(x|\theta_j^*)$ and $d_nB_n(x)/D_n$ converges to $\sum_{j=1}^{k^*}\beta_j^T Df(x|\theta_j^*)$, and $d_nC_n(x)/D_n$ converges to $\sum_{j=1}^{k^*}\gamma_j D^2f(x|\theta_j^*)\gamma_j$, for some finite α_j, β_j and γ_j, not all of them vanishing (in fact, at least one is them is 1). We have

$$d_n|p_{G_n}(x) - p_{G'_n}(x)|/d_\rho^2(G_n, G'_n) \to \Big|\sum_{j=1}^{k^*}\alpha_j f(x|\theta_j^*) + \beta_j^T Df(x|\theta_j^*) + \gamma_j^T D^2f(x|\theta_j^*)\gamma_j\Big|. \tag{22}$$

for all x. This entails that the right side of the preceeding display must be 0 for almost all x. By strong identifiability, all coefficients must be 0, which leads to contradiction.

With respect to $\psi_1(G, G')$, suppose that the claim is not true, which implies the existence of a subsequence G_n, G'_n such that that $\psi_1(G_n, G'_n) \to 0$. Going through the same argument as above, we have $\alpha_j, \beta_j, \gamma_j$, not all of which are zero, such that Eq.(22) holds. An application of Fatou's lemma yields $\int|\sum_{j=1}^{k^*}\alpha_j f(x|\theta_j) + \beta_j^T Df(x|\theta_j) + \gamma_j^T D^2f(x|\theta_j)\gamma_j|d\mu = 0$. Thus the integrand must be 0 for almost all x, leading to contradiction. □

Proof of Theorem 2.

Proof. To obtain an upper bound of $W_2^2(G, G') = d_{\rho^2}(G, G')$ in terms of $V(p_G, p_{G'})$ under the condition that $V(p_G, p_{G'}) \to 0$, our strategy is approximate G and G' by convolving these with some mollifier K_δ. By triangular inequality, $W_2(G, G') \leq W_2(G, G * K_\delta) + W_2(G', G' * K_\delta) + W_2(G * K_\delta, G' * K_\delta)$. The first two terms are simple to bound, while the last term can be handled by expressing $G * K_\delta$ as the convolution the mixture density p_G with another function. We also need the following elementary lemma (whose proof is given Section 5.3).

Lemma 6. *Assume that p and p' are two probability density functions on $\mathbb{R}^d$ with bounded s-moments.*

(a) For t such that $0 < t < s$,

$$\int |p(x) - p'(x)| \|x\|^t dx \leq 2\|p - p'\|_{L_1}^{(s-t)/s} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)^{t/s}.$$

(b) Let $V_d = \pi^{d/2}\Gamma(d/2 + 1)$ denote the volume of the d-dimensional unit sphere. Then,

$$\|p - p'\|_{L_1} \leq 2V_d^{s/(d+2s)} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)^{\frac{d}{d+2s}} \|p - p'\|_{L_2}^{\frac{2s}{d+2s}}.$$

Take any $s > 0$, and let $K : \mathbb{R}^d \to (0, \infty)$ be a symmetric density function on $\mathbb{R}^d$ whose Fourier transform $\tilde{K}$ is a continuous function whose support is bounded in $[-1, 1]^d$. Moreover, K has bounded moments up to order s. Consider molifiers $K_\delta(x) = \frac{1}{\delta^d} K(x/\delta)$ for $\delta > 0$. Let $\tilde{K}_\delta$ and $\tilde{f}$ be the Fourier transforms for K_δ and f, respectively. Define g_δ to be the inverse Fourier transform of $\tilde{K}_\delta / \tilde{f}$:

$$g_\delta(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\langle \omega, x \rangle} \frac{\tilde{K}_\delta(\omega)}{\tilde{f}(\omega)} d\omega.$$

Note that function $\tilde{K}_\delta(\omega)/\tilde{f}(\omega)$ has bounded support. So, $g_\delta \in L_1(\mathbb{R})$, and $\tilde{g}_\delta := \tilde{K}_\delta(\omega)/\tilde{f}(\omega)$ is the Fourier transform of g_δ. By the convolution theorem, $f * g_\delta = K_\delta$. As a result,

$$G * K_\delta = G * f * g_\delta = p_G * g_\delta.$$

From the definition of K_δ, the second moment under K_δ is $O(\delta^2)$. Consider a coupling G and $G * K_\delta$ under which we have a pair of random variables $(\theta, \theta + \epsilon)$ where ϵ is independent of θ, the marginal distributions of θ and ϵ are G K_δ, respectively. Under this coupling, $\mathbb{E}\|(\theta + \epsilon) - \theta\|^2 = O(\delta^2)$, which entails that $W_2^2(G, G * K_\delta) = O(\delta^2)$.

