Electronic Journal of Statistics
Vol. 10 (2016) 608–627
ISSN: 1935-7524
DOI: 10.1214/16-EJS1119

Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures

Fengnan Gao* **and Aad van der Vaart†**

Mathematical Institute
Leiden University
Niels Bohrweg 1
2333 CA Leiden, Netherlands
e-mail: gaof@math.leidenuniv.nl; avdvaart@math.leidenuniv.nl

Abstract: We study nonparametric Bayesian inference with location mixtures of the Laplace density and a Dirichlet process prior on the mixing distribution. We derive a contraction rate of the corresponding posterior distribution, both for the mixing distribution relative to the Wasserstein metric and for the mixed density relative to the Hellinger and L_q metrics.

MSC 2010 subject classifications: Primary 62G20; secondary 62G05.
Keywords and phrases: Bayesian inference, contraction rate, Dirichlet process, minimax rate, Wasserstein metric.

Received July 2015.

1. Introduction

Consider statistical inference using the following hierarchical Bayesian model for observations $X_1, \ldots, X_n$:

(i) A probability distribution G on $\mathbb{R}$ is generated from the Dirichlet process prior $\mathrm{DP}(\alpha)$ with base measure α.
(ii) An i.i.d. sample $Z_1, \ldots, Z_n$ is generated from G.
(iii) An i.i.d. sample $e_1, \ldots, e_n$ is generated from a known density f, independent of the other samples.
(iv) The observations are $X_i = Z_i + e_i$, for $i = 1, \ldots, n$.

In this setting the conditional density of the data $X_1, \ldots, X_n$ given G is a sample from the convolution

$$p_G = f * G$$

of the density f and the measure G. The scheme defines a conditional distribution of G given the data $X_1, \ldots, X_n$, the *posterior distribution* of G, and consequently also posterior distributions for quantities that derive from G, including the convolution density p_G. We are interested in whether this posterior

*Research supported by the Netherlands Organization for Scientific Research.
†The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

distribution can recover a "true" mixing distribution G_0 if the observations $X_1, \ldots, X_n$ are in reality a sample from the mixed distribution p_{G_0}, for some given probability distribution G_0.

The main contribution of this paper is for the case that f is the Laplace density $f(x) = e^{-|x|}/2$. For distributions on the full line Laplace mixtures seem the second most popular class next to mixtures of the normal distribution, with applications in for instance speech recognition or astronomy (Kotz et al. (2001)) and clustering problem in genetics (Bailey et al. (1994)). For the present theoretical investigation the Laplace kernel is interesting as a test case of a non-supersmooth kernel.

We consider two notions of recovery. The first notion measures the distance between the posterior of G and G_0 through the *Wasserstein metric*

$$W_k(G, G') = \inf_{\gamma \in \Gamma(G,G')} \Big(\int |x - y|^k \, d\gamma(x, y) \Big)^{1/k},$$

where $\Gamma(G, G')$ is the collection of all *couplings* γ of G and G' into a bivariate measure with marginals G and G' (i.e. if $(x, y) \sim \gamma$, then $x \sim G$ and $y \sim G'$), and $k \geq 1$. The Wasserstein metric is a classical metric on probability distributions, which is well suited for use in obtaining rates of estimation of measures. It is weaker than the total variation distance (which is more natural as a distance on densities), can be interpreted through transportation of measure (see Villani (2009)), and has also been used in applications such as as comparing the color histograms of digital images. Recovery of the posterior distribution relative to the Wasserstein metric was considered by Nguyen (2013), within a general mixing framework. We refer to this paper for further motivation of the Wasserstein metric for mixtures, and to Villani (2009) for general background on the Wasserstein metric. In the present paper we improve the upper bound on posterior contraction rates given in Nguyen (2013), at least in the case of the Laplace mixtures, obtaining a rate of nearly $n^{-1/8}$ for W_1 (and slower rates for $k > 1$). Apparently the minimax rate of contraction for Laplace mixtures relative to the Wasserstein metric is currently unknown. Recent work on recovery of a mixing distribution by non-Bayesian methods is given in Zhang (1990). It is not clear from our result whether the upper bound $n^{-1/8}$ is sharp.

The second notion of recovery measures the distance of the posterior of G to G_0 indirectly through the Hellinger or L_q-distances between the mixed densities p_G and p_{G_0}. This is equivalent to studying the estimation of the true density p_{G_0} of the observations through the density p_G under the posterior distribution. As the Laplace kernel f has Fourier transform

$$\tilde{f}(\lambda) = \frac{1}{1 + \lambda^2},$$

it follows that the mixed densities p_G have Fourier transforms satisfying

$$|\tilde{p}_G(\lambda)| \leq \frac{1}{1 + \lambda^2}.$$

Estimation of a density with a polynomially decaying Fourier transform was first considered in Watson and Leadbetter (1963). According to their Theorem in Section 3A, a suitable kernel estimator possesses a root mean square error of $n^{-3/8}$ with respect to the L_2-norm for estimating a density with Fourier transform that decays exactly at the order 2. This rate is the "usual rate" $n^{-\alpha/(2\alpha+1)}$ of nonparametric estimation for smoothness $\alpha = 3/2$. This is understandable as $|\tilde{p}(\lambda)| \lesssim 1/(1+|\lambda|^2)$ implies that $\int (1+|\lambda|^2)^\alpha |\tilde{p}(\lambda)|^2\, d\lambda < \infty$, for every $\alpha < 3/2$, so that a density with Fourier transform decaying at square rate belongs to any Sobolev class of regularity $\alpha < 3/2$. Indeed in Golubev (1992), the rate $n^{-\alpha/(2\alpha+1)}$ is shown to be minimax for estimating a density in a Sobolev ball of functions on the line. In the present paper we show that the posterior distribution of Laplace mixtures p_G contracts to p_{G_0} at the rate $n^{-3/8}$ up to a logarithm factor, relative to the L_2-norm and Hellinger distance, and also establish rates for other L_q-metrics. Thus the Dirichlet posterior (nearly) attains the minimax rate for estimating a density in a Sobolev ball of order 3/2. It may be noted that the Laplace density itself is Hölder of exactly order 1, which implies that Laplace mixtures are Hölder smooth of at least the same order. This insight would suggest a rate $n^{-1/3}$ (the usual nonparametric rate for $\alpha = 1$), which is slower than $n^{-3/8}$, and hence this insight is misleading.

Besides recovery relative to the Wasserstein metric and the induced metrics on p_G, one might consider recovery relative to a metric on the distribution function on G. Frequentist recovery rates for this problem were obtained in Fan (1991) under some restrictions. There is no simple relation between these rates and rates for the other metrics. The same is true for the rates for deconvolution of densities, as in Fan (1991). In fact, the Dirichlet prior and posterior considered here are well known to concentrate on discrete distributions, and hence are useless as priors for recovering a density of G.

Contraction rates for Dirichlet mixtures of the normal kernel were considered in Ghosal and Vaart (2001); Ghosal and van der Vaart (2007); Kruijer et al. (2010); Shen et al. (2011); Scricciolo (2011). The results in these papers are driven by the smoothness of the Gaussian kernel, whence the same approach will fail for the Laplace kernel. Nevertheless we borrow the idea of approximating the true mixed density by a finite mixture, albeit that the approximation is constructed in a different manner. Because more support points than in the Gaussian case are needed to obtain a given quality of approximation, higher entropy and lower prior mass concentration result, leading to a slower rate of posterior contraction. To obtain the contraction rate for the Wasserstein metrics we further derive a relationship of these metrics with a power of the Hellinger distance, and next apply a variant of the contraction theorem in Ghosal et al. (2000), which is included in the appendix of the paper. Contraction rates of mixtures with other priors than the Dirichlet were considered in Scricciolo (2011). Recovery of the mixing distribution is a deconvolution problem and as such can be considered an inverse problem. A general approach to posterior contraction rates in inverse problems can be found in Knapik and Salomond (2014), and results specific to deconvolution can be found in Donnet et al. (2014). These authors are interested in deconvolving a (smooth) mixing density rather than a

mixing distribution, and hence their results are not directly comparable to the results in the present paper.

