CONVERGENCE OF LATENT MIXING MEASURES IN FINITE AND INFINITE MIXTURE MODELS

By XuanLong Nguyen[1]

University of Michigan

This paper studies convergence behavior of latent mixing measures that arise in finite and infinite mixture models, using transportation distances (i.e., Wasserstein metrics). The relationship between Wasserstein distances on the space of mixing measures and f-divergence functionals such as Hellinger and Kullback–Leibler distances on the space of mixture distributions is investigated in detail using various identifiability conditions. Convergence in Wasserstein metrics for discrete measures implies convergence of individual atoms that provide support for the measures, thereby providing a natural interpretation of convergence of clusters in clustering applications where mixture models are typically employed. Convergence rates of posterior distributions for latent mixing measures are established, for both 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 mixing measures, which help to combine relatively simple models into richer classes of statistical models [24, 26]. 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., [13, 27, 29, 30, 33] for several examples of such models, as well as a recent book [19]).

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 $\mathbf{p} = (p_1, \ldots, p_k)$ 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 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 mixing measure $\hat{G}$ estimated on the basis of available data. An important work in this direction is by Chen [7], 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 [20]. 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 multidimensional or even an abstract space.

The analysis of consistency and convergence rates of posterior distributions for Bayesian estimation has seen much progress in the past decade. Key recent references include [2, 16, 17, 32, 38, 39]. Analysis of specific mixture models in a Bayesian setting has also been studied [14, 15, 18, 21]. All these works primarily focus on the convergence behavior of the posterior distribution of the data density p_G. On the other hand, results concerned with the convergence behavior of latent mixing measures G are quite rare. Notably, the analysis of convergence for mixing (smooth) densities often arises in the context of frequentist estimation for deconvolution problems, mainly within the kernel density estimation method (e.g., [6, 11, 40]). We also note recent progress on consistent parameter estimation for certain finite mixture models, for example, in an overfitted setting [31] or with an emphasis on computational efficiency [3, 22].

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 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 arose in the problem of optimal transportation [36]. Although not as popular as well-known divergence functionals such as Kullback–Leibler, total variation and Hellinger distances, Wasserstein distances have been utilized in a number of statistical contexts (e.g., [4, 9, 10, 25]). 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, for a given $r \geq 1$ the L_r Wasserstein metric on the space of discrete probability measures with support in Θ, namely, $\bar{\mathcal{G}}(\Theta)$, is

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

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, 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 the W_r 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 CDFs of univariate random variables, as studied by Chen [7], is in fact a special case of the W_1 metric when $\Theta = \mathbb{R}$.

The plan for the paper is as follows. Section 2 investigates 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 W_r 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 a Wasserstein metric. We present two key theorems, which provide upper bounds on $W_2(G, G')$ 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 [7]. Theorem 2 is applicable to mixing measures with an 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 W_2 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 the W_2 metric. These results are proved using the standard approach of Ghosal, Ghosh and van der Vaart [16]. 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, possibly for the first time, 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 a bounded

number of atomic support in $\mathbb{R}^d$, the posterior convergence rate for mixing measures is $(\log n)^{1/4} n^{-1/4}$ under suitable identifiability conditions. This rate is optimal up to a logarithmic factor in the minimax sense. 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 $(\log n)^{-1/\beta}$.

Notation. For ease of notation, 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: $V(p_G, p_{G'}) = \frac{1}{2} \int |p_G(x) - p_{G'}(x)| \, d\mu(x)$, Hellinger distance:

$$h^2(p_G, p_{G'}) = \frac{1}{2} \int \left(\sqrt{p_G(x)} - \sqrt{p_{G'}(x)}\right)^2 d\mu(x)$$

and 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 $h^2 \leq K/2$. $N(\varepsilon, \Theta, \rho)$ denotes the covering number of the metric space (Θ, ρ), that is, the minimum number of ε-balls needed to cover the entire space Θ. $D(\varepsilon, \Theta, \rho)$ denotes the packing number of (Θ, ρ), that is, the maximum number of points that are mutually separated by at least ε in distance. They are related by $N(\varepsilon, \Theta, \rho) \leq D(\varepsilon, \Theta, \rho) \leq N(\varepsilon/2, \Theta, \rho)$. $\mathrm{Diam}(\Theta)$ denotes the diameter of Θ.

2. Transportation distances for mixing measures.

2.1. *Definition and a basic inequality.* Let (Θ, ρ) be a space equipped with a nonnegative distance function $\rho : \Theta \times \Theta \to \mathbb{R}_+$, that is, a 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 $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ denotes the proportion vector, while $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are the associated atoms in Θ. $\mathbf{p}$ has to satisfy $0 \leq p_i \leq 1$ and $\sum_{i=1}^{k} p_k = 1$. [With a bit abuse of notation, 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 $\mathbf{p}$.] 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 $\mathcal{G}(\Theta) = \bigcup_{k \in \mathbb{N}_+} \mathcal{G}_k(\Theta)$, 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 $\mathbf{q} = (q_{ij})_{i \le k; j \le k'} \in [0, 1]^{k \times k'}$ 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 $\mathbf{q}$ a coupling of $\mathbf{p}$ and $\mathbf{p}'$. Let $\mathcal{Q}(\mathbf{p}, \mathbf{p}')$ denote the space of all such couplings. We start with the general transportation distance:

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

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

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 notation of Wasserstein metrics, $W_1(G, G') \equiv d_\rho(G, G')$ and $W_2^2(G, G') \equiv d_{\rho^2}(G, G')$. However, d_ρ will be employed when ρ may be a general or a nonstandard distance function or metric.

From here on, probability 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 notation, 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', for example, $d_\rho(G, G')$, and divergences of mixture densities $p_G, p_{G'}$. Divergences that play important roles in this paper are the total variational distance, the Hellinger distance and the Kullback–Leibler distance. All these are in fact instances of a broader class of divergences known as the f-divergences (Csiszár [8]; Ali and Silvey [1]):

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 transportation distance a *composite* transportation distance:

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

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 transportation 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 transportation 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 transportation distances. To ensure this property, additional conditions are needed on the space of probability 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 [35] 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 [7], 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 $\theta_1, \dots, \theta_k$, the equality

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

implies that $\alpha_i = 0$, $\beta_i = \gamma_i = \mathbf{0} \in \mathbb{R}^d$ for $i = 1, \dots, k$. 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 [34]; see also Theorem 1 of [21]. Chen identified a broad class of families, including the Gaussian family, for which the strong identifiability condition holds [7].

