Wasserstein distances for discrete measures and convergence in nonparametric mixture models [1]

XuanLong Nguyen
xuanlong@umich.edu

Technical Report 527
Department of Statistics
University of Michigan

September 15, 2011

Abstract

We consider Wasserstein distance functionals for comparing between and assessing the convergence of latent discrete measures, which serve as mixing distributions in hierarchical and nonparametric mixture models. We explore the space of discrete probability measures metrized by Wasserstein distances, 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. The convergence in Wasserstein metrics has a useful interpretation of the convergence of individual atoms that provide support for the discrete measure. It can be shown to be stronger than the weak convergence induced by standard f-divergence metrics, while the conditions for establishing the convergence can be formulated in terms of the metric space of the supporting atoms. These results are applied to establish rates of convergence of posterior distributions for latent discrete measures in several mixture models, including finite mixtures of multivariate distributions, finite mixtures of Gaussian processes 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 [25, 28]. In recent years the mixture modeling methodology has been significantly extended, by many authors taking the mixing distribution to be random and infinite dimensional via suitable priors constructed in a nested, hierarchical and nonparametric manner, resulting in rich models that can fit more complex and high dimensional data (see, e.g., [13, 37, 33, 32, 29] for several examples of

[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, Gaussian processes, Dirichlet processes.

This work was supported in part by NSF grants CCF-1115769 and OCI-1047871.

such models, as well as a recent book by [19]). Although initially viewed as a modeling device for density estimation [11, 10], discrete measures increasingly play an important role in the interpretation of the data population that they help to model, particularly in clustering applications.

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 yield the mixture density: $p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i)$. 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 [6], who used the L_1 metric on the cumulative distribution functions on the real line to study convergence rates of the mixing distribution G. Building upon Chen's work, Ishwaran, James and Sun [20] established the posterior consistency of a finite dimensional Dirichlet prior for Bayesian mixture models. Their analysis is specific to univariate mixture models, while our interest is when Θ has high or infinite dimensions. For instance, Θ may be a subset of a function space as in the work of [13, 29], or a space of probability measures [33].

Divergences such as the Hellinger distance, total variational distance and Kullback-Leibler distance are often employed to measure the distance between probability measures. They play a fundamental role in asymptotic statistics [23, 42]. In a mixture model, divergences applied to the data distributions (via density p_G) induce a weak topology on the space of discrete measures G. Moreover, they also helped to drive the development of minimum distance methods for estimating mixing distributions [3, 36, 8] As mixture models evolve into high dimensional and hierarchical models, the use of divergence functionals seems inadequate, because there is an increased disconnect between the convergence of the densities at the data level and the strong convergence of latent discrete measures further up in the model hierarchy, assuming that a notion of strong convergence for the latter can be formalized. From a computational viewpoint, divergence functionals applied to the data distribution p_G's are typically difficult to calculate when p_G's are complex, rendering many estimation methods using such distances difficult to implement.

In this paper, we consider a class of distance functionals known as the Wasserstein distance for discrete probability measures, and analyze the convergence of discrete mixing distributions in the setting of nonparametric mixture models. The Wasserstein (also known as Kantorovich) distances have been utilized in a number of statistical contexts (e.g., [27, 4, 9]). For discrete probability measures, they can be obtained by a "minimum matching" 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:

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

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 $\sum_i q_{ij} = p'_j$. The Wasserstein distances inherit directly the metric of the space of atomic support Θ, suggesting that they can be useful for comparing and assessing estimation procedures for discrete measures in hierarchical models. Indeed, a hint for it usefulness can be drawn from an observation that the L_1 distance for the CDF's of univariate random variables, as studied by [6], is in fact a special case of the L_1 Wasserstein distance when $\Theta = \mathbb{R}$. Moreover, 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 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, the convergence of mixing distribution G may be interpreted as the convergence of distinct typical behavior θ_i's that characterize the heterogeneous data population.

The plan for the paper is as follows. Sec. 2.1 provides a formal definition and reviews basic properties of the Wasserstein distances for discrete measures. In Sec. 2.2 we study the metric space $(\bar{\mathcal{G}}(\Theta), d_\rho)$ in some detail, investigating compactness and approximation properties. Since the Wasserstein distance is defined in terms of metric ρ, it is possible to obtain estimates of the covering number for $\bar{\mathcal{G}}(\Theta)$ and various subsets of interest in terms of the covering number for the space of atoms Θ. These results are useful in establishing rates of convergence in the mixture setting in the subsequent sections.

Section 3 explores the relationship between Wasserstein distances and well-known divergence functionals in the setting of a mixture model. Viewing the latter as specific instances of f-divergence functionals [7, 1], we first show that the an f-divergence between mixture densities p_G and $p_{G'}$ is dominated by the *composite* Wasserstein distance $d_\rho(G,G')$ in which $\rho(\theta_1,\theta_2)$ is also taken to be the same f-divergence between the likelihood densities $f(\cdot|\theta_1)$ and $f(\cdot|\theta_2)$. This implies that the d_ρ topology can be stronger than those induced by divergences between mixture densities. In addition, this result can be used to obtain bounds on small ball probabilities in the space of mixture densities p_G in terms of small ball probabilities in the metric space (Θ,ρ) [24]. In Sec. 3.2 we study identifiability conditions under which the convergence of mixture densities entails the convergence of mixing distributions in the Wasserstein metric. A simple consequence is Theorem 2, which establishes the consistency and convergence rates for a class of minimum distance estimators for the mixing distribution. In addition, we present a number of results on the existence of tests for discriminating a discrete measure with finite support against a class of discrete measures using an iid sample from a mixture distribution of data.

Section 4 focuses on the convergence of posterior distributions of latent discrete measures in a Bayesian nonparametric setting. Here, the mixing distribution 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. Consistency and convergence analysis for general Bayesian estimation procedures were initiated by the work of [34, 23]. A fairly complete general theory has been established – key recent references include [2, 16, 35, 43, 17, 44]. Analysis of specific mixture models in a Bayesian setting have also been studied extensively [15, 14, 21, 18]. All these work primarily focus on the convergence in the topology of Hellinger or a comparable distance metric in the space of densities p_G (an exception is [20], as noted previously).

In Theorems 3 and 4 of Section 4, we establish the strong convergence rates for the posterior distribution for G in terms of the d_ρ metric. These results are obtained using a general framework of Ghosal et al [16], and have several notable features. They rely on conditions on the support of the prior distribution Π, which can be formulated directly in terms of the metric space (Θ,ρ). This is convenient especially when Θ is of infinite dimensions. The claim of convergence in the Wasserstein metric is stronger than the weak convergence induced by the Hellinger metric in the existing work mentioned above. The general theory is applied to a number of well-known mixture models in the literature, including the finite mixture of multivariate distributions, finite mixture of Gaussian processes, and Dirichlet process mixtures, for which posterior consistency and convergence rates are derived.

Notations. By discrete measures we always mean discrete probability measures. $\mathcal{G}_k(\Theta),\mathcal{G}(\Theta)$ denote the spaces of discrete measure with k and finite number of atoms in Θ, respectively, while $\bar{\mathcal{G}}(\Theta)$ is the space of all discrete measures on Θ. $G(\boldsymbol{p},\boldsymbol{\theta}) = \sum_{i=1}^{k} p_i\delta_{\theta_i}$ and $G'(\boldsymbol{p'},\boldsymbol{\theta'}) = \sum_{j=1}^{k'} p_j\delta_{\theta'_j}$, resp., are two discrete measures in $\mathcal{G}_k(\Theta)$ and $\mathcal{G}_{k'}(\Theta)$. $f(x|\theta)$ denotes the conditional density, with respect to some dominating measure, of random variable X taking value in $\mathcal{X}$. G and G' yield mixture distributions given by the following densities, respectively: $p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i)$ and $p_{G'}(x) = \sum_{j=1}^{k'} p'_j f(x|\theta'_j)$. For ease of notations, we also use f_i in place of $f(\cdot|\theta_i)$, and f'_j in place of $f(\cdot|\theta'_j)$. Divergences studied in the paper include the total variational distance: $d_V(p_G,p_{G'}) := \frac{1}{2}\int |p_G(x)-p_{G'}(x)|d\mu$, the Hellinger distance: $d_h^2(p_G,p_{G'}) := \frac{1}{2}\int(\sqrt{p_G(x)} - \sqrt{p_{G'}(x)})^2 d\mu$, and the Kullback-Leibler divergence: $d_K(p_G,p_{G'}) = \int p_G(x)\log(p_G(x)/p_{G'}(x))d\mu$. $N(\epsilon,\Theta,\rho)$ 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. Diam(Θ) is the diameter of Θ.

2 Wasserstein distance metrics for discrete measures

2.1 Definition and basic properties

Let (Θ,ρ) be a space equiped with a non-negative distance function ρ such that $\rho(\theta_1,\theta_2) = 0$ 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 following 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 $\sum_{i=1}^{k} p_k = 1$. 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 $\mathcal{G}(\Theta) = \limsup \mathcal{G}_k(\Theta)$, 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 $k \times k'$ matrix whose entries satisfy the following "marginal" constraints: $\sum_{i=1}^{k} q_{ij} = p'_j$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, \ldots, k; j = 1, \ldots, k'$. Because the definition of $\boldsymbol{q}$ involves both $\boldsymbol{p}$ and $\boldsymbol{p'}$, we use $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ to denote the space of all such matrices. We start with the L_1 Wasserstein distance:

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

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

Remarks. (i) $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ is always non-empty, by consider the element $\boldsymbol{q}$ by taking $q_{ij} = p_i p'_j$. In this definition, if both k and k' are finite, the infimum can be replaced by minimum. (ii) Discrete measure G may admit different representations in terms of $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ be allowing equal atoms. It is easy to see that d_ρ is independent of such different representations. Unless specified otherwise, we shall always work with the representation of $G(G')$ in which all atoms θ_i's (resp. $(\theta'_j$'s) are distinct, so that $k(k')$ is the number of atoms associated with G (resp. G').

In some situations we also consider the L_2 Wasserstein distance, which corresponds to the square root of d_{ρ^2} in our definition, where $\rho(\theta_i, \theta'_j)$ is replaced by $\rho^2(\theta_i, \theta'_j)$. Note that $d_\rho^2(G, G') \leq d_{\rho^2}(G, G')$ by an application of Cauchy-Schwarz inequality.

In the following we state several basic and useful facts about Wasserstein distances. [2]

Lemma 1. *(a) $G = G'$ if and only if $d_\rho(G, G') = 0$.*

(b) If ρ admits triangular inequality on Θ, then d_ρ admits triangular inequality on $\bar{\mathcal{G}}(\Theta)$.

(c) For any $G \in \bar{\mathcal{G}}(\Theta)$, the ball $B(G, r) = \{G' : d_\rho(G, G') \leq r\}$ is a convex set.

[2] For completeness proofs of these facts are included in the Appendix.

In this paper we shall always assume that ρ is a metric on Θ, unless specifically noted otherwise. Lemma 1 implies that d_ρ is a valid metric on $\bar{\mathcal{G}}(\Theta)$. When $\Theta = \mathbb{R}$, ρ is taken to be the usual metric on the real line, the Wasserstein distance metric reduces to the L_1 metric on the cumulative distribution function for G and G':

Proposition 1. *Suppose that $\Theta \subseteq \mathbb{R}$. Define the CDF $G_c(x) = Pr(\theta \leq x|G)$ and $G'_c(x) = Pr(\theta' \leq x|G')$. Then $d_\rho(G, G') = \int |G_c(x) - G'_c(x)|dx$.*

In practice, $d_\rho(G, G')$ can be easily computed given G and G'. When both set of atoms and proportional probabilities associated with G and G' are given, the evaluation hinges on optimizing over matrices $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$. This is a linear programming problem for which efficient optimization algorithms are readily available.

Example 1. Let $G = \delta_{\theta_1}$, and $G' = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$. By definition, it is simple to derive that $d_\rho(G, G') = \sum_{j=1}^{k'} p'_j \rho(\theta_1, \theta'_j)$. By comparison, suppose that all atoms are distinct, so that G and G' have disjoint support. Then, both variational and Hellinger distances are constant: $d_V(G, G') = d_h^2(G, G') = 1$.

Example 2. Let $G = p\delta_{\theta_1} + (1-p)\delta_{\theta_2}$ and $G' = p'\delta_{\theta_1} + (1-p')\delta_{\theta_2}$. These two discrete measures share the same support at two atoms, and differ only in the proportional probabilities. By definition, $d_\rho(G, G') = \min_{\boldsymbol{q}}(q_{12} + q_{21})\rho(\theta_1, \theta_2)$, where the mimimization is subject to constraints that $\boldsymbol{q} \in [0, 1]^{2\times 2}$ and $q_{11} + q_{12} = p$; $q_{21} + q_{22} = 1-p$; $q_{11} + q_{21} = p'$; $q_{12} + q_{22} = 1-p'$. An easy calculation yields $q_{11} = \min(p, q)$, $q_{22} = \min(1-p, 1-p')$, $q_{12} + q_{21} = 1 - \min(p, p') - \min(1-p, 1-p') = |p - p'|$. Therefore, $d_\rho(G, G') = |p - p'|\rho(\theta_1, \theta_2)$. By comparison, the varitional distance $d_V(G, G') = |p - p'|$, while the Hellinger distance takes the form $d_h^2(G, G') = \frac{1}{2}[(\sqrt{p} - \sqrt{p'})^2 + (\sqrt{1-p} + \sqrt{1-p'})]^{1/2}$. Evidently, divergence functionals are rarely useful when being applied directly to discrete measures, because they do not take into account the geometry of the space of atoms Θ.

Example 3. Let $G_N = \sum_{n=1}^{N} \frac{1}{N}\delta_{Y_n}$, and $G'_N = \sum_{n=1}^{N} \frac{1}{N}\delta_{Y'_n}$, be the empirical distributions of G and G'. Here, $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ and $\boldsymbol{Y'} = (Y'_1, \ldots, Y'_N)$ are two iid samples from G and G', respectively. By definition,

$$d_\rho(G_N, G'_N) = \min_{\boldsymbol{q}} \sum_{1 \leq i,j \leq N} q_{ij}\rho(Y_i, Y'_j), \tag{2}$$

where the marginal constraints on $\boldsymbol{q}$ become $\sum_{i=1}^{N} q_{ij} = \sum_{j=1}^{N} q_{ij} = 1/N$ for any $i, j = 1, \ldots, N$. Note that $d_\rho(G_N, G'_N)$ is random due to the randomness of the atoms.