The papers such as Fan (1993); Lepski and Willer (2015) consider recovery of a mixing density relative to the L_p-norm in the frequentist setting. If the smoothness of the mixing density degenerates to 0, then the minimax rate decreases to a constant and it is not possible to find a consistent estimator. In the present paper we show that in the same problem but viewed as a deconvolution problem on distributions, endowed with the weaker Wasserstein distance, we may obtain polynomial rates for the mixing distribution without any smoothness assumption on the distribution. In particular, for any mixing distribution it is possible to construct a consistent estimator.

The paper is organized as follows. In the next section we state the main results of the paper, which are proved in the subsequent sections. In Section 3 we establish suitable finite approximations relative to the L_q- and Hellinger distances. The L_q-approximations also apply to other kernels than the Laplace, and are in terms of the tail decay of the kernel's characteristic function. In Sections 4 and 5 we apply these approximations to obtain bounds on the entropy of the mixtures relative to the L_q, Hellinger and Wasserstein metrics, and a lower bound on the prior mass in a neighbourhood of the true density. Sections 6 and 7 contain the proofs of the main results.

1.1. Notation and preliminaries

Throughout the paper integrals given without limits are considered to be integrals over the real line $\mathbb{R}$. The L_q-norm is denoted

$$\|g\|_q = \left(\int |g(x)|^q \, dx \right)^{1/q},$$

with $\|\cdot\|_\infty$ being the uniform norm. The *Hellinger distance* on the space of densities is given by

$$h(f,g) = \left(\int (f^{1/2}(x) - g^{1/2}(x))^2 \, dx \right)^{1/2}.$$

It is easy to see that $h^2(f,g) \leq \|f-g\|_1 \leq 2h(f,g)$, for any two probability densities f and g. Furthermore, if the densities f and g are uniformly bounded by a constant M, then $\|f-g\|_2 \leq 2\sqrt{M}h(f,g)$. The Kullback-Leiber discrepancy and corresponding variance are denoted by

$$K(p_0,p) = \int \log(p_0/p) \, dP_0, \qquad K_2(p_0,p) = \int (\log(p_0/p))^2 \, dP_0$$

with P_0 the measure corresponding to the density p_0.

We are primarily interested in the Laplace kernel, but a number of results are true for general kernels f. The Fourier transform of a function f and the

inverse Fourier transform of a function $\tilde{f}$ are given by

$$\tilde{f}(\lambda) = \int e^{\imath\lambda x} f(x)dx, \qquad f(x) = \frac{1}{2\pi}\int e^{-\imath\lambda x}\tilde{f}(\lambda)d\lambda.$$

For $\frac{1}{p}+\frac{1}{q}=1$ and $1 \le p \le 2$, *Hausdorff-Young's inequality* gives that $\|f\|_q \le (2\pi)^{-1/p}\|\tilde{f}\|_p$.

The covering number $N(\varepsilon, \Theta, \rho)$ of a metric space (Θ, ρ) is the minimum number of ε-balls needed to cover the entire space Θ.

Throughout the paper $\lesssim$ denotes inequality up to a constant multiple, where the constant is universal or fixed within the context. Furthermore $a_n \asymp b_n$ means $c \le \liminf_{n\to\infty} a_n/b_n \le \limsup_{n\to\infty} a_n/b_n \le C$, for some positive constants c and C.

We denote by $\mathcal{M}[-a,a]$ the set of all probability measures on a given interval $[-a,a]$.

2. Main results

Let $\Pi_n(\cdot|X_1,\ldots,X_n)$ be the posterior distribution for G in the scheme (i)-(iv) introduced at the beginning of the paper. We study this random distribution under the assumption that $X_1,\ldots,X_n$ are an i.i.d. sample from the mixture density $p_{G_0} = f * G_0$, for a given probability distribution G_0. We assume that G_0 is supported in a compact interval $[-a,a]$, and that the base measure α of the Dirichlet prior in (i) is concentrated on this interval with a Lebesgue density bounded away from 0 and ∞.

Theorem 1. *If G_0 is supported on $[-a,a]$ with f being Laplace kernel and α has support $[-a,a]$ with Lebesgue density bounded away from 0 and ∞, then for every $k \ge 1$, there exists a constant M such that*

$$\Pi\big(G : W_k(G,G_0) \ge Mn^{-3/(8k+16)}(\log n)^{(k+7/8)/(k+2)}|X_1,\ldots,X_n\big) \to 0, \quad (2.1)$$

in P_{G_0}-probability.

The rate for the Wasserstein metric W_k given in the theorem deteriorates with increasing k, which is perhaps not unreasonable as the Wasserstein metrics increase with k. The fastest rate is $n^{-1/8}(\log n)^{5/8}$, and is obtained for W_1.

Theorem 2. *If G_0 is supported on $[-a,a]$ with f being Laplace kernel and α has support $[-a,a]$ with Lebesgue density bounded away from 0 and ∞, then there exists a constant M such that*

$$\Pi_n\big(G : h(p_G, p_{G_0}) \ge M(\log n/n)^{3/8}|X_1,\ldots,X_n\big) \to 0, \quad (2.2)$$

in P_{G_0}-probability. Furthermore, for every $q \in [2,\infty)$ there exists M_q such that

$$\Pi_n\big(G : \|p_G - p_{G_0}\|_q \ge M_q(\log n/n)^{(q+1)/(q(q+2))}|X_1,\ldots,X_n\big) \to 0, \quad (2.3)$$

in P_{G_0}-probability.

The rate for the L_q-distance given in (2.3) deteriorates with increasing q. For $q = 2$ it is the same as the rate $(\log n/n)^{3/8}$ for the Hellinger distance.

In both theorems the mixing distributions are assumed to be supported on a fixed compact set. Without a restriction on the tails of the mixing distributions, no rate is possible. The assumption of a compact support ensures that the rate is fully determined by the complexity of the mixtures, and not their tail behaviour.

3. Finite approximation

In this section we show that a general mixture p_G can be approximated by a mixture with finitely many components, where the number of components depends on the accuracy of the approximation, the distance used, and the kernel f. We first consider approximation with respect to the L_q-norm, which applies to mixtures $p_G = f * G$, for a general kernel f, and next approximation with respect to the Hellinger distance for the case that f is the Laplace kernel. The first result generalizes the result of Ghosal and Vaart (2001) for normal mixtures. Also see Scricciolo (2011).

The result splits in two cases, depending on the tail behaviour of the Fourier transform $\tilde{f}$ of f:

-*ordinary smooth* f : $\limsup_{|\lambda|\to\infty} |\tilde{f}(\lambda)||\lambda|^{\beta} < \infty$, for some $\beta > 1/2$.

-*supersmooth* f : $\limsup_{|\lambda|\to\infty} |\tilde{f}(\lambda)|e^{|\lambda|^{\beta}} < \infty$, for some $\beta > 0$.

Lemma 1. *Let $\varepsilon < 1$ be sufficiently small and fixed. For a probability measure G on an interval $[-a, a]$ and $2 \le q \le \infty$, there exists a discrete measure G' on $[-a, a]$ with at most N support points in $[-a, a]$ such that*

$$\|p_G - p_{G'}\|_q \lesssim \varepsilon,$$

where

(i) $N \lesssim \varepsilon^{-(\beta - p^{-1})^{-1}}$ if f is ordinary smooth of order β, for p and q being conjugate ($p^{-1} + q^{-1} = 1$).

