Bernoulli **22**(3), 2016, 1535–1571
DOI: 10.3150/15-BEJ703

Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure

XUANLONG NGUYEN

[1]*Department of Statistics, University of Michigan, 456 West Hall, Ann Arbor, MI 48109-1107, USA. E-mail: xuanlong@umich.edu*

This paper studies posterior concentration behavior of the base probability measure of a Dirichlet measure, given observations associated with the sampled Dirichlet processes, as the number of observations tends to infinity. The base measure itself is endowed with another Dirichlet prior, a construction known as the hierarchical Dirichlet processes (Teh *et al.* [*J. Amer. Statist. Assoc.* **101** (2006) 1566–1581]). Convergence rates are established in transportation distances (i.e., Wasserstein metrics) under various conditions on the geometry of the support of the true base measure. As a consequence of the theory, we demonstrate the benefit of "borrowing strength" in the inference of multiple groups of data – a powerful insight often invoked to motivate hierarchical modeling. In certain settings, the gain in efficiency due to the latent hierarchy can be dramatic, improving from a standard nonparametric rate to a parametric rate of convergence. Tools developed include transportation distances for nonparametric Bayesian hierarchies of random measures, the existence of tests for Dirichlet measures, and geometric properties of the support of Dirichlet measures.

Keywords: Bayesian asymptotics; Dirichlet processes; geometry of support; posterior concentration; random measures; transportation distances; Wasserstein metrics

1. Introduction

Ferguson's Dirichlet process is a fundamental building block in nonparametric Bayesian statistics [3,8,23]. Recent advances in modeling and computation have seen Dirichlet processes routinely built into hierarchical probabilistic structures in innovative ways [16]. A particularly useful and interesting structure that is also the focus of this paper, is the hierarchical Dirichlet processes [25, 26] – a construction in which the base probability measure of the Dirichlet becomes an object of inference, which is endowed with yet another Dirichlet prior. The hierarchical Dirichlet processes have been successfully applied to the problem of clustering for grouped data in a vast array of domains.[1]

This paper investigates the asymptotic behavior of measure-valued latent variables that arise in the hierarchical Dirichlet processes. The basic question that we address is the convergence of an estimate of the base probability measure (hereafter "base measure") of a Dirichlet measure, given observations associated with the Dirichlet processes sampled by the Dirichlet. Let Θ be a complete separable metric space equipped with the Borel sigma algebra, $\mathscr{P}(\Theta)$ the space

[1]The Google scholar page shows more than 1400 citations of [26].

1350-7265 © 2016 ISI/BS

of probability measures on Θ, and let $G \in \mathscr{P}(\Theta)$ and $\alpha > 0$. Recall from [8] that a Dirichlet process Q is a random measure taking value in $\mathscr{P}(\Theta)$ and distributed by a Dirichlet measure $\mathscr{D}_{\alpha G}$, if for any measurable partition $(B_1, \ldots, B_k)$ of Θ for some $k \in \mathbb{N}$, $(Q(B_1), \ldots, Q(B_k))$ is a random vector distributed according to the k-dimensional Dirichlet distribution with parameters $(\alpha G(B_1), \ldots, \alpha G(B_k))$.

Questions. Let $Q_1, \ldots, Q_m$ be an i.i.d. m-sample from a Dirichlet measure $\mathscr{D}_{\alpha G}$, where $\alpha > 0$ is given and the base measure $G = G_0$ is unknown. By a basic property of Dirichlet processes, $Q_1, \ldots, Q_m$ are almost surely discrete probability measures on Θ. They will *not* be observed directly. Instead, for each $i = 1, \ldots, m$, we shall be given an i.i.d. n-sample $Y^i_{[n]} = (Y_{i1}, \ldots, Y_{in})$ from a mixture distribution in which Q_i serves as a mixing measure. This mixture distribution admits the density function $p_{Q_i}(x) := Q_i * f(x) := \int f(x|\theta) Q_i(\mathrm{d}\theta)$, where $f(\cdot|\cdot)$ is a known kernel density function defined with respect to a dominating measure on Θ.

To estimate G_0 by taking a Bayesian approach, the base measure G is endowed with a prior on the space of measures $\mathscr{P}(\Theta)$, yielding a hierarchical model specification as follows:

$$G \sim \Pi_G, \qquad Q_1, \ldots, Q_m | G \overset{\text{i.i.d.}}{\sim} \mathscr{D}_{\alpha G}, \tag{1}$$

$$Y_{i1}, \ldots, Y_{in} | Q_i \overset{\text{i.i.d.}}{\sim} Q_i * f \qquad \text{for } i = 1, \ldots, m. \tag{2}$$

For the choice of prior $\Pi_G := \mathscr{D}_{\gamma H}$, where $\gamma > 0$ and $H \in \mathscr{P}(\Theta)$ is nonatomic and known, this construction is called the hierarchical Dirichlet processes model [26]. Fast computational methods have been developed to collect samples from the posterior distributions of interest, such as those for the latent G and Q_i, given the $m \times n$ data set $Y^{[m]}_{[n]} := (Y^1_{[n]}, \ldots, Y^m_{[n]})$. The first question considered in this paper is the following:

(I) How fast does the posterior distribution of the base measure G concentrate toward the true G_0, as m and n tend to infinity?

An appealing aspect well appreciated by (Bayesian) modelers and practioners of hierarchical modeling is the notion of "borrowing strength". Latent variables shared higher up in a conditional independence probabilistic hierarchy provide an infrastructure through which one may improve the inference of a parameter of interest by borrowing from information on other related data and parameters that are also part of the model. For the hierarchical Dirichlet processes, the "borrowing" has a particularly concrete meaning: according to the model, the Dirichlet processes Q_i for all $i = 1, \ldots, m$ share the same set of supporting atoms as that of the base measure G. It is intuitive that the inference of the supporting atoms of, say, Q_1 for group 1, should benefit from the information given by other groups of data associated with Q_2, Q_3 and so on. To quantify this intuition, we ask the following:

(II) What is the posterior concentration behavior of a mixture distribution, denoted by $Q * f$, as Q is attached to the Bayesian hierarchy in the same way as the Q_i, in comparison to a "stand-alone" mixture model $Q * f$, where Q is endowed with an independent prior distribution?

By resolving question (I), we can demonstrate situations in which the Bayesian hierarchy has the effect of translating the posterior concentration behavior of base measure G to improved posterior concentration of each individual group of data in the setting of question (II). Both questions will be addressed using the tools that we develop with transportation distances [29].

Related work. The only work known to us about the inference of the Dirichlet base measure is by [17], who show that it is possible to obtain a consistent estimate (in some sense) of a base measure G_0, given an i.i.d. n-sample from $m = 1$ Dirichlet process Q_1 distributed by $\mathscr{D}_{\alpha G_0}$. This curious result is due to two crucial assumptions made in their work: the true base measure G_0 is nonatomic, *and* Q_1 is observed directly. Due to the fact that two Dirichlet measures with different nonatomic base measures are orthogonal, the estimation of nonatomic base measures becomes somewhat simple if the sampled Dirichlet processes Q_i are observed directly. Changing at least one of the two assumptions makes the question considerably more difficult, which leads to different answers and requires new proof techniques. In this paper, we study the case G_0 is an atomic measure with either finite or infinite support, and the Q_i are *not* observed directly. To get a sense of the challenge, consider the simplest case, that the base measure G_0 has a finite number of support points, say $G_0 = \sum_{i=1}^{k} \beta_i \delta_{\theta_i}$, where $\theta_1, \ldots, \theta_k$ are *known*. Having a single observation Q_1 distributed by $\mathscr{D}_{\alpha G_0}$ is equivalent to being given a single draw from a k-dim Dirichlet distribution with parameter $(\alpha\beta_1, \ldots, \alpha\beta_k)$. It is clearly impossible to obtain a consistent estimate of G_0 by setting $m = 1$ (or finite), and $n \to \infty$. In addition, the assumption that $Q_1, \ldots, Q_m$ are *not* observed directly makes the analysis considerably more delicate, due to the fact that we would no longer have access to a simple point estimate of the Dirichlet base measure, as allowed in [17]. We leave open the setting where G_0 is nonatomic *and* the Q_i are not observed directly. For this setting, the choice of Dirichlet prior in the hierarchical Dirichlet processes may not be appropriate, due to the discreteness of Dirichlet processes. On the other hand, there is no known practical estimation method available for this setting at the moment.

The convergence theory of posterior distributions has received much attention in the past decade. Recent references include [1,13,14,24,30,31]. See [12] for a concise overview. This theory when applied to density estimation problem has become quite mature – the dominant theme is a Hellinger theory of density estimation for observed data. On the other hand, asymptotic behaviors of latent variable models remain poorly understood. When the inference of a latent variable is of primary concern, the Hellinger theory alone is not adequate; moreover, the underlying geometry of the variables of interest has to be taken into account. There are some examples of such theory that have been developed recently, for example, for models of random functions [15,27], mixture models [11,19,22], models of random polytopes [20]. In a prior work, the author demonstrated the usefulness of Wasserstein distances in analyzing the convergence of latent mixing measures in mixture models [19]. This viewpoint will be deepened and generalized in this work to a canonical class of hierarchical models equipped with optimal transport distances for hierarchies for random measures.

Latent hierarchies of random variables have long been a versatile and highly effective modeling tool for Bayesian modelers (see, e.g., [2]). They can also be viewed as a device for frequentist concepts of shrinkage and random effects (see, e.g., Chapter 5 of [18]). Due to their wide usages, it is of interest to characterize the roles of latent hierarchies and their effects on posterior inference in a rigorous manner. Examples of hierarchical *and* parametric models that have been explored recently include the work by [10], who studied hidden Markov models, and by the author [20], who studied the finite admixtures for categorical data. Theoretical work addressing hierarchical *and* nonparametric models, remains scarce in the literature.

Overview of results. The contributions of this paper include: (1) an analysis of convergence for the estimation of the base measure (mean measure) of a Dirichlet measure, as well as the

convergence behavior of the induced marginal density of observed data; (2) a theoretical analysis of the effect of "borrowing of strength" in the latent nonparametric hierarchy of variables; and (3) as part of the proofs of these two results we develop new tools that help to explain the geometry of the support of Dirichlet measures, and the geometry of test sets that discriminate among different Dirichlet measures. As mentioned earlier, our geometric theory is equipped with Wasserstein distances, and a new class of transportation distances that we will introduce.

Recall that for $r \geq 1$, the L_r Wasserstein distance between two probability measures $G, G' \in \mathscr{P}(\Theta)$ is given as

$$W_r(G, G') = \inf_{\kappa \in \mathcal{T}(G,G')} \left[\int \|\theta - \theta'\|^r \, \mathrm{d}\kappa(\theta, \theta') \right]^{1/r}. \tag{3}$$

Here, $\mathcal{T}(G, G')$ is the space of all joint distributions on $\Theta \times \Theta$ whose marginal distributions are G and G'. Such a joint distribution κ is also called a coupling between G and G' [29].

There are three main theorems summarized in Section 2. Our first main result (Theorem 2.1) establishes the posterior concentration behavior for the marginal density $P_{Y_{[n]}|G}$ of a generic n-vector $Y_{[n]} = (Y_1, \ldots, Y_n)$, which is obtained by integrating out the latent variable Q (see the formulae of the density in equation (11)). Suppose that the $m \times n$ data set $Y_{[n]}^{[m]} := (Y_{[n]}^1, \ldots, Y_{[n]}^m)$ are generated by the model specified by equations (1) and (2), according to $G = G_0$ for some unknown $G_0 \in \mathscr{P}(\Theta)$, where Θ is taken to be a bounded subset of $\mathbb{R}^d$. For each fixed n, as $m \to \infty$, there is a vanishing sequence $\varepsilon_{mn} = C[(n^{3d}) \log(mn)/m]^{1/(2d+2)}$ such that the posterior probability

$$\Pi_G\big(h(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq \varepsilon_{mn} | Y_{[n]}^{[m]}\big) \longrightarrow 1 \tag{4}$$

in $P^m_{Y_{[n]}|G_0}$-probability. Here, $P^m_{Y_{[n]}|G_0}$ denotes the true probability measure that generates the data set, C is a constant independent of m and n, and h denotes the Hellinger distance. Moreover, equation (4) continues to hold if we allow $n := n(m)$ to increase (e.g., to infinity) as well. This concentration rate holds under minimum assumptions on the kernel density f of the mixture distributions. In fact, improved rates can be achieved when more is assumed about either f or G_0. For instance, if f is a standard Gaussian kernel, then $\varepsilon_{mn} \asymp [n^{2d}(\log n)(\log m)^{2d+1}/m]^{1/2}$, which is optimal in terms of m (up to a logarithmic quantity). This is quite noteworthy since G_0 may have infinite support. On the other hand, if we consider a hierarchical parametric setting, that is, G_0 has finite and known number of support points, while f is an arbitrary kernel satisfying some mild conditions, then we obtain parametric rate $\varepsilon_{mn} \asymp [\log(mn)/m]^{1/2}$.

Our second main result (Theorem 2.2 in Section 2) turns to the posterior concentration behavior of base measure G. In numerous applications of the hierarchical Dirichlet processes to biomedical and machine learning problems [26], the practitioners are usually not interested in the marginal densities of the observed groups of data per se, but rather the inference of the latent variables Q_i and G, as they represent specific information about the underlying heterogeneity in data population. In admixed modeling of population genetics, for instance, G encodes the population structures responsible for diverse genotypic patterns. In the topic modeling of documents and images, G may represent topics and objects, respectively, of the observed texts and visual scenes.

As we shall see, the posterior concentration of the marginal densities of the data can be shown to entail the concentration of the base measure G, provided (again) that the data are generated according to some true base measure $G = G_0$. In this asymptotic result, we work in the regime where $m \to \infty$, while $n := n(m)$ is also taken to increase at an arbitrary rate relative to m. We will show that

$$\Pi_G\big(W_1(G, G_0) \leq \varepsilon_{mn} + \Delta_n | Y_{[n]}^{[m]}\big) \longrightarrow 1 \tag{5}$$

in $P^m_{Y_{[n]}|G_0}$-probability, where ε_{mn} is the posterior concentration rate of the marginal densities as established in the previous theorem (cf. equation (4)). Quantity $\Delta_n \to 0$ as $n \to \infty$, and can be defined as a function of the *demixing* rate δ_n of a deconvolution problem (cf. [4,7,19,33]). [To be clear, δ_n is the rate of convergence – in W_2 in our case – for estimating a mixing measure Q given an i.i.d. n-sample of a mixture density $Q * f$.] The nature of the dependence of Δ_n on δ_n is interesting, as it hinges on the geometry of the support of the true base measure G_0. We can establish a sequence of gradually deteriorating rates as the support of G_0 becomes less sparse:

(i) If G_0 has a finite and known number of support points on a bounded subset of $\mathbb{R}^d$, then $\Delta_n \asymp \delta_n^{\alpha^*}$. In fact, we obtain the overall parametric rate of convergence under some conditions that $\varepsilon_{mn} + \Delta_n \asymp [\log(mn)/m]^{1/2} + [(\log n)^{1/2}/n^{1/4}]^{\alpha^*}$, where constant $\alpha^* = \inf_{\theta \in \operatorname{spt} G_0} \alpha G_0(\{\theta\})$.

(ii) If G_0 has a finite and unknown number of support points on a bounded subset of $\mathbb{R}^d$, then $\Delta_n \asymp \delta_n^{\alpha^*/(\alpha^*+1)}$.

(iii) If G_0 has an infinite number of geometrically sparse support points on a bounded subset of $\mathbb{R}^d$, then $\Delta_n \asymp \exp -[\log(1/\delta_n)]^{1/(1\vee\gamma_0+\gamma_1)}$ for *supersparse* measures, or $\Delta_n \asymp [\log(1/\delta_n)]^{-1/(\gamma_0+\gamma_1)}$ for *ordinary sparse* measures.

The notion of ordinary and supersparse measures mentioned in (iii) will be defined in Section 2. At a high level, they refer to probablity measures that have geometrically sparse support on Θ, where the sparseness is characterized in terms of parameters γ_0 and γ_1, which are, respectively, analogous to the Hausdorff dimension and the packing dimension that arise in fractal geometry [6,9].