By triangular inequality, $W_2(G, G') \leq W_2(G * K_\delta, G' * K_\delta) + O(\delta)$, so for some constant $C(K) > 0$ dependent only on kernel K,

$$W_2^2(G, G') \leq 2W_2^2(G * K_\delta, G' * K_\delta) + C(K)\delta^2. \tag{23}$$

Theorem 6.15 of [32] provides an upper bound for the Wasserstein distance: for any two probability measures μ and ν, $W_2^2(\mu,\nu) \leq 2\int \|x\|^2 d|\mu-\nu|(x)$, where $|\mu-\nu|$ is the total variation of measure $|\mu-\nu|$. Thus,

$$W_2^2(G*K_\delta, G'*K_\delta) \quad \leq \quad 2\int \|x\|^2 |G*K_\delta(x) - G'*K_\delta(x)| dx. \tag{24}$$

We note that since density function K has bounded s-th moment, $\int \|x\|^s G*K_\delta(dx) \leq 2^s\Big[\int \|\theta\|^s dG(\theta) + \int \|x\|^s K_\delta(x)dx\Big] = 2^s\Big[\int \|\theta\|^s dG(\theta) + \delta^s \int \|x\|^s K(x)dx\Big] < \infty$, because G's support points lie in a compact set. Applying Lemma 6 to Eq.(24), we obtain that for $\delta < 1$,

$$\begin{aligned} W_2^2(G*K_\delta, G'*K_\delta) &\leq C(d,K,s)\|G*K_\delta - G'*K_\delta\|_{L_1}^{(s-2)/s} \\ &\leq C(d,K,s)\|G*K_\delta - G'*K_\delta\|_{L_2}^{2(s-2)/(d+2s)}. \end{aligned} \tag{25}$$

Here, constants $C(d,K,s)$ are different in each line, and they are dependent only on d, s and the s-th moment of density function K.

Next, we use a known fact that for arbitrary (signed) measure μ on $\mathbb{R}^d$ and function $g \in L_2(\mathbb{R}^d)$, there holds $\|\mu * g\|_{L_2} \leq |\mu|\|g\|_{L_2}$, where $|\mu|$ denotes the total variation of μ:

$$\|G*K_\delta - G'*K_\delta\|_{L_2} = \|p_G*g_\delta - p_{G'}*g_\delta\|_{L_2} = \|(p_G - p_{G'})*g_\delta\|_{L_2} \leq 2V(p_G,p_{G'})\|g_\delta\|_{L_2}. \tag{26}$$

By Plancherel's identity,

$$\|g_\delta\|_{L_2}^2 = \frac{1}{(2\pi)^d}\int \frac{\tilde{K}_\delta(\omega)^2}{\tilde{f}(\omega)^2}d\omega = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{\tilde{K}(\omega\delta)^2}{\tilde{f}(\omega)^2}d\omega \leq C\int_{[-1/\delta,1/\delta]^d} \tilde{f}(\omega)^{-2}d\omega.$$

The last bound holds because $\tilde{K}$ has support in $[-1,1]^d$, and is bounded by a constant.