(ii) $N \lesssim (\log \varepsilon^{-1})^{\max(1,\beta^{-1})}$ if f is supersmooth of order β.

Proof. The Fourier transform of p_G is given by $\tilde{f}\tilde{G}$, for $\tilde{G}(\lambda) = \int e^{\imath\lambda z}\, dG(z)$. Determine G' so that it possesses the same moments as G up to order $k-1$, i.e.

$$\int z^j d(G - G')(z) = 0, \quad \forall\ 0 \le j \le k-1.$$

By Lemma A.1 in Ghosal and Vaart (2001) G' can be chosen to have at most k support points.

Then for G and G' supported on $[-a, a]$, we have

$$|\tilde{G}(\lambda) - \tilde{G}'(\lambda)| = \left| \int \left(e^{\imath\lambda z} - \sum_{j=0}^{k-1} \frac{(\imath\lambda z)^j}{j!} \right) d(G - G')(z) \right|$$

$$\leq \int \frac{|\imath\lambda z|^k}{k!} d(G+G')(z) \leq \Big(\frac{ae|\lambda|}{k}\Big)^k.$$

The inequality comes from $|e^{iy} - \sum_{j=0}^{k-1} (iy)^j/j!| \leq |y|^k/k! \leq (e|y|)^k/k^k$, for every $y \in \mathbb{R}$.

Therefore, by Hausdorff-Young's inequality,

$$\begin{aligned}\|p_G - p_{G'}\|_q^p &\leq \frac{1}{2\pi} \int |\tilde f(\lambda)|^p |\tilde G(\lambda) - \tilde G'(\lambda)|^p d\lambda \\ &\lesssim \int_{|\lambda|>M} |\tilde f(\lambda)|^p d\lambda + \int_{|\lambda|\leq M} \Big(\frac{ea|\lambda|}{k}\Big)^{pk} d\lambda.\end{aligned}$$

We denote the first term in the preceding display by I_1 and the second term by I_2. It is easy to bound I_2 as:

$$I_2 \asymp \Big(\frac{ea}{k}\Big)^{kp} \frac{M^{kp+1}}{kp+1} \lesssim \Big(\frac{eaM}{k}\Big)^{kp+1} \frac{1}{p}.$$

For I_1 we separately consider the supersmooth and ordinary smooth cases.

In the supersmooth case with parameter β, we note that the function $(t^{\beta^{-1}-1})/e^{\delta t}$ is monotonely decreasing for $t \geq pM^\beta$, when $\delta \geq (\beta^{-1}-1)/(pM^\beta)$. Thus, for large M,

$$\begin{aligned}I_1 &\lesssim \int_{|\lambda|>M} e^{-p|\lambda|^\beta} d\lambda = \frac{2}{\beta p^{\beta^{-1}}} \int_{t>pM^\beta} e^{-t} t^{\beta^{-1}-1} dt \\ &\leq \frac{2}{\beta p^{\beta^{-1}}} \int_{t>pM^\beta} e^{-(1-\delta)t}\, dt \frac{(pM^\beta)^{\beta^{-1}-1}}{e^{\delta pM^\beta}} = \frac{2}{1-\delta} \frac{1}{\beta p} e^{-pM^\beta} M^{1-\beta},\end{aligned}$$

where the bound is sharper if δ is smaller. Choosing the minimal value of δ, we obtain

$$I_1 \lesssim \frac{1}{1-(\beta^{-1}-1)/(pM^\beta)} \frac{1}{\beta p} e^{-pM^\beta} M^{1-\beta} \lesssim M^{1-\beta} e^{-pM^\beta},$$

for M sufficiently large. We next choose $M = 2\,(\log(1/\varepsilon))^{\frac{1}{\beta}}$ in order to ensure that $I_1 \leq \varepsilon^p$. Then $I_2 \lesssim \varepsilon^p$ if $k \geq 2eaM$ and $2^{-kp} \leq \varepsilon^p$. This is satisfied if $k = 2(\log \varepsilon^{-1})^{\max(\beta^{-1},1)}$.

In the ordinary smooth case with smoothness parameter β, we have the bound

$$I_1 \lesssim \int_{\lambda>M} |\lambda|^{-\beta p} d\lambda \lesssim \left(\frac{1}{M}\right)^{\beta p-1}.$$

We choose $M = (1/\varepsilon)^{-(\beta-1/p)^{-1}}$ to render the right side equal to ε^p. Then $I_2 \lesssim \varepsilon^p$ if $k = 2\varepsilon^{-(\beta-1/p)^{-1}}$. □

The number of support points in the preceding lemma is increasing in q and decreasing in β. For approximation in the L_2-norm ($q = 2$), the number of

support points is of order $\varepsilon^{-1/(\beta-1/2)}$, and this reduces to $\varepsilon^{-2/3}$ for the Laplace kernel (ordinary smooth with $\beta = 2$). The exponent $\beta - 1/2$ can be interpreted as (almost) the Sobolev smoothness of p_G, since, for $\alpha < \beta - 1/2$,

$$\int (1+|\lambda|^2)^\alpha |\tilde{p}_G(\lambda)|^2 d\lambda \lesssim \int (1+|\lambda|^2)^\alpha |\tilde{f}(\lambda)|^2 d\lambda < \infty.$$

We do not have a compelling intuition for this correspondence.

The Hellinger distance is more sensitive to areas where the densities are close to zero. This causes that the approach in the preceding lemma does not give sharp results. The following lemma does, but is restricted to the Laplace kernel.

Lemma 2. *For a probability measure G supported on $[-a,a]$ there exists a discrete measure G' with at most $N \asymp \varepsilon^{-2/3}$ support points such that for $p_G = f * G$ and f the Laplace density*

$$h(p_G, p_{G'}) \le \varepsilon.$$

Proof. Since $p_G(x) \ge f(|x|+a) = e^{-a}e^{-|x|}/2$, for every x and probability measure G supported on $[-a,a]$, the Hellinger distance between Laplace mixtures satisfies

$$h^2(p_G, p_{G'}) \le \int \frac{(p_G - p_{G'})^2}{p_G + p_{G'}}(x)\, dx \le e^a \int (p_{G'}(x) - p_G(x))^2 e^{|x|}\, dx.$$

If we write $q_G(x) = p_G(x)e^{|x|/2}$, and $\tilde{q}_G$ for the corresponding Fourier transform, then the integral in the right side is equal to $(1/2\pi)\int |\tilde{q}_{G'} - \tilde{q}_G|^2(\lambda)\, d\lambda$, by Plancherel's theorem. By an explicit computation we obtain

$$\tilde{q}_G(\lambda) = \frac{1}{2}\int\int e^{\imath\lambda x} e^{-|x-z|+|x|/2}\, dx\, dG(z) = \frac{1}{2}\int r(\lambda, z)\, dG(z),$$

where $r(\lambda, z)$ is given by

$$\begin{aligned} r(\lambda, z) &= \frac{e^{-z}}{\imath\lambda + 1/2} + e^{-z}\frac{e^{(\imath\lambda+3/2)z} - 1}{\imath\lambda + 3/2} - \frac{e^{(\imath\lambda+1/2)z}}{\imath\lambda - 1/2} \\ &= \frac{e^{-z}}{(\imath\lambda+1/2)(\imath\lambda+3/2)} - \frac{2e^{\imath\lambda z}e^{z/2}}{(\imath\lambda+3/2)(\imath\lambda-1/2)}. \end{aligned} \tag{3.1}$$

Now let G' be a discrete measure on $[-a,a]$ such that

$$\int e^{-z}\, d(G'-G)(z) = 0, \qquad \int e^{z/2} z^j\, d(G'-G)(z) = 0, \quad \forall\, 0 \le j \le k-1.$$

By Lemma A.1 in Ghosal and Vaart (2001) G' can be chosen to have at most $k+1$ support points.