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 a 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*

$$(3) \qquad \left| \gamma^T (D^2 f(x|\theta_1) - D^2 f(x|\theta_2))\gamma \right| \leq C\|\theta_1 - \theta_2\|^\delta \|\gamma\|^2$$

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 $k < \infty$,

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

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 $W_2(G_0, G_n) \to 0$, then it is simple to show that there is a subsequence of G_n that also has k distinct atoms, which converge in the 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 equation (4) can be deduced. This practically means that the conditions specified by (2) and (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 $G, G' \in \bar{\mathcal{G}}(\Theta)$, that is, mixing measures with a 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 $p_{G'}(x) = G' * f(x) = \sum_{j=1}^{k'} p'_j f(x - \theta'_j)$.

The key assumption is concerned 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).

THEOREM 2. *Suppose that G, G' are probability measures that place full support on a bounded subset $\Theta \subset \mathbb{R}^d$. f is a density function on $\mathbb{R}^d$ that is symmetric (around 0), that is, $\int_A f \, dx = \int_{-A} f \, dx$ 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*

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

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

(2) Supersmooth likelihood. *Suppose that $|\tilde{f}(\omega) \prod_{j=1}^{d} \exp(|\omega_j|^{\beta}/\gamma)| \geq d_0$ 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*

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

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

REMARK. The theorem does not actually require that mixing measures G, G' be discrete. Moreover, from the proof of the theorem, the condition that the support points of G and G' lie in a bounded subset of $\mathbb{R}^d$ can be removed and replaced by the boundedness of a given moment of the mixing measures. The upper bound remains the same for the supersmooth likelihood case. For the ordinary smooth case, we obtain a slightly weaker upper bound for $W_2(G, G')$.

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 $W_2^2(G, G') \lesssim (-\log V(p_G, p_{G'}))^{-1}$ as $W_2(G, G') \to 0$ [so that $V(p_G, p_{G'}) \to 0$, by Lemma 1]. For a Laplace density on $\mathbb{R}$, for example, $\tilde{f}(\omega) = \frac{1}{1+\omega^2}$, then $W_2^2(G, G') \lesssim V(p_G, p_{G'})^m$ for any $m < 4/9$, as $W_2(G, G') \to 0$.

3. Convergence of posterior distributions of mixing measures.

We turn to a study of convergence of mixing measures in a Bayesian setting. Let $X_1, \ldots, X_n$ be an i.i.d. 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 of support points for G may be unknown (and/or unbounded). 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,

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

We shall study conditions under which the posterior distribution is consistent, that is, 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 [16], 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 the mixture model setting (which can be viewed as counterparts of Theorems 2.1 and 2.4 of [16]). 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.

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

DEFINITION 5. Fix $G_0 \in \bar{\mathcal{G}}(\Theta)$. Let $\mathcal{G} \subset \bar{\mathcal{G}}(\Theta)$. Define the Hellinger information of the W_2 metric for subset $\mathcal{G}$ as a real-valued function on the real line $\Psi_{\mathcal{G}}: \mathbb{R} \to \mathbb{R}$:

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

Note the dependence of $\Psi_{\mathcal{G}}$ on the (fixed) G_0, but this is suppressed for ease of notation. It is obvious that $\Psi_{\mathcal{G}}$ is a nonnegative and nondecreasing function. The following characterizations of $\Psi_{\mathcal{G}}$ are simple consequences of Theorems 1 and 2:

PROPOSITION 1.   (a) *Suppose that $G_0 \in \mathcal{G}_k(\Theta)$, and both $\mathcal{G}_k(\Theta)$ and $\mathcal{G}$ are compact in the Wasserstein topology. In addition, assume that the family of likelihood functions is finitely identifiable. Then, $\Psi_{\mathcal{G}}(r) > 0$ for all $r > 0$.*

(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 $\Psi_{\mathcal{G}_k(\Theta)}(r) \geq c(k, G_0) r^4$ for all $r > 0$.*

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

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 i.i.d. sample $X_1, \ldots, X_n$. For a fixed pair of measures (G_0, G_1) such that $G_1 \in \mathcal{G}$, 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 $G_0 \in \bar{\mathcal{G}}(\Theta)$ and consider $\mathcal{G} \subset \bar{\mathcal{G}}(\Theta)$. 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, let $M(\mathcal{G}, G_1, r) = 1$.*

(II) *$\mathcal{G}$ is nonconvex, 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*

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

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

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

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

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. Hence, this packing number can be upper bounded by exploiting a lower bound of the Hellinger information.

Next, the existence of the 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 nonincreasing function* $D(\varepsilon)$, *some* $\varepsilon_n \geq 0$ *and every* $\varepsilon > \varepsilon_n$,

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

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

(10) $$P_{G_0}\varphi_n \leq D(\varepsilon) \sum_{t=t_0}^{\lceil \mathrm{Diam}(\Theta)/\varepsilon \rceil} \exp[-n\Psi_{\mathcal{G}}(t\varepsilon)/8],$$

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

REMARK.   It is interesting to observe that function $D(\varepsilon)$ is used to control the packing number of thin layers of Wasserstein balls (a similar quantity that also arises via the peeling argument in [16] (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 the space of discrete measures $\bar{\mathcal{G}}(\Theta)$ and the family of likelihood functions $f(x|\theta)$. The Kullback–Leibler neighborhood is defined as

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

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

THEOREM 3. *Fix $G_0 \in \bar{\mathcal{G}}(\Theta)$, and assume that the family of likelihood functions is finitely identifiable. Suppose that for a sequence $(\varepsilon_n)_{n \geq 1}$ that tends to a constant (or 0) such that $n\varepsilon_n^2 \to \infty$, and a constant $C > 0$, and* convex *sets $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$, we have*

$$\text{(13)} \qquad \log D(\varepsilon_n, \mathcal{G}_n, W_2) \leq n\varepsilon_n^2,$$

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

$$\text{(15)} \qquad \Pi(B_K(\varepsilon_n)) \geq \exp(-n\varepsilon_n^2 C).$$

Moreover, suppose M_n is a sequence such that

$$\text{(16)} \qquad \Psi_{\mathcal{G}_n}(M_n \varepsilon_n) \geq 8\varepsilon_n^2(C+4),$$

$$\text{(17)} \qquad \exp(2n\varepsilon_n^2) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/8] \to 0.$$

Then, $\Pi(G : W_2(G_0, G) \geq M_n \varepsilon_n | X_1, \ldots, X_n) \to 0$ 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 nonconvex sieves $\mathcal{G}_n$.

