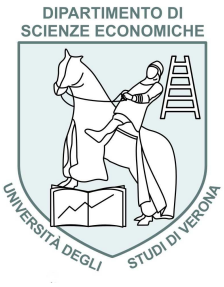

Working Paper Series
Department of Economics
University of Verona

Bayes and maximum likelihood for L1-Wasserstein deconvolution of Laplace mixtures

Catia Scricciolo

WP Number: 18 October 2016

ISSN: 2036-2919 (paper), 2036-4679 (online)

Bayes and maximum likelihood for L^1-Wasserstein deconvolution of Laplace mixtures

Catia Scricciolo ⋆

Received: October 2016

Abstract We consider the problem of recovering a distribution function on the real line from observations additively contaminated with errors having a Laplace distribution. Assuming the mixing distribution to be completely unknown leads to a nonparametric deconvolution problem. We begin by studying the rates of convergence in the Hellinger or the L^2-metric for the direct problem of estimating the sampling density, which is a Laplace mixture with a possibly unbounded set of locations: the rate of convergence for the Bayes' density estimator corresponding to a Dirichlet process prior over the space of all mixing distributions on the real line matches, up to a logarithmic factor, with the $n^{-3/8}\log^{1/8} n$ rate for the maximum likelihood estimator (MLE). Then, appealing to an inversion inequality translating the Hellinger or the L^2-distance between kernel mixture densities, with characteristic function of the kernel having polynomially decaying tails, into any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, provided they have finite Laplace transforms in a neighborhood of zero, we derive the rates of convergence in the L^1-Wasserstein metric for the MLE and the Bayes' estimator of the mixing distribution. Merging in the L^1-Wasserstein metric between Bayes and maximum likelihood follows as a by-product, together with an upper bound on the stochastic order of the discrepancy between the two estimation procedures.

Key words Deconvolution · Dirichlet process · entropy · Hellinger distance · Laplace mixture · maximum likelihood · posterior distribution · rate of convergence · sieve · Wasserstein metric

1 Introduction

The problem of recovering a distribution function from observations additively contaminated with measurement errors is the object of interest of this note. Assuming data are sampled from a convolution kernel mixture, the interest is in reconstructing the mixing or latent distribution from contaminated observations. The general statement of the problem is as follows. Let X be a random variable (r.v.) with probability distribution P_0. Let $p_0 := \mathrm{d}P_0/\mathrm{d}\lambda$ be the density of P_0 with respect to Lebesgue measure λ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Suppose that

$$X = Y + Z,$$

where Y and Z are independent, unobservable random variables, Z having known Lebesgue density f. We shall examine the case where the error has a Laplace distribution, with density

$$f(z) = \frac{1}{2}e^{-|z|}, \quad z \in \mathbb{R}.$$

⋆ C. Scricciolo
Department of Economics, University of Verona, Via Cantarane 24, I-37129 Verona, Italy
e-mail: catia.scricciolo@univr.it

The r.v. Y has an unknown distribution G_0 on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, with $\mathscr{Y} \subseteq \mathbb{R}$ and $\mathscr{B}(\mathscr{Y})$ the Borel σ-field on $\mathscr{Y}$. The density p_0 is then the convolution of f and G_0,

$$p_0(x) = \int_{\mathscr{Y}} f(x-y)\,\mathrm{d}G_0(y), \quad x \in \mathbb{R}.$$

We shall also write $p_0 \equiv p_{G_0}$ to express the dependence of p_0 on G_0. Letting $\mathscr{G}$ be the set of all probability measures G on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, the parameter space

$$\mathscr{P} := \left\{ p_G(\cdot) := \int_{\mathscr{Y}} f(\cdot - y)\,\mathrm{d}G(y),\ G \in \mathscr{G} \right\}$$

is the collection of all convolution Laplace mixtures and the model is nonparametric.