By the choice of G' the first term of $r(\lambda, z)$ gives no contribution to the difference $\int r(\lambda, z)\, d(G'-G)(z)$. As the second term of $r(\lambda, z)$ is for large $|\lambda|$ bounded in absolute value by a multiple of $|\lambda|^{-2}$, it follows that

$$I_2 := \int_{|\lambda|>M} \left| \int r(\lambda, z)\, d(G'-G)(z) \right|^2 d\lambda \lesssim \int_{\lambda>M} \lambda^{-4} d\lambda \asymp M^{-3}.$$

By the choice of G' in the second term of $r(\lambda, z)$ we can replace $e^{\imath\lambda z}$ by $e^{\imath\lambda z} - \sum_{j=0}^{k}(\imath\lambda z)^j/j!$ again without changing the integral $\int r(\lambda, z)\, d(G' - G)(z)$. It follows that

$$\begin{aligned} I_1 &:= \int_{|\lambda|\le M} \left| \int r(\lambda, z)\, d(G' - G)(z) \right|^2 d\lambda \\ &\le \int_{|\lambda|\le M} \left| \frac{2}{(\imath\lambda + 1/2)(\imath\lambda + 3/2)} \right|^2 \left| \int e^{z/2} \Big[e^{\imath\lambda z} - \sum_{j=0}^{k} (\imath\lambda z)^j \Big]\, d(G' - G)(z) \right|^2 d\lambda \\ &\lesssim \int_0^M \frac{(z\lambda)^{2k}}{(k!)^2}\, d\lambda \lesssim \frac{(aeM)^{2k+1}}{k^{2k+1}}. \end{aligned}$$

It follows, by a similar argument as in the proof of Lemma 1, that we can reduce both I_1 and I_2 to ε^2 by choosing and $M \asymp \varepsilon^{-2/3}$ and $k = 2aeM$. □

4. Entropy

We study the covering numbers of the class of mixtures $p_G = f * G$, where G ranges over the collection $\mathcal{M}[-a, a]$ of all probability measures on $[-a, a]$. We present a bound for any L_q-norm and general kernels f, and a bound for the Hellinger distance that is specific to the Laplace kernel.

Proposition 1. *If both $\|f\|_q$ and $\|f'\|_q$ are finite and $\tilde f$ has ordinary smoothness β, then, for $p_G = f * G$, and any $q \ge 2$,*

$$\log N\big(\varepsilon, \{p_G : G \in \mathcal{M}[-a, a]\}, \|\cdot\|_q\big) \lesssim \left(\frac{1}{\varepsilon}\right)^{\frac{1}{\beta - 1 + 1/q}} \log\left(\frac{1}{\varepsilon}\right). \tag{4.1}$$

Proof. Consider an ε-net of $\mathcal{P}_a = \{p_G : G \in \mathcal{M}[-a, a]\}$ by constructing $\mathcal{I}$ the collection of all p_G's such that the mixing measure $G \in \mathcal{M}[-a, a]$ is discrete and has at most $N \le D\varepsilon^{-(\beta - 1 + q^{-1})^{-1}}$ support points for some proper constant D.

In light of the approximation Lemma 1, the set of all mixtures p_G with G a discrete probability measure with $N \lesssim \varepsilon^{-(\beta - 1 + q^{-1})^{-1}}$ support points forms an ε-net over the set of all mixtures p_G as in the lemma. It suffices to construct an ε-net of the given cardinality over this set of discrete mixtures.

By Jensen's inequality and Fubini's theorem,

$$\|f(\cdot - \theta) - f\|_q = \left(\int \left| \theta \int_0^1 f'(x - \theta s)\, ds \right|^q dx \right)^{1/q} \le \|f'\|_q \theta.$$

Furthermore, for any probability vectors p and p' and locations θ_i,

$$\left\| \sum_{i=1}^N p_i f(\cdot - \theta_i) - \sum_{i=1}^N p_i' f(\cdot - \theta_i) \right\|_q \le \sum_{i=1}^N |p_i - p_i'| \|f(\cdot - \theta_i)\|_q = \|f\|_q \|p - p'\|_1.$$

Combining these inequalities, we see that for two discrete probability measures $G = \sum_{i=1}^N p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^N p_i' \delta_{\theta_i'}$,

$$\|p_G - p_{G'}\|_q \leq \|f'\|_q \max_i |\theta_i - \theta_i'| + \|f\|_q \|p - p'\|_1. \tag{4.2}$$

Thus we can construct an ε-net over the discrete mixtures by relocating the support points $(\theta_i)_{i=1}^N$ to the nearest points $(\theta_i')_{i=1}^N$ in a ε-net on $[-a, a]$, and relocating the weights p to the nearest point p' in an ε-net for the l_1-norm over the N-dimensional l_1-unit simplex. This gives a set of at most

$$\left(\frac{2a}{\varepsilon}\right)^N \left(\frac{5}{\varepsilon}\right)^N \sim \left(\frac{10a}{\varepsilon^2}\right)^N$$

measures p_G (cf. Lemma A.4 of Ghosal and van der Vaart (2007) for the entropy of the l_1-unit simplex). This gives the bound of the lemma. □

Proposition 2. *For f the Laplace kernel and $p_G = f * G$,*

$$\log N\big(\varepsilon, \{p_G : G \in \mathcal{M}[-a, a]\}, h\big) \lesssim \varepsilon^{-3/8} \log(1/\varepsilon). \tag{4.3}$$

Proof. Because the function $\sqrt{f}$ is absolutely continuous with derivative $x \mapsto -2^{-3/2} e^{-|x|/2} \operatorname{sgn}(x)$, we have by Jensen's inequality and Fubini's theorem that

$$\begin{aligned} h^2\big(f, f(\cdot - \theta)\big) &= \int \Big(\theta \int_0^1 -2^{-3/2} e^{-|x-\theta s|/2} \operatorname{sgn}(x - \theta s)\, ds\Big)^2 dx \\ &\leq \theta^2 \int_0^1 \int e^{-|x-\theta s|}\, dx\, ds = 2\theta^2. \end{aligned}$$

It follows that $h\big(f, f(\cdot - \theta)\big) \lesssim \theta$.

By convexity of the map $(u, v) \mapsto (\sqrt{u} - \sqrt{v})^2$, we have

$$\Big|\sqrt{\sum_i p_i f(\cdot - \theta_i)} - \sqrt{\sum_i p_i f(\cdot - \theta_i')}\Big|^2 \leq \sum_i p_i \big[\sqrt{f(\cdot - \theta_i)} - \sqrt{f(\cdot - \theta_i')}\big]^2.$$

By integrating this inequality we see that the densities p_G and $p_{G'}$ with mixing distributions $G = \sum_{i=1}^N p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^N p_i \delta_{\theta_i'}$ satisfy $h^2(p_G, p_{G'}) \lesssim \sum p_i |\theta_i - \theta_i'|^2 \leq \|\theta - \theta'\|_\infty^2$.

Furthermore, for distributions $G = \sum_{i=1}^N p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^N p_i' \delta_{\theta_i}$ with the same support points, but different weights, we have

$$\begin{aligned} h^2(p_G, p_{G'}) &\leq \int \frac{\big(\sum_{i=1}^N (p_i - p_i') f(x - \theta_i)\big)^2}{\sum_{i=1}^N (p_i + p_i') f(x - \theta_i)} dx \\ &\leq \int \Big(\sum_{i=1}^N |p_i - p_i'|\Big)^2 \frac{f^2(|x| - a)}{2 f(|x| + a)} dx \lesssim \|p - p'\|_1^2. \end{aligned}$$

Therefore the bound follows by arguments similar as in the proof of Proposition 1, where presently we use Lemma 2 to determine suitable finite approximations. □

The map $G \mapsto p_G = f * G$ is one-to-one as soon as the characteristic function of f is never zero. Under this condition we can also view the Wasserstein distance on the mixing distribution as a distance on the mixtures. Obviously the covering numbers are then free of the kernel.