The following proposition explores the asymptotic behavior of of $d_\rho(G_N, G'_N)$ when both G and G' are discrete measures themselves, in addition to offering another perspective on the Wasserstein distance:

Proposition 2. *For a permutation π_N of the set $(1, \ldots, N)$, define a "matching distance":*

$$D_{\pi_N}(G_N, G'_N) = \frac{1}{N}\sum_{n=1}^{N} \rho(Y_n, Y'_{\pi(n)}).$$

(a) *For arbitrary distributions G and G' (including non-atomic ones), we have $d_\rho(G_N, G'_N) = \min_{\pi_N} D_{\pi_N}(G_N, G'_N)$, where the minimization is taken over all permutations of $(1, \ldots, N)$.*

(b) *If $G, G' \in \mathcal{G}_\Theta$ with finite number of atoms k, k', then $d_\rho(G_N, G'_N) \overset{G \times G' a.s.}{\longrightarrow} d_\rho(G, G')$, as $N \to \infty$.*

2.2 Compactness and entropy estimates

In the following, several aspects of the metric space $(\bar{\mathcal{G}}(\Theta), d_\rho)$ are explored, which will prove useful in our development in the sequel. We start with the compactness or the lack thereof of several key subsets in $\bar{\mathcal{G}}(\Theta)$.

Lemma 2. *(a) If Θ is compact, then both $\mathcal{G}_k(\Theta)$ and $\bar{\mathcal{G}}(\Theta)$ are compact with respect to the d_ρ topology for any $k < \infty$. Moreover, $\bar{\mathcal{G}}(\Theta)$ is the closure of $\mathcal{G}(\Theta)$.*

(b) *$\mathcal{G}(\Theta)$ is not complete (and consequentially neither compact nor closed).*

(c) *The open balls in $\mathcal{G}(\Theta)$ are not open relative to $\bar{\mathcal{G}}(\Theta)$. Specifically, for every $G \in \mathcal{G}(\Theta)$ there is a sequence of $G_n \in \bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}(\Theta)$, i.e., discrete measures with infinite support in Θ, that converge to G.*

Next, we shall quantify the size of classes of discrete measures by estimating their entropy number under d_ρ metric. Recall that the entropy number of a set is the logarithm of its covering number. Because d_ρ inherits directly the ρ metric in Θ, it is possible to derive bounds for the covering for $\mathcal{G}_\Theta$ in terms of the covering number for Θ.

Lemma 3. *(a) $\log N(2\epsilon, \mathcal{G}_k(\Theta), d_\rho) \leq k(\log N(\epsilon, \Theta, \rho) + \log(e + e\textit{Diam}(\Theta)/\epsilon))$.*

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

(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,*

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

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

The above result shows that $\mathcal{G}_k(\Theta)$'s are "nice" classes, for compactness and having a relatively small entropy number which grows linearly with respect to the number of atoms k and the entropy of Θ. $\mathcal{G}(\Theta)$ is a non-compact set, while its closure set $\bar{\mathcal{G}}(\Theta)$ is compact with an entropy number that may grow exponentially with respect to the entropy of Θ. This motivates us to consider more tractable approximating classes. A class of discrete measures $\mathcal{G} \subset \bar{\mathcal{G}}(\Theta)$ is called ϵ-tight with respect to $\mathcal{G}_k(\Theta)$ if

$$\sup_{G \in \mathcal{G}} \inf_{G' \in \mathcal{G}_k(\Theta)} d_\rho(G, G') \leq \epsilon.$$

Lemma 4. *If $\mathcal{G} \subset \bar{\mathcal{G}}(\Theta)$ is ϵ-tight with respect to $\mathcal{G}_k(\Theta)$, then*

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

Proof. Any discrete measure $G \in \mathcal{G}$ can be approximated by an element in $\mathcal{G}_k(\Theta)$ with an approximation error ϵ. That element is again approximated by another element in the ϵ-covering. Thus $N(2\epsilon, \mathcal{G}, d_\rho) \leq N(\epsilon, \mathcal{G}_k(\Theta), d_\rho)$. Combining with Lemma 3(a) to conclude. □

Example 4. (Stick-breaking processes). Let $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$ be the support of a Dirichlet process DP(ν, P_0), which is a probability measure on $\bar{\mathcal{G}}(\Theta)$ [12]. Any $G \in \mathcal{G}$ takes the form $G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$, where θ_i's are iid draws from P_0, and p_i's are distributed according to a "stick-breaking" process: $p_1 = v_1, p_2 = v_2(1 - v_1), \ldots, p_k = v_k \prod_{i=1}^{k-1}(1 - v_i)$ for any $k = 1, 2, \ldots$, where $v_1, v_2, \ldots$ are iid beta random variables Beta$(1, \nu)$.

Observe that, $1 - \sum_{i=1}^{k} p_i = \prod_{i=1}^{k}(1 - v_i)$. By Markov's inequality, for any $\epsilon > 0$, $P(1 - \sum_{i=1}^{k} p_i \geq \epsilon) = P(\prod_{i=1}^{k}(1 - v_i) \geq \epsilon) \geq \inf_{m>0} \prod_{i=1}^{k} \mathbb{E}(1 - v_i)^m/\epsilon^m = \inf_{m>0}[\nu/(\nu + m)]^k/\epsilon^m = \exp[-\nu \log(1/\epsilon) - k \log k + k \log(e\nu) + k \log\log(1/\epsilon)] =: \Delta(\epsilon, k)$.

Whenever $1 - \sum_{i=1}^{k} p_i < \epsilon$, it is simple to see that $\inf_{G' \in \mathcal{G}_k(\Theta)} d_\rho(G, G') \leq \epsilon$Diam$(\Theta)$. Hence, under the Dirichlet measure, for a given k and ϵ, there exists a subset $\mathcal{G}' \subset \mathcal{G}$ that is ϵ-tight with respect to $\mathcal{G}_k(\Theta)$, and Pr$(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}') \leq \Delta(k, \epsilon/$Diam$(\Theta))$.

3 Inequalities and properties in a mixture model

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 value 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:

$$p_G(x) = \sum_{i=1}^{k} p_i f_i(x).$$

3.1 f-divergences and composite Wasserstein distances

Divergence functionals for measuring the distance between probability measures play a fundamental role in asymptotic statistics. In the setting of mixture models, divergences applied to the space of densities p_G of the data X can be used to define weak topologies on the space of discrete measures $\bar{\mathcal{G}}(\Theta)$. Chief among these are the total variational distance, Hellinger distance, and the Kullback-Leibler distance. The variational distance takes the form: $d_V(p_G, p_{G'}) = \frac{1}{2} \int |p_G(x) - p_{G'}(x)| d\mu$, the Hellinger distance, $d_h^2(p_G, p_{G'}) = \frac{1}{2} \int (\sqrt{p_G(x)} - \sqrt{p_{G'}(x)})^2 d\mu$, and the Kullback-Leibler divergence, $d_K(p_G, p_{G'}) = \int p_G \log(p_G/p_{G'}) d\mu$.

All these are in fact instances of a broader class of divergence functionals 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 $d_\phi(f_i, f'_j) = \int \phi(f'_j/f_i) f_i d\mu$. Likewise, the f-divergence between p_G and $p_{G'}$ is $d_\phi(p_G, p_{G'}) = \int \phi(p_{G'}/p_G) p_G d\mu$.*

For $\phi(u) = \frac{1}{2}(\sqrt{u} - 1)^2$ we obtain the squared Hellinger, e.g., $d_h^2(f_i, f'_j), d_h^2(p_G, p_{G'})$. For $\phi(u) = \frac{1}{2}|u - 1|$ we obtain the variational distance, e.g., $d_V(f_i, f'_j), d_V(p_G, p_{G'})$. For $\phi(u) = -\log u$, we obtain the Kullback-Leiber divergence, e.g., $d_K(f_i, f'_j)$ and $d_K(p_G, p_{G'})$.

Typically divergence functionals on the likelihood densities such as $d_h(f_i, f'_j)$ and $d_K(f_i, f'_j)$ are simple to calculate. Moreover, they may be used as a distance function or metric on Θ.

Definition 3. *When ρ is taken to be an f-divergence, $\rho(\theta_i, \theta'_j) = d_\phi(f_i, f'_j)$, for a convex function ϕ, the corresponding Wasserstein distance functional is called a* composite *Wasserstein distance:*

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

In particular, d_V, d_h, d_K induce the composite Wasserstein distances $d_{\rho V}, d_{\rho h}, d_{\rho K}$, respectively.

For technical convenience, we also introduce composite Wasserstein distance function $d_{\rho h^2}(G, G')$ obtained by taking $\rho(\theta_i, \theta'_j) := d_h^2(f_i, f'_j)$. The following inequalities between variational and Hellinger distances are well-known: $d_V^2(p_G, p_{G'})/2 \leq d_h^2(p_G, p_{G'}) \leq d_V(p_G, p_{G'})$ and $d_V^2(f_i, f'_j)/2 \leq d_h^2(f_i, f'_j) \leq d_V(f_i, f'_j)$. This entails the following inequalities between the composite Wasserstein distances:

$$d_{\rho h^2}(G, G') \leq d_{\rho V}(G, G') \leq \sqrt{2} d_{\rho h}(G, G').$$

For any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$, we have $(\sum_{ij} q_{ij} d_h(f_i, f'_j))^2 \leq \sum_{ij} q_{ij} \sum_{ij} q_{ij} d_h^2(f_i, f'_j) = \sum_{ij} q_{ij} d_h^2(f_i, f'_j)$ by Cauchy-Schwarz inequality. This and the fact that $d_h(f_i, f'_j) \leq 1$ yield

$$d_{\rho h}^2(G, G') \leq d_{\rho h^2}(G, G') \leq d_{\rho h}(G, G').$$

The main result in this section is that the composite Wasserstein distance $d_{\rho\phi}(G, G')$ is generally a stronger distance metric (function) than $d_\phi(p_G, p_{G'})$.

Lemma 5. *(a) For any $G, G' \in \bar{\mathcal{G}}(\Theta)$, $0 \leq d_h^2(p_G, p_{G'}) \leq d_{\rho h^2}(G, G') \leq 1$.*

(b) Suppose that there is a constant $c > 0$ such that for any $G, G' \in \mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$, and any $x \in \mathcal{X}$,

$$\sum_i p_i f_i^{1/2} \geq c f_j^{'1/2} (p_G/p_{G'})^{1/2}.$$

Then, $1 - d_{\rho h^2}(G, G') \geq c(1 - d_h^2(p_G, p_{G'}))$.

Remarks. This lemma claims that the composite Wasserstein distance $d_{\rho h^2}(G, G')$ dominates the Hellinger distance $d_h^2(p_G, p_{G'})$. Under an additional assumption stated in part (b), if $d_{\rho h^2}(G, G')$ is close to 1, then so is $d_h^2(p_G, p_{G'})$. The assumption is satisfied when either one of the following holds: (i) The likelihood function $f(x|\theta)$ is bounded away from both 0 and infinity for any $x \in \mathcal{X}$ and any $\theta \in \Theta$, or (ii) All mixing probabilities p_i, p'_j are bounded away from 0. Condition (i) does not pose any restriction on the discrete measures G, G'. Condition (ii) entails that $\mathcal{G} \subseteq \mathcal{G}_k(\Theta)$ for some finite k, while imposes no restriction on the likelihood function $f(x|\theta)$.

Proof. (a) The second inequality is trivial, because Hellinger distances are bounded from above by 1. To show the first inequality, for any $i = 1, \ldots, k$ and $j = 1, \ldots, k'$, we express the Hellinger distance $d_h^2(f_i, f'_j) = 1 - \int f_i^{1/2} {f'}_j^{1/2} d\mu = 1 - \inf_{\varphi_{ij}} \frac{1}{2} \int e^{\varphi_{ij}} f_i + e^{-\varphi_{ij}} f'_j d\mu$, where the infimum over all measurable functions φ_{ij} on $\mathcal{X}$, and can be achieved by setting $\varphi_{ij} = \log((f'_j/f_i)^{1/2})$ [30]. We have (to avoid cluttering, $d\mu$ is dropped from the integrals in what follows):

$$1 - d_{\rho h^2}(G, G') = 1 - \inf_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \sum_{ij} q_{ij} d_h^2(f_i, f'_j) \quad = \quad \sup_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \inf_{\varphi_{ij}} \frac{1}{2} \sum_{ij} \int q_{ij}(e^{\varphi_{ij}} f_i + e^{-\varphi_{ij}} f'_j).$$

For any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$, we have (the infimum is taken over all measurable functions φ on $\mathcal{X}$):

$$\begin{aligned} 1 - d_h^2(p_G, p_{G'}) &= \inf_{\varphi} \frac{1}{2} \int e^{\varphi} p_G + e^{-\varphi} p_{G'} = \inf_{\varphi} \frac{1}{2} \int \sum_i p_i f_i e^{\varphi} + \sum_j p'_j f'_j e^{-\varphi} \\ &= \inf_{\varphi} \frac{1}{2} \int \sum_{ij} q_{ij}(e^{\varphi} f_i + e^{-\varphi} f'_j) \geq \inf_{\varphi_{ij}} \frac{1}{2} \int \sum_{ij} q_{ij}(e^{\varphi_{ij}} f_i + e^{-\varphi_{ij}} f'_j), \end{aligned}$$

where the last inequality holds because of the space over which the infimum operation is performed is enlarged. Since the above inequality holds for all $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$, it follows that

$$1 - d_h^2(p_G, p_{G'}) \geq \sup_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \inf_{\varphi_{ij}} \frac{1}{2} \int \sum_{ij} q_{ij}(e^{\varphi_{ij}} f_i + e^{-\varphi_{ij}} f'_j) = \sup_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \inf_{\varphi_{ij}} \frac{1}{2} \sum_{ij} \int q_{ij}(e^{\varphi_{ij}} f_i + e^{-\varphi_{ij}} f'_j),$$

where the interchange between integral and sum is due to the monotone convergence theorem. This completes the proof.