Suppose we observe n independent copies $X_1, \ldots, X_n$ of X. The interest is in recovering the mixing distribution $G_0 \in \mathscr{G}$ from indirect observations. Inverse problems arise when the object of interest is only indirectly observed. The deconvolution problem may arise in a wide variety of contexts, where the error density is typically modelled using a Gaussian kernel, but also Laplace mixtures have found relevant applications. Full density deconvolution, together with the related many normal means problem, has drawn attention in the literature since the late 1950's and different estimation methods have been developed since then, especially taking the frequentist approach, the most popular being based on nonparametric maximum likelihood and on kernel methods. Rates of convergence have been mostly considered for *density* deconvolution: Fan (1991) shows that the kernel density estimator achieves global optimal rates; for a recent and comprehensive account on the subject the reader may refer to the monograph of Meister (2009). In this note, however, we shall not assume that the distribution to be recovered has Lebesgue density, for which case Wasserstein metrics are particularly well-suited. Dedecker *et al*. (2015) have obtained a lower bound on the L^p-Wasserstein risk when no smoothness assumption is imposed on the distribution to be recovered and the error distribution is ordinary smooth.

The deconvolution problem has only recently been studied from a Bayesian perspective: the typical scheme considers the mixing distribution generated from a Dirichlet process prior. Posterior contraction rates for recovering the mixing distribution in L^p-Wasserstein metrics have been investigated in Nguyen (2013) and Gao and van der Vaart (2016): convergence in Wasserstein metrics for discrete mixing distributions has a natural interpretation in terms of convergence of the single atoms supporting the probability measures. Adaptive recovery rates for deconvolving a density in a Sobolev space, which automatically rules out the case of a Laplace mixture, are derived in Donnet *et al*. (2014) for the fully as well as for the empirical Bayes approaches, the latter employing a data-driven choice of the prior hyperparameters of the Dirichlet process baseline measure.

In this note, we study nonparametric Bayes and maximum likelihood recovering of the mixing distribution G_0, when no smoothness assumption is imposed on it. The analysis starts with the estimation of the sampling density p_0: estimating the *mixed* density p_0 is the first step for recovering the *mixing* distribution G_0. Taking a Bayesian approach, if the random density p_G is modelled as a Dirichlet-Laplace mixture, then p_0 is consistently estimated a rate $n^{-3/8}$, up to a $(\log n)$-factor, if only G_0 has tails matching with those of the baseline measure of the Dirichlet process, see Propositions 1 and 2. This requirement allows to extend to a possibly unbounded set of locations the results of Gao and van der Vaart (2016), which take into account only the case of compactly supported mixing distributions. Taking instead a frequentist approach, p_0 can be estimated by the MLE again at a rate $n^{-3/8}$, up to a logarithmic factor. As far as we are aware, the result on the rate of convergence in the Hellinger metric for the MLE of a Laplace mixture is new and is obtained adopting a convenient approach proposed by Van de Geer (1996), according to which it is the dimension of the class of kernels and the behaviour of p_0 near zero that determine the rate of convergence in the Hellinger metric for the MLE. As previously mentioned, results on density estimation of p_0 are interesting in view of the fact that, appealing to an inversion inequality translating the Hellinger or the L^2-distance between kernel mixture densities with Fourier transform of the kernel having polynomially decaying tails into any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, rates of convergence in the L^1-Wasserstein metric for the MLE and the Bayes' estimator of the mixing distribution can be assessed. Merging in the L^1-Wasserstein metric between Bayes and maximum likelihood for deconvolving Laplace mixtures follows as a by-product.

Organization. The note is organized as follows. Convergence rates in the Hellinger metric for Bayes and maximum likelihood density estimation of Laplace convolution mixtures are studied in Sections 2 and 3, respectively, in view of the subsequent L^1-Wasserstein accuracy assessment for the two estimation procedures in recovering the true mixing distribution. Merging between Bayes and maximum likelihood follows, as shown in Section

4. Remarks and suggestions for possible refinements and extensions of the exposed results are presented in Section 5. Auxiliary lemmas, together with the proofs of the main results, are deferred to Appendices A–D.