Our last main theorem establishes the effect of "borrowing strength" of hierarchical modeling. Suppose that an i.i.d. $\tilde{n}$-sample $Y_{[\tilde{n}]}^0$ drawn from a mixture model $Q_0 * f$ is available, where $Q_0 = Q_0^* \in \mathscr{P}(\Theta)$ is unknown:

$$Y_{[\tilde{n}]}^0 | Q_0 \overset{\text{i.i.d.}}{\sim} Q_0 * f. \tag{6}$$

In a stand-alone setting Q_0 is endowed with a Dirichlet prior: $Q_0 \sim \mathscr{D}_{\alpha_0 H_0}$ for some known $\alpha_0 > 0$ and nonatomic base measure $H_0 \in \mathscr{P}(\Theta)$. Under mild conditions on the Dirichlet process mixture, it can be shown that in Hellinger metric, the posterior probability

$$\Pi_Q\big(h\big(Q_0 * f, Q_0^* * f\big) \geq C(\log \tilde{n}/\tilde{n})^{1/(d+2)} | Y_{[\tilde{n}]}^0\big) \longrightarrow 0 \tag{7}$$

in $P_{Y_{[\tilde{n}]}^0|Q_0^*}$-probability for some constant $C > 0$ (see [19]). Alternatively, suppose that Q_0 is attached to the hierarchical Dirichlet process in the same way as the $Q_1, \ldots, Q_m$, that is,

$$G \sim \mathscr{D}_{\gamma H}, \qquad Q_0, Q_1, \ldots, Q_m | G \overset{\text{i.i.d.}}{\sim} \mathscr{D}_{\alpha G}. \tag{8}$$

Implicit in this specification, due to a standard property of the Dirichlet, is the assumption that Q_0 shares the same set of supporting atoms as $Q_1, \ldots, Q_m$, as they share with the (latent) discrete base measure G.

Theorem 2.3 in Section 2 establishes the posterior concentration rate $\delta_{m,n,\tilde{n}}$ for the mixture density $Q_0 * f$, under the hierarchical model given by equation (8), as $\tilde{n} \to \infty$ and $m, n \to \infty$ at suitable rates. Specifically, suppose that the true base measure G_0 has a finite number of support points, if m and n grow sufficiently fast relatively to $\tilde{n}$ so that the base measure G converges to G_0 at a sufficiently fast rate, then the "borrowing of strength" from the $m \times n$ data set $Y_{[n]}^{[m]}$ to the inference about the data set $Y_{[\tilde{n}]}^{0}$ has a striking effect: In particular, if f is an ordinary smooth kernel density, we obtain $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/2}$. If f is a supersmooth kernel density with smoothness $\beta > 0$, then $\delta_{m,n,\tilde{n}} \asymp (1/\tilde{n})^{1/(\beta+2)}$. (The formal definition of smoothness conditions is given in Section 2.) These present sharp improvements from nonparametric rate $(\log \tilde{n}/\tilde{n})^{1/(d+2)}$ in equation (7). Thus, the hierarchical models are particularly beneficial to groups of data with small sample sizes, as the convergence of the latent variable further up in the hierarchy can be translated into faster (e.g., parametric) rates of convergence of these small-sample groups. This appears to be the first result that establishes the benefits of the latent hierarchy in a concrete manner.

Technical approach. The major part of the proof of the main theorems lies in our attempt to understand the identifiability of the Dirichlet base measure based on the marginal densities of the data. This is achieved by establishing suitable inequalities relating the three quantities: (1) a Wasserstein distance between two base measures, $W_r(G, G')$, (2) a suitable notion of distance between Dirichlet measures $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$, and (3) the variational distance or Kullback–Leibler divergence between the densities of n-vector $Y_{[n]}$, which are obtained by integrating out the (latent) Dirichlet process Q that is distributed by Dirichlet measures $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$. In fact, the establishment of these inequalities takes up the most space of this paper (Sections 3, 4 and 5). To this end, we define a notion of optimal transport distance between Dirichlet measures $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$ (see equation (21)), which is the optimal cost of moving the mass of atoms lying in the support of measure $\mathscr{D}_{\alpha G}$ to that of $\mathscr{D}_{\alpha' G'}$, where the cost of moving from an atom (i.e., a measure) $P_1 \in \mathscr{P}(\Theta)$ to another measure $P_2 \in \mathscr{P}(\Theta)$ is again defined as a Wasserstein distance $W_r(P_1, P_2)$ given by equation (3). In general, one can define distances of measures of measures and so on in a recursive way. This provides means for comparing between Bayesian hierarchies of random measures for an arbitrary number of hierarchy levels (see Section 3).

In order to derive inequalities for the aforementioned distances, our approach boils down to establishing the existence of a subset of $\mathscr{P}(\Theta)$ which can be used to distinguish one Dirichlet measure from a class of Dirichlet measures. Because we do not have direct access to the samples Q_i of a Dirichlet measure, only the estimates of such samples, the test set has to be robust. By robustness, we require that the measure of a tube-set constructed along the boundary of the test set be *regular*, by which we mean that it is possible to control the rate at which such measure vanishes, as the radius in Wasserstein metric of such tube-set tends to zero. Interestingly, the precise vanishing rates are closely linked to the geometrically sparse structure of the support of the true Dirichlet base measure. These results are developed in Section 4 and Section 5.

The proof of Theorem 2.3 requires results concerning the geometry of the support of a single Dirichlet measure. Although the support of a Dirichlet measure is very large, that is, the entire space $\mathscr{P}(\Theta)$ (cf. [8]), we show that most of the mass of a Dirichlet measure concentrates on a

very small set as measured by the covering number of Wasserstein balls defined on $\mathscr{P}(\mathbb{R}^d)$. Our result generalizes to higher dimensions the behavior of tail probabilities chosen from a Dirichlet measure on $\mathscr{P}(\mathbb{R})$ [5].

Limitations of our results. The asymptotic results established in this paper are distinguished by the nonstandard roles of two quantities m and n simultaneously present in the model. Although both determine the size of observed data, they play asymmetric roles in the model hierarchy: m is the number of groups of data, and n is the sample size for each group. When n is fixed and m increases, the concentration rates established for marginal densities of n-vectors in Theorem 2.1 are optimal up to some logarithmic terms in several settings. However, when n is allowed to increase, the rate gets worse. For parametric models, the $\log n$ term may be ignored. Unfortunately, for nonparametric models, the presence of a polynomial quantity of n in the numerator may be suboptimal. Such presence of n in the rate is due to the fact that the space of the marginal densities on n-vector $Y_{[n]}$ data appears to get larger with n. This explanation appears reasonable, but we should be quickly reminded that the n elements of $Y_{[n]}$ are in fact exchangeable – they carry a special dependence structure among themselves. In short, having explained the role of n in its appearance in the posterior concentration rate's upper bound, we do not know whether this appearance is optimal. A more definitive conclusion on the optimal nature of convergence rates of the marginal density can only be achieved by directly tackling a minimax theory of density estimation for exchangeable sequences. Such a theory is not available at the moment.

On the more difficult question regarding the inference of base measure G, our result given by Theorem 2.2 exhibits some notable weaknesses. First of all, the posterior concentration rate (5) is meaningful only in the regime that both m and n increase. The intuition behind our analysis for G is quite natural: as n increases, one should get a better handle on individual parameter Q_i in each group. And with m increasing as well, one should be able to improve the quality of the inference of the base measure G on the basis of the Q_i's. Unfortunately, if n grows too fast relatively to m, the upper bound (5) gets worse (and eventually becomes useless). Note that in this paper we are still unable to establish posterior concentration behavior for G in the case where n is fixed, and m grows (except the case $n = 1$). Our present techniques are probably not powerful enough to address this interesting and arguably more practical asymptotic regime. The limitations seems to have their roots in a decoupling technique employed in the development of Theorem 5.1 in Section 5, which derives an upper bound for the Wasserstein distances of Dirichlet base measures in terms of the corresponding marginal densities on n-vector $Y_{[n]}$. These issues will be elaborated further in the paper.

Organization of the paper. Section 2 describes the model setting and provides a full statement of the main theorems. Section 2.3 elaborates on the components of the proofs and the tools that we develop. Section 3 defines transportation distances for hierarchies of random measures. Section 4 analyzes regular boundaries of test sets that arise in the support of various classes of Dirichlet measures of interest. Section 5 gives upper bounds for Wasserstein distances of base measures. The proof of Theorem 2.1 is given in Section 3, the proof of Theorem 2.2 is given later in Section 5, which draws from the machinery developed in Sections 3, 4 and 5. The proof of Theorem 2.3 is given in Section 6, which also draws on the results on the geometry of the support of a single Dirichlet measure.

Notation. W_r denotes the L_r Wasserstein distance. $N(\varepsilon, \mathscr{G}, W_r)$ denotes the covering number of $\mathscr{G}$ in metric W_r. $D(\varepsilon, \mathscr{G}, W_r)$ is the packing number of the same metric [28]. $\operatorname{spt} G$ denotes

the support of probability measure G. Several divergence functionals of probability densities are employed: $K(p,q), h(p,q), V(p,q)$ denote the Kullback–Leibler divergence, Hellinger and variational distance between two densities p and q defined with respect to a measure on a common space: $K(p,q) = \int p\log(p/q)$, $h^2(p,q) = \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2$ and $V(P,Q) = \frac{1}{2}\int|p-q|$. In addition, we define $K_2(p,q) = \int p[\log(p/q)]^2$, $\chi(p,q) = \int p^2/q$. $A \lesssim B$ means $A \leq C \times B$ for some positive constant C that is either universal or specified otherwise. Similarly, for $A \gtrsim B$.

2. Main theorems and tools

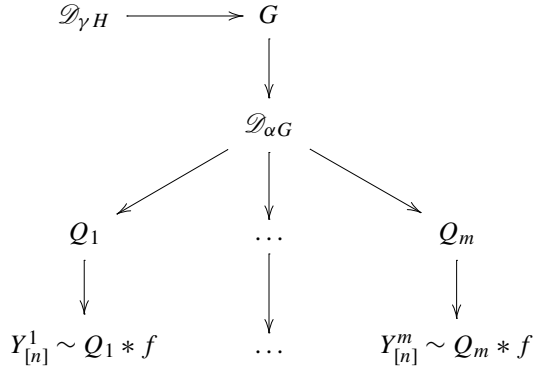
2.1. Model setting and definitions

Consider the following hierarchical probabilistic model:

$$G \sim \mathscr{D}_{\gamma H}, \qquad Q_1, \ldots, Q_m | G \overset{\text{i.i.d.}}{\sim} \mathscr{D}_{\alpha G}, \tag{9}$$

$$Y^i_{[n]} := (Y_{i1}, \ldots, Y_{in}) | Q_i \overset{\text{i.i.d.}}{\sim} Q_i * f \qquad \text{for } i = 1, \ldots, m. \tag{10}$$

The relationship among quantities of interest can be illustrated by the following diagram:



Dropping the index i, $Y_{[n]} := (Y_1, \ldots, Y_n)$ denotes the generic i.i.d. random n-vector according to the generic mixture density $Q * f$, where Q is sampled from Dirichlet measure $\mathscr{D}_{\alpha G}$. The marginal density of $Y_{[n]}$ takes the form:

$$p_{Y_{[n]}|G}(Y_{[n]}) = \int \prod_{j=1}^{n} Q * f(Y_j) \mathscr{D}_{\alpha G}(\mathrm{d}Q). \tag{11}$$

Given an $m \times n$ data set $Y^{[m]}_{[n]} := (Y^1_{[n]}, \ldots, Y^m_{[n]})$, the posterior distribution of G given $Y^{[m]}_{[n]}$ takes the form, for any measurable $\mathscr{B} \subset \mathscr{P}(\Theta)$:

$$\Pi_G\big(G \in \mathscr{B} | Y^{[m]}_{[n]}\big) = \frac{\int_{\mathscr{B}} \prod_{i=1}^m p_{Y_{[n]}|G}(Y^i_{[n]}) \mathscr{D}_{\gamma H}(\mathrm{d}G)}{\int \prod_{i=1}^m p_{Y_{[n]}|G}(Y^i_{[n]}) \mathscr{D}_{\gamma H}(\mathrm{d}G)}. \tag{12}$$

There are three main theorems. The first is concerned with the concentration behavior of the posterior distribution of marginal density $p_{Y_{[n]}|G}$ given the data $Y_{[n]}^{[m]}$, as $m \to \infty$, assuming that the data is generated according to $G = G_0$ for some fixed $G_0 \in \mathscr{P}(\Theta)$. The second deduces the posterior contraction of the base measure G, reposing upon that of $p_{Y_{[n]}|G}$. The third theorem is concerned with the concentration behavior of an individual mixing measure Q_i given the data.

Geometric sparseness conditions for G_0. Our theory is developed for a class of atomic base measure G_0. A simple example is the case G_0 has a finite number of support points. We also consider the case G_0 has infinite support, which admits a geometrically sparse structure that we now define.

Definition 2.1. *Given $c_1 \in (0,1)$, $c_2 > 0$ and a nonincreasing function $K : \mathbb{R}_+ \to \mathbb{R}_+$. A subset S of metric space Θ is (c_1, c_2, K)-sparse if for any sufficiently small $\delta > 0$ there is $\varepsilon \in (c_1\delta, \delta)$ according to which S can be covered by at most $K(\varepsilon)$ closed balls of radius ε, and every pair of such balls is separated by a distance at least $c_2\varepsilon$.*

Probability measure G_0 is said to be sparse, if its support is a (c_1, c_2, K)-sparse for a valid combination of c_1, c_2 and K. A *gauge function* for a sparse measure G_0, denoted by $g : \mathbb{R}_+ \to \mathbb{R}$, is defined as the maximal function such that for each sufficiently small ε, there is a valid ε-covering specified by the definition and that the G_0 measure on each of the covering ε-balls is bounded from below by $g(\varepsilon)$. g is clearly a nondecreasing function.

We say G_0 is *supersparse* with nonnegative parameters (γ_0, γ_1), if function K satisfies $K(\varepsilon) \lesssim [\log(1/\varepsilon)]^{\gamma_0}$, and function g satisfies $g(\varepsilon) \gtrsim [\log(1/\varepsilon)]^{-\gamma_1}$. G_0 is *ordinary sparse* with parameters (γ_0, γ_1) if $K(\varepsilon) \lesssim (1/\varepsilon)^{\gamma_0}$, and $g(\varepsilon) \gtrsim \varepsilon^{\gamma_1}$.

Examples. If $\Theta = [0,1]$ and $S = \{1/2^k | k \in \mathbb{N}, k \geq 1\} \cup \{0\}$, then S is (c_1, c_2, K)-sparse with $c_1 = 1/2$, $c_2 = 2$ and $K(\varepsilon) = \log(1/2\varepsilon)/\log 2$. If S is the support of G_0, and $G_0(\{1/2^k\}) \propto k^{-\gamma_1}$ for any $k \in \mathbb{N}$ and some $\gamma_1 > 1$, then G_0 is clearly a supersparse measure with parameters $\gamma_0 = 1$ and γ_1. Ordinary sparse measures as we defined typically arise in fractal geometry [6], where parameter γ_0 is analogous to the Hausdorff dimension of a set, while γ_1 is analogous to the packing dimension (see, e.g., [9]). Now, if $\Theta = [0,1]$ and S is the classical Cantor set, then S is (c, K)-sparse with $c_1 = 1/3$, $c_2 = 2$ and $K(\varepsilon) = \exp[\log(1/2\varepsilon)\log 2/\log 3]$. Set S has Hausdorff dimension equal $\gamma_0 = \log 2/\log 3$. Let G_0 be the γ_0-dimension Hausdorff measure on set S, then G_0 is ordinary sparse with $\gamma_0 = \gamma_1 = \log 2/\log 3$.

Conditions on kernel density f. The main theorems in this paper are established independently of the specific choices of kernel density f except some minor assumptions (A1), (A2) in the sequel. However, to obtain concrete rates in m and n, we will make additional assumptions on the smoothness of f when needed. Such assumptions are chosen mainly so we can make use of the concrete rates of demixing in a deconvolution problem, that is, the convergence rate of a point estimate of a mixing measure Q given an i.i.d. sample from the mixture density $Q * f$.