Proposition 3. *For any $k \geq 1$, and any sufficiently small $\varepsilon > 0$,*

$$\log N\big(\varepsilon, \mathcal{M}[-a,a], W_k\big) \lesssim \Big(\frac{1}{\varepsilon}\Big) \log(1/\varepsilon). \tag{4.4}$$

The proposition is a consequence Lemma 4, below, which applies to the set of all Borel probability measures on a general metric space (Θ, ρ) (cf. Nguyen (2013)).

Lemma 3. *For any probability measure G concentrated on countably many disjoint sets $\Theta_1, \Theta_2, \ldots$ and probability measure G' concentrated on disjoint sets $\Theta_1', \Theta_2', \ldots$,*

$$W_k(G, G') \leq \sup_i \sup_{\theta_i \in \Theta_i, \theta_i' \in \Theta_i'} \rho(\theta_i, \theta_i') + \operatorname{diam}(\Theta) \Big(\sum_i |G(\Theta_i) - G'(\Theta_i')| \Big)^{1/k}.$$

In particular,

$$W_k\Big(\sum_i p_i \delta_{\theta_i}, \sum_i p_i' \delta_{\theta_i'} \Big) \leq \max_i \rho(\theta_i, \theta_i') + \operatorname{diam}(\Theta) \|p - p'\|_1^{1/k}.$$

Proof. For $p_i = G(\Theta_i)$ and $p_i' = G'(\Theta_i')$ divide the interval $[0, \sum_i p_i \wedge p_i']$ into disjoint intervals I_i of lengths $p_i \wedge p_i'$. We couple variables $\bar\theta$ and $\bar\theta'$ by an auxiliary uniform variable U. If $U \in I_i$, then generate $\bar\theta \sim G(\cdot|\Theta_i)$ and $\bar\theta' \sim G'(\cdot|\Theta_i')$. Divide the remaining interval $[\sum_i p_i \wedge p_i', 1]$ into intervals J_i of lengths $p_i - p_i \wedge p_i'$ and, separately, intervals J_i' of length $p_i' - p_i \wedge p_i'$. If $U \in J_i$, then generate $\bar\theta \sim G(\cdot|\Theta_i)$ and if $U \in J_i'$, then generate $\bar\theta' \sim G'(\cdot|\Theta_i')$. Then $\bar\theta$ and $\bar\theta'$ have marginal distributions G and G', and

$$\mathbb{E}\rho^k(\bar\theta, \bar\theta') \leq \mathbb{E}\big[\rho^k(\bar\theta, \bar\theta') 1_{U \leq \sum_i p_i \wedge p_i'}\big] + \operatorname{diam}(\Theta)^k \mathbb{P}\big(U > \sum_i p_i \wedge p_i'\big).$$

The first term is bounded by the k-th power of the first term of the lemma, while the probability in the second term is equal to $1 - \sum_i p_i \wedge p_i' = \sum_i |p_i - p_i'|/2$. □

Lemma 4. *For the set $\mathcal{M}(\Theta)$ of all Borel probability measures on a metric space (Θ, ρ), any $k \geq 1$, and $0 < \varepsilon < \min\{2/3, \operatorname{diam}(\Theta)\}$,*

$$N\big(\varepsilon, \mathcal{M}(\Theta), W_k\big) \leq \Big(\frac{4 \operatorname{diam}(\Theta)}{\varepsilon}\Big)^{kN(\varepsilon, \Theta, \rho)}.$$

Proof. For a minimal ε-net over Θ of $N = N(\varepsilon, \Theta, \rho)$ points, let $\Theta = \cup_i \Theta_i$ be the partition obtained by assigning each θ to a closest point. For any G let $G_\varepsilon = \sum_i G(\Theta_i) \delta_{\theta_i}$, for arbitrary but fixed $\theta_i \in \Theta_i$. Since $W_k(G, G_\varepsilon) \leq \varepsilon$

by Lemma 3, we have $N(2\varepsilon, \mathcal{M}(\Theta), W_k) \leq N(\varepsilon, \mathcal{M}_\varepsilon, W_k)$, for $\mathcal{M}_\varepsilon$ the set of all G_ε. We next form the measures $G_{\varepsilon,p} = \sum_i p_i \delta_{\theta_i}$ for $(p_1, \ldots, p_N)$ ranging over an $(\varepsilon/\operatorname{diam}(\Theta))^k$-net for the l_1-distance over the N-dimensional unit simplex. By Lemma 3 every G_ε is within W_k-distance of some $G_{\varepsilon,p}$. Thus $N(\varepsilon, \mathcal{M}_\varepsilon, W_k)$ is bounded from above by the number of points p, which is bounded by $(4\operatorname{diam}(\Theta)/\varepsilon)^{kN}$ (cf. Lemma A.4 in Ghosal et al. (2000)). □

5. Prior mass

This main result of this section is the following proposition, which gives a lower bound on the prior mass of the prior (i)-(iv) in a neighbourhood of a mixture p_{G_0}.

Proposition 4. *If* Π *is the Dirichlet process* $\mathrm{DP}(\alpha)$ *with base measure* α *that has a Lebesgue density bounded away from 0 and* ∞ *on its support* $[-a, a]$*, and* f *is the Laplace kernel, then for every sufficiently small* $\varepsilon > 0$ *and every probability measure* G_0 *on* $[-a, a]$*,*

$$\log \Pi\Big(G : K(p_G, p_{G_0}) \leq \varepsilon^2, K_2(p_G, p_{G_0}) \leq \varepsilon^2\Big) \gtrsim \left(\frac{1}{\varepsilon}\right)^{2/3} \log\left(\frac{1}{\varepsilon}\right).$$

Proof. By Lemma 2 there exists a discrete measure G_1 with $N \lesssim \varepsilon^{-2/3}$ support points such that $h(p_{G_0}, p_{G_1}) \leq \varepsilon$. We may assume that the support points of G_1 are at least $2\varepsilon^2$-separated. If not, we take a maximal $2\varepsilon^2$-separated set in the support points of G_1, and replace G_1 by the discrete measure obtained by relocating the masses of G_1 to the nearest points in the $2\varepsilon^2$-net. Then $h(p_{G_1}, p_{G_1'}) \lesssim \varepsilon^2$, as seen in the proof of Proposition 2.

Now by Lemmas 6 and 5, if $G_1 = \sum_{i=1}^N p_j \delta_{z_j}$, with the support points z_j at least $2\varepsilon^2$-separated,

$$\begin{aligned}
\big\{G : \max(K, K_2)(p_{G_0}, p_G) < d_1\varepsilon^2\big\} &\supset \big\{G : h(p_{G_0}, p_G) \leq 2\varepsilon\big\} \\
&\supset \big\{G : h(p_{G_1}, p_G) \leq \varepsilon\big\} \\
&\supset \big\{G : \|p_G - p_{G_1}\|_1 \leq d_2\varepsilon^2\big\} \\
&\supset \Big\{G : \sum_{j=1}^N \big|G[z_j - \varepsilon^2, z_j + \varepsilon^2] - p_j\big| \leq \varepsilon^2\Big\}.
\end{aligned}$$

By Lemma A.2 of Ghosal and Vaart (2001), since the base measure α has density bounded away from zero and infinity on $[-a, a]$ by assumption, we have

$$\log \Pi\left(G : \sum_{j=1}^N \big|G[z_j - \varepsilon^2, z_j + \varepsilon^2] - p_j\big| \leq \varepsilon^2\right) \gtrsim -N \log\left(\frac{1}{\varepsilon}\right)$$

The lemma follows upon combining the preceding. □

Lemma 5. *If $G' = \sum_{j=1}^N p_i \delta_{z_j}$ is a probability measure supported on points $z_1, \ldots, z_N$ in $\mathbb{R}$ with $|z_j - z_k| > 2\varepsilon$ for $j \neq k$, then for any probability measure G on $\mathbb{R}$ and kernel f,*

$$\|p_G - p_{G'}\|_1 \leq 2\|f'\|_1 \varepsilon + 2\sum_{j=1}^N \big|G[z_j - \varepsilon, z_j + \varepsilon] - p_j\big|.$$

Lemma 6. *If G and G' are probability measures on $[-a, a]$, and f is the Laplace kernel, then*

$$h^2(p_G, p_{G'}) \lesssim \|p_G - p_{G'}\|_2, \tag{5.1}$$

$$\max\big(K(p_G, p_{G'}), K_2(p_G, p_{G'})\big) \lesssim h^2(p_G, p_{G'}). \tag{5.2}$$

Proofs. The first lemma is a generalization of Lemma 4 in Ghosal and van der Vaart (2007) from normal to general kernels, and is proved in the same manner.