Notation. We introduce the notation and give some definitions used throughout the manuscript.

The random variables $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) according to a density $p_0 \equiv p_{G_0}$ on the real line; all probability density functions are meant to be with respect to Lebesgue measure λ on $\mathbb{R}$ or on some subset thereof.

Probability spaces and random variables

- We use the same symbol, say G, to denote a probability measure on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, $\mathscr{Y} \subseteq \mathbb{R}$, or the corresponding cumulative distribution function (cdf).
- The distribution degenerate at a point $\theta \in \mathbb{R}$ is denoted by δ_θ.
- The notation Pf will abbreviate the expected value $\int f \, \mathrm{d}P$.
- Given a r.v. Y with distribution G, the *moment generating function* of Y or the *Laplace transform of* G is defined as

$$M_G(s) := E[e^{sY}] = \int_{\mathscr{Y}} e^{sy} \, \mathrm{d}G(y) \quad \text{for all } s \text{ for which the integral is finite.}$$

- For real $1 \leq p < +\infty$, let $L^p(\mathbb{R}) := \{f \,|\, f : \mathbb{R} \to \mathbb{R},\ f \text{ is Borel measurable},\ \int |f|^p \, \mathrm{d}\lambda < +\infty\}$. For $f \in L^p(\mathbb{R})$, the L^p-norm of f is defined as $\|f\|_p := (\int |f|^p \, \mathrm{d}\lambda)^{1/p}$. The supremum norm of a function f is defined as $\|f\|_\infty := \sup_{x \in \mathbb{R}} |f(x)|$.
- For $f \in L^1(\mathbb{R})$, the complex-valued function $\hat{f}(t) := \int_{-\infty}^{+\infty} e^{itx} f(x) \, \mathrm{d}x$, $t \in \mathbb{R}$, is called the *Fourier transform of* f.

Calculus

- We write "$\lesssim$" and "$\gtrsim$" for inequalities valid up to a constant multiple that is universal or fixed within the context, but anyway inessential for our purposes.
- For sequences of real numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, we write $a_n \sim b_n$ to mean that $(a_n / b_n) \to 1$ as $n \to \infty$. Analogously, for real-valued functions f and g, the notation $f \sim g$ means that $f/g \to 1$ in an asymptotic regime that is clear from the context.

Stochastic order symbols

Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. Let $Z_n : \Omega \to \mathbb{R}$, $n = 1, 2, \ldots$, be a sequence of random variables and $\{k_n\}_{n \geq 1}$ a sequence of positive real numbers. We write

- $Z_n = O_{\mathbf{P}}(k_n)$ if $\lim_{T \to +\infty} \limsup_{n \to +\infty} \mathbf{P}(|Z_n| > T k_n) = 0$. Then, $Z_n / k_n = O_{\mathbf{P}}(1)$,
- $Z_n = o_{\mathbf{P}}(k_n)$ if, for all $\epsilon > 0$, $\lim_{n \to +\infty} \mathbf{P}(|Z_n| > \epsilon k_n) = 0$. Then, $Z_n / k_n = o_{\mathbf{P}}(1)$.

Unless otherwise specified, in all stochastic order symbols used throughout the manuscript, the probability measure $\mathbf{P}$ is understood to be P_0^n, the joint law of the first n coordinate projections of the infinite product probability measure $P_0^{\mathbb{N}}$.