THEOREM 4. *Fix $G_0 \in \bar{\mathcal{G}}(\Theta)$. 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 $M(\mathcal{G}_n, G_1, \varepsilon)$ is defined by (6).*

(c) *There is a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2$ is bounded away from 0 or tending to infinity, and a sequence M_n such that*

$$\text{(18)} \qquad \begin{aligned} &\log D(\varepsilon/2, \mathcal{G}_n \cap B_W(G_0, 2\varepsilon) \setminus B_W(G_0, \varepsilon), W_2) \\ &\quad + \sup_{G_1 \in \mathcal{G}_n} \log M(\mathcal{G}_n, G_1, \varepsilon) \leq n\varepsilon_n^2 \qquad \forall \varepsilon \geq \varepsilon_n, \end{aligned}$$

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

$$\text{(20)} \qquad \frac{\Pi(B_W(G_0, 2j\varepsilon_n) \setminus B_W(G_0, j\varepsilon_n))}{\Pi(B_K(\varepsilon_n))} \leq \exp[n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/16] \qquad \forall j \geq M_n,$$

$$\text{(21)} \qquad \exp(2n\varepsilon_n^2) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/16] \to 0.$$

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

REMARK.   (i) From the theorem's proof, the above statement continues to hold if conditions (20) and (21) are replaced by the following condition:

$$(22) \qquad \exp(2n\varepsilon_n^2) / \Pi(B_K(\varepsilon_n)) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/16] \to 0.$$

(ii) In Theorem 4 and in Theorem 3 augmented with condition (a) of Theorem 4, it is simple to deduce the posterior convergence rates for the mixture density p_G. We obtain that for a sufficiently large constant $M > 0$,

$$\Pi(G : h(p_{G_0}, p_G) \geq M\varepsilon_n | X_1, \ldots, X_n) \to 0$$

in P_{G_0}-probability.

Before moving to specific examples, we state a simple lemma which provides estimates of the entropy under the W_r metric for a number of classes of discrete measures of interest. Because W_r inherits directly the ρ metric in Θ, the entropy for classes in $(\bar{\mathcal{G}}(\Theta), W_r)$ can typically be bounded in terms of the covering number for subsets of (Θ, ρ).

LEMMA 4.   *Suppose that Θ is bounded. Let $r \geq 1$.*

(a) $\log N(2\varepsilon, \mathcal{G}_k(\Theta), W_r) \leq k(\log N(\varepsilon, \Theta, \rho) + \log(e + e\,\mathrm{Diam}(\Theta)^r/\varepsilon^r))$.

(b) $\log N(2\varepsilon, \bar{\mathcal{G}}(\Theta), W_r) \leq N(\varepsilon, \Theta, \rho) \log(e + e\,\mathrm{Diam}(\Theta)^r/\varepsilon^r)$.

(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,*

$$\log N(\varepsilon/2, \{G \in \mathcal{G}_k(\Theta) : W_r(G_0, G) \leq 2\varepsilon\}, W_r)$$
$$\leq k\Big(\sup_{\Theta'} \log N(\varepsilon/4, \Theta', \rho) + \log(2^{2+3r} k\,\mathrm{Diam}(\Theta)/m)\Big),$$

where the supremum in the right-hand side is taken over all bounded subsets $\Theta' \subseteq \Theta$ such that $\mathrm{Diam}(\Theta') \leq 4M^{1/r}\varepsilon$.

4. Examples.   In this section we derive posterior contraction rates for two classes of mixture models, for example, 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 the following:

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 $G \in \mathrm{supp}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 < C_2 K(p_{G_0}, p_G)$ 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.

REMARK. 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 behaves 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* $(\log n)^{1/2}n^{-1/4}$.

PROOF. Let $G = \sum_{i=1}^k p_i\delta_{\theta_i}$. Combining Lemma 1 with assumption (A2), if $\|\theta_i - \theta_i^*\| \leq \varepsilon$ and $|p_i - p_i^*| \leq \varepsilon^2/(k\,\mathrm{Diam}(\Theta)^2)$ 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 $\mathbf{q} \in \mathcal{Q}$. Thus, $K(p_{G_0}, p_G) \leq C_1 W_2^2(G_0, G) \leq C_1 \sum_{i=1}^k (p_i^* \wedge p_i)\|\theta_i^* - \theta_i\|^2 + C_1 \sum_{i=1}^k |p_i - p_i^*|\,\mathrm{Diam}(\Theta)^2 \leq 2C_1\varepsilon^2$. Hence, under prior Π,

$$\Pi\big(G : K(p_{G_0}, p_G) \leq \varepsilon^2\big)$$
$$\geq \Pi\big(G : \|\theta_i - \theta_i^*\| \leq \varepsilon, |p_i - p_i^*| \leq \varepsilon^2/(k\,\mathrm{Diam}(\Theta)^2), i = 1, \ldots, k\big).$$

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

$$\Pi\big(G : W_2^2(G_0, G) \leq \varepsilon^2\big) \leq \Pi\big(G : \|\theta_i - \theta_i^*\| \leq O(\varepsilon), |p_i - p_i^*| \leq O(\varepsilon^2), \forall i\big)$$
$$\lesssim \varepsilon^{k(d+2)}.$$

Let ε_n be a sufficiently large multiple of $(\log n/n)^{1/2}$. 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 equation (19) trivially holds.