In view of the shape of the Laplace kernel, it is easy to see that for G compactly supported on $[-a, a]$,

$$f(|x| + a) \leq p_G(x) \leq f(|x| - a),$$

We bound the squared Hellinger distance as follows:

$$\begin{aligned} h^2(p_G, p_{G'}) &\leq \int \frac{(p_G - p_{G'})^2}{p_G + p_{G'}} dx \\ &\leq \int_{|x| \leq A} e^{A+a}(p_G - p_{G'})^2 dx + \int_{|x|>A} (p_G + p_{G'}) dx \\ &\lesssim e^a \|p_G - p_{G'}\|_2^2 e^A + e^{-A}. \end{aligned}$$

By the elementary inequality $t + \frac{u}{t} \geq 2\sqrt{u}$, for $u, t > 0$, we obtain (5.1) upon choosing $A = \min(a, \log \|p_G - p_{G'}\|_2^{-1} - a/2)$.

For the proof of the second assertion we first note that, if both G and G' are compactly supported on $[-a, a]$,

$$\frac{p_G(x)}{p_{G'}(x)} \leq \frac{f(|x| - a)}{f(|x| + a)} \leq e^{2a}.$$

Therefore $\|p_G / p_{G'}\|_\infty \leq e^{2a}$, and (5.2) follows by Lemma 8 in Ghosal and van der Vaart (2007). □

6. Proof of Theorem 1

The proof is based on the following comparison between the Wasserstein and Hellinger metrics. The lemma improves and generalizes Theorem 2 in Nguyen (2013). Let C_k be a constant such that the map $\varepsilon \mapsto \varepsilon[\log(C_k/\varepsilon)]^{k+1/2}$ is monotone on $(0, 2]$.

Lemma 7. *For probability measures G and G' supported on $[-a,a]$, and $p_G = f * G$ for a probability density f with $\inf_\lambda(1+|\lambda|^\beta)|\tilde f(\lambda)| > 0$, and any $k \geq 1$,*

$$W_k(G,G') \lesssim h(p_G,p_{G'})^{1/(k+\beta)}\Big(\log\frac{C_k}{h(p_G,p_{G'})}\Big)^{(k+1/2)/(k+\beta)}.$$

Proof. By Theorem 6.15 in Villani (2009) the Wasserstein distance $W_k(G,G')$ is bounded above by a multiple of the kth root of $\int |x|^k\, d|G-G'|(x)$, where $|G-G'|$ is the total variation measure of the difference $G-G'$. We apply this to the convolutions of G and G' with the normal distribution Φ_δ with mean 0 and variance δ^2, to find, for every $M>0$,

$$\begin{aligned} W_k(G*\Phi_\delta, G'*\Phi_\delta)^k &\lesssim \int |x|^k \big|(G-G')*\phi_\delta(x)\big|\,dx \\ &\leq \Big(\int_{-M}^M x^{2k}\,dx \int_{-M}^M \big|(G-G')*\phi_\delta(x)\big|^2\,dx\Big)^{1/2} \\ &\qquad + e^{-M}\int_{|x|>M} |x|^k e^{|x|}\big|(G-G')*\phi_\delta(x)\big|\,dx \\ &\lesssim M^{k+1/2}\big\|(G-G')*\phi_\delta\big\|_2 + e^{-M}e^{2|a|}\mathrm{E}e^{2|\delta Z|}, \end{aligned}$$

where Z is a standard normal variable. The number $K_\delta := e^{2|a|}\mathrm{E}e^{2|\delta Z|}$ is uniformly bounded by if $\delta \leq \delta_k$, for some fixed δ_k.

By Plancherel's theorem,

$$\begin{aligned} \big\|(G-G')*\phi_\delta\big\|_2^2 &= \int |\tilde G - \tilde G'|^2(\lambda)\tilde\phi_\delta^2(\lambda)\,d\lambda = \int |\tilde f(\tilde G-\tilde G')|^2(\lambda)\frac{\tilde\phi_\delta^2}{|\tilde f|^2}(\lambda)\,d\lambda \\ &\lesssim \|p_G - p_{G'}\|_2^2 \sup_\lambda \frac{\tilde\phi_\delta^2}{|\tilde f|^2}(\lambda) \lesssim h^2(p_G,p_{G'})\delta^{-2\beta}, \end{aligned}$$

where we have again applied Plancherel's theorem, used that the L_2-metric on uniformly bounded densities is bounded by the Hellinger distance, and the assumption on the Fourier transform of f, which shows that $(\tilde\phi_\delta/|\tilde f|)(\lambda) \lesssim (1+|\lambda|^\beta)e^{-\delta^2\lambda^2/2} \lesssim \delta^{-\beta}$.

If $U \sim G$ is independent of $Z \sim N(0,1)$, then $(U, U+\delta Z)$ gives a coupling of G and $G*\Phi_\delta$. Therefore the definition of the Wasserstein metric gives that $W_k(G, G*\Phi_\delta)^k \leq \mathrm{E}|\delta Z|^k \lesssim \delta^k$.

Combining the preceding inequalities with the triangle inequality we see that, for $\delta \in (0,\delta_k]$ and any $M>0$,

$$W_k(G,G')^k \lesssim M^{k+1/2}h(p_G,p_{G'})\delta^{-\beta} + e^{-M} + \delta^k.$$

The lemma follows by optimizing this over M and δ. Specifically, for $\varepsilon = h(p_G,p_{G'})$, we choose $M = k/(k+\beta)\,\log(C_k/\varepsilon)$ and $\delta = (M^{k+1/2}\varepsilon)^{1/(k+\beta)}$. These are eligible choices for

$$\delta_k = \sup_{\varepsilon\in(0,2]}\Big[\frac{k}{k+\beta}\log\frac{C_k}{\varepsilon}\Big]^{(k+1/2)/(k+\beta)}\varepsilon^{1/(k+\beta)},$$

which is indeed a finite number. In fact the supremum is taken at $\varepsilon = 2$, by the assumption on C_k. □

For the Laplace kernel f we choose $\beta = 2$ in the preceding lemma, and then obtain that $d(p_G, p_{G'}) \le h(p_G, p_{G'})$, for the "discrepancy" $d = \gamma^{-1}(W_k)$, and $\gamma(\varepsilon) = D_k \varepsilon^{1/(k+\beta)} [\log(C_k/\varepsilon)]^{(k+1/2)/(k+\beta)}$ a multiple of the (monotone) transformation in the right side of the preceding lemma. For small values of $W_k(G_1, G_2)$ we have

$$d(p_{G_1}, p_{G_2}) \asymp W_k^{k+2}(G_1, G_2) \Big(\log \frac{1}{W_k(G_1, G_2)} \Big)^{-k-1/2}. \tag{6.1}$$

As $k + 2 > 1$ the discrepancy d may not satisfy the triangle inequality, but it does possess the properties (a)–(d) in the appendix, Section 9. The balls of the discrepancy d are convex, as the Wasserstein metrics are convex (see Villani (2009)).