Semi-metrics and divergences

- For any pair of density functions $p, q : \mathbb{R} \to [0, +\infty)$, let the *Hellinger distance* be defined as $h(p, q) := \{\int (p^{1/2} - q^{1/2})^2 \, \mathrm{d}\lambda\}^{1/2}$. The following inequalities relating the Hellinger and the L^1-metric hold:

$$h^2(p, q) \leq \|p - q\|_1 \tag{1}$$

and

$$\|p - q\|_1 \leq 2h(p, q). \tag{2}$$

- For any density p on $\mathbb{R}$, let $\mathrm{KL}(p_0;\, p) := \int p_0 \log(p_0/p) \, \mathrm{d}\lambda$ be the *Kullback-Leibler divergence* of p from p_0 and $\mathrm{V}_2(p_0;\, p) := \int p_0 |\log(p_0/p)|^2 \, \mathrm{d}\lambda$ the second moment of the log-ratio $\log(p_0/p)$. We define a Kullback-Leibler type neighborhood of p_0 of radius ε^2 as

$$B_{\mathrm{KL}}(p_0;\, \varepsilon^2) := \{p :\ \mathrm{KL}(p_0;\, p) \leq \varepsilon^2,\ \mathrm{V}_2(p_0;\, p) \leq \varepsilon^2\}.$$

– For any real number $p \geq 1$ and any pair of probability measures $G_1, G_2 \in \mathscr{G}$, with finite pth moments, the *L^p-Wasserstein distance* between G_1 and G_2 is defined as

$$W_p(G_1, G_2) := \left(\inf_{\pi \in \Gamma(G_1, G_2)} \int_{\mathscr{Y} \times \mathscr{Y}} |x - y|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p},$$

where $\Gamma(G_1, G_2)$ is the set of all joint probability measures on $(\mathscr{Y} \times \mathscr{Y}) \subseteq \mathbb{R} \times \mathbb{R}$, whose marginals are G_1 and G_2.

Covering and entropy numbers

– Let (T, d) be a (subset of a) semi-metric space. For $\delta > 0$, the *δ-covering number* $N(\delta, T, d)$ is defined as the minimum number of d-balls of radius δ needed to cover T. The logarithm of the δ-covering number, $\log N(\delta, T, d)$, is called the *δ-entropy*.

– Let (T, d) be a (subset of a) semi-metric space. For $\delta > 0$, the *δ-packing number* $D(\delta, T, d)$ is defined as the maximum number of points in T such that the distance between each pair is at least δ. The logarithm of the δ-packing number, $\log D(\delta, T, d)$, is called the *δ-entropy*.

Covering and packing numbers are related by the following inequalities: $N(\delta, T, d) \leq D(\delta, T, d) \leq N(\delta/2, T, d)$.

2 Rates of convergence for L^1-Wasserstein deconvolution of Dirichlet-Laplace mixtures

In this section, we present results on the Bayesian recovery of a probability distribution from data contaminated with an additive error having a Laplace distribution: we derive the rate of convergence for the L^1-Wasserstein deconvolution of Dirichlet-Laplace mixtures. The random density is modelled as a Dirichlet-Laplace mixture

$$p_G(\cdot) \equiv (G * f)(\cdot) = \int_{\mathscr{Y}} f(\cdot - y) \, \mathrm{d}G(y),$$

with kernel f the Laplace density and mixing distribution G any probability measure on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, with $\mathscr{Y} \subseteq \mathbb{R}$. As a prior for G, we consider a Dirichlet process, with base measure α on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, denoted by $\mathscr{D}_\alpha$. We recall that a Dirichlet process on a measurable space $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, with finite and positive base measure α on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, is a random probability measure G on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ such that, for every finite partition $(B_1, \ldots, B_k)$ of $\mathscr{Y}$, $k \geq 1$, the vector of probabilities $(G(B_1), \ldots, G(B_k))$ has Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. A Dirichlet process mixture of Laplace densities can be structurally described as follows:

- $G \sim \mathscr{D}_\alpha$,
- given G, the r.v.'s $Y_1, \ldots, Y_n$ are i.i.d. according to G,
- given $(G, Y_1, \ldots, Y_n)$, the r.v.'s $Z_1, \ldots, Z_n$ are i.i.d. according to f,
- sampled values from p_G are defined as $X_i := Y_i + Z_i$ for $i = 1, \ldots, n$.