Next, we provide upper bounds for $D(\varepsilon/2, S, W_2)$, where $S = \{G \in \mathcal{G}_n : W_2(G_0, G) \leq 2\varepsilon\}$, and $M(\mathcal{G}_n, G_1, \varepsilon)$ so that (18) is satisfied. Indeed, for any $\varepsilon > 0$, $\log D(\varepsilon/2, S, W_2) \leq \log N(\varepsilon/4, S, W_2)$. By Lemma 4(c) and assumption (A4), $N(\varepsilon/4, S, W_1)$ is bounded in terms of $\sup_{\Theta'} \log N(\varepsilon/8, \Theta', \rho)$, which is bounded

above by a constant when Θ''s are subsets of Θ whose diameter is bounded by a multiple of ε. Turning to $M(\mathcal{G}_n, G_1, \varepsilon)$, due to strong identifiability and assumption (A2), $\Psi_{\mathcal{G}_n}(\varepsilon) \geq c\varepsilon^4$ for some constant $c > 0$. By Lemma 4(a), for some constant $c_1 > 0$, $\log M(\mathcal{G}_n, G_1, \varepsilon) \leq \log N(c_1\varepsilon^2, \mathcal{G}_k(\Theta) \cap B_W(G_1, \varepsilon/2), W_2) \leq k(\log N(c_1\varepsilon^2/2, \Theta, \rho) + \log(e + 4e\,\mathrm{Diam}(\Theta)^2/c_1^2\varepsilon^4)) \leq n\varepsilon_n^2/2$. Thus, equation (18) holds.

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

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

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

$$\exp(2n\varepsilon_n^2) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/16] \lesssim \exp(2n\varepsilon_n^2) \sum_{j \geq M_n} \exp[-nc(j\varepsilon_n)^4]$$

$$\lesssim \exp(2n\varepsilon_n^2 - ncM_n^4\varepsilon_n^4).$$

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

Under the assumptions specified above,

$$\Pi\big(G : j\varepsilon_n < W_2(G, G_0) \leq 2j\varepsilon_n\big)/\Pi\big(B_K(\varepsilon_n)\big) = O(1).$$

On the other hand, for $j \geq M_n$, we have $\exp[n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/16] \geq \exp[nc(j\varepsilon_n)^4/16]$ which is bounded below by an arbitrarily large constant by choosing M_n to be a large multiple of $\varepsilon_n^{-1/2}$, thereby ensuring (20).

Thus, by Theorem 4, rate of contraction for the posterior distribution of G under the W_2 distance metric is $(\log n)^{1/4}n^{-1/4}$, which is up to a logarithmic factor the minimax optimal rate $n^{-1/4}$ as proved for the univariate finite mixtures by [7]. □

4.2. *Dirichlet process mixtures.* Given the "true" discrete mixing 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 $\mathrm{DP}(\nu, P_0)$ that centers at P_0 with concentration parameter $\nu > 0$ [12]. Here, parameter P_0 is a probability measure on Θ. For any $r \geq 1$, the following lemma provides a lower bound of small ball probabilities of metric space $(\bar{\mathcal{G}}(\Theta), W_r)$ in terms of small ball P_0-probabilities of metric space (Θ, ρ).

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

$$\Pi\big(G : W_r^r(G_0, G) \leq (2^r + 1)\varepsilon^r\big)$$

$$\geq \frac{\Gamma(\nu)\nu^D}{(2D)^{D-1}} \left(\frac{\varepsilon}{\mathrm{Diam}(\Theta)}\right)^{r(D-1)} \sup_S \prod_{i=1}^D P_0(S_i).$$

Here, $S := (S_1, \ldots, S_D)$ denotes the D disjoint $\varepsilon/2$-balls that form a maximal ε-packing of Θ. The supremum is taken over all such packings. $\Gamma(\cdot)$ 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 $\mathrm{Diam}(S'_i) \leq 2\varepsilon$ 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, $\mathbf{m} = (m_1, \ldots, m_D) \sim \mathrm{Dir}(\nu\mu_1, \ldots, \nu\mu_D)$. To obtain an upper bound for $d_{\rho^r}(G_0, G)$, 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 $\|\mathbf{m} - \hat{\mathbf{p}}\|$) for both measures are coupled in an arbitrary way. The expectation under this coupling of the ρ^r distance provides one such upper bound, that is,

$$d_{\rho^r}(G_0, G) \leq (2\varepsilon)^r + \|\mathbf{m} - \hat{\mathbf{p}}\|_1 [\mathrm{Diam}(\Theta)]^r.$$

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

$$\mathrm{Pr}(d_{\rho^r}(G_0, G) \leq (2^r + 1)\varepsilon^r)$$

$$\geq \mathrm{Pr}(\|\mathbf{m} - \hat{\mathbf{p}}\|_1 \leq \delta^r)$$

$$\geq \mathrm{Pr}(|m_i - \hat{p}_i| \leq \delta^r/2D, i = 1, \ldots, D-1)$$

$$= \frac{\Gamma(\nu)}{\prod_{i=1}^{D} \Gamma(\nu\mu_i)} \int_{\Delta_{D-1} \cap |m_i - \hat{p}_i| \leq \delta^r/2D} \prod_{i=1}^{D-1} m_i^{\nu\mu_i - 1} \left(1 - \sum_{i=1}^{D-1} m_i\right)^{\nu\mu_D - 1} dm_i$$

$$\geq \frac{\Gamma(\nu)}{\prod_{i=1}^{D} \Gamma(\nu\mu_i)} \prod_{i=1}^{D-1} \int_{\max(\hat{p}_i - \delta^r/2D, 0)}^{\min(\hat{p}_i + \delta^r/2D, 1)} m_i^{\nu\mu_i - 1} dm_i$$

$$\geq \Gamma(\nu)(\delta^r/2D)^{D-1} \prod_{i=1}^{D} (\nu\mu_i).$$

The second inequality in the previous display is due to the fact that $\|\mathbf{m} - \hat{\mathbf{p}}\|_1 \leq 2\sum_{i=1}^{D-1} |m_i - \hat{p}_i|$. 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. $\quad\square$

Assumptions B. (B1) The nonatomic base measure P_0 places full support on a bounded 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*

$$\beta_n \asymp (\log n/n)^{2/((d+2)(4+(2\beta+1)d'))}$$

for any constant $d' > d$.