Collecting Eqs.(23)(24)(25)(26) and the preceeding display, we have:

$$W_2^2(G,G') \leq C(d,K,s)\Big\{\inf_{\delta\in(0,1)} \delta^2 + V(p_G,p_{G'})^{\frac{2(s-2)}{d+2s}}\Big[\int_{[-1/\delta,1/\delta]^d} \tilde{f}(\omega)^{-2}d\omega\Big]^{\frac{s-2}{d+2s}}\Big\}.$$

If $|\tilde{f}(\omega)\prod_{j=1}^d |\omega_j|^\beta| \geq d_0$ as $\omega_j \to \infty (j=1,\ldots,d)$ for some positive constant d_0, then

$$\begin{aligned} W_2^2(G,G') &\leq C(d,K,s,\beta)\Big\{\inf_{\delta\in(0,1)} \delta^2 + V(p_G,p_{G'})^{\frac{2(s-2)}{d+2s}}(1/\delta)^{\frac{(2\beta+1)d(s-2)}{d+2s}}\Big\} \\ &\leq C(d,K,s,\beta)V(p_G,p_{G'})^{\frac{4(s-2)}{2(d+2s)+(2\beta+1)d(s-2)}}. \end{aligned}$$

The exponent tends to $4/(4+(2\beta+1)d)$ as $s \to \infty$, we obtain that $W_2^2(G,G') \leq C(d,\beta,r)V(p_G,p_{G'})^r$, for any constant $r < 4/(4+(2\beta+1)d)$, as $V(p_G,p_{G'}) \to 0$.

If $|\tilde{f}(\omega)\prod_{j=1}^{d}\exp(|\omega_j|^\beta) \geq d_0$ as $\omega_j \to \infty (j=1,\ldots,d)$ for some positive constants β, d_0, then

$$W_2^2(G,G') \quad \leq \quad C(d,K,s,\beta)\Big\{\inf_{\delta\in(0,1)} \delta^2 + V(p_G,p_{G'})^{2(s-2)/(d+2s)} \exp -2d\delta^{-\beta}\frac{s-2}{d+2s}\Big\}.$$

Taking $\delta^{-\beta} = -\frac{1}{d}\log V(p_G,p_{G'})$, we obtain that

$$d_{\rho^2}(G,G') \leq C(d,\beta)(-\log V(p_G,p_{G'}))^{-2/\beta}.$$

□

Proof of Lemma 1.

Proof. We exploit the variational characterization of f-divergences [24],

$$\rho_\phi(f_i,f_j') = \sup_{\varphi_{ij}} \int \varphi_{ij} f_j' - \phi^*(\varphi_{ij}) f_i d\mu,$$

where the infimum is taken over all measurable functions defined on $\mathcal{X}$. ϕ^* denotes the Legendre-Fenchel conjugate dual of convex function ϕ. (ϕ^* is again a convex function on $\mathbb{R}$ and is defined by $\phi^*(v) = \sup_{u\in\mathbb{R}}(uv - \phi(u))$.)

By the variational characterization, ρ_ϕ is a convex functional (jointly of its two arguments). Thus, for any coupling Q of two mixing measures G and G', $\rho_\phi(p_G,p_{G'}) = \rho_\phi(\int f(\cdot|\theta)dG, \int f(\cdot|\theta')dG') = \rho_\phi(\int f(\cdot|\theta)dQ, \int f(\cdot|\theta)dQ) \leq \int \rho_\phi(f(\cdot|\theta), f(\cdot|\theta'))dQ$, where the inequality is obtained via Jensen's inequality. Since this holds for any Q, the desired bound follows. □

Proof of Proposition 1.

Proof. (a) Suppose that the claim is not true, there is a sequence of $(G_0,G) \in \mathcal{G}_k(\Theta)\times\mathcal{G}$ such that $W_2(G_0,G_2) \geq r/2 > 0$ always holds, and that converges in W_2 metric to $G_0^* \in \mathcal{G}_k$ and $G^* \in \mathcal{G}$, respectively. This is due to the compactness of both $\mathcal{G}_k(\Theta)$ and $\mathcal{G}$. We must have $W_2(G_0^*,G^*) \geq r/2 > 0$, so $G_0^* \neq G^*$. At the same time, $h(p_{G_0^*},p_{G^*}) = 0$, which implies that $p_{G_0^*} = p_{G^*}$ for almost all $x \in \mathcal{X}$. By finite identifiability condition, $G_0^* = G^*$, which is a contradiction.