Let the sampling density p_0 be itself a Laplace mixture with mixing distribution G_0, that is, $p_0 \equiv p_{G_0} = G_0 * f$. In order to assess the rate of convergence in the L^1-Wasserstein metric for the Bayes' estimator of the true mixing distribution G_0, we appeal to an inversion inequality relating the Hellinger or the L^2-distance between Laplace mixture densities to any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, see Lemma 4 in Appendix D. Therefore, we first derive rates of contraction in the Hellinger or the L^2-metric for the posterior distribution of a Dirichlet-Laplace mixture prior: convergence of the posterior distribution at a rate ε_n, in fact, implies the existence of Bayes' point estimators that converge in the frequentist sense at least as fast as ε_n. The same indirect approach to the problem has been taken by Gao and van der Vaart (2016), who deal with the case of compactly supported mixing distributions, while here we extend the results to mixing distributions possibly supported on the whole real line. We present two results on posterior contraction rates for a Dirichlet-Laplace mixture prior: the first one is relative to the Hellinger or the L^1-metric, the second one to the L^2-metric. Proofs are deferred to Appendix C.

Proposition 1 *Let $X_1, \ldots, X_n$ be i.i.d. observations from a density $p_0 \equiv p_{G_0} = G_0 * f$, with kernel f the Laplace density and mixing distribution G_0 such that, for some finite constant $c_0 > 0$,*

$$G_0([-T, T]^c) \lesssim \exp(-c_0 T) \quad \textit{for large } T > 0. \tag{3}$$

If the baseline measure α of the Dirichlet process is symmetric around zero and possesses density α' such that, for some constants $0 < b < +\infty$ and $0 < \delta \leq 1$,

$$\alpha'(y) \propto \exp(-b|y|^{\delta}), \quad y \in \mathbb{R}, \tag{4}$$

then there exists a sufficiently large constant $M > 0$ such that

$$\Pi(d(p_G,\, p_0) \geq Mn^{-3/8} \log^{1/2} n \mid X^{(n)}) = O_{\mathbf{P}}(1),$$

where d can be either the Hellinger or the L^1-metric.

Remark 1 In virtue of the inequality,

$$\forall\, G,\, G' \in \mathscr{G}, \quad \|p_G - p_{G'}\|_2^2 \leq 4\|f\|_{\infty} h^2(p_G,\, p_{G'}),$$

where $\|f\|_{\infty} = 1/2$ for the Laplace kernel density, see (20) in Lemma 3, the L^2-metric posterior contraction rate of a Dirichlet-Laplace mixture prior could, in principle, be derived from Proposition 1, which relies on Theorem 2.1 of Ghosal *et al*. (2000), page 503, or Theorem 2.1 of Ghosal and van der Vaart (2001), page 1239, but this would impose stronger conditions on the tail behaviour of G_0 and on the density α' of the baseline measure than those required in Proposition 2 below, which is, instead, based on Theorem 3 of Giné and Nickl (2011), page 2892, that is tailored for assessing posterior contraction rates in L^r-metrics, $1 < r < +\infty$, taking an approach that can only be used if one has sufficiently fine control of the approximation properties of the support of the prior law in the L^r-metric considered.

Proposition 2 *Let $X_1, \ldots, X_n$ be i.i.d. observations from a density $p_0 \equiv p_{G_0} = G_0 * f$, with kernel f the Laplace density and mixing distribution G_0 such that, for some decreasing function $A_0 : (0, +\infty) \to [0, 1]$,*

$$G_0([-T,\, T]^c) \leq A_0(T) \quad \textit{for large } T > 0. \tag{5}$$

If the baseline measure α of the Dirichlet process possesses density α' such that

$$\alpha'(y) \gtrsim A_0(|y|), \quad y \in \mathbb{R}, \tag{6}$$

then there exists a sufficiently large constant $M > 0$ such that

$$\Pi(\|p_G - p_0\|_2 \geq Mn^{-3/8} \log^{1/2} n \mid X^{(n)}) = O_{\mathbf{P}}(1). \tag{7}$$