For that purpose, f is a density function on $\mathbb{R}^d$ that is symmetric around 0, that is, $f(x|\theta) := f(x-\theta)$ such that $\int_B f(x)\,\mathrm{d}x = \int_{-B} f(x)\,\mathrm{d}x$ for any Borel set $B \subset \mathbb{R}^d$. In addition, the Fourier transform of f satisfies $\tilde{f}(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$. We say f is *ordinary smooth* with parameter $\beta > 0$ if $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}(\omega)^{-2}\,\mathrm{d}\omega \lesssim (1/\delta)^{2d\beta}$ as $\delta \to 0$. Say f is *supersmooth* with parameter $\beta > 0$ if $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}(\omega)^{-2}\,\mathrm{d}\omega \lesssim \exp(2d\delta^{-\beta})$ as $\delta \to 0$. These definitions are somewhat simpler

and more general than what is employed in [19]. Depending on the form of f, it was shown by [19] that there is a strictly increasing function $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ that there holds

$$W_2(Q, Q') \lesssim \Psi\big(V\big(Q * f, Q' * f\big)\big) \tag{13}$$

for any pair $Q, Q' \in \mathscr{P}(\Theta)$, provided that Θ is a bounded subset of $\mathbb{R}^d$, and $W_2(Q, Q)$ is sufficiently small. In particular, if f is ordinary smooth with parameter β, then $\Psi(u) = u^{1/(2+\beta d')}$ for any $d' > d$. If f is supersmooth, then $\Psi(u) = (-\log u)^{-1/\beta}$ (cf. Theorem 2 of [19]).

2.2. Main theorems

The following list of assumptions are required throughout the paper:

(A1) For some $r \geq 1, C_1 > 0$, $h(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1\|\theta - \theta'\|^r$ and $K(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1\|\theta - \theta'\|^r$ $\forall \theta, \theta' \in \Theta$.

(A2) There holds $M = \sup_{\theta,\theta' \in \Theta} \chi(f(\cdot|\theta), f(\cdot|\theta')) < \infty$.

(A3) $H \in \mathscr{P}(\Theta)$ is nonatomic, and for some constant $\eta_0 > 0$, $H(B) \geq \eta_0 \varepsilon^d$ for any closed ball B of radius ε.

It is simple to observe that (A1) holds for $r = 2$ for the Gaussian kernel density f, and holds for $r = 1$ for almost all standard kernel densities in the modeling literature (Laplace, Cauchy, Gamma, etc.). (A2) holds naturally for most choices of kernel densities, as long as Θ is bounded. (A3) is often satisfied by almost all (noninformative) prior choices made in practice.

We are ready to state the first theorem, which establishes the posterior concentration of the marginal density of n-vector $Y_{[n]}$ under the above assumptions.

Theorem 2.1. *Let Θ be a bounded subset of $\mathbb{R}^d$ and $G_0 \in \mathscr{P}(\Theta)$. Given assumptions* (A1)–(A3), *parameters $\alpha > 0, \gamma > 0$ and $H \in \mathscr{P}(\Theta)$ are known. Let m tend to infinity, while n can be either fixed to a constant, or n tending to infinity at a rate relatively to m. Then there is a large constant C independent of both m and n such that the posterior induced by the model of equations* (9) *and* (10) *satisfies*

$$\Pi_G\left(h(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \geq C\left[\frac{n^{3d}\log(mn)}{m}\right]^{1/(2d+2)} \Big| Y_{[n]}^{[m]}\right) \longrightarrow 0$$

in $P^m_{Y_{[n]}|G_0}$-probability. Moreover,

(i) *If f is a Gaussian kernel with a fixed variance, then the rate is improved to*

$$\varepsilon_{mn} = \left[\frac{n^{2d}(\log m)^{2d+1}\log n}{m}\right]^{1/2}.$$

(ii) *If G_0 has a finite* and *known number of support points, then the rate is improved to*

$$\varepsilon_{mn} = \left[\frac{\log(mn)}{m}\right]^{1/2}.$$

Remarks. 1. When n is fixed, the dependence of the rate on n carries no consequence. The theorem establishes in several cases that the concentration rate with respect to m is the optimal $m^{-1/2}$ up to a logarithmic quantity. This includes the parametric case (i.e., G_0 is assumed to have a known finite number of support points). But the much more interesting case is when one uses a Gaussian density kernel f, despite the possibility that G_0 may still have infinite support. In the general setting, where almost nothing is assumed of G_0 and f (except relatively mild assumptions in (A1)–(A3)), the nonparametric rate of $m^{-1/(2d+2)}$ appears quite natural.

2. When n is allowed to vary along with m, increasing n has the effect of worsening our upper bound for the posterior concentration rate. An explanation for this phenomenon is that as n gets large, the marginal density $p_{Y_{[n]}|G}$ may become more degenerate. More concretely, in the calculations that we shall present later, the (estimate of the) entropy of the space of marginal densities $\{p_{Y_{[n]}|G} | G \in \mathscr{P}(\Theta)\}$ under Hellinger metric is shown to increase with n (cf. Lemma 3.3). Only in the case of a parametric model (i.e., the number of support points of G_0 is known) do we observe that the effect of n is the negligible $(\log n)$. We do not know whether the presence of n in the rate's numerator is optimal – a definitive answer regarding the optimality of these rates may be settled by a minimax analysis, which is beyond the scope of this paper.

Next, we turn to the posterior concentration of the base measure G per se. An easy bound can be deduced for the case $n = 1$ from Theorem 2.1. Due the basic property of the Dirichlet measure that $\int Q(\mathrm{d}\theta)\mathscr{D}_{\alpha G}(\mathrm{d}Q) = G(\mathrm{d}\theta)$, and by an application of Fubini's theorem, the marginal density for a single data point takes the form:

$$\begin{aligned} p_{Y_{[1]}|G}(Y_{[1]}) &= \int\int f(Y_1 - \theta)Q(\mathrm{d}\theta)\mathscr{D}_{\alpha G}(\mathrm{d}Q) \\ &= \int f(Y_1 - \theta)G(\mathrm{d}\theta) = G * f(Y_1). \end{aligned}$$

Provided that all conditions stated in Theorem 2.1 hold, so that the posterior concentrate rate $\varepsilon_{m1} \asymp [\log(m)/m]^{1/(2d+2)}$ is attained for the marginal density $p_{Y_{[1]}|G}$, as $n = 1$ and $m \to \infty$. Combining this concentration rate with equation (13) gives the following:

$$\Pi_G\big(W_2(G, G_0) \leq \Psi(\varepsilon_{m1}) | Y_{[1]}^{[m]}\big) \longrightarrow 1$$

in $P^m_{Y_{[1]}|G_0}$-probability, as $m \to \infty$.

Unfortunately, we do not know how to extend this bound to the case where n is fixed to a constant greater than 1. In the following, we shall work in a regime where both m and $n = n(m)$ tend to infinity. Let $(\varepsilon_n, \delta_n)_{n\geq 1}$ be two nonnegative vanishing sequences, where $\delta_n = \Psi(\varepsilon_n)$ such that $\exp -n\varepsilon_n^2 = \mathrm{o}(\delta_n)$ and that the following holds: for any $Q \in \mathscr{P}(\Theta)$, there exists a point estimate $\hat{Q}_n$ given an n-i.i.d. sample from the mixture distribution $Q * f$, such that the following inequality holds:

$$\mathbb{P}\big(W_2(\hat{Q}_n, Q) \geq \delta_n\big) \leq 5\exp\big(-cn\varepsilon_n^2\big), \tag{14}$$

where constant c is universal, the probability measure $\mathbb{P}$ is given by the mixture density $Q * f$. We refer to δ_n as the demixing rate. The exact nature of $(\varepsilon_n, \delta_n)$ is not of concern at this point. In

addition, define

$$\alpha^* := \alpha \inf_{\theta \in \operatorname{spt} G_0} G_0\big(\{\theta\}\big).$$

Note that $\alpha^* > 0$ if G has finite support, and $\alpha^* = 0$ otherwise.

Theorem 2.2. *Let Θ be a bounded subset of $\mathbb{R}^d$ and $G_0 \in \mathscr{P}(\Theta)$. Given assumptions* (A1)–(A3), *parameters $\alpha \in (0, 1]$, $\gamma > 0$ and $H \in \mathscr{P}(\Theta)$ are known. Then, as $m \to \infty$ and $n = n(m) \to \infty$, there is a sequence ε_{mn} and Δ_n dependent on m and n such that under the model given equations* (9) *and* (10), *there holds*:

$$\Pi_G\big(W_1(G, G_0) \leq C(\varepsilon_{mn} + \Delta_n) | Y_{[n]}^{[m]}\big) \longrightarrow 1$$

in $P^m_{Y_{[n]}|G_0}$-probability for a large constant C independent of m and n. In particular, ε_{mn} is any posterior concentration rate for the marginal densities such as the ones established by Theorem 2.1. *Regarding the nature of Δ_n,*

(i) *If G_0 has finite* (*but unknown*) *number of support points, then*

$$\Delta_n \asymp \delta_n^{\alpha^*/(\alpha^*+1)}.$$

(ii) *If G_0 has infinite and supersparse support with parameters (γ_0, γ_1), then*

$$\Delta_n \asymp \exp - \big[\log(1/\delta_n)\big]^{1/(1 \vee \gamma_0 + \gamma_1)}.$$

(iii) *If G_0 has infinite and ordinary sparse support with parameters (γ_0, γ_1), then*

$$\Delta_n \asymp \big[\log(1/\delta_n)\big]^{-1/(\gamma_0+\gamma_1)}.$$

Remarks. 1. Section 5 establishes the existence of a point estimate which admits the finite-sample probability bound (14). In particular, ε_n is given as follows: $\varepsilon_n \asymp (\log n/n)^{r/2d}$, if $d > 2r$; $\varepsilon_n \asymp (\log n/n)^{r/(d+2r)}$ if $d < 2r$, and $\varepsilon_n \asymp (\log n)^{3/4}/n^{1/4}$ if $d = 2r$. Constant r is from assumption (A1). The rate of demixing δ_n is determined according to an additional condition on the smoothness of the kernel density f:

(a) If f is ordinary smooth with parameter $\beta > 0$, then $\delta_n = \varepsilon_n^{1/(2+\beta d')}$ for any $d' > d$.
(b) If f is supersmooth with parameter $\beta > 0$, then $\delta_n = [-\log \varepsilon_n]^{-1/\beta}$.

2. In the parametric case, the number of support points of G_0 is $k < \infty$ and k is known, H is taken to be a probability measure with k support points. Then we obtain the following parametric rate of posterior concentration for a finite admixture model for continuous data:

$$\varepsilon_{mn} + \Delta_n = \big[\log(mn)/m\big]^{1/2} + \delta_n^{\alpha^*}.$$

Under identifiability conditions for kernel density f, such as those considered by [19] (Theorem 1), one has $\varepsilon_n = (\log n)n^{-1/2}$ and $\delta_n = \varepsilon_n^{1/2} = (\log n)^{1/2} n^{-1/4}$. Finite admixtures for categorical data exhibit a quite different kind of geometry, and were investigated in [20].

3. The above theorem establishes that the posterior concentration rate is bounded from above by two quantities ε_{mn} and Δ_n. The former captures the contraction of the marginal density of observed data, while the latter captures the demixing (deconvolution) aspect of each individual mixing measure Q_i. It is natural to expect that $\Delta_n \gg \delta_n$, to account for the fact that the mixing measures Q_i are not observed directly. It is interesting how quantity Δ_n depends on the geometric sparsity of the support of the true base measure G_0: as G_0 becomes less sparse, Δ_n gets slower:

$$\delta_n \ll \delta_n^{\alpha^*} \ll \delta_n^{\alpha^*/(\alpha^*+1)} \ll \exp -\big[\log(1/\delta_n)\big]^{1/(1\vee\gamma_0+\gamma_1)} \ll \big[\log(1/\delta_n)\big]^{-1/(\gamma_0+\gamma_1)}.$$

Our final main result is about the posterior concentration behavior of the latent mixing measures Q_i, as the base measure G is integrated out, and the amount of data increases. For the ease of presentation, we isolate a particular mixing measure to be denoted by Q_0, and we shall assume that Q_0 is attached to the hierarchical Dirichlet process in the same way as the $Q_1, \ldots, Q_m$, that is,

$$G \sim \mathscr{D}_{\gamma H}, \qquad Q_0, Q_1, \ldots, Q_m | G \overset{\text{i.i.d.}}{\sim} \mathscr{D}_{\alpha G}. \tag{15}$$

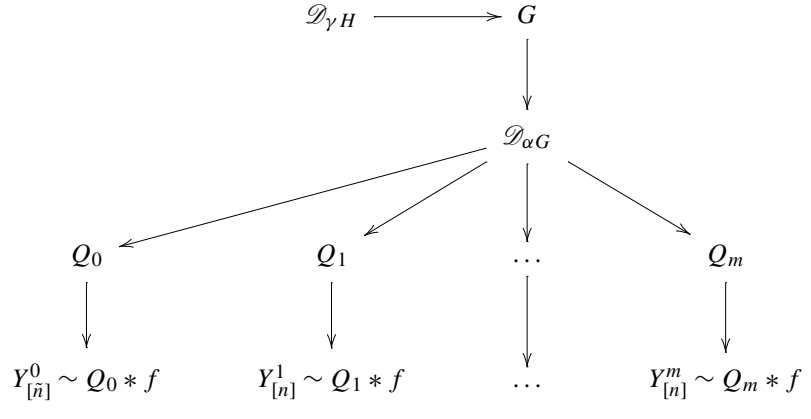
Suppose that an i.i.d. $\tilde{n}$-sample $Y^0_{[\tilde{n}]}$ drawn from a mixture model $Q_0 * f$ is available, where $Q_0 = Q_0^* \in \mathscr{P}(\Theta)$ is unknown:

$$Y^0_{[\tilde{n}]} | Q_0 \overset{\text{i.i.d.}}{\sim} Q_0 * f. \tag{16}$$

In addition, as before, $m \times n$ data set is available:

$$Y^i_{[n]} := (Y_{i1}, \ldots, Y_{in}) | Q_i \overset{\text{i.i.d.}}{\sim} Q_i * f \qquad \text{for } i = 1, \ldots, m. \tag{17}$$

The relationship among quantities of interest is illustrated by the following diagram:



The following theorem shows that the posterior distribution $\Pi(Q_0 | Y^0_{[\tilde{n}]}, Y^{[m]}_{[n]})$, defined with respect to specifications (15), (16) and (17), concentrates most its mass toward Q_0^*, as n, m and

$\tilde{n} \to \infty$ appropriately. The intuition for this result is rather simple. As the data size $m \times n$ grows appropriately, the posterior distribution for base measure G concentrates around the true G_0, which shall be assumed to be a discrete measure with a finite, but unknown number of support point. This benefits the inference of density $Q_0 * f$. Indeed, the (conditional) Dirichlet prior on the mixing measure Q_0 (given the $m \times n$ data) can be shown to be very thick, due to the fact that its base measure G_0 is conditionally close to a measure with a finite number of support points. In addition, one can identify subsets of the support of the (conditional) Dirichlet prior for Q_0 which take up most of its probability mass, while remaining small in size, as evaluated by the entropy/covering number. A combination of these two facts result in very favorable posterior concentration for the marginal density $Q_0 * f$. In fact, the rates become parametric, as they are independent of the parameter dimensionality d. By contrary, if we do not have the concentration of base measure G, there is very little control of the space over which Q_0 may vary. As a result, one can only establish the standard nonparametric rate of convergence under general conditions.

A complete statement of the theorem is the following. Motivated by the conclusion of Theorem 2.2 we shall assume that the posterior distribution of G concentrates at a certain rate δ_{mn} toward the true base measure G_0, which is now assumed to have a finite (but unknown) number of support points. This concentration behavior can in turn be translated to a sharp concentration behavior for the mixture density $Q_0 * f$.

Theorem 2.3. *Let Θ be a bounded subset of $\mathbb{R}^d$, $G_0, Q_0^* \in \mathscr{P}(\Theta)$. Suppose that assumptions* (A1) *and* (A2) *hold for some $r \geq 1$. Given parameters $\alpha \in (0, 1]$, $\gamma > 0$, and $H \in \mathscr{P}(\Theta)$ known. Assume further that*:

(a) *G_0 has $k < \infty$ support points in Θ; $Q_0^* \in \mathscr{P}(\Theta)$ such that* $\operatorname{spt} Q_0^* \subseteq \operatorname{spt} G_0$.