(b) Set $\varphi_{ij} = \log(f'_j/f_i)^{1/2}$ for any i, j, and $\varphi = \log(p_{G'}/p_G)^{1/2}$. From the proof of part (a),

$$\begin{aligned} 1 - d_{\rho h^2}(G, G') &= \sup_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \frac{1}{2} \sum_{ij} \int q_{ij}(e^{\varphi_{ij}} f_i + e^{-\varphi_{ij}} f'_j) \\ &= \sup_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \frac{1}{2} \int \sum_i f_i \sum_j q_{ij} e^{\varphi_{ij}} + \sum_j f'_j \sum_i q_{ij} e^{-\varphi_{ij}} \\ &\geq \frac{1}{2} \int \sum_i \biggl(p_i f_i \sum_j p'_j e^{\varphi_{ij}} \biggr) + \sum_j \biggl(p'_j f'_j \sum_i p_i e^{-\varphi_{ij}} \biggr), \end{aligned}$$

where the last step was obtained by setting $q_{ij} = p_i p'_j$, an admissible element in $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$, while the interchange between the integral and the sum is applicable due to the monotone convergence theorem. Compare this with $1 - d_h^2(p_G, p_{G'}) = \frac{1}{2}\int \sum_i p_i f_i e^{\varphi} + \sum_j p'_j f'_j e^{-\varphi}$, where $e^{\varphi} = (p_{G'}/p_G)^{1/2}$. We conclude that $1 - d_{\rho h^2}(G, G') \geq c(1 - d_h^2(p_G, p_{G'}))$, if for almost all $x \in \mathcal{X}$ the following inequalities hold, for any $i = 1, \ldots, k$ and $j = 1, \ldots, k'$: $\sum_j p'_j e^{\varphi_{ij}} \geq c e^{\varphi}$ and $\sum_i p_i e^{-\varphi_{ij}} \geq c e^{-\varphi}$. These inequalities are precisely given in the hypothesis. □

Part (a) in the previous lemma holds generally in the following sense.

Lemma 6. *Suppose that $d_\phi(p_G, p_{G'}) < \infty$ for some convex function ϕ.*
Then, $d_\phi(p_G, p_{G'}) \leq d_{\rho\phi}(G, G')$.

Proof. We exploit the variational characterization of f-divergences[31], $d_\phi(f_i, f'_j) = \sup_{\varphi_{ij}} \int \varphi_{ij} f'_j - \phi^*(\varphi_{ij}) f_i d\mu$. Here, the infimum is taken over all measurable function 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 $\phi^*(v) = \sup_{u \in \mathbb{R}}(uv - \phi(u))$.) Thus, $d_{\rho\phi}(G, G') = \inf_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \sum_{ij} q_{ij} \sup_{\varphi_{ij}} \int \varphi_{ij} f'_j - \phi^*(\varphi_{ij}) f_i$. On the other hand, for any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$,

$$\begin{aligned}
d_\phi(p_G, p_{G'}) &= \sup_{\varphi} \int \varphi p_{G'} - \phi^*(\varphi) p_G = \sup_{\varphi} \int \varphi \sum_j p'_j f'_j - \phi^*(\varphi) \sum_i p_i f_i \\
&= \sup_{\varphi} \int \varphi \sum_{ij} q_{ij} f'_j - \phi^*(\varphi) \sum_{ij} q_{ij} f_i = \sup_{\varphi} \int \sum_{ij} q_{ij}(\varphi f'_j - \phi^*(\varphi) f_i) \\
&\leq \sum_{ij} q_{ij} \sup_{\varphi_{ij}} \int (\varphi f'_j - \phi^*(\varphi) f_i),
\end{aligned}$$

where the last inequality holds because the supremum is taken over a larger set of functions. Moreover, the bound holds for any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$, so $d_\phi(p_G, p_{G'}) \leq d_{\rho\phi}(G, G')$. □

Collecting the results from the previous two lemmas we obtain the following inequalities for various distances between discrete measures G and G':

$$\begin{aligned}
0 \leq d_V^2(p_G, p_{G'})/2 \leq d_h^2(p_G, p_{G'}) \leq \{d_V(p_G, p_{G'}), d_{\rho h^2}(G, G')\} \leq d_{\rho V}(G, G') \leq \sqrt{2} d_{\rho h}(G, G'), \\
\text{and} \quad \{d_h^2(p_G, p_{G'}), d_{\rho h}^2(G, G')\} \leq d_{\rho h^2}(G, G') \leq d_{\rho h}(G, G') \leq 1.
\end{aligned}$$

Example 5. 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 $d_h^2(f_i, f'_j) = 1 - \exp -\frac{1}{8}\|\theta_i - \theta'_j\|^2 \leq \frac{1}{8}\|\theta_i - \theta'_j\|^2$. This entails that $d_{\rho h}(G, G') \leq d_\rho(G, G')/2\sqrt{2}$, and $d_{\rho h^2}(G, G') \leq d_{\rho^2}(G, G')/8$. Similarly for the Kullback-Leibler divergence, since $d_K(f_i, f'_j) = \frac{1}{2}\|\theta_i - \theta'_j\|^2$, by Lemma 6, $d_K(p_G, p_{G'}) \leq d_{\rho K}(G, G') = \frac{1}{2} d_{\rho^2}(G, G')$. Figure 1 illustrates the relationship between the squared Hellinger distance d_h^2, Wasserstein distance $d_{\rho h^2}$, and variational distance d_V. In this simulation, 5000 random pairs of mixture of two Gaussian distributions of the form $pN(\theta_1, 1) + (1-p)N(\theta_2, 1)$ were generated, where p is chosen uniformly in [0,1], while θ_1 and θ_2 are chosen uniformly at random in [-5,5].

Remarks. When Θ is a subset in an infinite dimensional space, it is not uncommon that p_G and $p_{G'}$ have disjoint support in $\mathcal{X}$. This happens when the collection of $\{f(\cdot|\theta_i) : i = 1, \ldots, k\}$ and the collection of $\{f(\cdot|\theta'_j) : j = 1, \ldots, k'\}$ do have disjoint supports on $\mathcal{X}$. Under this scenario, f-divergences such as variational, Hellinger distance and Kullback-Leibler distance are insensitive, since $d_h(p_G, p_{G'}) \equiv d_h(f_i, f'_j) \equiv d_V(p_G, p_{G'}) \equiv d_V(f_i, f'_j) = 1$, and $d_K(f_i, f'_j) = d_K(p_G, p_{G'}) = \infty$. One the other hand, Wasserstein distances may still be utilized by adopting suitable choices for distance (metric) ρ. This is illustrated in the following example.

Example 6. Let $\Theta = l_\infty(T)$ be a separable Banach space of uniformly bounded functions $\theta : T \to \mathbb{R}$ equipped with the uniform norm $\|\theta\| = \sup\{|\theta(t)| : t \in T\}$. Given $\theta \in \Theta$, define random function $X : T \to \mathbb{R}$ by taking $X(t) = \theta(t) + W(t)$, where $(W_t)_{t\in T}$ is distributed according to a zero-mean Gaussian stochastic process given by a continuous covariance kernel $K : T \times T \to \mathbb{R}$. We write $W \sim GP(0, K)$. This specifies the conditional distribution of X given θ. Suppose that $\theta_1, \ldots, \theta_k$ are iid draws from the same Gaussian process $GP(0, K)$, while $\theta'_1, \ldots, \theta'_{k'}$ are iid draws from GP$(0, K')$, using a different continuous covariance kernel K'. Thus, we obtain "mixture of Gaussian processes" distributions p_G and $p_{G'}$, a type of models that have been studied recently in the modeling literature (e.g., [13, 29]). We note that the support of Gaussian process $GP(0, K)$ is the closure in the uniform norm of the reproducing kernel Hilbert space $\mathbb{H}(K)$ given by K [22], see also [41]. Moreover, given a fixed θ'_j such that $\theta'_j \notin \mathbb{H}(K)$, then the distributions for $\theta'_j + W$ and W are orthogonal. Lukić and Beder [26] specify non-pathological conditions under which sample paths of $GP(0, K')$ do not belong to $\mathbb{H}(K)$, so that $\theta'_j \notin \mathbb{H}(K)$ with probability one. This entails the orthogonality between the distributions for $\theta_i + W$ and $\theta'_j + W$ for any pair of i, j with probability one. Thus, divergence functionals such as $d_V \equiv d_h \equiv 1$ or $d_K \equiv \infty$ are insensitive in this setting. This is because the RKHS norm which controls the the support of Gaussian processes is too strong to be useful in this context. Wasserstein distance $d_\rho(G, G')$ using a suitably weaker metric, e.g., $\rho(\theta, \theta') = \sup_{t\in T} |\theta(t) - \theta'(t)|$ may be still used to distinguish G and G' in a meaningful way.

3.2 Wasserstein metric identifiability in mixture models

We have shown in the previous section that for many choices of ρ, d_ρ yields a stronger topology on $\bar{\mathcal{G}}(\Theta)$ than the weak 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 $\mathcal{G}(\Theta)$, along with identifiability conditions for the family of likelihood functions $\{f(\cdot|\theta), \theta \in \Theta\}$.

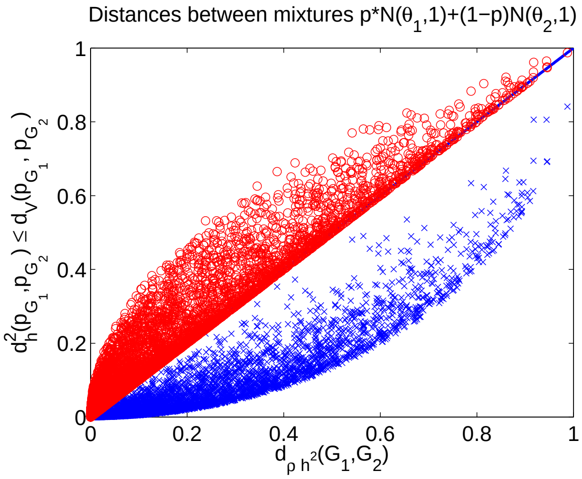

Figure 1: Scatter plot comparing Wasserstein distance functional $d_{\rho h^2}(G_1, G_2)$ against squared Hellinger distance $d_h^2(p_{G_1}, p_{G_2})$ (in x's) and the variational distance $d_V(p_{G_1}, p_{G_2})$ (in o's).

The classical definition of [39] 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 shall need to a slightly stronger version, allowing for the inclusion for discrete measures with infinite support:

Definition 4. *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 [6], herein adapted to a multivariate setting.

Definition 5. *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*

$$\operatorname*{ess\,sup}_{x \in \mathcal{X}} \left| \sum_{i=1}^{k} \alpha_i f(x|\theta_i) + \beta_i^T Df(x|\theta_i) + \gamma_i^T D^2 f(x|\theta_i)\gamma_i \right| = 0 \tag{3}$$

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 [38], see also Theorem 1 of [21]. Chen identified a broad class of families, including the Gaussian family, for which the strong identifiablity condition holds [6]. The notion of strong identifiability may also be extended to infinite dimensional settings in a standard way via first and second order Fréchet derivatives in normed spaces.

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

Lemma 7. *Suppose that Θ is compact, 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*

$$|\gamma^T (D^2 f(x|\theta_1) - D^2 f(x|\theta_2))\gamma| \leq C\rho(\theta_1, \theta_2)^\delta \|\gamma\|^2, \tag{4}$$

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 $k < \infty$:

$$\lim_{\epsilon \to 0} \inf_{G, G' \in \mathcal{G}_k(\Theta)} \biggl\{ \psi(G, G') : d_\rho(G_0, G) \vee d_\rho(G_0, G') \leq \epsilon \biggr\} > 0. \tag{5}$$

The assertion also holds with ψ being replaced by ψ_1.

Proof. Suppose that Eq. (5) is not true, then there will be sequences of G_n and G'_n tending to G_0 in d_ρ 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 $d_{\rho^2}(G_n, G'_n) = \sum_{ij} q_{ij}^{(n)} \rho^2(\theta_{n,i}, \theta'_{n,j})$. Note that $d_\rho^2(G_n, G'_n) \leq d_{\rho^2}(G_n, G'_n) = O(d_\rho(G_n, G'_n))$, while the latter inequality is due to the boundedness of Θ.

Let $\mathcal{O}_n = \{(i,j) : \rho^2(\theta_{n,i}, \theta'_{n,j}) \leq d_{\rho^2}^{1-\delta'}(G_n, G'_n)\}$ for some $\delta' \in (0,1)$. Then, $\sum_{(i,j)\notin \mathcal{O}_n} q_{ij}^{(n)} \leq d_{\rho^2}(G_n, G'_n)/d_{\rho^2}^{1-\delta'}(G_n, G'_n) \to 0$ as $n \to \infty$. Since $\boldsymbol{q}^{(n)} \in \mathcal{Q}(\boldsymbol{p}_n, \boldsymbol{p'}_n)$, we can express

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

and, by Taylor's expansion,

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

where

$$R_n(x) = O\Big(\sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}\rho^{2+\delta}(\theta_{n,i},\theta'_{n,j})\Big) = O\Big(\sum_{(i,j)\in\mathcal{O}_n} q_{ij}^{(n)}\rho^2(\theta_{n,i},\theta'_{n,j})d_{\rho^2}^{(1-\delta')\delta/2}(G_n,G'_n)\Big)$$

due to Eq. (4) and the definition of $\mathcal{O}_n$. So $R_n(x)/d_{\rho^2}(G_n,G'_n) \to 0$. The quantities $A_n(x), B_n(x)$ and $C_n(x)$ are linear functionals 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 properly rescaled, the limits of $A_n(x), B_n(x)$ and $C_n(x)$ are still linear functionals with constant coefficients not depending on x. In particular, $C_n(x)/D_n \to \sum_{j=1}^{k^*}\gamma_j^T D^2 f(x|\theta^*_j)\gamma_j$ for some γ_j's and not all these coefficients vanishing, since $\sum_{j=1}^{k^*}\|\gamma_j\|^2 = 1$. The coefficients in $A_n(x)/D_n$ and $B_n(x)/D_n$ can go either to infinity or to a constant by further selecting the subsequences of G_n and G'_n. If they go to infinity, a sequence $d_n = O(1)$ can be found such that $d_nA_n(x)/D_n$ converges to $\sum_{j=1}^{k^*}\alpha_j f(x|\theta^*_j)$ and $d_nB_n(x)/D_n$ converges to $\sum_{j=1}^{k^*}\beta_j^T Df(x|\theta^*_j)$ for some finite α_j and β_j. Thus, we have d_n and $\alpha_j,\beta_j,\gamma_j$, not all being zero, such that

$$d_n|p_{G_n}(x) - p_{G'_n}(x)|/d_\rho^2(G_n,G'_n) \to \Bigg|\sum_{j=1}^{k^*}\alpha_j f(x|\theta_j) + \beta_j^T Df(x|\theta_j) + \gamma_j^T D^2 f(x|\theta_j)\gamma_j\Bigg|. \tag{6}$$

for all x. This entails that the right side of the preceeding display must be 0 for all 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.(6) 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. □

Remarks. (i) Suppose that G_0 has exactly k support points in Θ. 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 using an argument in the first paragraph of the proof of Lemma 3(c), 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 for the Lemma then proceeds as before.