As previously mentioned, convergence of the posterior distribution at a rate ε_n implies the existence of point estimators, which are Bayes in the sense that they are based on the posterior distribution, that converge at least as fast as ε_n in the frequentist sense, see, for instance, Theorem 2.5 in Ghosal *et al*. (2000), page 506, for the construction of a point estimator that applies to general statistical models and posterior distributions. The posterior expectation of the density p_G, which we refer to as the Bayes' density estimator,

$$\hat{p}_n^{\mathrm{B}}(\cdot) := \int_{\mathscr{G}} p_G(\cdot) \Pi(\mathrm{d}G \mid X^{(n)}),$$

has a similar property when considered jointly with bounded metrics that are convex or whose square is convex. When the mixing distribution G is distributed according to a Dirichlet process, its posterior expectation has expression

$$\hat{G}_n^{\mathrm{B}} := E[G \mid X^{(n)}] = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n} \bar{\alpha} + \frac{n}{\alpha(\mathbb{R}) + n} \mathbb{P}_n,$$

so that the Bayes' density estimator takes the form

$$\hat{p}_n^{\mathrm{B}} = \hat{G}_n^{\mathrm{B}} * f = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n} (\bar{\alpha} * f) + \frac{n}{\alpha(\mathbb{R}) + n} (\mathbb{P}_n * f),$$

where $\bar{\alpha} := \alpha/\alpha(\mathbb{R})$ is the probability measure corresponding to the baseline measure α of the Dirichlet process and $\mathbb{P}_n$ is the empirical measure associated with the sample $X_1, \ldots, X_n$, namely, the discrete uniform measure on the sample values.

Corollary 1 *Under the assumptions of Propositions 1 and 2,*

$$d(\hat{p}_n^{\mathrm{B}},\, p_0) = O_{\mathbf{P}}(n^{-3/8}\log^{1/2} n),$$

where d is the Hellinger or the L^1-metric in the case of Proposition 1 and the L^2-metric in the case of Proposition 2.

Proof The proof follows standard arguments as, for instance, in Ghosal *et al*. (2000), pages 506–507. Let $\tilde{d}$ stand for either the L^r-metric, $r = 1,\, 2$, or the squared Hellinger distance. By convexity of the map $p \mapsto \tilde{d}(p,\, p_0)$ and Jensen's inequality,

$$\begin{aligned}\tilde{d}(\hat{p}_n^{\mathrm{B}},\, p_0) \leq \int_{\mathscr{G}} \tilde{d}(p_G,\, p_0)\Pi(\mathrm{d}G \mid X^{(n)}) &= \left(\int_{\tilde{d}(p_G,\, p_0)<M\psi_n} + \int_{\tilde{d}(p_G,\, p_0)\geq M\psi_n}\right)\tilde{d}(p_G,\, p_0)\Pi(\mathrm{d}G \mid X^{(n)}) \\ &\lesssim M\psi_n + \tilde{d}_\infty \Pi(\tilde{d}(p_G,\, p_0) \geq M\psi_n \mid X^{(n)}),\end{aligned}$$

where $M > 0$ is a sufficiently large constant,

$$\tilde{d}_\infty := \sup_{g,\, z} \tilde{d}(g,\, z) = \begin{cases} 2 & \text{for } \tilde{d} = \|\cdot\|_1 \text{ or } \tilde{d} = h^2, \\ 2\|f\|_2 & \text{for } \tilde{d} = \|\cdot\|_2, \end{cases}$$

and, for $\varepsilon_n = n^{-3/8}\log^{1/2} n$,

$$\psi_n := \begin{cases} \varepsilon_n^2 & \text{for } \tilde{d} = h^2, \\ \varepsilon_n & \text{for } \tilde{d} = \|\cdot\|_r,\ r = 1,\, 2. \end{cases}$$