(b) is an immediate consequence of Theorem 1, by noting that under the given hypothesis, there is $c(k) > 0$ depending on k, such that $d_h^2(p_{G_0},p_G) \geq V^2(p_{G_0},p_G)/2 \geq c(k,G_0)W_2^4(G_0,G) \geq c(k,G_0)W_2^4(G_0,G)$ for sufficiently small $W_2(G_0,G)$. The boundedness of Θ implies the boundedness of $W_2(G_0,G)$, thereby extending the claim for the entire admissible range of $W_2(G_0,G)$. (c) is obtained in a similar way from Theorem 2. □

5.2 Proofs of Theorems 3 and 4

We outline in this section the proofs of Theorems 3 and 4. Our proof follows the same steps as in [13], with suitable modifications for the inclusion of the Hellinger information function, and the conditions involving latent mixing measures. The proof consists of results on the existence of tests, which are turned into probability bounds on the posterior contraction.

Proof of Lemma 2.

Proof. We first consider Case (I). Define $\mathcal{P}_1 = \{p_G | G \in \mathcal{G} \cap B_W(G_1, r/2)\}$. Since ρ is a metric in Θ, it is a standard fact of Wasserstein metrics that $B_W(G_1, r/2)$ is a convex set. Since $\mathcal{G}$ is also convex, so is the set $\mathcal{G} \cap B_W(G_1, r/2)$. It follows that $\mathcal{P}_1$ is a convex set of mixture distributions. Now, applying a result from Birgé [3] and Le Cam ([19], Lemma 4, pg. 478) there exist tests φ_n that discriminate between P_{G_0} and convex set $\mathcal{P}_1$ such that:

$$P_{G_0}\varphi_n \leq \exp[-n \inf h^2(P_{G_0}, P_1)/2]. \tag{27}$$

$$\sup_{G \in \mathcal{G} \cap B_W(G_1, r/2)} P_G(1 - \varphi_n) \leq \exp[-n \inf h^2(P_{G_0}, P_1)/2], \tag{28}$$

where the exponent in the upper bounds are given by the infimum Hellinger distance among *all* $P_1 \in \operatorname{conv} \mathcal{P}_1 = \mathcal{P}_1$. Due to the triangle inequality, if $W_2(G_0, G_1) = r$ and $W_2(G_1, G) \leq r/2$ then $W_2(G_0, G) \geq r/2$. So,

$$C_k(\mathcal{G}, r) = \inf_{G \in \mathcal{G}: W_2(G_0, G) \geq r/2} h^2(p_{G_0}, p_G) \leq \inf h^2(p_{G_0}, P_1).$$

This completes the proof of Case (I).

Turning to Case (II), for a constant $c_0 > 0$ to be determined, consider a maximal $c_0 r$-packing in W_2 metric in set $\mathcal{G} \cap B_W(G_1, r/2)$. This yields a set of $M(\mathcal{G}, G_1, r) = D(c_o r, \mathcal{G} \cap B_W(G_1, r/2), W_2)$ points $\tilde{G}_1, \ldots, \tilde{G}_M \in \mathcal{G} \cap B_W(G_1, r/2)$. (In the following we drop the subscripts of $M(\cdot)$).

We note the following fact: For any $t = 1, \ldots, M$, if $G \in \mathcal{G} \cap B_W(G_1, r/2)$ and $W_2(G, \tilde{G}_t) \leq c_0 r$, by Lemma 1 we have $h^2(p_G, p_{\tilde{G}_t}) \leq d_{\rho_h^2}(G, \tilde{G}_t) \leq C_1 d_{\rho^{2\alpha}}(G, \tilde{G}_t) \leq C_1 \text{Diam}(\Theta)^{2(\alpha-1)} W_2^2(G, \tilde{G}_t) \leq C_1 \text{Diam}(\Theta)^{2(\alpha-1)} c_0^2 r^2$ (the second inequality is due to the condition on the likelihood functions); and so it follows that $h(p_{G_0}, p_G) \geq h(p_{G_0}, p_{\tilde{G}_t}) - h(p_G, p_{\tilde{G}_t}) \geq C_k(\mathcal{G}, r)^{1/2} - C_1 \text{Diam}(\Theta)^{2(\alpha-1)} c_0^2 r^2$. Choose $c_0 = \frac{C_k(\mathcal{G}, r)^{1/4}}{r \text{Diam}(\Theta)^{\alpha-1} \sqrt{2C_1}}$ so that the previous bounds become $h(p_G, p_{\tilde{G}_t}) \leq C_k(\mathcal{G}, r)^{1/2}/2 \leq h(p_{G_0}, p_{\tilde{G}_t})/2$ and $h(p_{G_0}, p_G) \geq C_k(\mathcal{G}, r)^{1/2}/2$.