(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 (13), (14) and (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 $\beta_n = M_n \varepsilon_n$ 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 $\Pi(G \in B_K(\varepsilon_n)) \geq \Pi(G : d_{\rho^{m_1}}(G_0, G) \leq C_3 \varepsilon_n^{2 \vee 2/m_2})$ 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 $D \asymp [\text{Diam}(\Theta)/\varepsilon_n]^d$. Combining these facts with Lemma 5, we have $\log \Pi(G \in B_K(\varepsilon_n)) \gtrsim (D-1) \log(\varepsilon_n/\text{Diam}(\Theta)) + (2D-1) \log(1/D) + D \log \nu$, where the approximation constant is dependent on m_1, m_2. It is simple to check that condition (15) holds, $\log \Pi(G \in B_K(\varepsilon_n)) \geq -Cn\varepsilon_n^2$, 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 $\log N(2\varepsilon_n, \bar{\mathcal{G}}(\Theta), W_2) \leq N(\varepsilon_n, \Theta, \rho) \log(e + e \, \text{Diam}(\Theta)^2/\varepsilon_n^2) \leq (\text{Diam}(\Theta)/\varepsilon_n)^d \log(e + e \, \text{Diam}(\Theta)^2/\varepsilon_n^2) \leq n\varepsilon_n^2$ by the specified rate of ε_n.

Step 2: For any $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$, let $R_{\mathcal{G}}(r)$ be the inverse of the Hellinger information function of the W_2 metric. Specifically, for any $t \geq 0$,

$$R_{\mathcal{G}}(t) = \inf\{r \geq 0 | \Psi_{\mathcal{G}}(r) \geq t\}.$$

Note that $R_{\mathcal{G}}(0) = 0$. $R_{\mathcal{G}}(\cdot)$ is nondecreasing because $\Psi_{\mathcal{G}}(\cdot)$ is.

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

$$\exp(2n\varepsilon_n^2) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/8]$$

$$\leq \mathrm{Diam}(\Theta)/\varepsilon_n \exp(2n\varepsilon_n^2 - n\Psi_{\mathcal{G}_n}(M_n\varepsilon_n)/8) \to 0.$$

Hence, Theorem 3 can be applied to conclude that

$$\Pi(W_2(G_0, G) \geq \beta_n | X_1, \ldots, X_n) \to 0$$

in P_{G_0}-probability. Under the ordinary smoothness condition (as specified in Theorem 2), $R_{\bar{\mathcal{G}}(\Theta)}(t) = t^{1/(4+(2\beta+1)d+\delta)}$, where δ is an arbitrarily positive constant. So,

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

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

5. Proofs.

5.1. *Proofs of Wasserstein identifiability results.*

PROOF OF THEOREM 1. Suppose that equation (4) is not true, then there will be sequences of G_n and G_n' tending to G_0 in the 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 a finite number of atoms, there is $\mathbf{q}^{(n)} \in \mathcal{Q}(\mathbf{p}_n, \mathbf{p}_n')$ 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 $\mathbf{q}^{(n)} \in \mathcal{Q}(\mathbf{p}_n, \mathbf{p}_n')$, we can express

$$\psi(G_n, G_n') = \sup_x \left| \sum_{i=1}^{k_n} p_{n,i} f(x|\theta_{n,i}) - \sum_{j=1}^{k_n'} p_{n,j}' f(x|\theta_{n,j}') \right| \bigg/ W_2^2(G_n, G_n')$$

$$= \sup_x \left| \sum_{ij} q_{ij}^{(n)} (f(x|\theta_{n,i}) - f(x|\theta_{n,j}')) \right| \bigg/ W_2^2(G_n, G_n')$$

and, by Taylor's expansion,

$$\psi(G_n, G'_n)$$

$$= \sup_x \left| \sum_{(i,j)\notin\mathcal{O}_n} q_{ij}^{(n)}(f(x|\theta'_{n,j}) - f(x|\theta_{n,i})) \right.$$

$$+ \sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}(\theta'_{n,j} - \theta_{n,i})^T Df(x|\theta_{n,i})$$

$$+ \left. \sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}(\theta'_{n,j} - \theta_{n,i})^T D^2 f(x|\theta_{n,i})(\theta'_{n,j} - \theta_{n,i}) + R_n(x) \right|$$

$$\Big/ W_2^2(G_n, G'_n)$$

$$=: \sup_x |A_n(x) + B_n(x) + C_n(x) + R_n(x)|/D_n,$$

where

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

due to (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 $D^2 f(x|\theta)$ 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, $d_n A_n(x)/D_n$ converges to $\sum_{j=1}^{k^*} \alpha_j f(x|\theta_j^*)$ and $d_n B_n(x)/D_n$ converges to $\sum_{j=1}^{k^*} \beta_j^T Df(x|\theta_j^*)$, and $d_n C_n(x)/D_n$ converges to

$\sum_{j=1}^{k^*} \gamma_j D^2 f(x|\theta_j^*)\gamma_j$, for some finite α_j, β_j and γ_j, not all of them vanishing (in fact, at least one of them is 1). We have

(23)
$$d_n |p_{G_n}(x) - p_{G_n'}(x)| / W_2^2(G_n, G_n')$$

$$\to \left| \sum_{j=1}^{k^*} \alpha_j f(x|\theta_j^*) + \beta_j^T Df(x|\theta_j^*) + \gamma_j^T D^2 f(x|\theta_j^*)\gamma_j \right|$$

for all x. This entails that the right-hand side of the preceding 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 equation (23) holds. An application of Fatou's lemma yields $\int |\sum_{j=1}^{k^*} \alpha_j f(x|\theta_j) + \beta_j^T Df(x|\theta_j) + \gamma_j^T D^2 f(x|\theta_j)\gamma_j| d\mu = 0$. Thus, the integrand must be 0 for almost all x, leading to contradiction. $\square$

PROOF OF THEOREM 2. 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 the triangular inequality, $W_2(G, G') \le W_2(G, G * K_\delta) + W_2(G', G' * K_\delta) + W_2(G * K_\delta, G' * K_\delta)$. The first two terms are simple to bound, while the last term can be handled by expressing $G * K_\delta$ as the convolution of 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 $0 < t < s$,*

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

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

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

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 mollifiers

$K_\delta(x) = \frac{1}{\delta^d} K(x/\delta)$ 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 $(\theta, \theta + \varepsilon)$ where ε is independent of θ, the marginal distributions of θ and ε are G K_δ, respectively. Under this coupling, $\mathbb{E}\|(\theta + \varepsilon) - \theta\|^2 = O(\delta^2)$, which entails that $W_2^2(G, G * K_\delta) = O(\delta^2)$.