(b) *For each $\tilde{n}$, there is a net $\delta_{mn} = \delta_{mn}(\tilde{n}) \downarrow 0$ indexed by m, n such that under the model specifications* (15), (16) *and* (17), *there holds*: $\Pi_G(W_1(G, G_0) \geq C\delta_{mn} | Y_{[n]}^{[m]}, Y_{[\tilde{n}]}^0) \longrightarrow 0$ *in $P_{Y_{[n]}|G_0}^m \times P_{Y_{[\tilde{n}]}|Q_0^*}^0$-probability, as $m \to \infty$ and $n = n(m) \to \infty$ at a suitable rate with respect to m. Here, C is a constant independent of $\tilde{n}, m, n$.*

Then, as $\tilde{n} \to \infty$ and then m and $n = (m) \to \infty$, we have

$$\Pi_Q\big(h\big(Q_0 * f, Q_0^* * f\big) \geq \delta_{m,n,\tilde{n}} | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}\big) \longrightarrow 0$$

in $P_{Y_{[\tilde{n}]}|Q_0^*}^0 \times P_{Y_{[n]}|G_0}^m$-probability, where the rates $\delta_{m,n,\tilde{n}}$ are given as follows:

(i) $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/(d+2)} + \delta_{mn}^{r/2} \log(1/\delta_{mn})$.

(ii) $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/2}$ *if f is ordinary smooth with smoothness $\beta > 0$, and n and m grow sufficiently fast so that δ_{mn} is sufficiently small relatively to $\tilde{n}$ (see details in the remarks below).*

(iii) $\delta_{m,n,\tilde{n}} \asymp (1/\tilde{n})^{1/(\beta+2)}$, *if f is supersmooth with smoothness $\beta > 0$, n and m grow sufficiently fast so that δ_{mn} is sufficiently small relatively to $\tilde{n}$.*

Remarks. 1. Condition (a) that $\operatorname{spt} Q_0^* \subseteq \operatorname{spt} G_0$ motivates the incorporation of mixture distribution $Q_0 * f$ into the Bayesian hierarchy as specified by equation (15). According to the model,

Q_0 shares the same supporting atoms with $Q_1, \ldots, Q_m$, as they all inherit from random base measure G. Note also that the condition on the posterior of G as stated in (b) is closely related to but nonetheless different from the conclusion reached by Theorem 2.2, due to the additional conditioning on $Y^0_{[\tilde{n}]}$. This condition may be proved directly under additional assumptions on Q_0^* and G_0, by a technically cumbersome (but conceptually simple) modification of the proof of Theorem 2.2. We avoid this unnecessary complication as it is not central to the main message of the present theorem.

2. In the statement of part (ii), m and n are required to grow at a rate so that $\delta_{mn} \lesssim \tilde{n}^{-(\alpha+k+M_0)}(\log \tilde{n})^{-(\alpha+k-2)}$, for some constant $M_0 > 0$ depending only on d, k, β and $\operatorname{diam}(\Theta)$. In part (iii), we require $\delta_{mn} \lesssim \tilde{n}^{-2(\alpha+k)/(\beta+2)}(\log \tilde{n})^{-2(\alpha+k-1)} \exp(-4\tilde{n}^{\beta/(\beta+2)})$.

3. To appreciate the statistical content of this theorem, recall a stand-alone setting in which Q_0 is endowed with an independent Dirichlet prior: $Q_0 \sim \mathscr{D}_{\alpha_0 H_0}$ for some known $\alpha_0 > 0$ and nonatomic base measure $H_0 \in \mathscr{P}(\Theta)$. Combining with the model specification expressed by (16), we obtain the posterior distribution for mixture density $Q_0 * f$, which admits the following concentration behavior under some mild conditions (cf. [19]):

$$\Pi_Q\big(h\big(Q_0 * f, Q_0^* * f\big) \geq (\log \tilde{n}/\tilde{n})^{1/(d+2)} | Y^0_{[\tilde{n}]}\big) \longrightarrow 0 \tag{18}$$

in $P_{Y^0_{[\tilde{n}]}|Q_0^*}$-probability. Now, the rate in the above display should be compared to the general rate given by claim (i) of Theorem 2.3: $(\log \tilde{n}/\tilde{n})^{1/(d+2)} + \delta_{mn}^{r/2} \log(1/\delta_{mn})$. The extra quantity $\delta_{mn}^{r/2} \log(1/\delta_{mn})$ can be viewed as the general "overhead cost" for maintaining the latent hierarchy involving the random Dirichlet prior $\mathscr{D}_{\alpha G}$ in the hierarchical model.

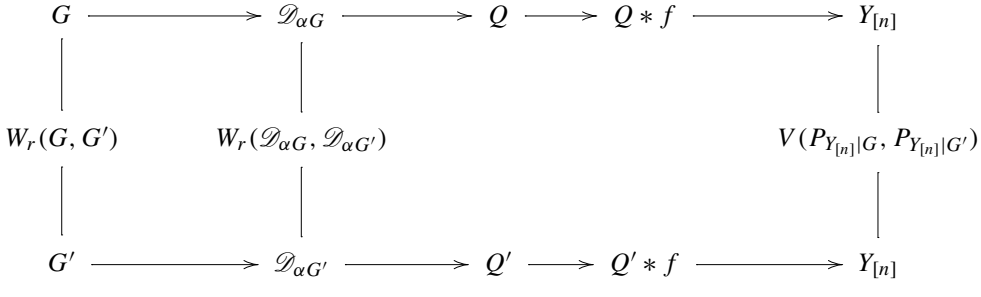
4. Claims (ii) and (iii) demonstrate the benefits of hierarchical modeling for groups of data with relatively small sample size: when $n \gg \tilde{n}$ (and $m = m(n) \to \infty$ suitably) so that δ_{mn} is sufficiently small, we obtain parametric rates for the mixture density $Q_0 * f$: $(\log \tilde{n}/\tilde{n})^{1/2}$ for ordinary smooth kernels, and $(1/\tilde{n})^{1/(\beta+2)}$ for supersmooth kernels. This is a sharp improvement over the standard rate $(\log \tilde{n}/\tilde{n})^{1/(d+2)}$ one would get for fitting a stand-alone mixture model $Q_0 * f$ using a Dirichlet process prior. Technically, this improvement is due to the confluence of two factors: By attaching Q_0 to the Bayesian hierarchy one is able to exploit the assumption that random measure Q_0 shares the same supporting atoms as the random base measure G. This is translated to a favorable level of thickness of the conditional prior for Q_0 (given the $m \times n$ data $Y^{[m]}_{[n]}$), as measured by small Kullback–Leibler neighborhoods. The second factor is due to our new construction of a sieves (subsets of) $\mathscr{P}(\Theta)$ over which the Dirichlet measure concentrates most its mass on, but which have suitably small entropy numbers. These details will be elaborated in Section 6.

Summarizing our results: Theorem 2.1 establishes posterior concentration of the marginal densities generating the observed data, while Theorem 2.2 establishes posterior concentration of the latent Dirichlet base measure in a hierarchical setting. Theorem 2.3 demonstrates dramatic gains in the efficiency of statistical inference of individual groups of data with relatively small sample size. For groups with relatively large sample size, the concentration rate appears to be weaken due to the overhead of maintaining the latent hierarchy. This quantifies the effects of "borrowing

of strength", from large groups of data to smaller groups. This is arguably a good virtue of hierarchical models: it is the populations with smaller sample sizes that need improved inference the most.

2.3. Method of proof

The major part of the proof of Theorem 2.1 and 2.2 lies in our attempt to establish the relationship between the three important quantities: (1) a Wasserstein distance between two base measures, $W_r(G, G')$, (2) a suitable notion of distance between Dirichlet measures $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha G'}$, and (3) the variational distance/Kullback–Leibler divergence between the marginal densities of n-vector $Y_{[n]}$, which are obtained by integrating out the mixing measure Q, which is a Dirichlet process distributed by $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha G'}$, respectively. The link from G (resp., G') to the induced $P_{Y_{[n]}|G}$ (resp., $P_{Y_{[n]}|G'}$) is illustrated by the following diagram:



In order to establish the relationship among the aforementioned distances, we need to investigate the geometry of the support of individual Dirichlet measures, and the geometry of test sets that arise when a given Dirichlet measure is tested (discriminated) against a large class of Dirichlet measures. This study forms the bulk of the paper in Section 3, Section 4 and Section 5.

Transportation distances for Bayesian hierarchies. To begin, in Section 3 we develop a general notion of transportation distance of Bayesian hierarchies of random measures. This notion plays a fundamental role in our theory, and we believe is also of independent interest. Using transportation distances, it is possible to compare between not only two probability measures defined on Θ, but also two probability measures on the space of measures on Θ, and so on. Transportation distances are natural for comparing between Bayesian hierarchies, because the geometry of the space of support of measures is inherited directly in the definition of the transportation distances between the measures. In particular, $W_r(\mathscr{D}_{\alpha G}, \mathscr{D}_{\alpha G'})$ is defined as the Wasserstein distance on the Polish space $\mathscr{P}(\mathscr{P}(\Theta))$, by inheriting the Wasserstein distance on the Polish space of measures $\mathscr{P}(\Theta)$. (The notation W_r is reused as a harmless abuse of notation.) It can be shown that

$$W_r(\mathscr{D}_{\alpha G}, \mathscr{D}_{\alpha' G'}) \geq W_r\big(G, G'\big).$$

The above inequality holds generally if $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$ are replaced by any pair of probability measures on $\mathscr{P}(\Theta)$ that admit a suitable notion of mean measures G, and G', respectively. Moreover, the Dirichlet measures allow a remarkable identity: when $\alpha = \alpha'$, we have

$$W_r(\mathscr{D}_{\alpha G}, \mathscr{D}_{\alpha G'}) = W_r\big(G, G'\big).$$

Repeated applications of Jensen's inequality yield the following upper bound for the KL divergence:[2]

$$h^2(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \leq K(P_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \lesssim n W_r^r(\mathscr{D}_{\alpha G}, \mathscr{D}_{\alpha G'}) = n W_r^r\big(G, G'\big).$$

Bounds on Wasserstein distances. The most demanding part of the paper lies in establishing an upper bound of the Wasserstein distance $W_r(G, G')$ in terms of the variational distance $V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$. This is ultimately achieved by Theorem 5.1 in Section 5, which states that for a fixed $G \in \mathscr{P}(\Theta)$ and *any* $G' \in \mathscr{P}(\Theta)$,

$$W_r^r\big(G, G'\big) \lesssim V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) + A_n\big(G, G'\big), \tag{19}$$

where $A_n(G, G')$ is a quantity that tends to 0 as $n \to \infty$. The rate at which $A_n(G, G')$ tends to zero depends only on the geometrically sparse structure of G, not G'. The proof of this result hinges on the existence of a suitable set $\mathscr{B}_n \subset \mathscr{P}(\Theta)$ measurable with respect to (the sigma algebra induced by) the observed variables $Y_{[n]}$, which can then be used to distinguish G' from G, in the sense that

$$W_r^r\big(G, G'\big) \lesssim P_{Y_{[n]}|G'}(\mathscr{B}_n) - P_{Y_{[n]}|G}(\mathscr{B}_n) + A_n\big(G, G'\big). \tag{20}$$

We develop two main lines of attack to arrive at a construction of $\mathscr{B}_n$.

First, we establish the existence of a point estimate for the mixing measure on the basis of the observed $Y_{[n]}$. Moreover, such point estimates have to admit a finite-sample probability bound of the following form: given $Y_{[n]} \sim Q * f$, there exist a point estimate $\hat{Q}_n$ such that under the $Q * f$ probability, there holds

$$\mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta_n\big) \lesssim \exp -n\varepsilon_n^2,$$

where δ_n and ε_n are suitable vanishing sequences. These finite-sample bounds are presented in Section 5. The existence of $\hat{Q}_n$ will then be utilized in the construction of a suitable set $\mathscr{B}_n$. In particular, one may pretend to have direct observations from the Dirichlet measures to construct the test sets, with a possible loss of accuracy captured by the demixing rate δ_n.

Regular boundaries in the support of Dirichet measures. Now, to control $A_n(G, G')$, we need the second piece of the argument, which establishes the existence of a robust test that can be used to distinguish a Dirichlet measure $\mathscr{D}_{\alpha G}$ from a class of Dirichet measures $\mathscr{C} = \{\mathscr{D}_{\alpha' G'} | G' \in$

[2]Within this subsection, the details on the constants underlying $\lesssim$ and $\gtrsim$ are omitted for the sake of brevity.

$\mathscr{P}(\Theta)\}$, where the robustness here is measured by Wasserstein metric W_r on $\mathscr{P}(\Theta)$. The robustness is needed to account for the possible loss of accuracy δ_n incurred by demixing, as alluded to in the previous paragraph. A formal theory of robust tests is developed in Section 4. Central to this theory is a notion of regularity for a given class of Dirichlet measures $\mathscr{C}$ with respect to a fixed Dirichlet measure $\mathcal{D} := \mathscr{D}_{\alpha G}$. In particular, we say that $\mathscr{C}$ has regular boundary with respect to $\mathcal{D}$ if for each element $\mathcal{D}' = \mathscr{D}_{\alpha' G'} \in \mathscr{C}$ there is a measurable subset $\mathscr{B} \subset \mathscr{P}(\Theta)$ for which the following holds: (i) $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) \gtrsim W_r^r(G, G')$ and (ii)

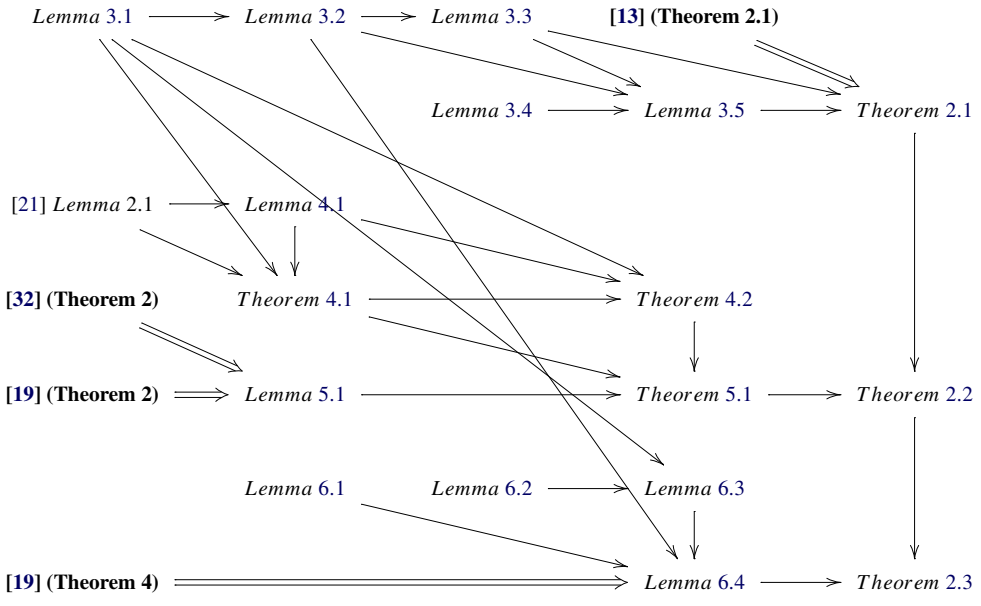
$$\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) \to 0$$

as $\delta \to 0$. Set $\mathscr{B}$ can be thought of as a test set which is used to approximate the variation distance between a fixed $\mathcal{D}$ and an arbitrary $\mathcal{D}'$ which varies in $\mathscr{C}$. $\mathscr{B}_\delta$ is defined to be the set of all $P \in \mathscr{P}(\Theta)$ for which there is a $Q \in \mathscr{B}$ and $W_r(Q, P) \leq \delta$. Various forms of regularity are developed, which specifies how fast the quantity in the previous display tends to 0. Thus, the achievement of this section is to show that the regularity behavior is closely tied to the geometry of the support of base measure G. Theorems 4.1 and 4.2 provide a complete picture of regularity for the case G has finite support, and the case G has infinite and geometrically sparse support. Now, by controlling the rate at which $\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B})$ tends to 0, we can control the rate at which $A_n(G, G')$ tends to 0, completing the proof of (19).

Posterior concentration proofs. With the tools and inequalities established in Section 3 at our disposal, the proof of Theorem 2.1 is easily available by appealing to a general theorem for establishing posterior concentration of a density [13], and verifying the sufficient conditions in terms of entropy numbers, the prior thickness in Kullback–Leibler divergence, and so on. The proof of Theorem 2.2 follows by combining the result from Theorem 2.1 with Theorem 5.1 described above.