For each pair of $G_0, \tilde{G}_t$, there exist tests $\omega_n^{(t)}$ of p_{G_0} versus the closed Hellinger ball $\{p_G : h(p_G, p_{\tilde{G}_t}) \leq h(p_{G_0}, p_{\tilde{G}_t})/2\}$ such that

$$P_{G_0}\omega_n^{(t)} \leq \exp[-nh^2(P_{G_0}, P_{\tilde{G}_t})/8],$$

$$\sup_{G \in \bar{\mathcal{G}}(\Theta): h(p_G, p_{\tilde{G}_t}) \leq h(p_{G_0}, p_{\tilde{G}_t})/2} P_G(1 - \omega_n^{(t)}) \leq \exp[-nh^2(P_{G_0}, P_{\tilde{G}_t})/8].$$

Consider the test $\varphi_n = \max_{t \le M} \omega_n^{(t)}$, then

$$\begin{aligned} P_{G_0}\varphi_n &\le M\exp[-nC_k(\mathcal{G}, r)/8], \\ \sup_{G \in \mathcal{G} \cap B_W(G_1, r/2)} P_G(1-\varphi_n) &\le \exp[-nC_k(\mathcal{G}, r)/8]. \end{aligned}$$

The first inequality is due to $\varphi_n \le \sum_{t=1}^M \omega_n^{(t)}$, and the second is due to the fact that for any $G \in \mathcal{G} \cap B_W(G_1, r/2)$ there is some $t \le M$ such that $W_2(G, \tilde{G}_t) \le c_0 r$, so that $h(p_G, p_{\tilde{G}_t}) \le h(p_{G_0}, p_{\tilde{G}_t})/2$. □

Proof of Lemma 3.

Proof. For a given $t \in \mathbb{N}$ choose a maximal $t\epsilon/2$-packing for set $S_t = \{G : t\epsilon < W_2(G_0, G) \le (t+1)\epsilon\}$. This yields a set S_t' of at most $D(t\epsilon/2, S_t, W_2)$ points. Moreover, every $G \in S_t$ is within distance $t\epsilon/2$ of at least one of the points in S_t'. For every such point $G_1 \in S_t'$, there exists a test ω_n satisfying Eqs. (7) and (8). Take φ_n to be the maximum of all tests attached this way to some point $G_1 \in S_t'$ for some $t \ge t_0$. Then, by union bound, and the fact that $D(\epsilon)$ is non-increasing,

$$P_{G_0}\varphi_n \le \sum_{t \ge t_0} \sum_{G_1 \in S_t'} M(\mathcal{G}, G_1, t\epsilon)\exp[-nC_k(\mathcal{G}, t\epsilon)/8] \le D(\epsilon) \sum_{t \ge t_0} \exp[-nC_k(\mathcal{G}, t\epsilon)/8]$$

$$\sup_{G \in \cup_{u \ge t_0} S_u} P_G(1-\varphi_n) \le \sup_{u \ge t_0} \exp[-nC_k(\mathcal{G}, u\epsilon)/8] \le \exp[-nC_k(\mathcal{G}, t_0\epsilon)/8],$$

where the last inequality is due the monotonicity of $C_k(\mathcal{G}, \cdot)$. □

Proof of Theorem 3 and 4 By a result of Ghosal et al [13] (Lemma 8.1, pg. 524), for every $\epsilon > 0$ and probability measure Π on the set $B_K(\epsilon)$ defined by Eq. (12), we have, for every $C > 0$,

$$P_{G_0}\left(\int \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} d\Pi(G) \le \exp(-(1+C)n\epsilon^2) \right) \le \frac{1}{C^2 n\epsilon^2}.$$

This entails that, for a fixed $C \ge 1$, there is an event A_n with P_{G_0}-probability at least $1 - (Cn\epsilon_n^2)^{-1}$, for which there holds:

$$\int \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i) d\Pi(G) \ge \exp(-2n\epsilon_n^2)\Pi(B_K(\epsilon_n)). \tag{29}$$