(ii) When $f(x|\theta)$ is a multivariate normal density (as in Example 5) we also know that $d_{\rho^2}(G,G') \geq 8d_{\rho h^2}(G,G')$. Thus, $d_V(p_G,p_{G'}) > 8cd_{\rho h^2}(G,G')$ for some small constant $c > 0$. This clarifies the relationship between $d_V(p_G,p_{G'})$ and $d_{\rho h^2}(G,G')$ reported empirically in Fig 1.

(iii) For the rest of this paper, by strong identifiability we always mean conditions specified in Lemma 7 so that Eq. (5) can be deduced. This practically means that the conditions specified by Eq. (3) and Eq. (4) be given, while the compactness of Θ may sometimes be required (see Remark (i) in the preceeding paragraph).

The following notion plays a central role in the sequel.

Definition 6. *Let $\mathcal{G}$ be a subset of $\bar{\mathcal{G}}(\Theta)$. For each $k < \infty$, define the Hellinger information of d_ρ metric as a real-valued function on the real line $C_k(\mathcal{G}, \cdot) : \mathbb{R} \to \mathbb{R}$:*

$$C_k(\mathcal{G}, r) = \inf_{G_0 \in \mathcal{G}_k(\Theta), G \in \mathcal{G} : d_\rho(G_0, G) \geq r/2} d_h^2(p_{G_0}, p_G)/2. \tag{7}$$

It is obvious that C_k is a non-negative and non-decreasing function.

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

(b) Suppose that Θ is compact, and the family of likelihood functions is strongly identifiable. Then, for some constant $c > 0$, $C_k(\mathcal{G}_k(\Theta), r) \geq cr^4$ for all $r \geq 0$.

Furthermore, the compactness of Θ can be replaced by the assumption that G_0 has exactly k distinct atoms, and that Θ is bounded.

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 $d_\rho(G_0, G_2) \geq r/2 > 0$ always holds, and that converges in d_ρ 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 $d_\rho(G_0^*, G^*) \geq r/2 > 0$, so $G_0^* \neq G^*$ by Lemma 1. At the same time, $d_h(p_{G_0^*}, p_{G^*}) = 0$, 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 consequences of Lemma 7, by noting that under the given hypothesis, there is $c > 0$ such that $d_h^2(p_{G_0}, p_G) \geq d_V^2(p_{G_0}, p_G)/2 \geq cd_{\rho^2}^2(G_0, G) \geq cd_\rho^4(G_0, G)$ for sufficiently small $d_\rho(G_0, G)$. The boundedness of Θ implies the boundedness of $d_\rho(G_0, G)$, thereby extending the claim for the entire admissible range of $d_\rho(G_0, G)$. □

3.3 Minimum distance nonparametric estimator and existence of tests

We shall present several immediate consequences of the results in the previous section on the estimation and testing in a mixture model. In both settings, let $X_1, \ldots, X_n$ be an iid sample from the mixture distribution $p_{G_0}(x) = \int f(x|\theta) G_0(\theta) d\mu$, where G_0 is the "true" discrete measure in $\bar{\mathcal{G}}(\Theta)$. The likelihood function $f(\cdot|\theta)$ is given.

Estimation. First, we consider a class of minimum distance estimation method for discrete measure G, and discuss the convergence rates. Let $\hat{p}_n$ denotes a nonparametric estimate of the density function p_G, using any well-known density estimation methods (e.g., the penalized maximum likelihood method, methods of sieves, or kernel density estimation). Then, an estimate $\hat{G}_n$ for G_0 is obtained by solving the following optimization:

$$\hat{G}_n = \inf_{G \in \mathcal{G}} d_h(\hat{p}_n, p_G),$$

where the infimum is taken over the compact set $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$.

Theorem 2. *Suppose that $G_0 \in \mathcal{G}(\Theta)$, i.e., G_0 has finite support, $G_0 \in \mathcal{G}$, and $d_h(\hat{p}_n, p_{G_0}) = O_P(\epsilon_n) = o_P(1)$.*

(a) Under the condition of finite identifiablity, $d_\rho(\hat{G}_n, G_0) = o_P(1)$.

(b) Suppose that that $\mathcal{G} = \mathcal{G}_k(\Theta)$ for some known $k < \infty$, Θ is compact and the family of likelihood functions is strongly identifiable. Then,

$$d_\rho(\hat{G}_n, G_0) = O_P(\epsilon_n^{1/2}).$$

Proof. (a) By the definition of $\hat{G}_n$, and the assumption that $G_0 \in \mathcal{G}$, $d_h(\hat{p}_n, p_{\hat{G}_n}) \leq d_h(\hat{p}_n, p_{G_0})$. By triangle inequality, $d_h(p_{G_0}, p_{\hat{G}_n}) \leq d_h(p_{G_0}, \hat{p}_n) + d_h(\hat{p}_n, p_{\hat{G}_n)} \leq 2d_h(p_{G_0}, \hat{p}_n) = O_P(\epsilon_n)$. Suppose that $\hat{G}_n$ does *not* converge to G_0 in d_ρ metric. Due to the compactness of $\mathcal{G}$, there is a subsequence of $\hat{G}_n$ which converges to $G_1 \neq G_0$. This leads to $d_h(p_{G_1}, p_{G_0}) = 0$, so that $p_{G_1} = p_{G_0}$ for almost all $x \in \mathcal{X}$. Noting that $G_0 \in \mathcal{G}(\Theta)$ by assumption, and $G_1 \in \bar{\mathcal{G}}(\Theta)$. By finite identifiability, $G_1 = G_0$, which leads to contradiction.

(b) From Theorem 1(b), there is a constant $c > 0$ such that $d_h(p_{G_0}, p_{\hat{G}_n}) \geq c d_\rho^2(G_0, \hat{G}_n)$. Combining with part (a) to obtain the desired claim. □

Existence of tests. Next, we explore the existence of tests for discriminating a discrete measure with finite support against a another class of discrete measures using data samples from a mixture distribution. These types of result are also useful for the subsequential analysis on the convergence of posterior distributions of the mixing distribution.

A test ϕ_n is an indicator function of the iid sample $X_1, \ldots, X_n$. For each pair of discrete measures G_0, G_1 we consider tests for discriminating $G_0 \in \mathcal{G}(\Theta)$ against a closed ball $B(G_1, d_\rho(G_0, G_1)/2) = \{G \in \bar{\mathcal{G}}(\Theta) : d_\rho(G_1, G) \leq d_\rho(G_1, G_0)/2\}$. In the following P_G denotes the expectation under the mixture distribution given by density p_G.

Lemma 8. *For some fixed $k < \infty$, suppose that $\mathcal{G}_k(\Theta) \subseteq \mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$. Then, for every pair of discrete measures $(G_0, G_1) \in \mathcal{G}_k(\Theta) \times \mathcal{G}$ there exist tests $\{\varphi_n\}$ that have the following properties:*

$$\begin{aligned} P_{G_0}\varphi_n &\leq \exp[-nC_k(\mathcal{G}, d_\rho(G_0, G_1))] && (8) \\ \sup_{G \in \bar{\mathcal{G}}(\Theta): d_\rho(G, G_1) < d_\rho(G_0, G_1)/2} P_G(1 - \varphi_n) &\leq \exp[-nC_k(\mathcal{G}, d_\rho(G_0, G_1))]. && (9) \end{aligned}$$

Proof. Let $r = d_\rho(G_0, G_1) > 0$. Consider pairs of sets $\mathcal{P}_0$ and $\mathcal{P}_1$, where $\mathcal{P}_0 = \{p_{G_0}\}$, a singleton set of only one mixture distribution given by $G_0 \in \mathcal{G}_k(\Theta)$, and $\mathcal{P}_1 = \{p_G | d_\rho(G, G_1) \leq r/2\}$. According to Lemma 1, the closed ball $\{G : d_\rho(G, G_1) \leq r/2\}$ is a convex set. This implies that $\mathcal{P}_1$ is a convex set of mixture distributions. Now, applying a result from Birgé

[5] and Le Cam ([23], Lemma 4, pg. 478) there exist tests φ_n that discriminating between two convex sets $\mathcal{P}_0$ and $\mathcal{P}_1$ such that:

$$\begin{aligned} P_{G_0}\varphi_n &\leq \exp[-n \inf d_h^2(P_0, P_1)/2]. \\ \sup_{G \in \bar{\mathcal{G}}(\Theta): d_\rho(G, G_1) \leq r/2} P_G(1-\varphi_n) &\leq \exp[-n \inf d_h^2(P_0, P_1)/2], \end{aligned}$$

where the exponent in the upper bounds are given by the minimum Hellinger distance among *all* such pairs of $(P_0, P_1) \in (\mathcal{P}_0, \mathcal{P}_1)$. Due to the triangle inequality, if $d_\rho(G_0, G_1) = r$ and $d_\rho(G_1, G) \leq r/2$ then $d_\rho(G_0, G) \geq r/2$. So $C_k(\mathcal{G}, r) \leq \inf d_h^2(P_0, P_1)/2$. This completes the proof. □

Next, we show the existence of test for discriminating G_0 against the complement of a closed ball, by adapting a result of [16] (Theorem 7.1):

Lemma 9. *Let $\mathcal{G} \subseteq \bar{\mathcal{G}}(\Theta)$. $G_0 \in \mathcal{G}_k(\Theta) \subseteq \mathcal{G}$ for some $k < \infty$. Suppose that for some non-increasing function $D(\epsilon)$, some $\epsilon_n \geq 0$ and every $\epsilon > \epsilon_n$,*

$$D(\epsilon/2, \{G \in \mathcal{G} : \epsilon \leq d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq D(\epsilon). \tag{10}$$

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

$$\begin{aligned} P_{G_0}\varphi_n &\leq D(\epsilon) \sum_{t=1}^{\lceil \mathit{Diam}(\Theta)/\epsilon \rceil} \exp[-nC_k(\mathcal{G}, t\epsilon)] && (11) \\ \sup_{G \in \mathcal{G}: d_\rho(G_0, G) > t\epsilon} P_G(1-\varphi_n) &\leq \exp[-nC_k(\mathcal{G}, t\epsilon)]. && (12) \end{aligned}$$

Proof. For a given $t \in \mathbb{N}$ choose a maximal $t\epsilon/2$-packing in set $S_t = \{G : t\epsilon < d_\rho(G_0, G) \leq (t+1)\epsilon\}$. This yields a set S'_t of at most $D(t\epsilon)$ points, by assumption. 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. (8) and (9). Take φ_n to be the maximum of all tests attached this way to some point $G_1 \in S'_t$ for some $t \in \mathbb{N}$. Then, by union bound,

$$P_{G_0}\varphi_n \leq \sum_t \sum_{G_1 \in S'_t} \exp[-nC_k(\mathcal{G}, t\epsilon)] \leq D(\epsilon) \sum_t \exp[-nC_k(\mathcal{G}, t\epsilon)]$$

$$\sup_{G \in \cup_{u \geq t} S_u} P_G(1-\varphi_n) \leq \sup_{u \geq t} \exp[-nC_k(\mathcal{G}, u\epsilon)] \leq \exp[-nC_k(\mathcal{G}, t\epsilon)],$$

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

4 Convergence of posterior distributions of discrete measures

In this section we study the convergence of discrete mixing distributions in a Bayesian setting. Let $X_1, \ldots, X_n$ be an iid sample according to the mixture distribution $P_G(x) =$

$\int f(x|\theta)G(\theta)d\mu$, where f is known, while $G = G_0$ for some unknown discrete measure in $\mathcal{G}_k(\Theta)$. In the Bayesian 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 d_ρ neighborhoods of G_0, and derive the rates of the convergence. The analysis is based upon the general framework of Ghosal, Ghosh and van der Vaart [16], who analyzed the convergence of posterior distributions in terms of f-divergences such as Hellinger and variational distances on the mixture densities of the data. By contrast, the following results are based on conditions formulated directly in terms of the Wasserstein distance, which are simpler to verify for mixture models. The claims of convergence rates are with respect to Wasserstein distance metrics, which are stronger than the f-divergence metrics, and directly related to the convergence behavior of the individual atoms that provide support for the mixing distribution G.

Our 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. The second 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 third is on the Hellinger information of d_ρ metric, function $C_k(\mathcal{G}, r)$. Define the Kullback-Leibler neighborhood:

$$B(\epsilon) = \left\{ G \in \bar{\mathcal{G}}(\Theta) : -p_{G_0}\left(\log \frac{p_G}{p_{G_0}}\right) \leq \epsilon^2, p_{G_0}\left(\log \frac{p_G}{p_{G_0}}\right)^2 \leq \epsilon^2 \right\}. \tag{13}$$

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 $n\epsilon_n^2 \to \infty$, sets $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$ and a constant $C > 0$, we have*

$$\log D(\epsilon_n, \mathcal{G}_n, d_\rho) \leq n\epsilon_n^2, \tag{14}$$

$$\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) \leq \exp[-n\epsilon_n^2(C+4)], \tag{15}$$

$$\Pi(B(\epsilon_n)) \geq \exp(-n\epsilon_n^2 C). \tag{16}$$

Moreover, suppose M_n is a sequence so that

$$C_k(\mathcal{G}_n, M_n\epsilon_n) \geq \epsilon_n^2(C+4), \tag{17}$$

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)] \to 0. \tag{18}$$

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

A stronger theorem (using a substantially weaker condition on the covering number) can be formulated as follows.