Finally, the proof of Theorem 2.3 follows from a posterior concentration result for the mixing measure Q, which is distributed by the prior $\mathscr{D}_{\alpha G}$, conditionally given the event that the base measure G is perturbed by a small Wasserstein distance W_1 from G_0 that has $k < \infty$ support points; see Lemma 6.4 in Section 6. The proof of this lemma also follows the standard strategy of the posterior concentration proof mentioned earlier. The main novelty lies in the construction of a sieves of subsets of $\mathscr{P}(\Theta)$ which yields favorable rates of posterior concentration. This construction is possible by showing that the Dirichlet measure places most its mass on subsets (of $\mathscr{P}(\Theta)$) which can be covered by a relatively small number of balls in W_r. Such results about the Wasserstein geometry of the support of a Dirichlet measure may be of independent interest, and are collected in Section 6.2.

Due to the large number of technical results, many of which are new and rather nonstandard, for the ease of the readers we include the following chart that illustrates the dependence structures of the main theorems and accompanying lemmas. Also included are several existing theorems (in bold) upon which our results are built in crucial ways.



2.4. Concluding remarks and further development

In this paper, we study posterior concentration behaviors for the base measure of a Dirichlet measure and related quantities, given observations associated with sampled Dirichlet processes, using tools developed with optimal transport distances. There are a number of open questions that remain. First, regarding Theorem 2.1, we still do not know whether the established (upper bound) of the concentration rate is optimal or not, with respect to the number m of groups, and more interestingly with respect to the sample size n per group. Perhaps a proper way to address this question is to directly develop a minimax optimal theory for the variables residing in latent hierarchies of models such as the one we have considered. Second, regarding Theorem 2.2, our result is applicable only in the setting where both m and n grow, not the case where m grows and n is fixed. Our proof method is not capable of saying much on the latter setting. Finally, it may be of interest to consider the problem of estimating a nonatomic base measure, while the Dirichlet processes are not directly observed.

3. Transportation distances of Bayesian hierarchies

Let Θ be a complete separable metric space (i.e., Θ is a Polish space) and $\mathscr{P}(\Theta)$ be the space of Borel probability measures on Θ. The weak topology on $\mathscr{P}(\Theta)$ (or narrow topology) is induced by convergence against $C_b(\Theta)$, that is, bounded continuous test functions on Θ. Since Θ is Polish, $\mathscr{P}(\Theta)$ is itself a Polish space. $\mathscr{P}(\Theta)$ is metrized by the W_r Wasserstein distance: for

$G, G' \in \mathscr{P}(\Theta)$ and $r \geq 1$,

$$W_r(G, G') = \inf_{\kappa \in \mathcal{T}(G,G')} \left[\int \|\theta - \theta'\|^r \, \mathrm{d}\kappa(\theta, \theta') \right]^{1/r}.$$

By a recursion of notation, $\mathscr{P}(\mathscr{P}(\Theta))$ is defined as the space of Borel probability measures on $\mathscr{P}(\Theta)$. This is a Polish space, and will be endowed again with a Wasserstein metric that is induced by metric W_r on $\mathscr{P}(\Theta)$:

$$W_r(\mathcal{D}, \mathcal{D}') = \inf_{\mathcal{K} \in \mathcal{T}(\mathcal{D},\mathcal{D}')} \left[\int W_r^r(G, G') \, \mathrm{d}\mathcal{K}(G, G') \right]^{1/r}. \tag{21}$$

We can safely reuse notation W_r as the context is clear from the arguments. Since the cost function $\|\theta - \theta'\|$ is continuous, the existence of an optimal coupling $\kappa \in \mathcal{T}(G, G')$ which achieves the infimum is guaranteed due to the tightness of $\mathcal{T}(G, G')$ (cf. Theorem 4.1 of [29]). Moreover, $W_r(G, G')$ is a continuous function and $\mathcal{T}(\mathcal{D}, \mathcal{D}')$ is again tight, so the existence of an optimal coupling in $\mathcal{T}(\mathcal{D}, \mathcal{D}')$ is also guaranteed.

Now we present a lemma on a monotonic property of Wasserstein metrics defined along the recursive construction for every pair of centered random measures on Θ. Part (b) highlights a very special property of the Dirichlet measure. In what follows, P denotes a generic measure-valued random variable. By $\int P \, \mathrm{d}\mathcal{D} = G$ we mean $\int P(A) \, \mathrm{d}\mathcal{D} = G(A)$ for any measurable subset $A \subset \Theta$.

Lemma 3.1. (a) *Let $\mathcal{D}, \mathcal{D}' \in \mathscr{P}(\mathscr{P}(\Theta))$ such that $\int P \, \mathrm{d}\mathcal{D} = G$ and $\int P \, \mathrm{d}\mathcal{D}' = G'$. For $r \geq 1$, if $W_r(\mathcal{D}, \mathcal{D}')$ is finite then $W_r(\mathcal{D}, \mathcal{D}') \geq W_r(G, G')$.*

(b) *Let $\mathcal{D} = \mathscr{D}_{\alpha G}$ and $\mathcal{D}' = \mathscr{D}_{\alpha G'}$. Then $W_r(\mathcal{D}, \mathcal{D}') = W_r(G, G')$ if both quantities are finite.*

Recall the generative process defined by equations (9) and (10): The marginal density $p_{Y_{[n]}|G}$ is obtained by integrating out random measures Q, which is distributed by $\mathscr{D}_{\alpha G}$; see equation (11). By a repeated application of Jensen's inequality, it is simple to establish upper bounds on Kullback–Leibler distance $K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$ and other related distances in terms of transportation distance between G and G'.

Lemma 3.2. (a) *Under assumption* (A1),

$$K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq C_1 n W_r^r(G, G'),$$
$$h^2(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq C_1 n W_{2r}^{2r}(G, G'),$$
$$h^2(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq \sqrt{1 - \left(1 - C_1 W_{2r}^{2r}(G, G')\right)^n}.$$

(b) *Under assumption* (A2), *we have* $\chi(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq M^n$.

The following lemma establishes an estimate of the entropy number for the space of marginal densities $\{p_{Y_{[n]}|G} | G \in \mathscr{P}(\Theta)\}$. Part (a) gives a very general entropy bound. Tightened bounds are

possible given when more is known either about the space of G, or the kernel density f. These entropy bounds have direct consequences on the kind of concentration rates that we will get in Theorem 2.1.

Lemma 3.3. (a) *Under assumption* (A1), *for any* $\varepsilon \in (0, 1/2)$,

$$\log N\big(\varepsilon, \big\{p_{Y_{[n]}|G} | G \in \mathscr{P}(\Theta)\big\}, h\big) \le \big(2C_1 n \operatorname{diam}(\Theta)/\varepsilon^2\big)^d \log\big(\mathrm{e} + 2\mathrm{e}C_1 n \operatorname{diam}(\Theta)/\varepsilon^2\big).$$

(b) *Under assumption* (A1), *for any* $\varepsilon \in (0, 1/2)$, $k \in \mathbb{N}$,

$$\begin{aligned}&\log N\big(\varepsilon, \{p_{Y_{[n]}|G} | G \text{ has } k \text{ support points on } \Theta\}, h\big)\\ &\quad\le kd \log\big(2C_1 n \operatorname{diam}(\Theta)/\varepsilon^2\big) + \log\big(\mathrm{e} + 2\mathrm{e}C_1 n \operatorname{diam}(\Theta)/\varepsilon^2\big).\end{aligned}$$

(c) *If* f *is a Gaussian kernel on* $\mathbb{R}^d$, $f(x) = \frac{1}{(2\pi)^{d/2}\sigma^d}\mathrm{e}^{-\|x\|^2/2\sigma^2}$, *for some* $\sigma > 0$, *then*

$$\log N\big(\varepsilon, \big\{p_{Y_{[n]}|G} | G \in \mathscr{P}(\Theta)\big\}, h\big) \lesssim \big(\log(1/\varepsilon)\big)^{2d+1} n^{2d} \log n,$$

where the multiplying constant depends only on d, σ, Θ (*and not on* n).

Next, define the Kullback–Leibler neighborhood of a given $G_0 \in \mathscr{P}(\Theta)$ with respect to n-vector $Y_{[n]}$ as follows:

$$B_K(G_0, \delta) = \big\{G \in \mathscr{P}(\Theta) | K(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \le \delta^2, K_2(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \le \delta^2\big\}. \tag{22}$$

The following result gives probability bound on small balls as defined by Wasserstein metric (Lemma 5 of [19]):

Lemma 3.4. *Suppose that* $\operatorname{law}(G) = \mathscr{D}_{\gamma H}$, *where* H *is a nonatomic probability measure on* Θ. *For a small* $\varepsilon > 0$, *let* $D = D(\varepsilon, \Theta, \|\cdot\|)$ *the packing number of* Θ *under* $\|\cdot\|$. *Then, for any* $G_0 \in \mathscr{P}(\Theta)$,

$$\mathbb{P}\big(G\colon\ W_r^r(G_0, G) \le \big(2^r + 1\big)\varepsilon^r\big) \ge \frac{\Gamma(\gamma)\gamma^D}{(2D)^{D-1}}\bigg(\frac{\varepsilon}{\operatorname{diam}(\Theta)}\bigg)^{r(D-1)} \sup_S \prod_{i=1}^D H(S_i).$$

Here, $(S_1, \ldots, S_D)$ *denotes the* D *disjoint* $\varepsilon/2$*-balls that form a maximal packing of* Θ. Γ *denotes the gamma function. The supremum is taken over all packings* $S := (S_1, \ldots, S_D)$.

Combine the previous lemmas to obtain an estimate of the thickness of the hierarchical Dirichlet prior:

Lemma 3.5. *Given assumptions* (A1)–(A3), Θ *a bounded subset of* $\mathbb{R}^d$.

(a) *Let* $D := (\mathrm{diam}(\Theta))^d (n^3/\delta^2)^{d/r}$ *and constants* c, C *depending only on* C_1, M, η_0, γ, $\mathrm{diam}(\Theta)$ *and* r. *Then, for any* $G_0 \in \mathscr{P}(\Theta)$, $\delta > 0$ *and* $n > C \log(1/\delta)$, *the following inequality holds under the probability measure* $\mathcal{D}_{\gamma H}$:

$$\log \mathbb{P}\big(G \in B_K(G_0, \delta)\big) \geq c \log\big[\gamma^D \big(\delta^2/n^3\big)^{(1+d/r)(D-1)+Dd/r}\big].$$

(b) *If in addition,* G_0 *has exactly* k *support points in* Θ, *then*

$$\log \mathbb{P}\big(G \in B_K(G_0, \delta)\big) \geq c \log\big[\gamma^k \big(\delta^2/n^3\big)^{kd/r+k/r} \big(1/k \,\mathrm{diam}\,\Theta^r\big)^k\big].$$

(c) *If* f *is the Gaussian kernel* (*given in Lemma* 3.3), *then for any* $G_0 \in \mathscr{P}(\Theta)$, *the bound in part* (b) *of the lemma continues to hold with* $k \lesssim (\log(1/\delta))^{2d} (nd)^{2d}$.

The proofs of all lemmas presented in this section are deferred to [21].

Proof of Theorem 2.1. The proof is a straightforward application of a standard result in Bayesian asymptotics for density estimation. In particular, we shall appeal to Theorem 2.1 of [13]. First, let n be fixed, so that n acts as the (fixed) dimensionality of the n-vector $Y_{[n]}$. According to this theorem, as sample size m tends to infinity, as long as the constructed rate sequence ε_{mn} satisfies the entropy condition on the class of marginal densities:

$$\log D\big(\varepsilon_{mn}, \big\{P_{Y_{[n]}|G} | G \in \mathscr{P}(\Theta)\big\}, h\big) \leq m\varepsilon_{mn}^2$$

and the condition on the prior thickness:

$$-\log \mathbb{P}\big(G \in B_K(G_0, \varepsilon_{mn})\big) \leq Mm\varepsilon_{mn}^2$$

for some universal constant $M > 0$, then the conclusion of Theorem 2.1 is established for some sufficiently large constant $C > 0$ not depending on m or n. Indeed, the entropy condition is an immediate consequence of Lemma 3.3, while the prior thickness condition is immediate from Lemma 3.5. Finally, an examination of the proof of [13] reveals that the conclusion also holds by allowing n to vary as a function of m. □

4. Regular boundaries in the support of Dirichlet measures

In this section, we study the property of the boundary of certain sets (of measures) which can be used to test one Dirichlet measure against another. Typically, such a test set can be defined via the variational distance between the two measures. However, for the purpose of subsequent development we need a more robust test in which the robustness can be expressed in terms of the measure of the test set's perturbation along its boundary. Recall the variational distance between $\mathcal{D}, \mathcal{D}' \in \mathscr{P}(\mathscr{P}(\Theta))$ is given by

$$V\big(\mathcal{D}, \mathcal{D}'\big) = \sup_{\mathscr{B} \subset \mathscr{P}(\Theta)} \big|\mathcal{D}(\mathscr{B}) - \mathcal{D}'(\mathscr{B})\big|.$$

Here, the supremum is taken over all Borel measurable sets $\mathscr{B} \subset \mathscr{P}(\Theta)$. In what follows, fix $r \geq 1$. For a subset $\mathscr{B} \subset \mathscr{P}(\Theta)$ the boundary set $\operatorname{bd}\mathscr{B}$ is defined as the set of all elements $P \in \mathscr{P}(\Theta)$ such that every W_r neighborhood for P has nonempty intersection with $\mathscr{B}$ as well as the complement set $\mathscr{B}^c = \mathscr{P}(\Theta) \setminus \mathscr{B}$.

The primary objects in consideration are a pair of $(\mathcal{D}, \mathscr{C})$, with $\mathcal{D} \in \mathscr{P}(\Theta)$, $\mathscr{C} \subset \mathscr{P}(\mathscr{P}(\Theta))$, where $\mathcal{D} = \mathscr{D}_{\alpha G}$ for some fixed $G \in \mathscr{P}(\Theta)$ and $\alpha > 0$. $\mathscr{C}$ is a class of Dirichlet measures $\mathscr{C} := \{\mathscr{D}_{\alpha' G'} | G' \in \mathscr{G}, \alpha' > 0\}$ for some fixed $\mathscr{G} \subset \mathscr{P}(\Theta)$.

Definition 4.1. *A class $\mathscr{C} \subset \mathscr{P}(\mathscr{P}(\Theta))$ of Dirichlet measures is said to have α^*-regular boundary with respect to $\mathcal{D} = \mathscr{D}_{\alpha G}$ for some constant $\alpha^* > 0$, if there are positive constants C_0, c_0 and c_1 dependent only on $\mathcal{D}$ such that for each $\mathcal{D}' = \mathscr{D}_{\alpha' G'} \in \mathscr{C}$ there exists a measurable subset $\mathscr{B} \subset \mathscr{P}(\Theta)$ for which the following hold*:

(i) $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) \geq c_0 W_r^r(G, G')$,
(ii) $\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) \leq C_0(\delta / W_r(G, G'))^{\alpha^*}$ *for any* $\delta \leq c_1 W_r(G, G')$.

$\mathscr{C}$ is said to have strong *α^*-regularity with respect to $\mathcal{D}$ if condition* (ii) *is replaced by*

(iii) $\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) \leq C_0 \delta^{\alpha^*}$ *for any* $\delta \leq c_1$.

$\mathscr{C}$ is said to have weak *regularity with respect to $\mathcal{D}$ if condition* (ii) *is replaced by*

(iv) $\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) = \mathrm{o}(1)$ *as* $\delta \to 0$.