Let $\mathcal{O}_n = \{G \in \bar{\mathcal{G}}(\Theta) : W_2(G_0, G) \ge M_n\epsilon_n\}$, $S_{n,j} = \{G \in \mathcal{G}_n : W_2(G_0, G) \in [j\epsilon_n, (j+1)\epsilon_n)\}$ for each $j \ge 1$. The conditions specified by Lemma 3 are satisfied by

setting $D(\epsilon) = \exp(n\epsilon_n^2)$ (constant in ϵ). Thus there exist tests φ_n for which Eq. (10) and (11) hold. Then,

$$\begin{aligned} & P_{G_0}\Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n) \\ = \; & P_{G_0}[\varphi_n \Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n)] + P_{G_0}[(1-\varphi_n)\Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n)] \\ \leq \; & P_{G_0}[\varphi_n \Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n)] + P_{G_0}\mathbb{I}(A_n^c) \\ & + P_{G_0}[(1-\varphi_n)\Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n)\mathbb{I}(A_n)]. \end{aligned}$$

Exploiting Lemma 3, all terms in the preceeding display can be shown to vanish as $n \to \infty$. The proof for Theorem 3 proceeds in a similar way to Theorem 2.1 of [13], while the proof for Theorem 4 is similar to their Theorem 2.4.

5.3 Proof of other auxiliary lemmas

Proof of Lemma 4.

Proof. (a) Suppose that $(\eta_1, \ldots, \eta_T)$ forms an ϵ-covering for Θ under metric ρ, where $T = N(\epsilon, \Theta, \rho)$ denotes the (minimum) covering number. Take any discrete measure $G = \sum_{i=1}^k p_i \delta_{\theta_i}$. For each θ_i there is an approximating θ_i' among the η_j's such that $\rho(\theta_i, \theta_i') < \epsilon$. Let $\boldsymbol{p'} = (p_1', \ldots, p_k')$ be a k-dim vector in the probability simplex that deviates from $\boldsymbol{p}$ by less than δ in l_1 distance: $\|\boldsymbol{p'} - \boldsymbol{p}\|_1 \leq \delta$. Define $G' = \sum_{i=1}^k p_i' \delta_{\theta_i'}$. We shall argue that

$$d_\rho(G, G') \leq \sum_{i=1}^k (p_i \wedge p_i')\rho(\theta_i, \theta_i') + \|\boldsymbol{p} - \boldsymbol{p'}\|_1 \text{Diam}(\Theta) \leq \epsilon + \delta \text{Diam}(\Theta).$$

(To see this, consider the following coupling between G and G': associating $p_i \wedge p_i'$ probability mass of θ_i (from G) with the same probability mass of θ_i' (from G'), while the remaining mass from G and G' (of probability $\|\boldsymbol{p} - \boldsymbol{p'}\|_1$) are coupled in an arbitrary way. The right hand side of the previous display is an upper bound of the expectation of the ρ distance between two random variables distributed according to the described coupling.)

It follows that a $(\epsilon + \delta\text{Diam}(\Theta))$-covering for $\mathcal{G}_k(\Theta)$ can be constructed by combining each element of a δ-covering in l_1 metric of the $k-1$-probability simplex and k ϵ-covering's of Θ. Now, the covering number of $k-1$-probability simplex is less than the number of cubes of length δ/k covering $[0,1]^k$ times the volume of $\{(p_1', \ldots, p_k') : p_j' \geq 0, \sum_j p_j' \leq 1 + \delta\}$, i.e., $(k/\delta)^k(1+\delta)^k/k! \sim (1+1/\delta)^k e^k/\sqrt{2\pi k}$. It follows that $N(\epsilon + \delta\text{Diam}(\Theta), \mathcal{G}_k(\Theta), d_\rho) \leq T^k(1+1/\delta)^k e^k/\sqrt{2\pi k}$. Take $\delta = \epsilon/\text{Diam}(\Theta)$ to achieve the claim.