Theorem 4. *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 $\epsilon_n \to 0$ such that $n\epsilon_n^2$ is bounded away from 0 or tending to infinity, and sets $\mathcal{G}_n \subset \bar{\mathcal{G}}(\Theta)$, we have*

$$\log D(\epsilon/2, \{G \in \mathcal{G}_n : \epsilon \leq d_\rho(G_0, G) \leq 2\epsilon\}, d_\rho) \leq n\epsilon_n^2 \text{ for every } \epsilon \geq \epsilon_n, \tag{19}$$

$$\frac{\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n)}{\Pi(B(\epsilon_n))} = o(\exp(-2n\epsilon_n^2)), \tag{20}$$

$$\frac{\Pi(G : j\epsilon_n < d_\rho(G, G_0) \leq 2j\epsilon_n)}{\Pi(B(\epsilon_n))} \leq \exp[nC_k(\mathcal{G}_n, j\epsilon_n)/2] \text{ for any } j \geq M_n, \tag{21}$$

where M_n is a sequence such that

$$\exp(2n\epsilon_n^2) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/2] \to 0. \tag{22}$$

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

Remarks. (i) The above statement continues to hold if conditions (21) and (22) are replaced by the following condition:

$$\exp(2n\epsilon_n^2)/\Pi(B(\epsilon_n)) \sum_{j \geq M_n} \exp[-nC_k(\mathcal{G}_n, j\epsilon_n)] \to 0. \tag{23}$$

(ii) The above theorems are stated for the L_1 Wasserstein metric d_ρ, but they also hold for L_2 Wasserstein metric $d_{\rho^2}^{1/2}$, with a slight modification of the definition of the Hellinger information function to $C_k(\mathcal{G}, r) = \inf_{d_{\rho^2}^{1/2}(G_0,G) \geq r/2} d_h^2(p_{G_0}, p_G)$.

In the sequel, the general theory is illustrated in specific examples of hierarchical mixture distributions.

4.1 Finite mixture of multivariate distributions

Let Θ be a compact 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) The family of likelihood functions $f(\cdot|\theta)$ is strongly identifiable. In addition, for some constant $C_1 > 0$, $d_K(f_i, f_j') \leq C_1\rho^2(\theta_i, \theta_j')$ for any $\theta_i, \theta_j' \in \Theta$.

(A2) For any $G \in \text{supp}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 < C_2 d_K(p_{G_0}, p_G)$ 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 $c_3\delta^{kd} \leq \Pi(\rho(\theta_i, \theta_i^*) \leq \delta, i = 1 \ldots, k) \leq C_3\delta^{kd}$ for some constants $c_3, C_3 > 0$.

Remarks. (A1) holds for the family of Gaussian densities with parameter mean θ. (A3) holds when the prior distribution on the relevant parameters behave like a uniform distribution, up to a multiplicative constant. (A2) holds when p_{G_0}/p_G is bounded from below for any $G \in \text{supp}(\Pi)$.

Let $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$. Combining Lemma 6 with Assumption (A1), if $\rho(\theta_i, \theta_i^*) \leq \epsilon$ and $|p_i - p_i^*| \leq \epsilon^2/(k\text{Diam}(\Theta)^2)$ for $i = 1, \ldots, k$, then $d_K(p_{G_0}, p_G) \leq d_{\rho K}(G_0, G) \leq C_1 \sum_{1 \leq i,j \leq k} q_{ij} \rho^2(\theta_i^*, \theta_j)$, for any $\boldsymbol{q} \in \mathcal{Q}$. Thus, $d_K(p_{G_0}, p_G) \leq C_1 d_{\rho^2}(G_0, G) \leq C_1 \sum_{i=1}^{k} (p_i^* \wedge p_i) \rho^2(\theta_i^*, \theta_i) + C_1 \sum_{i=1}^{k} |p_i - p_i^*| \text{Diam}(\Theta)^2 \leq 2C_1 \epsilon^2$. Hence, under prior Π,

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

In view of Assumptions (A2) and (A3), we have $\Pi(B(\epsilon)) \gtrsim \epsilon^{k(d+2)}$. Conversely, for sufficiently small ϵ, if $d_\rho^2(G_0, G) \leq \epsilon^2$ then by reordering the index of the atoms, we must have $\rho(\theta_i, \theta_i^*) = O(\epsilon)$ and $|p_i - p_i^*| = O(\epsilon^2)$ for all $i = 1, \ldots, k$ (see the argument in the proof of Lemma 3(c)). This entails that under the prior Π,

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

Let $\epsilon_n = n^{-1/2}$. We proceed by verifying conditions of Theorem 4, as this theorem provides the right rate for parametric mixture models under the L_2 Wasserstein distance metric $d_{\rho^2}^{1/2}$. Let $\mathcal{G}_n := \mathcal{G}_k(\Theta)$. Then $\Pi(\bar{\mathcal{G}}(\Theta) \setminus \mathcal{G}_n) = 0$, so Eq. (20) trivially holds.

Next, we show that $D(\epsilon/2, S, d_{\rho^2}^{1/2})$, where $S = \{G \in \mathcal{G}_n : d_\rho(G_0, G) \leq 2\epsilon\}$, is bounded above by a constant, so that (19) is satisfied. Indeed, for any $\epsilon > 0$, $\log D(\epsilon/2, S, d_{\rho^2}^{1/2}) \leq \log N(\epsilon/4, S, d_{\rho^2}^{1/2}) \leq N(\epsilon/4, S, d_\rho)$. By Lemma 3 (c), $N(\epsilon/4, S, d_\rho)$ 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. Thus, Eq. (19) holds.

By Theorem 1(b) and Assumption (A4), there is $C_k(\mathcal{G}_n, j\epsilon_n) = \inf_{d_{\rho^2}(G_0,G) \geq (j\epsilon/2)^2} d_h^2(p_{G_0}, p_G) \geq c(j\epsilon_n)^4$ for some constant $c > 0$. To ensure condition (22), note that:

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

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, $\Pi(G : j\epsilon_n < d_\rho(G, G_0) \leq 2j\epsilon_n)/\Pi(B(\epsilon_n)) = O(1)$. On the other hand, for $j \geq M_n$, we have $\exp[nC_k(\mathcal{G}_n, j\epsilon_n)/2] \geq \exp[nc(M_n\epsilon_n)^4/2]$ which is bounded below by arbitrarily large constant by choosing M_n to be a large multiple of $\epsilon_n^{-1/2}$, thereby ensuring (21).

Thus, by Theorem 4, rate of contraction for the posterior distribution of G under $d_{\rho^2}^{1/2}$ distance metric is $n^{-1/4}$, which is also the minimax rate $n^{-1/4}$ as proved in the univariate case by [6]. Our calculation is summarized by:

Proposition 3. *Under Assumptions (A1–A3), the contraction rate in the L_2 Wasserstein distance metric of the posterior distribution of G is $n^{-1/4}$.*

4.2 Finite mixture of Gaussian processes

Let Θ be the space of bounded functions $\theta : T \to \mathbb{R}$. Suppose that the "true" G_0 has k distinct atoms in Θ: $G_0 = \sum_{i=1}^{k} p_i^* \delta_{\theta_i^*}$, while the prior Π places with probability one the space of mixtures of k zero-mean Gaussian processes $GP(0, K)$ on Θ, where $k < \infty$ is known. K is the given continuous covariance function $K : T \times T \to \mathbb{R}$. A discrete measure G in the support of Π takes the form $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$, where θ_i are iid draws from $GP(0, K)$.

The Gaussian process can be viewed as a tight Borel measurable map in a separable Banach space $\Theta := \mathbb{B} = l_\infty(T)$, equipped with the uniform norm $\|\theta\| = \sup\{|\theta(t)| : t \in T\}$. The support of a centered (i.e., zero mean) version of such a process is equal to the closure of the reproducing kernel Hilbert space (RKHS), to be denoted by $\mathbb{H}(K)$, or simply $\mathbb{H}$, of the covariance kernel K of the process [22]. Assume that $G_0 \in \mathcal{G}_k(\bar{\mathbb{H}})$.

Our analysis of convergence rate of posterior distributions for mixture models based on Gaussian processes are built upon a recent work by van der Vaart and van Zanten [41]. Let ρ be the metric on Θ using the uniform norm, $\rho(\theta_i, \theta_j) = \|\theta_i - \theta_j\|$. For each $i = 1, \ldots, k$, define the concentration function:

$$\phi_{\theta_i^*}(\epsilon) = \inf_{h \in \mathbb{H} : \rho(h, \theta_i^*) < \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|\theta\| < \epsilon),$$

where Pr denotes the probability under the zero mean Gaussian measure. Let $(\epsilon_n)_{n \in \mathbb{N}}$ be a sequence of positive numbers that satisfy the condition that for any $i = 1, \ldots, k$ and any n,

$$\phi_{\theta_i^*}(\epsilon_n) \leq n\epsilon_n^2. \tag{24}$$

For a given constant $C_0 > 1$ with $e^{-C_0 n\epsilon_n^2} < 1/2$, define a sequence of measurable sets $(B_n)_{n \geq 1}$, $B_n = \epsilon_n \mathbb{B}_1 + C_n \mathbb{H}_1$, where $C_n = -2\Phi^{-1}(e^{-C_0 n\epsilon_n^2})$. By Theorem 2.1 of [40], the sequence of sets B_n admits the following useful properties:

$$\log N(3\epsilon_n, B_n, \rho) \leq 6C_0 n\epsilon_n^2, \tag{25}$$

$$\Pr(\theta \notin B_n) \leq e^{-C_0 n\epsilon_n^2}, \tag{26}$$

$$\Pr(\rho(\theta, \theta_i^*) < 2\epsilon_n) \geq e^{-n\epsilon_n^2} \text{ for each } i = 1, \ldots k. \tag{27}$$

We need additional assumptions:

Assumptions B.

(B1) The family of likelihood functions $\{f(\cdot|\theta), \theta \in \Theta\}$ is strongly identifiable for the infinite dimensional Θ. In addition, for some constant $C_1 > 0$, $d_K(f_i, f_j') \leq C_1 \rho^2(\theta_i, \theta_j')$ for any $\theta_i, \theta_j' \in \Theta$.

(B2) For any $G \in \text{supp}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 < C_2 d_K(p_{G_0}, p_G)$ for some constant $C_2 > 0$.

(B3) Under prior Π, for small $\delta > 0$, $\Pi(|p_i - p_i^*| \leq \delta, i = 1 \ldots, k) \geq c_3 \delta^{k\alpha}$ for some constants $c_3, \alpha > 0$.

(B4) Under prior Π, $C_4 = \mathbb{E}\|\theta\|^2 < \infty$.

Proposition 4. *Given Assumptions (B1–B4). Let ϵ_n be a sequence tending to 0 such that $\log n = o(n\epsilon_n^2)$ and Eq. (24) holds. Then, for a sufficiently large constant $M > 0$, $\Pi(d_\rho(G_0, G) \geq M\epsilon_n^{1/2}|X_1, \ldots, X_n) \to 0$ in P_{G_0}-probability.*

Remarks. The reader is referred to [41] for examples of the convergence rate ϵ_n that satisfy condition (24) for different choices of Θ. In particular, if under prior Π, the support of the prior distribution on Θ are functions on T with smoothness $\gamma_1 > 0$, while the "true" support points θ_i^*'s of G_0 are functions with smoothness $\gamma_2 > 0$, then concentration functions $\phi_{\theta_i^*}(\epsilon) = \epsilon^{-1/(\gamma_1 \wedge \gamma_2)}$ for each $i = 1, \ldots, k$. Accordingly, the rate ϵ_n for which Eq. (24) holds is $\epsilon_n \asymp n^{-\frac{\gamma_1 \wedge \gamma_2}{2\gamma_1 \wedge \gamma_2 + 1}}$. The contraction rate for the posterior distribution of G is $n^{-\frac{\gamma_1 \wedge \gamma_2}{2(2\gamma_1 \wedge \gamma_2 + 1)}}$.

4.3 Infinite mixtures based on the Dirichlet process

Given the "true" discrete measure $G_0 = \sum_{i=1}^k p_i^* \delta_{\theta_i^*} \in \mathcal{G}_k(\Theta)$, but $k < \infty$ 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 $\nu > 0$. Here, parameter P_0 is a probability measure on Θ.

The following lemma provides a lower bound of small ball probabilities of metric space $(\bar{\mathcal{G}}(\Theta), d_{\rho^2}^{1/2})$ in terms of small probabilities of metric space (Θ, ρ).

Lemma 10. *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,*

$$\Pi(G : d_{\rho^2}(G_0, G) \leq 5\epsilon^2) \geq \Gamma(\nu)(\epsilon^2 D^{-1}/\mathit{Diam}(\Theta))^{D-1}\nu^D \prod_{i=1}^{D} P_0(S_i).$$

Here $(S_1, \ldots, S_D)$ denotes the D disjoint $\epsilon/2$-balls whose centers form a maximal ϵ-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, $\boldsymbol{m} = (m_1, \ldots, m_D) \sim \text{Dir}(\nu\mu_1, \ldots, \nu\mu_D)$. Note that

$$d_{\rho^2}(G_0, G) \leq 4\epsilon^2 + \|\boldsymbol{m} - \hat{\boldsymbol{p}}\|_1[\text{Diam}(\Theta)]^2.$$

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,

$$\begin{aligned}
&\Pr(d_\rho(G_0, G) \leq 5\epsilon^2) \geq \Pr(\|\boldsymbol{m} - \hat{\boldsymbol{p}}\|_1 \leq \delta^2) \geq \Pr(|m_i - \hat{p}_i| \leq \delta^2/D, i = 1, \ldots, D) \\
&= \frac{\Gamma(\nu)}{\prod_{i=1}^D \Gamma(\nu\mu_i)} \int_{\Delta_{D-1} \cap |m_i - \hat{p}_i| \leq \delta^2/D} \prod_{i=1}^{D-1} m_i^{\nu\mu_i - 1}(1 - \sum_{i=1}^{D-1} m_i)^{\nu\mu_D - 1} dm_i \ldots dm_{D-1} \\
&\quad \geq \frac{\Gamma(\nu)}{\prod_{i=1}^D \Gamma(\nu\mu_i)} \prod_{i=1}^{D-1} \int_{\max(\hat{p}_i - \delta^2/D, 0)}^{\min(\hat{p}_i + \delta^2/D, 1)} m_i^{\nu\mu_i - 1} dm_i \geq \Gamma(\nu)(\delta^2/D)^{D-1} \prod_{i=1}^D (\nu\mu_i).
\end{aligned}$$

(The second 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 third inequality is due to the fact that $\Gamma(\alpha) \leq 1/\alpha$ for $0 < \alpha \leq 1$). This gives the desired claim. □

Assumptions C.