It follows that Theorem 3 applies to obtain a rate of posterior contraction relative to d and hence relative to $W_k \sim d^{1/(k+2)} \big(\log(1/d)\big)^{(k+1/2)/(k+2)}$. We apply the theorem with $\mathcal{P} = \mathcal{P}_n$ equal to the set of mixtures $p_G = f * G$, as G ranges over $\mathcal{M}[-a, a]$. Thus (9.3) is trivially satisfied.

For the entropy condition (9.1) we have, by Proposition 3,

$$\begin{aligned} \log N(\varepsilon, \mathcal{P}_n, d) &= \log N\Big(\varepsilon^{1/(k+2)} \Big(\log \frac{1}{\varepsilon} \Big)^{(k+1/2)/(k+2)}, \mathcal{M}[-a, a], W_k \Big) \\ &\lesssim \Big(\frac{1}{\varepsilon} \Big)^{1/(k+2)} \Big(\log \frac{1}{\varepsilon} \Big)^{1+(k+1/2)/(k+2)}. \end{aligned}$$

Thus (9.1) holds for the rate $\varepsilon_n \gtrsim n^{-\gamma}$, for every $\gamma < (k+2)/(2k+5)$.

The prior mass condition (9.2) is satisfied with the rate $\varepsilon_n \asymp (\log n/n)^{3/8}$, in view of Proposition 4.

Theorem 3 yields a rate of contraction relative to d equal to the slower of the two rates, which is $(\log n/n)^{3/8}$. This translates into the rate for the Wasserstein distance as given in Theorem 1.

7. Proof of Theorem 2

We apply Theorem 3, with $\mathcal{P} = \mathcal{P}_n$ the set of all mixtures p_G as G ranges over $\mathcal{M}[-a, a]$. For $d = h$ the rate follows immediately by combining Propositions 1 and 4.

Since the densities p_G are uniformly bounded by $1/2$, the L_q distance $\|p_G - p_{G'}\|_q$ is bounded above by a multiple of $h(p_G, p_{G'})^{2/q}$. We can therefore apply Theorem 3 with the discrepancy $d(p, p') = \|p - p'\|_q^{q/2}$. In view of Proposition 1

$$\log N\big(\varepsilon, \mathcal{P}_n, d\big) \lesssim \varepsilon^{-2/(q+1)} \log(1/\varepsilon).$$

Therefore the entropy condition (9.1) is satisfied with $\varepsilon_n \asymp (\log n/n)^{(q+1)/(2q+4)}$. By Proposition 4 the prior mass condition is satisfied for $\varepsilon_n \asymp (\log n/n)^{3/8}$. By

Theorem 3 the rate of contraction relative to d is the slower of these two rates, which is the first. The rate relative to the L_q-norm is the $(2/q)$th power of this rate.

8. Normal mixtures

We reproduce the results on normal mixtures from Ghosal and Vaart (2001), but in L_2-norm. Note the normal kernel is supersmooth with $\beta = 2$, by the approximation lemma, for any measure G_1 compactly supported on $[-a, a]$ we can always find a discrete measure G_2 with number of support points of order $N \asymp \log \varepsilon^{-1}$ such that $\|p_{G_1} - p_{G_2}\|_2 \leq \varepsilon$. It is easy to establish

$$h^2(p_{G_1}, p_{G_2}) \lesssim \|p_{G_1} - p_{G_2}\|_2.$$

Following the same procedure as before, assuming G_0 is the true measure, we obtain for prior mass condition

$$\log \Pi\left(G : \max\left(P_{G_0} \log \frac{p_{G_0}}{p_G}, P_{G_0}\left(\log \frac{p_{G_0}}{p_G}\right)^2\right) \leq \varepsilon^2\right) \gtrsim -\left(\log \frac{1}{\varepsilon}\right)^2,$$

Thus we obtain $\varepsilon_n = \log n/\sqrt{n}$.

By Lemma 1, we have the following estimate for entropy condition

$$\log N(\varepsilon, \mathcal{P}_a, \|\cdot\|_2) \lesssim \left(\log \frac{1}{\varepsilon}\right)^2,$$

this coincides with the estimate of prior mass condition, thus we obtain the rate of $\varepsilon_n = \log n/\sqrt{n}$ with respect to L_2-norm. This is the same with what is obtained in Ghosal and Vaart (2001), only in L_2-norm. However we lose a $\sqrt{\log n}$-factor comparing to Watson and Leadbetter (1963), which is $\sqrt{\log n/n}$.

9. Appendix: Contraction rates relative to non-metrics

The basic theorem of Ghosal et al. (2000) gives a posterior contraction rate in terms of a metric on densities that is bounded above by the Hellinger distance. In the present situation we would like to apply this result to a power smaller than one of the Wasserstein metric, which is not a metric. In this appendix we establish a rate of contraction which is valid for more general discrepancies.

We consider a general "discrepancy measure" d, which is a map $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ on the product of the set of densities on a given measurable space and itself, which has the properties, for some constant $C > 0$:

(a) $d(x, y) \geq 0$;
(b) $d(x, y) = 0$ if and only if $x = y$;
(c) $d(x, y) = d(y, x)$;
(d) $d(x, y) \leq C\big(d(x, z) + d(y, z)\big)$.

Thus d is a metric except that the triangle inequality is replaced with a weaker condition that incorporates a constant C, possibly bigger than 1. Call a set of the form $\{x : d(x, y) < c\}$ a d-ball, and define covering numbers $N(\varepsilon, \mathcal{P}, d)$ relative to d as usual.

Let $\Pi_n(\cdot|X_1, \ldots, X_n)$ be the posterior distribution of p given an i.i.d. sample $X_1, \ldots, X_n$ from a density p that is equipped with a prior probability distribution Π.

Theorem 3. *Suppose d has the properties as given, the sets $\{p : d(p, p') < \delta\}$ are convex, and satisfies $d(p_0, p) \le h(p_0, p)$, for every $p \in \mathcal{P}$. Then $\Pi_n\big(d(p, p_0) > M\varepsilon_n|X_1, \ldots, X_n\big) \to 0$ in P_0^n-probability for any ε_n such that $n\varepsilon_n^2 \to \infty$ and such that, for positive constants c_1, c_2 and sets $\mathcal{P}_n \subset \mathcal{P}$,*

$$\log N(\varepsilon_n, \mathcal{P}_n, d) \le c_1 n\varepsilon_n^2, \tag{9.1}$$

$$\Pi_n(p : K(p_0, p) < \varepsilon_n^2, K_2(p_0, p) < \varepsilon_n^2) \ge e^{-c_2 n\varepsilon_n^2}, \tag{9.2}$$

$$\Pi_n(\mathcal{P} - \mathcal{P}_n) \le e^{-(c_2+4)n\varepsilon_n^2}. \tag{9.3}$$

Proof. For every $\varepsilon > 4C\varepsilon_n$, we have $\log N(C^{-1}\varepsilon/4, \mathcal{P}_n, d) \le \log N(\varepsilon_n, \mathcal{P}_n, d) \le c_1 n\varepsilon_n^2$, take $N(\varepsilon) = \exp(c_1 n\varepsilon_n^2)$ and $\varepsilon = MC^{-1}\varepsilon_n$, $j = 1$ in Lemma 8, where $M > 4C$ is a large constant to be chosen later, there exist tests φ_n with errors

$$P_0^n\varphi_n \le e^{c_1 n\varepsilon_n^2} \frac{e^{-nM^2C^{-2}\varepsilon_n^2/32}}{1 - e^{-nM^2C^{-2}\varepsilon_n^2/32}},$$

$$\sup_{p\in\mathcal{P}_n : d(p,p_0) > M\varepsilon_n} P^n(1 - \varphi_n) \le e^{-nM^2C^{-2}\varepsilon_n^2/32}.$$