By the 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,

$$(24) \qquad W_2^2(G, G') \leq 2W_2^2(G * K_\delta, G' * K_\delta) + C(K)\delta^2.$$

Theorem 6.15 of [37] 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,

$$(25) \qquad W_2^2(G * K_\delta, G' * K_\delta) \leq 2 \int \|x\|^2 |G * K_\delta(x) - G' * K_\delta(x)| \, dx.$$

We note that since density function K has a bounded sth moment,

$$\int \|x\|^s G * K_\delta(dx) \leq 2^s \left[\int \|\theta\|^s \, dG(\theta) + \int \|x\|^s K_\delta(x) \, dx \right]$$

$$= 2^s \left[\int \|\theta\|^s \, dG(\theta) + \delta^s \int \|x\|^s K(x) \, dx \right] < \infty,$$

because G's support points lie in a bounded subset of $\mathbb{R}^d$. Applying Lemma 6 to (25), we obtain that for $\delta < 1$,

$$(26) \qquad \begin{aligned} W_2^2(G * K_\delta, G' * K_\delta) &\leq C(d, K, s) \|G * K_\delta - G' * K_\delta\|_{L_1}^{(s-2)/s} \\ &\leq C(d, K, s) \|G * K_\delta - G' * K_\delta\|_{L_2}^{2(s-2)/(d+2s)}. \end{aligned}$$

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

Next, we use a known fact that for an 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:

$$
\begin{aligned}
\|G * K_\delta - G' * K_\delta\|_{L_2} &= \|p_G * g_\delta - p_{G'} * g_\delta\|_{L_2} \\
&= \|(p_G - p_{G'}) * g_\delta\|_{L_2} \\
&\leq 2V(p_G, p_{G'}) \|g_\delta\|_{L_2}.
\end{aligned}
$$
(27)

By Plancherel's identity,

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

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

Collecting equations (24), (25), (26) and (27) and the preceding display, we have

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

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

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

The exponent tends to $4/(4 + (2\beta + 1)d)$ as $s \to \infty$, so 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

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

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

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

PROOF OF LEMMA 1. We exploit the variational characterization of f-divergences (e.g., [28]),

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

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. $\square$

PROOF OF PROPOSITION 1. (a) Suppose that the claim is not true, and 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 the 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

$$d_h^2(p_{G_0}, p_G) \geq V^2(p_{G_0}, p_G)/2$$
$$\geq c(k, G_0)W_2^4(G_0, G)$$

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 to Theorem 2. $\square$

5.2. *Proofs of posterior contraction theorems.* We outline in this section the proofs of Theorems 3 and 4. Our proof follows the same steps as in [16], 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. 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 [5] and Le Cam ([23], Lemma 4, page 478), there exist tests φ_n that discriminate between P_{G_0} and convex set $\mathcal{P}_1$ such that

(28) $$P_{G_0}\varphi_n \le \exp[-n \inf h^2(P_{G_0}, P_1)/2],$$

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

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 $W_2(G_1, G) \le r/2$, then $W_2(G_0, G) \ge r/2$. So,

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

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 the 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 $W_2(G, \tilde{G}_t) \le c_0 r$, by Lemma 1 we have $h^2(p_G, p_{\tilde{G}_t}) \le d_{\rho_{\tilde{h}}^2}(G, \tilde{G}_t) \le C_1 d_{\rho^{2\alpha}}(G, \tilde{G}_t) \le C_1 \operatorname{Diam}(\Theta)^{2(\alpha-1)} W_2^2(G, \tilde{G}_t) \le C_1 \operatorname{Diam}(\Theta)^{2(\alpha-1)} c_0^2 r^2$ (the second inequality is due to the condition on the likelihood functions); and so it follows that

$$h(p_{G_0}, p_G) \ge h(p_{G_0}, p_{\tilde{G}_t}) - h(p_G, p_{\tilde{G}_t}) \ge \Psi_{\mathcal{G}}(r)^{1/2} - C_1^{1/2} \operatorname{Diam}(\Theta)^{\alpha-1} c_0 r.$$

Choose $c_0 = \frac{\Psi_{\mathcal{G}}(r)^{1/2}}{2r \operatorname{Diam}(\Theta)^{\alpha-1} C_1^{1/2}}$ so that the previous bounds become $h(p_G, p_{\tilde{G}_t}) \le \Psi_{\mathcal{G}}(r)^{1/2}/2 \le h(p_{G_0}, p_{\tilde{G}_t})/2$ and $h(p_{G_0}, p_G) \ge \Psi_{\mathcal{G}}(r)^{1/2}/2$.

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

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

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

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

$$P_{G_0}\varphi_n \leq M \exp[-n\Psi_{\mathcal{G}}(r)/8],$$

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

The first inequality is due to $\varphi_n \leq \sum_{t=1}^{M} \omega_n^{(t)}$, 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 $t \leq M$ such that $W_2(G, \tilde{G}_t) \leq c_0 r$, so that $h(p_G, p_{\tilde{G}_t}) \leq h(p_{G_0}, p_{\tilde{G}_t})/2$. $\square$