Remark. The nontrivial requirement here is that constants C_0, c_0 and c_1 are independent of $\mathcal{D}' \in \mathscr{C}$. Consider the following example: $\mathscr{G} := \{G' \in \mathscr{P}(\Theta) | \operatorname{spt} G' \cap \operatorname{spt} G = \varnothing\}$. Take $\mathcal{D}' := \mathscr{D}_{\alpha' G'}$ for some $G' \in \mathscr{G}$. By a standard fact of Dirichlet measures (e.g., see Theorem 3.2.4 of [14]), $\operatorname{spt}\mathcal{D} = \{P\colon \operatorname{spt} P \subset \operatorname{spt} G\}$ and $\operatorname{spt}\mathcal{D}' = \{P\colon \operatorname{spt} P \subset \operatorname{spt} G'\}$. Thus, we also have $\operatorname{spt}\mathcal{D} \cap \operatorname{spt}\mathcal{D}' = \varnothing$. It follows that $V(\mathcal{D}, \mathcal{D}') = 1$. If we choose $\delta_1 = \inf_{\theta \in \operatorname{spt} G; \theta' \in \operatorname{spt} G'} \|\theta - \theta'\| > 0$, and let $\mathscr{B} = (\operatorname{spt}\mathcal{D}')_{\delta_1/2}$, then $\mathcal{D}'(\mathscr{B}) = 1$ and $\mathcal{D}(\mathscr{B}) = 0$. Moreover, for any $\delta \leq \delta_1/4$, $\mathcal{D}(\mathscr{B}_\delta) = 0$, so $\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) = 0$. At the first glance, this construction appears to suggest that $\mathscr{C} := \{\mathscr{D}_{\alpha' G'} | G' \in \mathscr{G}\}$ has (strong) α^*-regular boundary with $\mathcal{D}$ for any $\alpha^* > 0$. This is not the case, because it is not possible to guarantee that $\delta_1 > c_1 W_r(G, G')$ for some c_1 independent of G'. That is, δ_1 can be arbitrarily close to 0 even as $W_r(G, G')$ remains bounded away from 0.

4.1. The case of finite support

We study the regularity of boundaries for the pair $(\mathcal{D}, \mathscr{C})$, where the base measure G of $\mathcal{D} = \mathscr{D}_{\alpha G}$ has a finite number of support points, while class $\mathscr{C}$ consists of Dirichlet measures $\mathcal{D}' = \mathscr{D}_{\alpha G'}$ where G' may have infinite support in Θ. In the following subsection, we extend the theory to handle the case that G has infinite and geometrically sparse support.

Theorem 4.1. *Suppose that Θ is bounded. Let $\mathcal{D} = \mathscr{D}_{\alpha G}$, where $G = \sum_{i=1}^k \beta_i \delta_{\theta_i}$ for some $k < \infty$ and $\alpha \in (0, 1]$. Let $\alpha_1 > \alpha_0 > 0$ be given. Define*

$$\mathscr{C} = \left\{\mathscr{D}_{\alpha' G'} | G' \in \mathscr{P}(\Theta); \alpha' \in [\alpha_0, \alpha_1]\right\}.$$

Then $\mathscr{C}$ has $\alpha^* r$-regular boundary with respect to $\mathcal{D}$, where $\alpha^* = \min_i \alpha\beta_i$.

Proof. Take any $G' \in \mathscr{P}(\Theta)$. Let $\varepsilon := W_r(G, G')$. Choose constants c_1, c_2 such that $c_1^r + c_2 \operatorname{diam}(\Theta)^r \leq 1/2^r$ and $c_1 \operatorname{diam}(\Theta) < m := \min_{1 \leq i \neq j \leq k} \|\theta_i - \theta_j\|/4$. Let $S = \bigcup_{i=1}^k B_i$, where B_i's for $i = 1, \ldots, k$ are closed Euclidean balls of radius $c_1\varepsilon$ and centering at $\theta_1, \ldots, \theta_k$, respectively. Any $G' \in \mathscr{P}(\Theta)$ admits either (A) $G'(S^c) \geq c_2\varepsilon^r$, or (B) $G'(S^c) < c_2\varepsilon^r$.

Case (A). $G'(S^c) \geq c_2\varepsilon^r$. Let $\mathscr{B} = \{Q \in \mathscr{P}(\Theta) | Q(S^c) > 1/2\}$. Clearly, $\mathcal{D}(\mathscr{B}) = 0$. Moreover, for any $Q \in \mathscr{B}$ and $Q' \in \operatorname{spt} \mathcal{D}$, $W_r^r(Q, Q') \geq (1/2)(c_1\varepsilon)^r$. So for any $\delta < (1/2)^{1/r} c_1\varepsilon$, $\mathcal{D}(\mathscr{B}_\delta) = 0$. Condition (ii) of Definition 4.1 is satisfied.

It remains to verify condition (i). If $G'(S) = 0$, then $G'(S^c) = 1$ and $\mathcal{D}'(\mathscr{B}) = 1$. So, $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) = 1$. On the other hand, if $G'(S) > 0$ and suppose that $\operatorname{law}(Q) = \mathcal{D}'$, then $\operatorname{law}(Q(S)) = \operatorname{Beta}(\alpha' G'(S), \alpha' G'(S^c))$. So,

$$\begin{aligned}
\mathcal{D}'(\mathscr{B}) &= \int_0^{1/2} \frac{\Gamma(\alpha')}{\Gamma(\alpha' G'(S))\Gamma(\alpha' G'(S^c))} x^{\alpha' G'(S)-1}(1-x)^{\alpha' G'(S^c)-1}\,\mathrm{d}x \\
&\geq \frac{(1/2)^{\alpha'}\Gamma(\alpha')}{\Gamma(\alpha' G'(S))\Gamma(\alpha' G'(S^c))} \int_0^{1/2} x^{\alpha' G'(S)-1}\,\mathrm{d}x \\
&= \frac{(1/2)^{\alpha'}\Gamma(\alpha')}{\Gamma(\alpha' G'(S))\Gamma(\alpha' G'(S^c))} \times \frac{(1/2)^{\alpha' G'(S)}}{\alpha' G'(S)} \\
&= \frac{(1/2)^{\alpha' + \alpha' G'(S)}\Gamma(\alpha')\alpha' G'(S^c)}{\Gamma(\alpha' G'(S)+1)\Gamma(\alpha' G'(S^c)+1)} \\
&\geq \frac{(1/2)^{2\alpha'}\Gamma(\alpha')\alpha' G'(S^c)}{\max_{1 \leq x \leq \alpha'+1} \Gamma(x)^2} \geq \frac{(1/2)^{2\alpha'}\Gamma(\alpha')\alpha' c_2\varepsilon^r}{\max_{1 \leq x \leq \alpha'+1} \Gamma(x)^2}.
\end{aligned}$$

In the above display, the first inequality is due to $(1-x)^\gamma \geq 1$ if $\gamma \leq 0$, and $(1-x)^\gamma \geq (1/2)^\gamma$ if $\gamma > 0$ for $x \in [0, 1/2]$. The third equality is due to $x\Gamma(x) = \Gamma(x+1)$ for any $x > 0$. Condition (i) is verified.

Case (B). $\beta_0' := G'(S^c) < c_2\varepsilon^r$. Let $\beta_i' = G'(B_i)$ for $i = 1, \ldots, k$. Consider the map $\Phi : \mathscr{P}(\Theta) \to \Delta^{k-1}$, defined by

$$\Phi(Q) := \big(Q(B_1)/Q(S), \ldots, Q(B_k)/Q(S)\big).$$

Define $P_1 := \operatorname{Dir}(\alpha\beta_1, \ldots, \alpha\beta_k)$ and $P_2 := \operatorname{Dir}(\alpha'\beta_1', \ldots, \alpha'\beta_k')$. By a standard property of Dirichlet measures, P_1 and P_2 are push-forward measures of $\mathcal{D}$ and $\mathcal{D}'$, respectively, by Φ. (i.e., if $\operatorname{law}(Q) = \mathcal{D}$, then $\operatorname{law}(\Phi(Q)) = P_1$. If $\operatorname{law}(Q) = \mathcal{D}'$ then $\operatorname{law}(\Phi(Q)) = P_2$.) Define

$$B_1 := \left\{ \mathbf{q} \in \Delta^{k-1} \middle| \frac{\mathrm{d}P_2}{\mathrm{d}P_1}(\mathbf{q}) > 1 \right\}.$$

(This is exactly the same set defined by equation (4) of [21] in the proof of Lemma 4.1 that we shall encounter in the sequel.) Now let $\mathscr{B} = \Phi^{-1}(B_1)$. Then we have $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) = P_2(B_1) - P_1(B_1) = V(P_1, P_2)$.

To verify condition (ii) of Definition 4.1, recall that

$$\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) = \mathcal{D}\left(\left\{ Q = \sum_{i=1}^k q_i \delta_{\theta_i} \,\middle|\, Q \notin \mathscr{B};\, W_r(Q, Q') \le \delta \text{ for some } Q' \in \mathscr{B} \right\}\right).$$

For a measure of the form $Q = \sum_{i=1}^k q_i \delta_{\theta_i}$, $W_r(Q, Q') \le \delta$ entails $Q(B_i) - Q'(B_i) = q_i - Q'(B_i) \le \delta^r/(c_1\varepsilon)^r$, and $Q'(B_i) - q_i \le \delta^r/(m - c_1\varepsilon)^r < \delta^r/(c_1\varepsilon)^r$, for any $i = 1, \ldots, k$. As well, $Q'(S^c) \le \delta^r/(c_1\varepsilon)^r$. This implies that

$$\left|Q(B_i)/Q(S) - Q'(B_i)/Q'(S)\right| = \left| q_i - \frac{Q'(B_i)}{1 - Q'(S^c)} \right| \le \frac{2\delta^r/(c_1\varepsilon)^r}{1 - \delta^r/(c_1\varepsilon)^r} \le 4\delta^r/(c_1\varepsilon)^r,$$

where the last inequality holds as soon as $\delta \le c_1\varepsilon/2^{1/r}$. In short, $W_r(Q, Q') \le \delta$ implies that $\|\Phi(Q) - \Phi(Q')\|_\infty \le 4\delta^r/(c_1\varepsilon)^r$. We have

$$\begin{aligned}
\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) &\le \mathcal{D}\left(\left\{ Q | Q \notin \mathscr{B};\, \left\|\Phi(Q) - \Phi\left(Q'\right)\right\|_\infty \le 4\delta^r/(c_1\varepsilon)^r \text{ for some } Q' \in \mathscr{B} \right\}\right) \\
&= P_1\left(\left\{ \mathbf{q} | \mathbf{q} \notin B_1;\, \left\|\mathbf{q} - \mathbf{q}'\right\|_\infty \le 4\delta^r/(c_1\varepsilon)^r \text{ for some } \mathbf{q}' \in B_1 \right\}\right) \\
&\le C_0(\delta/\varepsilon)^{\alpha^* r}.
\end{aligned}$$

The equality in the previous display is due to the definition of $\mathscr{B}$, while the last inequality is essentially the proof of Lemma 4.1(b). C_0 is a positive constant dependent only on $\mathcal{D}$.

It remains to verify condition (i) in Definition 4.1. We have

$$\begin{aligned}
V(P_1, P_2) &= V(\mathscr{D}_{\sum_{i=1}^k \alpha\beta_i \delta_{\theta_i}}, \mathscr{D}_{\sum_{i=1}^k \alpha'\beta_i' \delta_{\theta_i}}) \\
&\ge \frac{1}{(2\operatorname{diam}(\Theta))^r} W_r^r(\mathscr{D}_{\alpha G}, \mathscr{D}_{\sum_{i=1}^k \alpha'\beta_i' \delta_{\theta_i}}) \qquad (23) \\
&\ge \frac{1}{(2\operatorname{diam}(\Theta))^r} W_r^r\left(G, \sum_{i=1}^k \frac{\beta_i'}{1 - \beta_0'} \delta_{\theta_i}\right).
\end{aligned}$$

The first inequality in the above display is due to Theorem 6.15 of [29], while the second inequality is due to Lemma 3.1(a). Now, we have

$$\begin{aligned}
W_r^r\left(G', \sum_{i=1}^k \frac{\beta_i'}{1 - \beta_0'} \delta_{\theta_i}\right) &\le (c_1\varepsilon)^r \sum_{i=1}^k \left(\beta_i' \wedge \frac{\beta_i'}{1 - \beta_0'}\right) + \operatorname{diam}(\Theta)^r \sum_{i=1}^k \left|\beta_i' - \frac{\beta_i'}{1 - \beta_0'}\right| \\
&\le (c_1\varepsilon)^r + \operatorname{diam}(\Theta)^r \sum_{i=1}^k \frac{\beta_i'\beta_0'}{1 - \beta_0'} \\
&\le \varepsilon^r\left(c_1^r + c_2 \operatorname{diam}(\Theta)^r\right) \le \varepsilon^r/2^r.
\end{aligned}$$

The last inequalities in the above display is due to the hypothesis that $\beta_0' < c_2\varepsilon^r$, and the choice of c_1, c_2. By triangle inequality,

$$W_r\left(G, \sum_{i=1}^{k} \frac{\beta_i'}{1-\beta_0'}\delta_{\theta_i}\right) \geq W_r\big(G, G'\big) - W_r\left(G', \sum_{i=1}^{k} \frac{\beta_i'}{1-\beta_0'}\delta_{\theta_i}\right) \geq \varepsilon - \varepsilon/2 = \varepsilon/2.$$

Combining with equation (23), we obtain that $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) = V(P_1, P_2) \geq \frac{1}{(2\,\mathrm{diam}(\Theta))^r}(\varepsilon/2)^r$. This concludes the proof. □

The following lemma, which establishes strong regularity for a restricted class of Dirichlet measures, supplies a key argument in the proof of the previous theorem. The proof of this lemma is quite technical and deferred to [21].

Lemma 4.1. *Let $\mathcal{D} = \mathscr{D}_{\alpha G}$, where $G = \sum_{i=1}^{k} \beta_i \delta_{\theta_i}$ for some $k < \infty$, $\alpha, \alpha' > 0$. Define*

$$\mathscr{C} = \big\{\mathscr{D}_{\alpha' G'} | G' \in \mathscr{P}(\Theta), \mathrm{spt}\, G' = \mathrm{spt}\, G\big\}.$$

(a) *If* $\min_i \alpha\beta_i \geq 1$, *then* $\mathscr{C}$ *has* strong r*-regular boundary with respect to* $\mathcal{D}$.
(b) *If* $\max_i \alpha\beta_i < 1$, *then* $\mathscr{C}$ *has* strong $\alpha^* r$*-regular boundary with respect to* $\mathcal{D}$, *where* $\alpha^* = \min_i \alpha\beta_i$.

4.2. The case of infinite and geometrically sparse support

In this subsection, we study a class of base measures G that have infinite support points, but that remain amenable to our analysis of regular boundaries. In particular, we consider the class of sparse measures on Θ (either ordinary sparse or supersparse) given by Definition 2.1.

Theorem 4.2. *Assume that $\mathcal{D} = \mathscr{D}_{\alpha G}$ for some $\alpha \in (0, 1]$. $\mathrm{spt}\, G$ is a (c_1, c_2, K)-sparse subset of a bounded space Θ and that G is a sparse measure equipped with gauge function g. Let $\alpha_1 \geq \alpha_0 > 0$. Then, for any $\mathcal{D}' \in \mathscr{C}$, where*

$$\mathscr{C} = \big\{\mathcal{D}' = \mathscr{D}_{\alpha' G'} | G' \in \mathscr{P}(\Theta), \alpha' \in [\alpha_0, \alpha_1]\big\}$$

there exists a measurable set $\mathscr{B} \subset \mathscr{P}(\Theta)$ for which

(i) $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) \gtrsim W_r^r(G, G')$,
(ii) *for any* $\delta \lesssim W_r(G, G')$,

$$\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) \lesssim 24^{K(c_0 W_r(G,G'))} \times \left(\frac{\delta}{W_r(G, G')}\right)^{\alpha r g(c_0 W_r(G,G'))}.$$

Here, c_0 and the multiplying constants in $\lesssim$ and $\gtrsim$ depend only on $\mathcal{D}$.

The proof of this result is similar to Theorem 4.1 and deferred to [21].

5. Upper bounds for Wasserstein distances of base measures

The main purpose of this section is to obtain an upper bound of distance of Dirichlet base measures $W_r(G, G')$ in terms of the variational distance of the marginal densities of observed data $V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$. In particular, we will establish an inequality of the form: for a fixed $G \in \mathscr{P}(\Theta)$ and *any* $G' \in \mathscr{P}(\Theta)$,

$$W_r^r\big(G, G'\big) \lesssim V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) + A_n\big(G, G'\big), \tag{24}$$

where $A_n(G, G')$ is a quantity that tends to 0 as $n \to \infty$. The rate at which $A_n(G, G')$ tends to 0 depends on the sparse structure of G, and the smoothness of the kernel density $f(x|\theta)$. The full details are given in the statement of Theorem 5.1. It is worth contrasting this to the relatively easier inequalities in the opposite direction, given by Lemma 3.2: $V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq h(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \lesssim n W_{2r}^{2r}(G, G')$ holds generally for any pair of G, G'.