Next the proof proceeds as in Ghosal et al. (2000). All terms should tend to zero for $M^2/(32C^2) > c_1$ and $M^2/(32C^2) > 2 + c_2$. □

Lemma 8. *Let d be a discrepancy measure in the sense of (a)–(d) whose balls are convex and which is bounded from above by the Hellinger distance h. If $N(C^{-1}\varepsilon/4, \mathcal{Q}, d) \le N(\varepsilon)$ for any $\varepsilon > C\varepsilon_n > 0$ and some non-increasing function $N : (0, \infty) \to (0, \infty)$, then for every $\varepsilon > C\varepsilon_n$ and n, there exists a test φ_n such that for all $j \in \mathbb{N}$,*

$$P^n\varphi_n \le N(\varepsilon)\frac{e^{-n\varepsilon^2/32}}{1 - e^{-n\varepsilon^2/32}}, \qquad \sup_{Q\in\mathcal{Q}, d(P,Q) > Cj\varepsilon} Q^n(1 - \varphi_n) \le e^{-n\varepsilon^2 j^2/32}.$$

Proof. For a given $j \in \mathbb{N}$, choose a maximal set $Q_{j,1}, Q_{j,2}, \ldots, Q_{j,N_j}$ in the set $\mathcal{Q}_j = \{Q \in \mathcal{Q} : Cj\varepsilon < d(P, Q) < 2Cj\varepsilon\}$ such that $d(Q_{j,k}, Q_{j,l}) \ge j\varepsilon/2$ for every $k \ne l$. By property (d) of the discrepancy every ball in a cover of $\mathcal{Q}_j$ by balls of radius $C^{-1}j\varepsilon/4$ contains at most one $Q_{j,k}$. Thus $N_j \le N(C^{-1}j\varepsilon/4, \mathcal{Q}_j, d) \le N(\varepsilon)$. Furthermore, the N_j balls $B_{j,l}$ of radius $j\varepsilon/2$ around $Q_{j,l}$ cover $\mathcal{Q}_j$, as otherwise the set of $Q_{j,l}$ would not be maximal. For any point Q in each $B_{j,l}$, we have

$$d(P, Q) \ge C^{-1}d(P, Q_{j,l}) - d(Q, Q_{j,l}) \ge j\varepsilon/2.$$

Since the Hellinger distance bounds d from above, also $h(P, B_{j,l}) \geq j\varepsilon/2$. By Lemma 9, there exist a test $\varphi_{j,l}$ of P versus $B_{j,l}$ with error probabilities bounded from above by $e^{-nj^2\varepsilon^2/32}$. Let φ_n be the supremum of all the tests $\varphi_{j,l}$ obtained in this way, for $j = 1, 2, \ldots$, and $l = 1, 2, \ldots, N_j$. Then,

$$P^n\varphi \leq \sum_{j=1}^{\infty}\sum_{l=1} N_j e^{-nj^2\varepsilon^2/32} \leq \sum_{j=1}^{\infty} N(C^{-1}j\varepsilon/4, \mathcal{Q}_j, d)e^{-nj^2\varepsilon^2/32}$$
$$\leq N(\varepsilon)\frac{e^{-n\varepsilon^2/32}}{1 - e^{-n\varepsilon^2/32}},$$

and for every $j \in \mathbb{N}$,

$$\sup_{Q\in\cup_{l>j}\mathcal{Q}_l} Q^n(1 - \varphi_n) \leq \sup_{l>j} e^{-nl^2\varepsilon^2/32} \leq e^{-nj^2\varepsilon^2/32},$$

by the construction of φ_n. □

The following lemma comes from the general results of Birgé (1984) and Le Cam (1986).

Lemma 9. *For any probability measure P and dominated, convex set of probability measures $\mathcal{Q}$ with $h(p, q) > \varepsilon$ for any $q \in \mathcal{Q}$ and any $n \in \mathbb{N}$, there exists a test ϕ_n such that*

$$P^n\phi_n \leq e^{-n\varepsilon^2/8}, \quad \sup_{Q\in\mathcal{Q}} Q^n(1 - \phi_n) \leq e^{-n\varepsilon^2/8}$$

Acknowledgements

The authors thank the referees and the (associate) editors for their help in improving the presentation of the paper and pointing out useful references.

References

T. L. Bailey and C. Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.*, 3(2):259–282, 1984. ISSN 0208-4147. MR0764150

S. Donnet, V. Rivoirard, J. Rousseau, and C. Scricciolo. Posterior concentration rates for empirical bayes procedures, with applications to dirichlet process mixtures. *arXiv preprint arXiv:1406.4406*, 2014.

J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, Sept. 1991. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1176348248. MR1126324

J. Fan. Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.*, 21(2):600–610, 06 1993. URL http://dx.doi.org/10.1214/aos/1176349139. MR1232507

S. Ghosal and A. W. v. d. Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233–1263, Oct. 2001. ISSN 0090-5364. URL http://www.jstor.org/stable/2699987. MR1873329

S. Ghosal and A. van der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, Apr. 2007. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1183667289. MR2336864

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, Apr. 2000. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1016218228. MR1790007

G. K. Golubev. Nonparametric estimation of smooth probability densities in l_2. *Problemy Peredachi Informatsii*, 28(1):52–62, 1992. MR1163140

B. Knapik and J.-B. Salomond. A general approach to posterior contraction in nonparametric inverse problems. *arXiv preprint arXiv:1407.0335*, 2014.

S. Kotz, T. J. Kozubowski, and K. Podgórski. *The Laplace Distribution and Generalizations*. Birkhäuser Boston, Inc., Boston, MA, 2001. ISBN 0-8176-4166-1. URL http://dx.doi.org/10.1007/978-1-4612-0173-1. A revisit with applications to communications, economics, engineering, and finance. MR1935481

W. Kruijer, J. Rousseau, and A. van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257, 2010. ISSN 1935-7524. URL http://dx.doi.org/10.1214/10-EJS584. MR2735885

L. M. Le Cam. *Asymptotic methods in statistical decision theory*. Springer, New York [N.Y.] [etc.], 1986. ISBN 0387963073 9780387963075. MR0856411

O. Lepski and T. Willer. Lower bounds in the convolution structure density model. Working paper or preprint, Nov. 2015. URL https://hal.archives-ouvertes.fr/hal-01226357.

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, Feb. 2013. ISSN 0090-5364. URL http://projecteuclid.org/euclid.aos/1364302747. MR3059422

C. Scricciolo. Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electron. J. Stat.*, 5:270–308, 2011. ISSN 1935-7524. URL http://dx.doi.org/10.1214/11-EJS604. MR2802044

W. Shen, S. T. Tokdar, and S. Ghosal. Adaptive Bayesian multivariate density estimation with dirichlet mixtures. arXiv e-print 1109.6406, Sept. 2011. URL http://arxiv.org/abs/1109.6406. Biometrika (2013) 100 (3): 623–640. MR3094441

C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 9783540710493 3540710493 9783540710509 3540710507. MR2459454

G. S. Watson and M. R. Leadbetter. On the estimation of the probability density, i. *The Annals of Mathematical Statistics*, 34(2):480–491, June

1963. ISSN 0003-4851. URL http://projecteuclid.org/euclid.aoms/1177704159. MR0148149

C.-H. Zhang. Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics*, 18(2):pp. 806–831, 1990. ISSN 00905364. URL http://www.jstor.org/stable/2242135. MR1056338