PROOF OF LEMMA 3.   For a given $t \in \mathbb{N}$ choose a maximal $t\varepsilon/2$-packing for set $S_t = \{G : t\varepsilon < W_2(G_0, G) \leq (t+1)\varepsilon\}$. This yields a set S'_t of at most $D(t\varepsilon/2, S_t, W_2)$ points. Moreover, every $G \in S_t$ is within distance $t\varepsilon/2$ 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 equations (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 $t \geq t_0$. Then, by the union bound and the fact that $D(\varepsilon)$ is nonincreasing,

$$P_{G_0}\varphi_n \leq \sum_{t \geq t_0} \sum_{G_1 \in S'_t} M(\mathcal{G}, G_1, t\varepsilon) \exp[-n\Psi_{\mathcal{G}}(t\varepsilon)/8]$$

$$\leq D(\varepsilon) \sum_{t \geq t_0} \exp[-n\Psi_{\mathcal{G}}(t\varepsilon)/8],$$

$$\sup_{G \in \bigcup_{u \geq t_0} S_u} P_G(1 - \varphi_n) \leq \sup_{u \geq t_0} \exp[-n\Psi_{\mathcal{G}}(u\varepsilon)/8] \leq \exp[-n\Psi_{\mathcal{G}}(t_0\varepsilon)/8],$$

where the last inequality is due the monotonicity of $\Psi_{\mathcal{G}}(\cdot)$. $\square$

PROOF OF THEOREMS 3 AND 4.   The proof for Theorem 3 proceeds in a similar way to Theorem 2.1 of [16], while the proof for Theorem 4 is similar to their Theorem 2.4. The main difference is that the posterior distribution statements are made with respect to mixing measure G rather than mixture density p_G. By a result of Ghosal et al. [16] (Lemma 8.1, page 524), for every $\varepsilon > 0$ and probability measure Π on the set $B_K(\varepsilon)$ defined by (12), we have, for every $C > 0$,

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

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

$$(30) \qquad \int \prod_{i=1}^{n} p_G(X_i)/p_{G_0}(X_i) d\Pi(G) \geq \exp(-2Cn\varepsilon_n^2)\Pi(B_K(\varepsilon_n)).$$

Let $\mathcal{O}_n = \{G \in \bar{\mathcal{G}}(\Theta) : W_2(G_0, G) \geq M_n \varepsilon_n\}$, $S_{n,j} = \{G \in \mathcal{G}_n : W_2(G_0, G) \in [j\varepsilon_n, (j+1)\varepsilon_n)\}$ for each $j \geq 1$. The conditions specified by Lemma 3 are satisfied by setting $D(\varepsilon) = \exp(n\varepsilon_n^2)$ (constant in ε). Thus, there exist tests φ_n for which equations (10) and (11) hold. Then,

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

Due to Lemma 3, the first term in the preceding display is bounded above by $P_{G_0}\varphi_n \leq D(\varepsilon_n) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/8] \to 0$, thanks to (21). The second term in the above display is bounded by $(C^2 n\varepsilon_n^2)^{-1}$ by the definition of A_n. If $n\varepsilon_n^2 \to \infty$, let $C = 1$. If $n\varepsilon_n^2$ tends to a positive constant away from 0, we let C be arbitrarily large so that this probability in the second term vanishes to 0. To show that the third term in the display also vanishes as $n \to \infty$, we exploit the following expression:

$$
\begin{aligned}
&\Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n) \\
&\quad = \int_{\mathcal{O}_n} \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\, d\Pi(G) \Big/ \int \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\, d\Pi(G),
\end{aligned}
$$

and then obtain a lower bound for the denominator by (30). For the nominator, by Fubini's theorem,

$$
\begin{aligned}
&P_{G_0} \int_{\mathcal{O}_n \cap \mathcal{G}_n} (1 - \varphi_n) \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\, d\Pi(G) \\
(31) \quad &= P_{G_0} \sum_{j \geq M_n} \int_{S_{n,j}} (1 - \varphi_n) \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\, d\Pi(G) \\
&= \sum_{j \geq M_n} \int_{S_{n,j}} P_G(1 - \varphi_n)\, d\Pi(G) \leq \sum_{j \geq M_n} \Pi(S_{n,j}) \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/8],
\end{aligned}
$$

where the last inequality is due to (11). In addition, by (19),

$$
\begin{aligned}
&P_{G_0} \int_{\mathcal{O}_n \setminus \mathcal{G}_n} (1 - \varphi_n) \prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\, d\Pi(G) \\
(32) \quad &= \int_{\mathcal{O}_n \setminus \mathcal{G}_n} P_G(1 - \varphi_n)\, d\Pi(G) \\
&\leq \Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) = o\big(\exp(-2n\varepsilon_n^2)\Pi(B_K(\varepsilon_n))\big).
\end{aligned}
$$

Combining bounds (31) and (32) and condition (20), we obtain

$$P_{G_0}(1 - \varphi_n)\Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n)\mathbb{I}(A_n)$$

$$\leq \frac{o(\exp(-2n\varepsilon_n^2)\Pi(B_K(\varepsilon_n))) + \sum_{j \geq M_n} \Pi(S_{n,j}) \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/8]}{\exp(-2n\varepsilon_n^2)\Pi(B_K(\varepsilon_n))}$$

$$\leq o(1) + \exp(2n\varepsilon_n^2) \sum_{j \geq M_n} \exp[-n\Psi_{\mathcal{G}_n}(j\varepsilon_n)/16].$$

The upper bound in the preceding display converges to 0 by (21), thereby concluding the proof of Theorem 4. The proof of Theorem 3 proceeds similarly. □

5.3. *Proof of other auxiliary lemmas.*

PROOF OF LEMMA 4.   To simplify notation, we give a proof for $W_1 \equiv d_\rho$. The general case for $W_r^r \equiv d_{\rho^r}$ can be carried out in the same way.

(a) Suppose that $(\eta_1, \ldots, \eta_T)$ forms an ε-covering for Θ under metric ρ, where $T = N(\varepsilon, \Theta, \rho)$ 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 $\rho(\theta_i, \theta_i') < \varepsilon$. Let $\mathbf{p}' = (p_1', \ldots, p_k')$ be a k-dim vector in the probability simplex that deviates from $\mathbf{p}$ by less than δ in l_1 distance: $\|\mathbf{p}' - \mathbf{p}\|_1 \leq \delta$. Define $G' = \sum_{i=1}^k p_i' \delta_{\theta_i'}$. We shall argue that