(b) As before, let $(\eta_1, \ldots, \eta_T)$ be an ϵ-covering of Θ. Take any $G = \sum_{i=1}^k p_i \delta_{\theta_i} \in \bar{\mathcal{G}}(\Theta)$, where k may be infinity. The collection of atoms $\theta_1, \ldots, \theta_k$ can be subdivided into disjoint subsets $S_1, \ldots, S_T$, some of which may be empty, so that for each $t = 1, \ldots, T$, $\rho(\theta_i, \eta_t) \leq \epsilon$ for any $\theta_i \in S_t$. Define $p_t' = \sum_{i=1}^k p_i \mathbb{I}(\theta_i \in S_t)$, and let $G' = \sum_{t=1}^T p_t' \delta_{\eta_t}$, then we are guaranteed that $d_\rho(G, G') \leq \sum_{i=1}^k \sum_{t=1}^T p_i \mathbb{I}(\theta_i \in S_t)\rho(\theta_i, \eta_t) \leq \epsilon$, by using a similar coupling argument as in part (a).

Let $\boldsymbol{p''} = (p''_1, \ldots, p''_T)$ be a T-dim vector in the probability simplex that deviates from $\boldsymbol{p'}$ by less than δ in l_1 distance: $\|\boldsymbol{p''} - \boldsymbol{p'}\|_1 \leq \delta$. Take $G'' = \sum_{t=1}^T p''_t \delta_{\eta_t}$. It is simple to observe that $d_\rho(G', G'') \leq \text{Diam}(\Theta)\delta$. By triangle inequality, $d_\rho(G, G'') \leq d_\rho(G, G') + d_\rho(G', G'') \leq \epsilon + \delta\text{Diam}(\Theta)$.

The foregoing arguments establish that an $(\epsilon + \delta\text{Diam}(\Theta))$-covering in the Wasserstein metric for $\bar{\mathcal{G}}(\Theta)$ can be constructed by combining each element of the δ-covering in l_1 of the $T-1$ simplex and a single covering of Θ. From the proof of part (a), $N(\epsilon + \delta\text{Diam}(\Theta), \bar{\mathcal{G}}(\Theta), d_\rho) \leq (1 + 1/\delta)^T e^T / \sqrt{2\pi T}$. Take $\delta = \epsilon/\text{Diam}(\Theta)$ to conclude.

(c) Consider a $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ such that $d_\rho(G_0, G) \leq 2\epsilon$. By definition, there is a coupling $q \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p^*})$ so that $\sum_{ij} q_{ij} \rho(\theta^*_i, \theta_j) \leq 2\epsilon$. Since $\sum_j q_{ij} = p^*_i$, this implies that $2\epsilon \geq \sum_{i=1}^k p^*_i \min_j \rho(\theta^*_i, \theta_j)$. Thus, for each $i = 1, \ldots, k$ there is a j such that $\rho(\theta^*_i, \theta_j) \leq 2\epsilon/p^*_i \leq 2M\epsilon$. Without loss of generality, assume that $\rho(\theta^*_i, \theta_i) \leq 2M\epsilon$ for all $i = 1, \ldots, k$. For sufficiently small ϵ, for any i, it is simple to observe that $d_\rho(G_0, G) \geq |p^*_i - p_i| \min_{j \neq i} \rho(\theta^*_i, \theta_j) \geq |p^*_i - p_i| \min_j \rho(\theta^*_i, \theta^*_j)/2$. Thus, $|p^*_i - p_i| \leq 4\epsilon/m$.

Now, an $\epsilon/4 + \delta\text{Diam}(\Theta)$ covering in d_ρ for $\{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}$ can be constructed by combining the $\epsilon/4$-covering for each of the k sets $\{\theta \in \Theta : \rho(\theta, \theta^*_i) \leq 2M\epsilon\}$ and the δ/k-covering for each of the k sets $[p^*_i - 4\epsilon/m, p^*_i + 4\epsilon/m]$. This entails that: $N(\epsilon/4 + \delta\text{Diam}(\Theta), \{G \in \mathcal{G}_k(\Theta) : d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq [\sup_{\Theta'} N(\epsilon/4, \Theta', \rho)]^k (8\epsilon k/m\delta)^k$. Take $\delta = \epsilon/(4\text{Diam}(\Theta))$ to conclude the proof.

□

Proof of Lemma 6.

Proof. (a) For arbitrary constant $R > 0$, we have $\int |p(x) - p'(x)| \|x\|^t dx \leq \int_{\|x\| \leq R} |p - p'| \|x\|^t + \int_{\|x\| \geq R} (p + p') \|x\|^t \leq R^t \|p - p'\|_{L_1} + R^{-(s-t)} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)$. Choosing $R = [(\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)/|p - p'|_{L_1}]^{1/s}$ to conclude.