The proof of inequality (24) hinges on the existence of a suitable set $\mathscr{B}_n \subset \mathscr{P}(\Theta)$ measurable with respect to (the sigma algebra induced by) the observed variables $Y_{[n]}$, which can then be used to distinguish G' from G, in the sense that

$$W_r^r\big(G, G'\big) \lesssim P_{Y_{[n]}|G'}(\mathscr{B}_n) - P_{Y_{[n]}|G}(\mathscr{B}_n) + A_n\big(G, G'\big).$$

In the previous section, we have already shown the existence of subset $\mathscr{B} \subset \mathscr{P}(\Theta)$ for which

$$W_r^r\big(G, G'\big) \lesssim \mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}).$$

To link up this result to the desired bound (24), the missing piece of the puzzle is the existence of a point estimate for the mixing measures on the basis of observed variables $Y_{[n]}$. In the following, we shall establish the existence of such point estimators, which admit finite-sample probability bounds that may also be of independent interest.

5.1. Finite-sample probability bounds for deconvolution problem

Let $\mathcal{Q}$ be a subset of $\mathscr{P}(\Theta)$, and $\mathcal{F} = \{Q * f | Q \in \mathcal{Q}\}$. Let $\mathcal{Q}_k \subset \mathscr{P}(\Theta)$ be subset of measures with at most k support points. $\mathcal{F}_k = \{Q * f | Q \in \mathcal{Q}_k\}$. Given an i.i.d. n-vector $Y_{[n]} = (Y_1, \ldots, Y_n)$ according to the convolution mixture density $Q_0 * f$ for some $Q_0 \in \mathcal{Q}$. Let η_n be a sequence of positive numbers converging to zero. Following [32], we consider an η_n-MLE (maximum likelihood estimator) $\hat{f}_n \in \mathcal{F}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_n(Y_i) \geq \sup_{g \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \log g(Y_i) - \eta_n.$$

By our construction, there exists $\hat{Q}_n \in \mathcal{Q}$ such that $\hat{f}_n = \hat{Q}_n * f$.

Lemma 5.1. *Suppose that assumption* (A1) *holds for some* $r \geq 1, C_1 > 0$. *Let* η_n *satisfy* $\eta_n \leq c_1 \varepsilon_n^2$, $\varepsilon_n \to 0$ *at a rate to be specified. Then the* η_n*-MLE satisfies the following bound under*

$Q_0 * f$-measure, for any $Q_0 \in \mathcal{Q}$:

$$\mathbb{P}\big(h(\hat{f}_n, Q_0 * f) \geq \varepsilon_n\big) \leq 5\exp\big(-c_2 n\varepsilon_n^2\big), \tag{25}$$

$$\mathbb{P}\big(W_2(\hat{Q}_n, Q_0) \geq \delta_n\big) \leq 5\exp\big(-c_2 n\varepsilon_n^2\big), \tag{26}$$

where c_1, c_2 are some universal positive constants. ε_n and δ_n are given as follows:

(a) $\varepsilon_n = C_2(\log n/n)^{r/2d}$, *if* $d > 2r$; $\varepsilon_n = C_2(\log n/n)^{r/(d+2r)}$ *if* $d < 2r$, *and* $\varepsilon_n = (\log n)^{3/4}/n^{1/4}$ *if* $d = 2r$.

(b) $\varepsilon_n = C_2 n^{-1/2}\log n$, *if* $\mathcal{Q} = \mathcal{Q}_k$ *and* $\mathcal{F} = \mathcal{F}_k$ *for some* $k < \infty$.

(c) *If f is ordinary smooth with parameter $\beta > 0$, then $\delta_n = C_3\varepsilon_n^{1/(2+\beta d')}$ for any $d' > d$.*

(d) *If f is supersmooth with parameter $\beta > 0$, then $\delta_n = C_3[-\log\varepsilon_n]^{-1/\beta}$.*

Here, C_2, C_3 are different constants in each case. C_2 depends only on d, r, Θ and C_1, while C_3 depends only on d, β, Θ and C_2.

Proof. Recall Theorem 2 of [32], which is restated as follows: Suppose that $\varepsilon = \varepsilon_n$ satisfies the following inequality:

$$\int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} \big[\log N(u/c_3, \mathcal{F}, h)\big]^{1/2}\,\mathrm{d}u \leq c_4 n^{1/2}\varepsilon^2, \tag{27}$$

where c_3 and c_4 are certain universal constants (cf. Theorem 1 of [32]). Then, for some universal constants $c_1, c_2 > 0$, if $\eta_n \leq c_1\varepsilon_n^2$, the following probability bound holds under $Q_0 * f$-measure, for any $Q_0 \in \mathcal{Q}$,

$$\mathbb{P}\big(h(\hat{f}_n, Q_0 * f) \geq \varepsilon_n\big) \leq 5\exp\big(-c_2 n\varepsilon_n^2\big).$$

It remains to verify the entropy condition (27) given the rates specified in the statement of the present lemma. We shall make use of the following entropy bounds (cf. Lemma 4 of [19]):

$$\log N(2\delta, \mathcal{Q}, W_r) \leq N\big(\delta, \Theta, \|\cdot\|\big)\log\big(e + e\,\mathrm{diam}(\Theta)^r/\delta^r\big), \tag{28}$$

$$\log(2\delta, \mathcal{Q}_k, W_r) \leq k\big(\log N\big(\delta, \Theta, \|\cdot\|\big) + \log\big(e + e\,\mathrm{diam}(\Theta)^r/\delta^r\big)\big). \tag{29}$$

By assumption (A2) and Lemma 3.2, we have $h^2(Q * f, Q' * f) \leq C_1 W_{2r}^{2r}(Q, Q')$. This implies that

$$N(u/c_3, \mathcal{F}, h) \leq N\big(\big(u^2/c_3^2 C_1\big)^{1/2r}, \mathcal{Q}, W_{2r}\big).$$

Since $\Theta \subset \mathbb{R}^d$, $N(\delta, \Theta, \|\cdot\|) \leq (\mathrm{diam}(\Theta)/\delta)^d$. So, by (28),

$$\begin{aligned}
&\int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} \big[\log N\big(\big(u^2/c_3^2 C_1\big)^{1/2r}, \mathcal{Q}, W_{2r}\big)\big]^{1/2}\,\mathrm{d}u \\
&\quad\leq \int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} \left[N\left(\frac{u^{1/r}}{2c_3^{1/r}C_1^{1/2r}}, \Theta, \|\cdot\|\right)\log\big(e + e\,\mathrm{diam}(\Theta)^{2r}2^{2r}c_3^2 C_1/u^2\big)\right]^{1/2}\mathrm{d}u
\end{aligned}$$

$$\leq \int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} \big(2\operatorname{diam}(\Theta)\big)^{d/2} c_3^{d/2r} C_1^{d/4r} u^{-d/2r} \big[\log\big(e + e\operatorname{diam}(\Theta)^{2r} 2^{2r} c_3^2 C_1/u^2\big)\big]^{1/2} \mathrm{d}u.$$

For equation (27) to hold, it suffices to have the right-hand side of the inequality in the above display bounded by $c_4 n^{1/2}\varepsilon^2$. Indeed, this is straightforward to check for the rates given in part (a) of the lemma.

Part (b) of the lemma is proved in the same way, by invoking a tighter bound on the covering number via equation (29). Parts (c) and (d) are immediate consequences of part (a) and (b) by invoking Theorem 2 of [19]. □

5.2. Key upper bound for the Wasserstein distance of base measures

We are ready to prove the key theorem of this section.

Theorem 5.1. *Suppose that Θ is a bounded subset of $\mathbb{R}^d$,* (A1) *holds for some $C_1 > 0$ and some $r \in [1, 2]$. Let δ_n and ε_n be vanishing sequences for which equation* (26) *holds. Fix $G \in \mathscr{P}(\Theta)$ and $\alpha \in (0, 1]$, while G' varies in $\mathscr{P}(\Theta)$. Let $\alpha^* = \alpha \inf_{\theta \in \operatorname{spt} G} G(\{\theta\})$. Then there are positive constants c_0, c_1, C_0 depending only on G, and $c_2 > 0$ a universal constant, such that for any $G' \in \mathscr{P}(\Theta)$, $\alpha' \in [\alpha_1, \alpha_0]$ given and n sufficiently large so that $\delta_n \lesssim W_r(G, G')$, the following holds*:

$$c_0 W_r^r\big(G, G'\big) \leq V\big(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}\big) + 10\exp\big(-c_2 n\varepsilon_n^2\big) + A_n\big(W_r\big(G, G'\big)\big), \tag{30}$$

where $A_n(W_r(G, G'))$ takes the form:

$$A_n(\omega) = \begin{cases} C_0(2\delta_n/\omega)^{\alpha^* r}, & \text{if } G \text{ has finite support,} \\ C_0 24^{K(c_1\omega)}(2\delta_n/\omega)^{\alpha r g(c_1\omega)}, & \text{if } G \text{ is } (\gamma_1, \gamma_2, K)\text{-sparse with gauge } g. \end{cases} \tag{31}$$

Proof. Suppose that G has finite support. By Theorem 4.1 (applied for W_r) there are positive constants C_0, c_0 independent of G' such that for some measurable set $\mathscr{B} \subset \mathscr{P}(\Theta)$, (i) $\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}) \geq c_0 W_r^r(G, G')$ and (ii) $\mathcal{D}(\mathscr{B}_\delta \setminus \mathscr{B}) \leq C_0(\delta/W_r(G, G'))^{\alpha^* r}$ for all $\delta \lesssim W_r(G, G')$.

Recall that $\hat{Q}_n$ is a point estimate of Q defined earlier in this section. By the definition of variational distance, for any $\delta > 0$

$$V\big(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}\big) \geq \mathbb{P}\big(\hat{Q}_n \in \mathscr{B}_\delta | G'\big) - \mathbb{P}(\hat{Q}_n \in \mathscr{B}_\delta | G).$$

Here, $\mathbb{P}(\cdot|G)$ is taken to mean the probability of an event given that the observations are generated according to the Dirichlet base measure G. Set $\mathscr{B}_\delta := \{Q \in \mathscr{P}(\Theta) |$ there is $Q' \in \mathscr{B}$ such that $W_r(Q, Q') \leq \delta\}$. We have

$$\begin{aligned} \mathbb{P}\big(\hat{Q}_n \in \mathscr{B}_\delta | G'\big) &\geq \mathbb{P}\big(\hat{Q}_n \in \mathscr{B}_\delta, W_r(\hat{Q}_n, Q) < \delta | G'\big) \\ &\geq \mathbb{P}\big(Q \in \mathscr{B}, W_r(\hat{Q}_n, Q) < \delta | G'\big) \\ &\geq \mathcal{D}'(\mathscr{B}) - \mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta | G'\big). \end{aligned}$$

We also have

$$\begin{aligned}\mathbb{P}(\hat{Q}_n \in \mathscr{B}_\delta | G) &\leq \mathbb{P}\big(\hat{Q}_n \in \mathscr{B}_\delta, W_r(\hat{Q}_n, Q) < \delta | G\big) + \mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta | G\big) \\ &\leq \mathbb{P}(Q \in \mathscr{B}_{2\delta} | G) + \mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta | G\big) \\ &= \mathcal{D}(\mathscr{B}_{2\delta}) + \mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta | G\big).\end{aligned}$$

Hence,

$$\begin{aligned}&V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \\ &\geq \mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B}_{2\delta}) - 2 \sup_{Q \in \mathcal{Q}} \mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta\big) \\ &\geq \big(\mathcal{D}'(\mathscr{B}) - \mathcal{D}(\mathscr{B})\big) - \mathcal{D}(\mathscr{B}_{2\delta} \setminus B) - 2 \sup_{Q \in \mathcal{Q}} \mathbb{P}\big(W_r(\hat{Q}_n, Q) \geq \delta\big).\end{aligned}$$

Since $r \in [1, 2]$, $W_r(\hat{Q}_n, Q) \leq W_2(\hat{Q}_n, Q)$. Choose $\delta := \delta_n$ such that equation (26) holds. Then, as soon as $2\delta_n \lesssim W_r(G, G')$, for some multiplying constant depending only on G, we have

$$V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \geq c_0 W_r^r\big(G, G'\big) - C_0\big(2\delta_n / W_r\big(G, G'\big)\big)^{\alpha^* r} - 10\exp\big(-c_2 n\varepsilon_n^2\big).$$

The case that G has infinite support proceeds in a similar way by invoking Theorem 4.2. □

Remark. As we shall see shortly, Theorem 5.1 is instrumental in the proof of Theorem 2.2: one can now deduce the convergence of the Dirichlet base measure G (toward G_0) from the convergence of the corresponding marginal density $p_{Y_{[n]}|G}$ (toward $p_{Y_{[n]}|G_0}$). We note that the bound represented by (30) is not sharp in certain regimes, which carry immediate consequences on the kind of posterior concentration rates that we can obtain for G. In particular, the right-hand side of inequality (30) increases as $n \to \infty$, due to the fact that $V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'})$ typically increases as n increases, while the left-hand side is independent of n.

The root of this unnatural feature is due to a simple technique employed in the proof of Theorem 5.1, which targets the regime that $n \to \infty$, so that one can build on the machinery of the existence of a robust test for Dirichlet base measures developed in Section 4. Ideally, one would like to construct a test for base measure G given n-vector data $Y_{[n]}$, by integrating out the latent variable Q. Instead, the bound (30) of Theorem (5.1) is derived by a decoupling approach: one can first obtain a point estimate for Q on the basis of the data $Y_{[n]}$, and then relies on the existence of a robust test for G based on the population of Q. Due to the decoupling approach, we necessarily require n to grow so that the quality of the point estimate for Q is sufficiently good. An artifact of this technique, however, is that the upper bound for $W_r(G, G')$ can only be derived as a summation of several quantities, two of which vanish as n increases (as desired), but the same cannot be said for the remaining quantity, that is, the variational distance of marginal densities of n-vector $Y_{[n]}$.

5.3. Proof of Theorem 2.2

Now we are ready to prove Theorem 2.2. By Theorem 2.1, as $m \to \infty$, while n either varies with m or is held fixed, we have

$$\Pi_G\big(V(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq \varepsilon_{mn} | Y_{[n]}^{[m]}\big) \to 1$$

in $P^m_{Y_{[n]}|G_0}$-probability. Here, we exploit the fact that $V \leq h$. Now, by Theorem 5.1 applied to the pair of G_0, G, with the latter allowed to vary in $\mathscr{P}(\Theta)$, there are positive constants c_0, c_1, C_0 depending on G_0 and $c_2 > 0$ a universal constant such that

$$c_0 W_1(G_0, G) \leq V(P_{Y_{[n]}|G_0}, P_{Y_{[n]}|G}) + 10\exp\big(-c_2 n\varepsilon_n^2\big) + A_n\big(W_1(G_0, G)\big), \tag{32}$$

for any $G \in \mathscr{P}(\Theta)$. So we have

$$\Pi_G\big(c_0 W_1(G_0, G) \leq \varepsilon_{mn} + 10\exp\big(-c_2 n\varepsilon_n^2\big) + A_n\big(W_1(G_0, G)\big) | Y_{[n]}^{[m]}\big) \to 1$$

in $P^m_{Y_{[n]}|G_0}$-probability.

To derive concrete concentration rates, consider the case G_0 has finite support. By Theorem 5.1 $A_n(W_1(G_0, G)) \asymp (2\delta_n / W_1(G_0, G))^{\alpha^*}$. Plugging to equation (32), we obtain

$$\begin{aligned} W_1(G_0, G) &\lesssim V(P_{Y_{[n]}|G_0}, P_{Y_{[n]}|G}) + \exp\big(-c_2 n\varepsilon^2\big) + \delta_n^{\alpha^*/(\alpha^*+1)} \\ &\lesssim V(P_{Y_{[n]}|G_0}, P_{Y_{[n]}|G}) + \delta_n^{\alpha^*/(\alpha^*+1)}, \end{aligned}$$

where we have exploited the fact that the term $\exp(-c_2 n\varepsilon^2)$ is negligible compared to the remaining terms. The conclusion of the theorem follows immediately.