(C1) The non-atomic base measure P_0 places full support on a compact set Θ. The family of the likelihood densities $f(\cdot|\theta)$ is finitely identifiable. Moreover, for some constant $C_1 > 0$, $d_K(f_i, f'_j) \leq C_1 \rho^2(\theta_i, \theta'_j)$ for any $\theta_i, \theta'_j \in \Theta$.

(C2) For any $G \in \text{supp}(\Pi)$, $\int p_{G_0}(\log(p_{G_0}/p_G))^2 \leq C_2 d_K(p_{G_0}, p_G)$ for some constant $C_2 > 0$.

(C3) P_0 places sufficient probability mass on all small balls that pack Θ. Specifically, there is a universal constant $c_3 > 0$ such that the probability of the D-partition $(S_1, \ldots, S_D)$ specified in Lemma 10 satisfy for any $\epsilon > 0$:

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

(C4) The packing number $D(\epsilon, \Theta, \rho) \asymp [\text{Diam}(\Theta)/\epsilon]^d$.

Theorem 5. *Under Assumptions (C1–C4), there is a sequence $\beta_n \searrow 0$ such that $\Pi(d_\rho(G_0, G) \geq \beta_n | X_1, \ldots, X_n) \to 0$ in P_{G_0} probability.*

Proof. The proof consists of two main steps. First, we shall prove that under Assumptions (C1–C4), conditions specified by Eqs. (14) (15) (16) in Theorem 3 are satisfied by taking $\mathcal{G}_n = \bar{\mathcal{G}}(\Theta)$, 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 subsequentially $\beta_n = M_n\epsilon_n$ for which Theorem 3 can be applied.

Step 1: By Lemma 6 and (C1), $d_K(p_{G_0}, p_G) \leq d_{\rho K}(G_0, G) \leq C_1 d_{\rho^2}(G_0, G)$. Combining with (C2), we obtain that $\Pi(G \in B(\epsilon_n)) \geq \Pi(G : d_{\rho^2}(G_0, G) \leq C_3\epsilon_n^2)$ for some constant C_3. Combining this bound with (C3) and (C4), which are applied to Lemma 10 we have: $\log \Pi(G \in B(\epsilon_n)) \gtrsim (D-1)\log(\epsilon_n^2/\text{Diam}(\Theta)) + (2D-1)\log(1/D) + D\log\nu$, where $D \asymp [\text{Diam}(\Theta)/\epsilon_n]^d$. It is simple to check that condition (16) holds, $\log \Pi(G \in B(\epsilon_n)) \geq -Cn\epsilon_n^2$, by the given rate of ϵ_n, for any constant $C > 0$.

Since $\mathcal{G}_n = \bar{\mathcal{G}}(\Theta)$, (15) trivially holds. Turning to condition (14), by Lemma 3(b), we have $\log N(2\epsilon_n, \bar{\mathcal{G}}(\Theta), d_\rho) \leq N(\epsilon_n, \Theta, \rho)\log(e + e\text{Diam}(\Theta)/\epsilon_n) \leq (\text{Diam}(\Theta)/\epsilon_n)^d \log(e + e\text{Diam}(\Theta)/\epsilon_n) \leq n\epsilon_n^2$ 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. Moreover, $R_k(\bar{\mathcal{G}}(\Theta), t) \searrow 0$ as $t \to 0$. Indeed, if this is not true, then there is a sequence of $t_m \to 0$ and $\delta > 0$ such that $R_k(\bar{\mathcal{G}}(\Theta), t_m) > \delta$ for all $m \geq 1$. This implies that $C_k(\bar{\mathcal{G}}(\Theta), r) = 0$ for any $r < \delta$, which leads to contradiction by Theorem 1(a), due to the compactness of $\bar{\mathcal{G}}(\Theta)$ and finite identifiability from (C1).

Let $(\epsilon_n)_{n\geq 1}$ be the sequence determined in the previous step of the proof. Let $M_n = R_k(\bar{\mathcal{G}}(\Theta), \epsilon_n^2(C+4))/\epsilon_n$, and $\beta_n = M_n\epsilon_n$. Thus, $\beta_n = R_k(\bar{\mathcal{G}}(\Theta), \epsilon_n^2(C+4)) \to 0$ as $n \to \infty$. Condition (17) holds by definition of R_k, i.e., $C_k(\mathcal{G}(\Theta), M_n\epsilon_n) \geq \epsilon_n^2(C+4)$. To verify (18), note that the running sum with respect to j cannot have more than $\text{Diam}(\Theta)/\epsilon_n$, and due to the monotonicity of C_k, we have

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

Hence, Theorem 3 can be applied to conclude the proof. □

Remark. (i) Assumption (C5) is typical when Θ is a compact subset of a finite dimensional space such as $\mathbb{R}^d$.

(ii) If Assumption (C5) is replaced by the assumption that $\log D(\epsilon, \Theta, \rho) \asymp (1/\epsilon)^{1/\gamma}$, the theorem continues to hold by choosing ϵ_n to be a large multiple of $1/(\log n)^\gamma$. This modified assumption is typical of infinite dimensional spaces of functions with smoothness γ. However, we do not have a concrete example of a probability measure P_0 that places positive or full support on compact set Θ, and that admits usable guarantees on small ball probabilities such as the one stated in (C4).

(iii) As shown in the proof, the rate of convergence of the mixing distribution G is given by $R_k(\mathcal{G}_n, \epsilon_n^2)$, the inverse of the Hellinger information function $C_k(\mathcal{G}_n, \cdot)$. Unlike the case where $\mathcal{G}_n = \mathcal{G}_k(\Theta)$, (see Theorem 1 (b)), the lack of understanding of the behavior of $C_k(\bar{\mathcal{G}}(\Theta), \cdot)$ prevents us from obtaining explicit rates of convergence. We hope to address this issue in a future work.

5 Proofs

Proof of Lemma 2.

Proof. (a) Let $(G_n)_{n\geq 1}$ be a sequence of discrete measures in $\bar{\mathcal{G}}(\Theta)$. 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. The space of the infinite dimensional simplex is compact in the sense that there is a subsequence of G_n such that $\boldsymbol{p}_n = (p_{n,1}, p_{n,2}, \ldots)$ tends to some $\boldsymbol{p} = (p_1, p_2, \ldots)$ in the l_1 distance, and $\sum_i p_i = 1$. Due to the compactness of Θ, within this subsequence G_n there is a subsequence such that $\sup_{i=1}^{\infty} \rho(\theta_{n,i}, \theta_i) \to 0$ for some $\theta_1, \theta_2, \ldots \in \Theta$. Let $G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$. Then, $G \in \bar{\mathcal{G}}(\Theta)$. It is simple to verify that $d_\rho(G, G_n) \leq \sum_{i=1}^{\infty}(p_{n,i} \wedge p_i)\rho(\theta_{n,i}, \theta_i) + \|\boldsymbol{p}_n - \boldsymbol{p}\|_1 \text{Diam}(\Theta) \to 0$ (where $\text{Diam}(\Theta)$ denotes the diameter of Θ). Thus, $\bar{\mathcal{G}}(\Theta)$ is compact. The compactness of $\mathcal{G}_k(\Theta)$ is be shown in a similar manner.

(b) To show that $\mathcal{G}(\Theta)$ is not complete, we will construct a Cauchy sequence of discrete measures with finite support that do not converge to an element in $\mathcal{G}(\Theta)$. Let $(\theta_1, \theta_2, \ldots)$ is a sequence converging to some $\theta^* \in \Theta$. Take any $\gamma \in (0, 1)$. Let $p_1 = \gamma$, $p_2 = (1-\gamma)\gamma, \ldots, p_n = (1-\gamma)^{n-1}\gamma$. Clearly, $\sum_{i=1}^n p_i = 1 - (1-\gamma)^n \to 1$. Let $G_n = \sum_{i=1}^n p_i \delta(\theta_i) + (1-\gamma)^n \delta_{\theta^*}$. Then $d_\rho(G_n, G_{n+1}) \leq [(1-\gamma)^n - (1-\gamma)^{n+1}]\rho(\theta_{n+1}, \theta^*) \to 0$. Thus, G_n tends to some $G \in \bar{\mathcal{G}}(\Theta)$. Let $S = \{\theta_1, \theta_2, \ldots\}_{n=1}^{\infty} \cup \{\theta^*\}$. It is clear that supp$(G)$ is a dense subset of S and vice versa. (Indeed, if some $\theta' \in$ supp(G) with probability mass c, but $\inf_i \rho(\theta', \theta_i) > \epsilon > 0$, then for any n, we have $d_\rho(G, G_n) > c\epsilon$). Thus, supp$(G) = S$, which entails that $G \notin \mathcal{G}(\Theta)$.

(c) Take any $G \in \mathcal{G}(\Theta)$, and arbitrary discrete measure with infinite support $G' \in \bar{\mathcal{G}}(\Theta)$. Take any positive sequence $\alpha_n \to 0$. Then the sequence of $G_n = \alpha_n G' + (1-\alpha_n)G$ clearly converges to G in d_ρ, because $d_\rho(G_n, G) \leq \alpha_n d_\rho(G', G) + (1-\alpha_n) d_\rho(G, G) = \alpha_n d_\rho(G', G) \to 0$ due to convexity of d_ρ (Lemma 1). □

Proof of Lemma 3.

Proof. (a) Suppose that $(\eta_1, \ldots, \eta_T)$ forms an ϵ-covering for Θ under metric ρ, where $T = N(\epsilon, S, \rho)$ denote the (minimum) covering number. Take any discrete measure $G(\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{i=1}^k p_i \delta_{\theta_i}$. 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 $G' = \sum_{i=1}^k p'_i \delta_{\theta'_i}$. Then $d_\rho(G, G') \leq \sum_{i=1}^k (p_i \wedge p'_i)\rho(\theta_i, \theta'_i) + \|\boldsymbol{p} - \boldsymbol{p'}\|_1 \text{Diam}(\Theta) \leq \epsilon + \delta \text{Diam}(\Theta)$. This implies 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.

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) Suppose that $(\eta_1, \ldots, \eta_T)$ forms an ϵ-covering for Θ under metric ρ, and $T = N(\epsilon, S, \rho)$. Take any discrete measure $G(\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{i=1}^k p_i \delta_{\theta_i} \in \bar{\mathcal{G}}_\Theta$, where k may be infinity. The collection of atoms $\theta_1, \ldots, \theta_k$ can be subdivide 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 $\theta_i \in S_t$. Thus, if we define $p'_t = \sum_{i=1}^k p_i \mathbb{I}(\theta_i \in S_t)$, and discrete measure $G'(\boldsymbol{p'}, \boldsymbol{\eta}) = \sum_{t=1}^T p'_t \delta_{\eta_t}$, then we are guaranteed that $d_\rho(G,G') \leq \sum_{i=1}^{k}\sum_{t=1}^{T} p_i \mathbb{I}(\theta_i \in S_t)\rho(\theta_i,\eta_t) \leq \epsilon$.

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 $(\epsilon + \delta\text{Diam}(\Theta))$-covering in the Wasserstein metric for the subset $\mathcal{G}_S \subseteq \mathcal{G}(\Theta)$ 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), $N(\epsilon+\delta\text{Diam}(\Theta),\mathcal{G}_S,d_\rho) \leq (1+1/\delta)^T e^T/\sqrt{2\pi T}$. 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 $q \in \mathcal{Q}(\boldsymbol{p},\boldsymbol{p^*})$ 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, $|p^*_i - p_i| \leq 4\epsilon/m$.

Thus, 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 Theorem 4.

Proof. By a result of Ghosal et al [16] (Lemma 8.1, pg. 524), for every $\epsilon > 0$ and probability measure Π on the set $B(\epsilon)$ defined by Eq. (13), we have, for every $C > 0$,

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

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

$$\int \prod_{i=1}^{n} p_G(X_i)/p_{G_0}(X_i) d\Pi(G) \geq \exp(-2n\epsilon_n^2)\Pi(B(\epsilon_n)). \tag{28}$$

Let $\mathcal{O}_n = \{G \in \bar{\mathcal{G}}(\Theta) : d_\rho(G_0,G) \geq M_n\epsilon_n\}$, $S_{n,j} = \{G \in \mathcal{G}_n : d_\rho(G_0,G) \in [j\epsilon_n,(j+1)\epsilon_n)\}$ for each $j \geq 1$. The conditions specified by Lemma 9 are satisfied by setting $D(\epsilon) = \exp(n\epsilon_n^2)$ (constant in ϵ). Thus there exist tests φ_n for which Eq. (11) and (12) hold. Then,

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

Due to Lemma 9, the first term in the preceeding display is bounded above by $P_{G_0}\varphi_n \leq D(\epsilon_n)\sum_{j\geq M_n}\exp[-nC_k(\mathcal{G}_n, j\epsilon_n)] \to 0$, thanks to Eq. (22). The second term in the above display is bounded by $(Cn\epsilon_n^2)^{-1}$ by the definition of A_n. Since $n\epsilon_n^2$ is bounded away from 0, C can be chosen arbitrarily large so that the second term can be made arbitrarily small. To show that third term in the display also vanishes as $n \to \infty$. we exploit the following expression:

$$\Pi(G \in \mathcal{O}_n|X_1,\ldots,X_n) = \int_{\mathcal{O}_n}\prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\Pi(G)\bigg/\int\prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\Pi(G),$$

and then obtain a lower bound for the denominator by Eq. (28). For the nominator, by Fubini's theorem:

$$\begin{aligned}
& P_{G_0}\int_{\mathcal{O}_n\cap\mathcal{G}_n}(1-\varphi_n)\prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\Pi(G) \\
= \; & P_{G_0}\sum_{j\geq M_n}\int_{S_{n,j}}(1-\varphi_n)\prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\Pi(G) \\
= \; & \sum_{j\geq M_n}\int_{S_{n,j}} P_G(1-\varphi_n)\Pi(G) \leq \sum_{j\geq M_n}\Pi(S_{n,j})\exp[-nC_k(\mathcal{G}_n, j\epsilon_n)], \qquad (29)
\end{aligned}$$

where the last inequality is due to Eq. (12), and by (20),

$$\begin{aligned}
& P_{G_0}\int_{\mathcal{O}_n-\mathcal{G}_n}(1-\varphi_n)\prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\Pi(G) = \int_{\mathcal{O}_n\setminus\mathcal{G}_n} P_G(1-\varphi_n)\Pi(G) \\
\leq \; & \Pi(\bar{\mathcal{G}}(\Theta)\setminus\mathcal{G}_n) = o(\exp(-2n\epsilon_n^2)\Pi(B(\epsilon_n))). \qquad (30)
\end{aligned}$$

Combining bounds (29) and (30) and condition (21), we obtain:

$$\begin{aligned}
& P_{G_0}(1-\varphi_n)\Pi(G\in\mathcal{O}_n|X_1,\ldots,X_n)\mathbb{I}(A_n) \\
\leq \; & \frac{o(\exp(-2n\epsilon_n^2)\Pi(B(\epsilon_n))) + \sum_{j\geq M_n}\Pi(S_{n,j})\exp[-nC_k(\mathcal{G}_n, j\epsilon_n)]}{\exp(-2n\epsilon_n^2)\Pi(B(\epsilon_n))} \\
\leq \; & o(1) + \exp(2n\epsilon_n^2)\sum_{j\geq M_n}\exp[-nC_k(\mathcal{G}_n, j\epsilon_n)/2]
\end{aligned}$$

The upper bound in the preceeding display converges to 0 by Eq. (22), thereby concluding the proof. □

Proof of Theorem 3.