(b) For any $R > 0$, we have $\int_{\|x\| \leq R} |p(x) - p'(x)| dx \leq V_d^{1/2} R^{d/2} [\int_{\|x\| \leq R} (p(x) - p'(x))^2 dx]^{1/2} \leq V_d^{1/2} R^{d/2} \|p - p'\|_{L_2}$. We also have $\int_{\|x\| \geq R} |p(x) - p'(x)| dx \leq \int_{\|x\| \geq R} p(x) + p'(x) dx \leq R^{-s} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)$. Thus, $\|p - p'\|_{L_1} \leq \inf_{R > 0} V_d^{1/2} R^{d/2} \|p - p'\|_{L_2} + R^{-s} (\mathbb{E}_p \|X\|^s + \mathbb{E}_{p'} \|X\|^s)$, which gives the desired bound. □

Acknowledgement. We wish to thank the Arash Amini, the Associate Editor, and in particular an anonymous reviewer for valuable comments and suggestions. The shortened proof of Lemma 1 is due to the reviewer, who also helped to point out a mistake in a previous version of Lemma 2. This work was supported in part by NSF grants CCF-1115769 and OCI-1047871.

References

[1] A. Barron, M. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist*, 27:536–561, 1999.

[2] P. Bickel and D. Freedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217, 1981.

[3] L. Birgé. Sur un théorèm de minimax et son application aux tests. *Probab. Math. Statist.*, 3:259–282, 1984.

[4] R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of American Statistical Association*, 83:1184–1186, 1988.

[5] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995.

[6] E. del Barrio, J. Cuesta-Albertos, C. Matrán, and J. Rodríguez-Rodríguez. Tests of goodness of fit based on the l_2-wasserstein distance. *Annals of Statistics*, 27(4):1230–1239, 1999.

[7] R. M. Dudley. *Probabilities and metrics: Convergence of laws on metric spaces, with a view to statistical testing*. Aarhus Universitet, 1976.

[8] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272, 1991.

[9] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[10] A.E. Gelfand, A. Kottas, and S.N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.*, 100:1021–1035, 2005.

[11] C. Genovese and L. Wasserman. Rates of convergence for the gaussian mixture sieve. *Annals of Statistics*, 28:1105–1127, 2000.

[12] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Ann. Statist.*, 27:143–158, 1999.

[13] S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[14] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.

[15] S. Ghosal and A. van der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *Ann. Statist.*, 35:697–723, 2007.

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

[17] H. Ishwaran, L. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of American Statistical Association*, 96(456):1316–1332, 2001.

[18] H. Ishwaran and M. Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963, 2002.

[19] L. LeCam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, 1986.

[20] B. Lindsay. *Mixture models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA, 1995.

[21] C. Mallows. A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43:508–515, 1972.

[22] G. McLachlan and K. Basford. *Mixture models: Inference and Applications to Clustering*. Marcel-Dekker, New York, 1988.

[23] X. Nguyen. Inference of global clusters from locally distributed data. *Bayesian Analysis*, 5(4):817–846, 2010.

[24] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[25] S. Petrone, M. Guidani, and A.E. Gelfand. Hybrid Dirichlet processes for functional data. *Journal of the Royal Statistical Society B*, 71(4):755–782, 2009.

[26] A. Rodriguez, D. Dunson, and A.E. Gelfand. The nested Dirichlet process. *J. Amer. Statist. Assoc.*, 103(483):1131–1154, 2008.

[27] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29:687–714, 2001.

[28] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

[29] H. Teicher. On the mixture of distributions. *Ann. Math. Statist.*, 31:55–73, 1960.

[30] H. Teicher. Identifiability of mixtures. *Ann. Math. Statist.*, 32:244–248, 1961.

[31] Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[32] Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.

[33] S. Walker. New approaches to bayesian consistency. *Ann. Statist.*, 32(5):2028–2043, 2004.

[34] S. Walker, A. Lijoi, and I. Prunster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35(2):738–746, 2007.

[35] C. Zhang. Fourier methods for estimating mixing densities and distributions. *Annals of Statistics*, 18(2):806–831, 1990.