Next, consider the case G_0 has infinite support, and in fact has geometrically sparse support. For the case that G_0 is super sparse with parameters (γ_0, γ_1), that is, $K(\varepsilon) \lesssim [\log(1/\varepsilon)]^{\gamma_0}$, and $g(\varepsilon) \gtrsim [\log(1/\varepsilon)]^{-\gamma_1}$. It is simple to verify that as long as $\varepsilon \gtrsim \delta_n$, the constraint

$$\varepsilon \lesssim A_n(\varepsilon) = 24^{K(c_1\varepsilon)} \times (2\delta_n/\varepsilon)^{c_1 g(c_1\varepsilon)}$$

implies that

$$\varepsilon \lesssim \exp -\big[\log(1/\delta_n)\big]^{1/(\gamma_1+1\vee\gamma_0)}.$$

Thus, equation (32) entails that

$$W_1(G_0, G) \lesssim V(P_{Y_{[n]}|G_0}, P_{Y_{[n]}|G}) + \exp -\big[\log(1/\delta_n)\big]^{1/(\gamma_1+1\vee\gamma_0)}.$$

For the case that G_0 is ordinary sparse with parameters (γ_0, γ_1), that is $K(\varepsilon) \lesssim (1/\varepsilon)^{\gamma_0}$, and $g(\varepsilon) \gtrsim \varepsilon^{\gamma_1}$. Similarly, note that the inequality

$$\varepsilon \lesssim A_n(\varepsilon)$$

entails that

$$\varepsilon \lesssim \big[\log(1/\delta_n)\big]^{-1/(\gamma_1+\gamma_0)}.$$

Thus we have shown that

$$\Pi_G\big(W_1(G_0, G) \lesssim \varepsilon_{mn} + \Delta_n | Y_{[n]}^{[m]}\big) \to 1$$

in $P_{Y_{[n]}|G_0}^m$-probability, for the choice of Δ_n given in the statement of the theorem.

Examples of ε_n and δ_n are given in Lemma 5.1: If f is an ordinary smooth kernel density, $\log(1/\delta_n) \asymp \frac{1}{2+\beta d'} \log(1/\varepsilon_n) \asymp \log n$. If f is a supersmooth kernel density, $\log(1/\delta_n) \asymp \frac{1}{\beta} \log\log(1/\varepsilon_n) \asymp \log\log n$.

6. Borrowing strength in hierarchical Bayes

This section is devoted to the proof of Theorem 2.3. The proof is a simple consequence from Lemma 6.4, which establishes the posterior concentration behavior for a mixture distribution $Q * f$, where Q is a Dirichlet process distributed by $\mathscr{D}_{\alpha G}$, given that the base measure G is a small perturbation from the true base measure G_0 that is now assumed to have finite support. A complete statement of Lemma 6.4 is given in Section 6.3. In the following we proceed to give a proof of Theorem 2.3.

6.1. Proof of Theorem 2.3

Recall that for each $\tilde{n}$, $\delta_{mn} = \delta_{mn}(\tilde{n})$ is a net of scalars indexed by m, n that tend to 0. Define $A_{mn}^{(\tilde{n})} := \{G\colon\ W_1(G, G_0) \geq \delta_{mn}\}$ and $B_{mn}^{(\tilde{n})} := \{Q_0\colon\ h(Q_0 * f, Q_0^* * f) \geq C((\log \tilde{n}/\tilde{n})^{1/(d+2)} + \delta_{mn}^{r/2} \log(1/\delta_{mn}))\}$ for some large constant C. Due to the conditional independence of $Y_{[\tilde{n}]}^0$ and $Y_{[n]}^{[m]}$ given G,

$$\begin{aligned}\Pi_Q\big(Q_0 \in B_{mn}^{(\tilde{n})} | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}\big) &= \int \Pi_Q\big(Q_0 \in B_{mn}^{(\tilde{n})} | G, Y_{[\tilde{n}]}^0\big)\, \mathrm{d}\Pi_G\big(G | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}\big) \\ &\leq \int_{\mathscr{P}(\Theta)\setminus A_{mn}^{(\tilde{n})}} \Pi_Q\big(Q_0 \in B_{mn}^{(\tilde{n})} | G, Y_{[\tilde{n}]}^0\big)\, \mathrm{d}\Pi_G\big(G | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}\big) \\ &\quad + \Pi_G\big(G \in A_{mn}^{(\tilde{n})} | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}\big).\end{aligned}$$

For each $\tilde{n}$, the second quantity in the upper bound tends to 0 in $P_{Y_{[\tilde{n}]}^0|Q_0^*} \times P_{Y_{[n]}|G_0}$-probability, as $m, n \to \infty$ at suitable rates by condition (b) of the theorem. Now, as $\tilde{n} \to \infty$, the first quantity tends to 0 as a consequence of Lemma 6.4. This completes the proof for (i). Parts (ii) and (iii) are proved in the same way.

6.2. Wasserstein geometry of the support of a single Dirichlet measure

Before proceeding to a proof for Lemma 6.4, we prepare three technical lemmas, which provide a detailed picture of the geometry of the support of a Dirichlet measure, and may be of independent interest. The first lemma demonstrates gains in the thickness of the conditional Dirichlet prior (given a perturbed base measure) compared to the unconditional Dirichlet prior. The second and third lemma show that Dirichlet measure concentrates most its mass on "small" sets, by which we mean sets that admit a small number of covering balls in Wasserstein metrics. This characterization enables the construction of a suitable sieves as required by the proof of Lemma 6.4.

Lemma 6.1. *Given* $G_0 = \sum_{i=1}^k \beta_i \delta_{\theta_i}$ *and small* $\varepsilon > 0$. *Let* $G \in \mathscr{P}(\Theta)$ *such that* $W_1(G, G_0) \le \varepsilon$. *Suppose that* $\mathrm{law}(Q) = \mathscr{D}_{\alpha G}$, *where* $\alpha \in (0, 1]$.

(a) *For any* $Q_0 \in \mathscr{P}(\Theta)$ *such that* $\operatorname{spt} Q_0 \subset \operatorname{spt} G_0$, *and any* δ *such that* $\delta \ge \max_{i \le k} 2\varepsilon/\beta_i$ *and* $\delta \le \min_{i,j \le k} \|\theta_i - \theta_j\|/2$, *any* $r \ge 1$, *there holds*

$$\mathbb{P}\big(W_r(Q_0, Q) \le 2^{1/r}\delta\big) \ge \Gamma(\alpha)(\alpha/2)^k \bigg(\frac{\delta^r}{2k \operatorname{diam}(\Theta)}\bigg)^{\alpha+k-1} \prod_{i=1}^k \beta_i.$$

(b) *In addition, suppose that* (A1)–(A2) *hold for some* $r \ge 1$. *Then, there are constants* $C, c > 0$ *depending only on* $\alpha, k, C_1, M, \operatorname{diam}(\Theta), r$ *and* β_i*'s such that for any* δ *such that* $\delta/\log(1/\delta) \ge C\varepsilon^{r/2}$,

$$\mathbb{P}\big(Q \in B_K(Q_0, \delta)\big) \ge c\big(\delta/\log(1/\delta)\big)^{2(\alpha+k-1)}.$$

This should be contrasted with the general small ball probability bound of Dirichlet process as stated by Lemma 3.4. In that lemma, the base measure is an arbitrary nonatomic measure, while the lower bound is applied to any small W_r ball centering at an arbitrary measure. The lower bound is exponentially small in the radius. In the present lemma, the base measure G is constrained to being close to a discrete measure G_0 with $k < \infty$ support points, while the lower bound is applied to small W_r balls centering at Q_0 that shares the same support as G_0. As a result, the lower bound is only polynomially small in the radius.

The following lemma relies on the intuition that the Dirichlet measure concentrates most its mass on probability measures which place most their mass on a "small" number of support points.

Lemma 6.2. *Let* $\mathcal{D} := \mathscr{D}_{\alpha G}$ *and* $r \ge 1$. *For any* $\delta > 0$, *and for any* $k \in \mathbb{N}_+$, *there is a measurable set* $\mathcal{B}_k \subset \mathscr{P}(\Theta)$ *satisfies the following properties*:

(a) $\sup_{Q \in \mathcal{B}_k} \inf_{Q' \in \mathcal{Q}_k} W_r(Q, Q') \le \delta$.
(b) $\log N(\delta, \mathcal{B}_k, W_r) \le k(\log N(\delta/4, \Theta, \|\cdot\|) + \log(e + 4e \operatorname{diam}(\Theta)^r/\delta^r))$.
(c) *There holds*

$$\mathcal{D}\big(\mathscr{P}(\Theta) \setminus \mathcal{B}_k\big) \le k^{-k}\big(\delta/\operatorname{diam}(\Theta)\big)^{\alpha r}\big[e\alpha r \log\big(\operatorname{diam}(\Theta)/\delta\big)\big]^k.$$

To see that the set $\mathcal{B}_k$ has small entropy relative to $\mathscr{P}(\Theta)$, we note a general estimate for $\mathscr{P}(\Theta)$, which gives an upper bound that is exponentially large in terms of the entropy of Θ (cf. equation (28)):

$$\log N\big(\delta, \mathscr{P}(\Theta), W_r\big) \leq N\big(\delta/2, \Theta, \|\cdot\|\big) \log\big(e + 2e \operatorname{diam}(\Theta)^r/\delta^r\big).$$

In Lemma 6.2, the bound on entropy of $\mathcal{B}_k$ increases only linearly in the entropy of Θ. However, it also increases with k, which controls the measure of the complement of $\mathcal{B}_k$. Next, we consider the additional assumption that the Dirichlet base measure is a small perturbation of a discrete measure with k support points. The strength of this result compared to the previous lemma is that the entropy estimate depends only linearly on the entropy of Θ, while k is fixed. The measure of the complement set of $\mathcal{B}$ is controlled only by the amount of perturbation.

Lemma 6.3. *Given $\varepsilon > 0$, $k < \infty$, $r \geq 1$. Let $G_0, G \in \mathscr{P}(\Theta)$ such that G_0 has k support points and $W_1(G, G_0) \leq \varepsilon$. Let $\mathcal{D} := \mathscr{D}_{\alpha G}$ for some $\alpha > 0$. For any $\delta > 0$, there is a measurable set $\mathcal{B} \subset \mathscr{P}(\Theta)$ that satisfies the following*:

(a) $\log N(\delta, \mathcal{B}, W_r) \leq k(\log N(\delta/4, \Theta, \|\cdot\|) + \log(e + 4e \operatorname{diam}(\Theta)^r/\delta^r))$.
(b) $\mathcal{D}(\mathscr{P}(\Theta) \setminus \mathcal{B}) \leq \varepsilon \operatorname{diam}(\Theta)^{r-1}/\delta^r$.

The proofs of all three lemmas are given in [21].

6.3. Posterior concentration under perturbation of base measure

Here, we state a key result that is needed in the proof of Theorem 2.3.

Lemma 6.4. *Let Θ be a bounded subset of $\mathbb{R}^d$. Assumptions* (A1)–(A2) *hold. Let $Q_0 \in \mathscr{P}(\Theta)$ such that* $\operatorname{spt} Q_0 \subset \operatorname{spt} G_0$, *where $G_0 = \sum_{i=1}^k \beta_i \delta_{\theta_i}$ for some $k < \infty$. Let Π_G be an arbitrary prior distribution on $\mathscr{P}(\Theta)$. Consider the following hierarchical model*:

$$G \sim \Pi_G,\ Q|G \ \sim\ \Pi_Q := \mathscr{D}_{\alpha G},$$

$$Y_{[n]} = (Y_1, \ldots, Y_n)|Q \overset{\text{i.i.d.}}{\sim} Q * f.$$

Let $\varepsilon_n \downarrow 0$ and define events $\mathcal{E}_n := \{W_1(G, G_0) \leq \varepsilon_n\}$. Then the posterior distribution of Q given $Y_{[n]}$ admits the following as $n \to \infty$:

$$\Pi_Q\big(h(Q * f, Q_0 * f) \geq \delta_n | Y_{[n]}, \mathcal{E}_n\big) \to 0, \tag{33}$$

$$\Pi_Q\big(W_2(Q, Q_0) \geq M_n \delta_n | Y_{[n]}, \mathcal{E}_n\big) \to 0 \tag{34}$$

in $(Q_0 * f) \times \Pi_G$-probability, where the rates δ_n and $M_n\delta_n$ are given as follows:

(i) $\delta_n \asymp (\log n/n)^{1/(d+2)} + \varepsilon_n^{r/2} \log(1/\varepsilon_n)$.
(ii) *If f is ordinary smooth with smoothness $\beta > 0$, $M_n\delta_n \asymp \delta_n^{1/(2+\beta d')}$ for any $d' > d$.*
(iii) *If f is supersmooth with smoothness $\beta > 0$, then $M_n\delta_n \asymp (-\log \delta_n)^{-1/\beta}$.*

If $\varepsilon_n \downarrow 0$ suitably fast, then the following rates for δ_n are valid:

(iv) *If f is ordinary smooth, and $\varepsilon_n \to 0$ sufficiently fast such that $\varepsilon_n \lesssim n^{-(\alpha+k+4M_0)} \times (\log n)^{-(\alpha+k-2)}$, where M_0 is some large constant, then $\delta_n \asymp (\log n/n)^{1/2}$.*

(v) *If f is supersmooth with smoothness $\beta > 0$, and $\varepsilon_n \to 0$ sufficiently fast such that $\varepsilon_n \lesssim n^{-2(\alpha+k)/(\beta+2)}(\log n)^{-2(\alpha+k-1)} \exp(-4n^{\beta/(\beta+2)})$, then $\delta_n \asymp (1/n)^{1/(\beta+2)}$.*

We defer the proof of this lemma to [21]. The basic structure contains of mostly standard calculations. The main novel part of the proof lies in the construction of suitable sieves that yield fast rates of convergence. The existence of such sieves is a direct consequence of the geometric lemmas presented in the previous subsection.

Acknowledgements

This research was supported in part by NSF grants CCF-1115769, NSF CAREER DMS-1351362, and NSF CNS-1409303. The author wishes to thank the Associate Editor and referees for many helpful comments.

Supplementary Material

Proofs of remaining results (DOI: 10.3150/15-BEJ703SUPP; .pdf). Due to space constraints, we provide the proofs of the remaining technical results of this paper in [21].

References

[1] Barron, A., Schervish, M.J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[2] Berger, J.O. (1993). *Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics*. New York: Springer. MR1234489

[3] Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

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

[5] Doss, H. and Sellke, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *Ann. Statist.* **10** 1302–1305. MR0673666

[6] Falconer, K.J. (1986). *The Geometry of Fractal Sets. Cambridge Tracts in Mathematics* **85**. Cambridge: Cambridge Univ. Press. MR0867284

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

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

[9] Garcia, I., Molter, U. and Scotto, R. (2007). Dimension functions of Cantor sets. *Proc. Amer. Math. Soc.* **135** 3151–3161. MR2322745

[10] Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli* **20** 2039–2075. MR3263098
[11] Gassiat, E. and van Handel, R. (2014). The local geometry of finite mixtures. *Trans. Amer. Math. Soc.* **366** 1047–1072. MR3130325
[12] Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics*. *Camb. Ser. Stat. Probab. Math*. 35–79. Cambridge: Cambridge Univ. Press. MR2730660
[13] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007
[14] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. *Springer Series in Statistics*. New York: Springer. MR1992245
[15] Giné, E. and Nickl, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395
[16] Hjort, N.L., Holmes, C., Müller, P. and Walker, S.G., eds. (2010). *Bayesian Nonparametrics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge: Cambridge Univ. Press. MR2722987
[17] Korwar, R.M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705–711. MR0350950
[18] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. New York: Springer. MR1639875
[19] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422
[20] Nguyen, X. (2015). Posterior contraction of the population polytope in finite admixture models. *Bernoulli* **21** 618–646. MR3322333
[21] Nguyen, X. (2015). Supplement to "Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure." DOI:10.3150/15-BEJ703SUPP.
[22] Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454
[23] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433
[24] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337
[25] Teh, Y.W. and Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*. *Camb. Ser. Stat. Probab. Math*. 158–207. Cambridge: Cambridge Univ. Press. MR2730663
[26] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480
[27] van der Vaart, A.W. and van Zanten, J.H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663
[28] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer. MR1385671
[29] Villani, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* **338**. Berlin: Springer. MR2459454
[30] Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043. MR2102501
[31] Walker, S.G., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866
[32] Wong, W.H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. MR1332570

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

Received November 2013 and revised October 2014