Proof. The proof proceeds in a similar manner to the proof of Theorem 4 up to the application of (12), where we now have: $P_{G_0}\int_{\mathcal{O}_n\cap\mathcal{G}_n}(1-\varphi_n)\prod_{i=1}^n p_G(X_i)/p_{G_0}(X_i)\Pi(G) \leq$

$\int_{\mathcal{O}_n \cap \mathcal{G}_n} P_G(1-\varphi_n)\Pi(G) \leq \exp[-nC_k(\mathcal{G}_n, M_n\epsilon_n)]$, and subsequently,

$$P_{G_0}(1-\varphi_n)\Pi(G \in \mathcal{O}_n | X_1, \ldots, X_n)\mathbb{I}(A_n) \leq \frac{\exp[-n\epsilon_n^2(C+4)] + \exp[-nC_k(\mathcal{G}_n, M_n\epsilon_n)]}{\exp(-2n\epsilon_n^2)\Pi(B(\epsilon_n))}$$

which tends to 0 by conditions (15),(16) and (17) of the theorem. The conclusion then follows. □

Proof of Proposition 4.

Proof. The proof proceeds by verifying that all conditions in Theorem 3 hold for some constant $C > 0$. Define the following sequence of subsets $\mathcal{G}_n := \mathcal{G}_k(B_n) \subset \mathcal{G}_k(\Theta)$. By Assumptions (B1) and (B4), there is $c > 0$ such $C_k(\mathcal{G}_n, r) \geq cr^4$ for sufficiently small $r \geq 0$.

Let $C_4 = \mathbb{E}\|\theta\|^2 < \infty$, then for any $h \in \mathbb{H}$, $\|h\| \leq C_4\|h\|_{\mathbb{H}}$ (cf. [40] pg. 203). Thus, Diam$(B_n) \leq 2(\epsilon_n + C_4 C_n) \leq 2(\epsilon_n + C_4\epsilon_n\sqrt{10C_0 n})$ (cf. [41] pg. 1454, for the second equality). By Lemma 3(a), $\log N(2\epsilon_n, \mathcal{G}_n, d_\rho) \leq k(\log N(\epsilon_n, B_n, \rho) + \log(e + e\text{Diam}(B_n)/\epsilon_n))$. Combined with (25), we have $\log D(4\epsilon_n, \mathcal{G}_n, d_\rho) \leq \log N(2\epsilon_n, \mathcal{G}_n, d_\rho) \leq O(kn\epsilon_n^2)$, due to (B4) and the assumption that $\log n = o(n\epsilon_n^2)$. If we replace ϵ_n by a sufficiently large multiple of ϵ_n, we shall obtain the bound (14) precisely.

Turning to (15), by the union bound, Pr$(G \notin \mathcal{G}_n) \leq \sum_{i=1}^k$ Pr$(\theta_i \notin B_n) \leq ke^{-C_0 n\epsilon_n^2} \leq \exp(-n\epsilon_n^2(C+4))$, by choosing constant C_0 sufficiently large (after C is fixed).

Next, we consider condition (16). Suppose that $G = \sum_{i=1}^k p_i\delta_{\theta_i}$ where $\rho(\theta_i, \theta_i^*) \leq \epsilon_n$ for all $i = 1, \ldots, k$, and $|p_i - p_i^*| \leq \epsilon_n^2/(k\text{Diam}(B_n)^2)$. Combining Lemma 6 with Assumption (B1) on the likelihood functions, we obtain that $d_K(p_{G_0}, p_G) \leq d_{\rho K}(G_0, G) \leq C_1 \sum_{1\leq i,j\leq k} q_{ij}\rho^2(\theta_i^*, \theta_j)$, for any $\boldsymbol{q} \in \mathcal{Q}$. It is simple to check that $\inf_{\boldsymbol{q}} \sum_{1\leq i,j \leq k} q_{ij}\rho^2(\theta_i^*, \theta_j) \leq \sum_{i=1}^k (p_i^* \wedge p_i)\rho^2(\theta_i^*, \theta_i) + \sum_{i=1}^k |p_i - p_i^*|\text{Diam}(B_n)^2 \leq 2\epsilon_n^2$. Hence,

$$\begin{aligned}
\Pi(d_K(p_{G_0}, p_G) \leq 2\epsilon_n^2) &\geq \Pi(\rho(\theta_i, \theta_i^*) \leq \epsilon_n; |p_i - p_i^*| \leq \epsilon_n^2/(k\text{Diam}(B_n)^2), i = 1, \ldots, k) \\
&\geq \exp(-kn\epsilon_n^2/4)c_3(\epsilon_n^2/(k\text{Diam}(B_n)^2))^{k\alpha} \\
&\geq c_3\exp(-kn\epsilon_n^2/4)\left(4k(1 + C_4\sqrt{10C_0 n})^2\right)^{-k\alpha} \\
&\geq \exp(-n\epsilon_n^2 C).
\end{aligned}$$

(The second inequality is due to Assumption (B3) and (27), the fourth inequality is due to Assumption (B5)). In view of Asumption (B2), this implies that condition (16) holds by choosing a sufficiently large constant C.

Finally, we shall choose M_n such that $M_n\epsilon_n \to 0$, and $C_k(\mathcal{G}_n, M_n\epsilon_n) \geq c(M_n\epsilon_n)^4 \geq \epsilon_n^2(C+4)$. This is possible by taking M_n to be a large multiple of $\epsilon_n^{-1/2}$. As a result, $M_n\epsilon_n \asymp \epsilon_n^{1/2}$. The proof is complete by invoking the conclusion of Thm 3. □

6 Appendix

Proof of Lemma 1.

Proof. (a) Take $\boldsymbol{q}$ to be $q_{ii} = p_i$ and $q_{ij} = 0$ if $i \neq j$. Then, $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p})$ and $\sum_{1 \leq i,j \leq k} q_{ij}\rho(\theta_i, \theta_j) = 0$. This implies that $d_\rho(G, G) = 0$. Conversely, suppose that $d_\rho(G, G') = 0$. We need to show that $G = G'$. By definition, there is a sequence of $(\{\boldsymbol{q}^n_{ij}\})_{n \geq 1}$ in $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ such that $\sum_{ij} q^n_{ij}\rho(\theta_i, \theta'_j) \to 0$. For any $j = 1, \ldots, k'$, and any i such that $\theta_i \neq \theta'_j$, $q^n_{ij} \to 0$. Note that $q^n_{ij} \leq p_i$ and $\sum_i p_i = 1$, so $\sum_{i:\theta_i \neq \theta'_j} q^n_{ij} \to 0$ by dominated convergence. But $\sum_i q^n_{ij} = p'_j$. So, either $p'_j = 0$, or there is one and only one $i \leq k$, to be denoted by $i(j)$ such that $\theta_i = \theta'_j$, in which case, we have $q^n_{i(j)j} \to p'_j$. The $i(j)$'s must be distinct for different j. Hence, $1 = \sum_{i=1}^k p_i \geq \sum_{j=1}^{k'} p_{i(j)} \geq \sum_{j=1}^{k'} q^n_{i(j)j} \geq \sum_{j=1}^{k'} \liminf_{n \to \infty} q^n_{i(j)j} = \sum_{j=1}^{k'} p'_j = 1$, where the last inequality is due to Fatou's lemma. This implies $p'_j = p_i$ for the i such that $\theta_i = \theta'_j$, or otherwise $p'_j = 0$. This entails that $G = G'$.

(b) Let $G(\boldsymbol{p}, \boldsymbol{\theta}), G'(\boldsymbol{p'}, \boldsymbol{\theta'})$ and $G''(\boldsymbol{\theta''})$ be three discrete measures with k, k', k'' number of atoms, respectively. We shall show that $d_\rho(G, G') + d_\rho(G', G'') \geq d_\rho(G, G'')$. Take any small $\epsilon > 0$. By definition, there are $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ and $\boldsymbol{q'} \in \mathcal{Q}(\boldsymbol{p'}, \boldsymbol{p''})$ such that $d_\rho(G, G') \geq \sum_{i,j} q_{ij}\rho(\theta_i, \theta'_j) - \epsilon$, and $d_\rho(G', G'') \geq \sum_{j,l} q'_{j,l}\rho(\theta'_j, \theta''_l) - \epsilon$. The marginal constraints enforced on $\boldsymbol{q}$ and $\boldsymbol{q'}$ imply that there exist a joint distribution of three discrete random variables Z, Z', Z'' taking values in $\{1, \ldots, k\}$, $\{1, \ldots, k'\}$, and $\{1, \ldots, k''\}$, respectively, such that $P(Z = i, Z' = j) = q_{ij}$ and $P(Z' = j, Z'' = l) = q'_{jl}$. Let us define a joint distribution for (Y, Y', Y'') taking value in Θ^3 through the conditional distribution $P(Y, Y', Y'' | Z, Z', Z'') = P(Y|Z)P(Y'|Z')P(Y''|Z'')$. That is, Y is conditionally independent of all other random variables given Z, and so on. Moreover, if $Z = i$, then $Y = \theta_i$. Likewise, if $Z' = j$ then $Y' = \theta'_j$. If $Z'' = l$ then $Y'' = \theta''_l$. Accordingly, $d_\rho(G, G') \geq \mathbb{E}\rho(Y, Y') - \epsilon$ and $d_\rho(G', G'') = \mathbb{E}\rho(Y', Y'') - \epsilon$, where the expectation is taken with respect to the joint distribution of Z, Z', Z'' and Y, Y', Y''.

By the lemma's hypothesis, $\rho(Y, Y') + \rho(Y', Y'') \geq \rho(Y, Y'')$ for any realizations of Y, Y' and Y''. Therefore, $d_\rho(G, G') + d_\rho(G', G'') \geq \mathbb{E}\rho(Y, Y'') - 2\epsilon$. On the other hand, $\mathbb{E}\rho(Y, Y'') = \sum_{i,l} P(Z = i, Z'' = l)\rho(\theta_i, \theta''_l)$. So $\mathbb{E}\rho(Y, Y') \geq d_\rho(G, G'')$ by definition. This implies that $d_\rho(G, G') + d_\rho(G', G'') \geq d_\rho(G, G'') - 2\epsilon$ for any $\epsilon > 0$. So the triangle inequality holds.

(c) Let $G'(\boldsymbol{p'}, \boldsymbol{\theta'})$ and $G''(\boldsymbol{p''}, \boldsymbol{\theta''})$ be two discrete probability measures in $\bar{\mathcal{G}}(\Theta)$. A convex combination of G' and G'' is another discrete probability measure in $\bar{\mathcal{G}}(\Theta)$, taking the following form, for some $\alpha \in (0, 1)$:

$$H = \alpha G' + (1 - \alpha)G'' = \sum_{j=1}^{k'} \alpha p'_j \delta_{\theta'_j} + \sum_{l=1}^{k''} (1 - \alpha) p''_l \delta_{\theta''_l}.$$

(Note that the collection of atoms associated with this representation of H may not be distinct).

Suppose that $G, G', G'' \in \bar{\mathcal{G}}(\Theta)$ and $G', G'' \in B(G, r)$. We need to show that $H \in B(G, r)$. For any $\epsilon > 0$, let $\boldsymbol{q} \in [0, 1]^{k \times k'}$ and $\boldsymbol{q'} \in [0, 1]^{k \times k''}$ be matrices of probabilities satisfying appropriate marginal constraints; and $d_\rho(G,G') \geq \sum_{i,j} q_{ij}\rho(\theta_i,\theta'_j) - \epsilon$, and $d_\rho(G,G'') = \sum_{i,l} q'_{il}\rho(\theta_i,\theta''_l) - \epsilon$. Construct a matrix of probabilities $\boldsymbol{q''} \in [0,1]^{k\times(k'+k'')}$ such that $q''_{ij} = \alpha q_{ij}$ for any $i = 1,\ldots,k$ and $j = 1,\ldots,k'$, and $q''_{ij} = (1-\alpha)q'_{i,j-k'}$ for any $i = 1,\ldots,k$ and $j = k'+1,\ldots,k'+k''$. It is clear that $\sum_{j=1}^{k'+k''} q''_{ij} = \alpha\sum_{j=1}^{k'} q_{ij} + (1-\alpha)\sum_{l=1}^{k''} q'_{il} = \alpha p_i + (1-\alpha)p_i = p_i$ for any $i \leq k$, and $\sum_{i=1}^{k} q''_{ij} = \sum_{i=1}^{k} \alpha q_{ij} = \alpha p'_j$ for any $j \leq k'$, and $\sum_{i=1}^{k} q''_{ij} = \sum_{i=1}^{k}(1-\alpha)q'_{i,j-k'} = (1-\alpha)p''_{j-k'}$ for any $j = k'+1,\ldots,k'+k''$. Thus, $\boldsymbol{q''}$ is a valid matrix of probability in the definition of $d_\rho(G,H)$. It follows that $d(G,H) \leq \alpha\sum_{i,j} q_{ij}\rho(\theta_i,\theta'_j) + (1-\alpha)\sum_{i.l} q'_{il}\rho(\theta_i,\theta''_l) \leq \alpha d_\rho(G,G') + (1-\alpha)d_\rho(G,G'') + \epsilon \leq r + \epsilon$. This holds for any $\epsilon > 0$, so $H \in B(G,r)$. □

Proof of Proposition 1.