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

[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 $\|\mathbf{p} - \mathbf{p}'\|_1$) 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 $(\varepsilon + \delta \mathrm{Diam}(\Theta))$-covering for $\mathcal{G}_k(\Theta)$ can be constructed by combining each element of a δ-covering in the l_1 metric of the $k - 1$-probability simplex and k ε-coverings of Θ. Now, the covering number of the $k-1$-probability simplex is less than the number of cubes of length δ/k covering $[0, 1]^k$ times the volume of $\{(p_1', \ldots, p_k') : p_j' \geq 0, \sum_j p_j' \leq 1 + \delta\}$, that is, $(k/\delta)^k (1 + \delta)^k/k! \sim (1 + 1/\delta)^k e^k/\sqrt{2\pi k}$. It follows that $N(\varepsilon + \delta \mathrm{Diam}(\Theta), \mathcal{G}_k(\Theta), d_\rho) \leq T^k(1 + 1/\delta)^k e^k/\sqrt{2\pi k}$. Take $\delta = \varepsilon/\mathrm{Diam}(\Theta)$ 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 $t = 1, \dots, T$, $\rho(\theta_i, \eta_t) \leq \varepsilon$ 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

$$d_\rho(G, G') \leq \sum_{i=1}^k \sum_{t=1}^T p_i \mathbb{I}(\theta_i \in S_t) \rho(\theta_i, \eta_t) \leq \varepsilon$$

by using a similar coupling argument as in part (a).

Let $\mathbf{p}'' = (p''_1, \dots, p''_T)$ be a T-dim vector in the probability simplex that deviates from $\mathbf{p}'$ by less than δ in the l_1 distance: $\|\mathbf{p}'' - \mathbf{p}'\|_1 \leq \delta$. 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 the triangle inequality,

$$d_\rho(G, G'') \leq d_\rho(G, G') + d_\rho(G', G'') \leq \varepsilon + \delta \text{Diam}(\Theta).$$

The foregoing arguments establish that an $(\varepsilon + \delta \text{Diam}(\Theta))$-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), $N(\varepsilon + \delta \text{Diam}(\Theta), \bar{\mathcal{G}}(\Theta), d_\rho) \leq (1 + 1/\delta)^T e^T / \sqrt{2\pi T}$. Take $\delta = \varepsilon / \text{Diam}(\Theta)$ to conclude.

(c) Consider a $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ such that $d_\rho(G_0, G) \leq 2\varepsilon$. By definition, there is a coupling $q \in \mathcal{Q}(\mathbf{p}, \mathbf{p}^*)$ so that $\sum_{ij} q_{ij} \rho(\theta_i^*, \theta_j) \leq 2\varepsilon$. Since $\sum_j q_{ij} = p_i^*$, this implies that $2\varepsilon \geq \sum_{i=1}^k p_i^* \min_j \rho(\theta_i^*, \theta_j)$. Thus, for each $i = 1, \dots, k$ there is a j such that $\rho(\theta_i^*, \theta_j) \leq 2\varepsilon/p_i^* \leq 2M\varepsilon$. Without loss of generality, assume that $\rho(\theta_i^*, \theta_i) \leq 2M\varepsilon$ for all $i = 1, \dots, k$. 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, $|p_i^* - p_i| \leq 4\varepsilon/m$.

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

PROOF OF LEMMA 6. (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

$$\int_{\|x\| \leq R} |p(x) - p'(x)| \, dx \leq V_d^{1/2} R^{d/2} \left[\int_{\|x\| \leq R} (p(x) - p'(x))^2 \, dx \right]^{1/2}$$

$$\leq V_d^{1/2} R^{d/2} \|p - p'\|_{L_2}.$$

We also have

$$\int_{\|x\|\geq R}\left|p(x)-p'(x)\right|dx\leq\int_{\|x\|\geq R}p(x)+p'(x)\,dx\leq R^{-s}\left(\mathbb{E}_p\|X\|^s+\mathbb{E}_{p'}\|X\|^s\right).$$

Thus,

$$\left\|p-p'\right\|_{L_1}\leq\inf_{R>0}V_d^{1/2}R^{d/2}\left\|p-p'\right\|_{L_2}+R^{-s}\left(\mathbb{E}_p\|X\|^s+\mathbb{E}_{p'}\|X\|^s\right),$$

which gives the desired bound. $\square$

REFERENCES

[1] ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **28** 131–142. MR0196777

[2] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[3] BELKIN, M. and SINHA, K. (2010). Polynomial learning of distribution families. In *FOCS*.

[4] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

[5] BIRGÉ, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282. MR0764150

[6] CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186. MR0997599

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

[8] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345

[9] DEL BARRIO, E., CUESTA-ALBERTOS, J. A., MATRÁN, C. and RODRÍGUEZ-RODRÍGUEZ, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.* **27** 1230–1239. MR1740113

[10] DUDLEY, R. M. (1976). *Probabilities and Metrics*: *Convergence of Laws on Metric Spaces, with a View to Statistical Testing. Lecture Notes Series* **45**. Matematisk Institut, Aarhus Universitet, Aarhus. MR0488202

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

[12] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[13] GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. MR2201028

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

[15] GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. MR1701105

[16] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[17] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[18] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[19] HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G., eds. (2010). *Bayesian Nonparametrics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge Univ. Press, Cambridge. MR2722987

[20] ISHWARAN, H., JAMES, L. F. and SUN, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96** 1316–1332. MR1946579

[21] ISHWARAN, H. and ZAREPOUR, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12** 941–963. MR1929973

[22] KALAI, A., MOITRA, A. and VALIANT, G. (2012). Disentangling Gaussians. *Communications of the ACM* **55** 113–120.

[23] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411

[24] LINDSAY, B. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. IMS, Hayward, CA.

[25] MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515. MR0298812

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

[27] NGUYEN, X. (2010). Inference of global clusters from locally distributed data. *Bayesian Anal.* **5** 817–845. MR2740158

[28] NGUYEN, X., WAINWRIGHT, M. J. and JORDAN, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inform. Theory* **56** 5847–5861. MR2808937

[29] PETRONE, S., GUINDANI, M. and GELFAND, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 755–782. MR2750094

[30] RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831

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

[32] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[33] TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480

[34] TEICHER, H. (1960). On the mixture of distributions. *Ann. Math. Statist.* **31** 55–73. MR0121825

[35] TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248. MR0120677

[36] VILLANI, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483

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

[38] WALKER, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043. MR2102501

[39] WALKER, S. G., LIJOI, A. and PRÜNSTER, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866

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

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