Proof. Take any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})$. Then, $\int |G_c(x) - G'_c(x)|dx = \int |\sum_{i=1}^{k} p_i \mathbb{I}(\theta_i \leq x) - \sum_{j=1}^{k'} \mathbb{I}(\theta'_j \leq x)|dx = \int |\sum_{i,j} q_{ij}(\mathbb{I}(\theta_i \leq x) - \mathbb{I}(\theta'_j \leq x))|dx \leq \int \sum_{ij} q_{ij}|\mathbb{I}(\theta_i \leq x) - \mathbb{I}(\theta'_j \leq x)|dx = \sum_{ij} q_{ij} \int |\mathbb{I}(\theta_i \leq x) - \mathbb{I}(\theta'_j \leq x)|dx = \sum_{ij} q_{ij}|\theta_i - \theta'_j|$. The inequality is obtained by applying Jensen's inequality to the absolute function. Since this holds for any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})$, this implies that $\|G_c - G'_c\|_1 \leq \inf_{\boldsymbol{q}\in\mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})} \sum_{ij} q_{ij}|\theta_i - \theta'_j| = d_\rho(G,G')$. It remains to show that there exists one such $\boldsymbol{q}$ for which equality holds.

Without loss of generality, assume that θ_i's and θ'_j's reordered so that $\theta_1 \leq \ldots \leq \theta_k$ and $\theta'_1 \leq \ldots \leq \theta'_{k'}$. Then, $0 \leq G_c(\theta_1) \leq \ldots \leq G_c(\theta_k) \leq 1$, and $0 \leq G'(\theta'_1) \leq \ldots \leq G'(\theta'_{k'}) \leq 1$. For notational convenience, add $\theta_0 = \theta'_0 = -\infty$ and $\theta_{k+1} = \theta'_{k'+1} = \infty$. For any $i \leq k$ and $j \leq k'$, define $q_{ij} = (G_c(\theta_i) \wedge G'_c(\theta'_j)) - (G_c(\theta_{i-1}) \vee G'_c(\theta'_{j-1}))_+$. Then, $\sum_{j=1}^{k'} q_{ij} = G_c(\theta_i) - G_c(\theta_{i-1}) = p_i$ and $\sum_{i=1}^{k} q_{ij} = G'_c(\theta'_j) - G'_c(\theta'_{j-1}) = p'_j$. So, $q \in \mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})$.

Consider the collection of all non-empty rectangles in $\mathbb{R} \times [0,1]$, taken over all pairs of (i,j), whose x-coordinates are θ_i and θ'_j (in either order), while the y-coordinates are $G_c(\theta_{i-1}) \vee G'_c(\theta'_{j-1})$ and $G_c(\theta_i) \wedge G'_c(\theta'_j)$, whenever the latter quantity is strictly greater than the former. Consider one such rectangle for a pair of i,j. Then, $G_c(\theta_i) > G'_c(\theta'_{j-1})$ and $G'_c(\theta'_j) > G_c(\theta_{i-1})$. Without loss of generality, suppose that $\theta_i < \theta'_j$. Let (θ^*, y) be any point inside the rectangle, so $\theta_i < \theta^* < \theta'_j$, then $y \leq G_c(\theta_i) \wedge G'_c(\theta_j) \leq G_c(\theta_i) \leq G_c(\theta^*)$, while $y > G_c(\theta_{i-1}) \vee G'_c(\theta'_{j-1}) \geq G'_c(\theta'_{j-1}) \geq G'_c(\theta^*)$, since $\theta^* < \theta'_j$. Thus, (θ^*, y) must lie between the two curves characterizing the cdf G_c and G'_c. This implies that the rectangle lies between these two curves as well. We can also verify that all these rectangles are mutually disjoint. Indeed, consider another rectangle whose x-coordinates are $\theta_{i'}$ and $\theta'_{j'}$. Without loss of generality, suppose that $\theta_{i'} < \theta_i$. Then the y-coordinate of the farthest corner from the origin of the rectangle associated with the (i',j') pair is $G_c(\theta_{i'}) \wedge G_c(\theta'_{j'}) \leq G_c(\theta_{i'}) \leq G_c(\theta_{i-1}) \leq G_c(\theta_{i-1}) \vee G'_c(\theta'_{j-1})$, which is the y-coordinate of the nearest corner from the origin of the rectangle associated with (i,j). Thus, the two rectangles are disjoint. This entails that the summation of the area of all defined rectangles is bounded from above by the total area of the region bordered by the two curves for G_c and G'_c. The area of this region is $\|G_c - G'_c\|_1$, so this entails that $\sum_{ij} q_{ij}|\theta_i - \theta'_j| \leq \|G_c - G'_c\|_1$, which concludes the proof. □

Proof of Proposition 2.

Proof. (a) We shall apply Lagrangian multiplier theory to characterize the constrained optimization that involves in Eq. (2). Using $\rho_{ij} = \rho(Y_i, Y'_j)$ for short, we express the primal form as follows:

$$\min_{\boldsymbol{q}} \sum_{1 \leq i,j \leq N} q_{ij}\rho_{ij} + \sum_{i=1}^{N} \lambda_i (\sum_j q_{ij} - 1/N) + \sum_{j=1}^{N} \gamma_j (\sum_i q_{ij} - 1/N) - \sum_{ij} \alpha_{ij} q_{ij},$$

where $\alpha_{ij} \geq 0, \lambda_i, \gamma_j$'s are Lagrangian multipliers associated with the relevant constraints. (Note that the inequality constraints $q_{ij} \leq 1$ become vacuous in the presence of other constraints). Take derivative of the Lagrangian functional with respect to q_{ij} and set it to 0, we obtain the following equivalent dual problem:

$$\begin{aligned} \max \quad & -\frac{1}{N}\sum_{i=1}^{N} \lambda_i - \frac{1}{N}\sum_{j=1}^{N} \gamma_j \\ & \lambda_i + \gamma_j + \rho_{ij} \geq 0 \text{ for all } i, j. \end{aligned}$$

The dual problem has a simple characterization for its optimum. For each $i = 1, \ldots, N$, we must have $\lambda_i = \max_j -(\gamma_j + \rho_{ij})$. Thus, there must be a j for which $\lambda_i + \gamma_j + \rho_{ij} = 0$. Likewise, for each $j = 1, \ldots, N$ there must be an i for which this identity holds. Thus, there is a permutation π of $(1, \ldots, N)$ so that $\lambda_i + \gamma_{\pi(i)} + \rho_{i,\pi(i)} = 0$ holds for any $i = 1, \ldots, N$. Plugging these identities back to the objective function in either the primal or the dual form, we obtain an equivalent optimization that defines precisely D_{π_N}.

(b) Due to part (a), it suffices to show that $\min_{\pi_n} D_{\pi_N}(G_N, G'_N) \to D_\rho(G, G')$ almost surely. Let $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_k)$ be the proportion of Y_i's taking value $\theta_1, \ldots, \theta_k$, respectively, and $\hat{\boldsymbol{p}}' = (\hat{p}'_1, \ldots, \hat{p}'_{k'})$ is the proportion of Y'_j's taking value $\theta'_1, \ldots, \theta'_{k'}$, respectively. Given a permutation π_N, let q_{ij} be the (random) proportion of times a θ_i is matched with a θ'_j. Clearly,

$$D_{\pi_N}(G_N, G'_N) = \sum_{i,j} q_{ij}\rho(\theta_i, \theta'_j). \tag{31}$$

Moreover, q_{ij}'s satisfy the marginal constraints $\sum_j q_{ij} = \hat{p}_i$ for any $i = 1, \ldots, k$, and $\sum_i q_{ij} = \hat{p}'_j$ for any $j = 1, \ldots, k'$. In other words, $\boldsymbol{q} \in \mathcal{Q}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}')$. Recall $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ as the space of admissible $\boldsymbol{q}$ as specified in the definition of d_ρ. By the law of large numbers, as $N \to \infty$ we have $\max_{i,j}\{|\hat{p}_i - p_i| \vee |\hat{p}'_j - p'_j|\} \to 0$ almost surely. Observe that $\mathcal{Q}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}')$ can approximate $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ arbitrarily well, in the sense that for any $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ and any small $\delta > 0$, with probability one there is $\boldsymbol{q}' \in Q_N$ such that $\|\boldsymbol{q} - \boldsymbol{q}'\|_1 \leq \delta$ for sufficiently large N.

Because a change to the value of any Y_i can affect $D_{\pi_N}(G_N, G'_N)$ by no more than an amount of $O(1/N)$, by McDiarmid's inequality, for any π_N we have the concentration inequality:

$$P(|D_{\pi_N}(G, G') - \mathbb{E}D_{\pi_N}(G, G')| \geq \epsilon) \leq C \exp -N\epsilon^2/C$$

for some universal constant C.

The next step is to establish the uniform convergence of D_{π_N}, i.e., $\max_{\pi_N} |D_{\pi_N} - \mathbb{E}D_{\pi_N}| \to 0$ almost surely. Although the space of permutations grows very fast as N grows, the effect on D_{π_N} is small and only through the corresponding distribution $\boldsymbol{q} \in \mathcal{Q}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}')$, which is a small space as measured by the covering number. Specifically, for each $\epsilon > 0$, let $\{\boldsymbol{q}^1, \ldots, \boldsymbol{q}^T\}$ be a ϵ-covering of under the l_1 metric of $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$. That is, for any $q \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ there exists a $\boldsymbol{q}^t$ such that $\|\boldsymbol{q} - \boldsymbol{q}^t\|_1 = \sum_{i,j} |q_{ij} - q^t_{ij}| \leq \epsilon$. This property also holds for for any $q \in \mathcal{Q}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}')$ with probability one (for sufficiently large N), because $\mathcal{Q}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}')$ can approximate $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ as discussed in the previous paragraph. Let $N(\epsilon, \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}'), l_1)$ be the minimum of T for which such a covering exists. For any permutation π_N, for N sufficiently large there is a permutation π'_N (equivalently $\boldsymbol{q}$) that can be approximated by one such $\boldsymbol{q}^t \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ in the covering. This implies that D_{π_N} differs by no more than ϵ from $D_{\pi'_N}$. So, by an application of the union bound, $P(\max_{\pi_N} |D_{\pi_N}(G, G') - \mathbb{E}D_{\pi_N}(G, G')| \geq 3\epsilon) \leq N(\epsilon, \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}'), l_1) \times C \exp -N\epsilon^2/C$. Because $N(\epsilon, \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}'), l_1) < \infty$, by Borel-Cantelli's lemma, we obtain that $\max_{\pi_N} |D_{\pi_N} - \mathbb{E}D_{\pi_N}| \to 0$ almost surely. It then follows that $\min_{\pi_N} D_{\pi_N} - \min_{\boldsymbol{q} \in Q_N} \mathbb{E}D_{\pi_N} \to 0$ almost surely. It remains to show that $\min_{\boldsymbol{q} \in Q_N} \mathbb{E}D_{\pi_N} \to d_\rho(G, G')$ almost surely.

Note that $\mathbb{E}D_{\pi_N} = \sum_{i,j} \tilde{q}_{ij} \rho(\theta_i, \theta'_j)$, where $\tilde{q}$ is defined by letting $\tilde{q}_{ij} = \mathbb{E}q_{ij}$. $\tilde{q}$ is a valid probability distribution on $\{1, \ldots, k\} \times \{1, \ldots, k'\}$ that has to satisfy the marginal constraints $\sum_j \tilde{q}_{ij} = \mathbb{E}\sum_j q_{ij} = \mathbb{E}\hat{p}_i = p_i$, and $\sum_i \tilde{q}_{ij} = p'_j$. Thus, $\tilde{\boldsymbol{q}} \in Q$. Because $\mathcal{Q}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}')$ can approximate $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p}')$ increasingly well, $\min_{\boldsymbol{q}_N} \mathbb{E}D_{\pi_N} \to \min_{\boldsymbol{q} \in Q} \sum_{ij} q_{ij} \rho(\theta_i, \theta'_j) = d_\rho(G, G')$ almost surely.

□

References

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc. Series B*, 28:131–142, 1966.

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

[3] R. Beran. Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5(3):445–463, 1977.

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

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

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

[7] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[8] A. Cutler and O. Cordero-Brana. Minimum hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

[9] 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.

[10] M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[11] T. Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, New York, 1983. Academic Press.

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

[13] 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.

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

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

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

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

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

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

[20] 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.

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

[22] G. Kallianpur. Abstract wiener processes and their reproducing kernel hilbert spaces. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 17:113–123, 1971.

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

[24] W. Li and Q. Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods (Handbook of Statistics)*, 19:533–597, 2001.

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

[26] M. Lukić and J. Beder. Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353:3945–3970, 2001.

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

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

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

[30] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009.

[31] 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.

[32] 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.

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

[34] L. Schwartz. On bayes procedures. *Z. Wahr. Verw. Gebiete*, 4:10–26, 1965.

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

[36] R. Tamura and D. Boos. Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

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

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

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

[40] A. van der Vaart and J. van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *Annals of Statistics*, 36(3):1435–1463, 2008.

[41] A. van der Vaart and J. van Zanten. Reproducing kernel hilbert spaces of gaussian priors. *IMS Collections: Pushing the Limits of Contemporary Statistics: Contributions in Honors of Jayanta K. Ghosh*, 3:200–222, 2008.

[42] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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

[44] 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.