It follows that

$$P_0^n \tilde{d}(\hat{p}_n^{\mathrm{B}},\, p_0) \lesssim M\psi_n + \tilde{d}_\infty P_0^n \Pi(\tilde{d}(p_G,\, p_0) \geq M\psi_n \mid X^{(n)}) \lesssim \psi_n$$

because both Propositions 1 and 2 yield that $\Pi(\tilde{d}(p_G,\, p_0) \geq M\psi_n \mid X^{(n)}) = o_{\mathbf{P}}(\psi_n)$. The assertion follows. □

We now state the result on the rate of convergence for the Bayes' estimator of G_0 for L^1-Wasserstein deconvolution of Dirichlet-Laplace mixtures.

Proposition 3 *Under the same assumptions as in Propositions 1 and 2, if $\bar{\alpha}$ and G_0 have finite moment generating functions on an interval $(-s_0,\, s_0)$ for some $s_0 > 1$, then*

$$W_1(\hat{G}_n^{\mathrm{B}},\, G_0) = O_{\mathbf{P}}(n^{-1/8}\log^{1/2} n). \tag{8}$$

Proof The assertion follows by combining Lemma 4 and Corollary 1. □

Some remarks are in order. There are two main reasons why we focus on deconvolution in the L^1-Wasserstein metric: the first one is related to the inversion inequality in (22), where the upper bound on the L^p-Wasserstein distance, as a function of p, increases as p gets larger, thus possibly leading to sub-optimal rates. In absence of any statement on the optimality of the bound in (22), it is advisable to begin the analysis from the smallest order $p = 1$. The second reason is related to the interpretation of the assertion in (8): the L^1-Wasserstein distance between any two probability measures G_1 and G_2 on $(\mathscr{Y},\, \mathscr{B}(\mathscr{Y}))$, $\mathscr{Y} \subseteq \mathbb{R}$, with finite first absolute moments, is by itself an interesting distance in probability limit theory because it metrizes weak convergence plus convergence of the first absolute moments, but it is even more interesting in view of the fact that, letting Q_1 and Q_2 denote the left-continuous inverse or quantile functions, that is, $Q_i(u) := \inf\{y \in \mathscr{Y} :\, G_i(y) \geq u\}$, $u \in (0,\, 1)$, $i = 1,\, 2$, it can be written as the L^1-distance between the cumulative distribution functions or the quantile functions,

$$W_1(G_1,\, G_2) = \|G_1 - G_2\|_1 = \int_{\mathscr{Y}} |G_1(y) - G_2(y)|\,\mathrm{d}y = \int_0^1 |Q_1(u) - Q_2(u)|\,\mathrm{d}u = \|Q_1 - Q_2\|_1,$$

cf. Shorack and Wellner (1986), page 64. Thus, by rewriting $W_1(\hat{G}_n^{\mathrm{B}},\, G_0)$ as the L^1-distance between the cdf's $\hat{G}_n^{\mathrm{B}}$ and G_0, the assertion of Proposition 3

$$W_1(\hat{G}_n^{\mathrm{B}},\, G_0) = \|\hat{G}_n^{\mathrm{B}} - G_0\|_1 = O_{\mathbf{P}}(n^{-1/8}\log^{1/2} n)$$

becomes more transparent and meaningful.

3 Rates of convergence for ML estimation and L^1-Wasserstein deconvolution of Laplace mixtures

In this section, we first study the rate of convergence in the Hellinger metric for the MLE $\hat{p}_n$ of a Laplace mixture density $p_0 \equiv p_{G_0} = G_0 * f$ with unknown mixing distribution $G_0 \in \mathscr{G}$. Then, applying an inversion inequality that relates the Hellinger distance between Laplace mixtures to any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, see Lemma 4 in Appendix D, we derive the rate of convergence in the L^1-Wasserstein metric for the MLE of the mixing distribution G_0, say $\hat{G}_n$, that corresponds to the MLE $\hat{p}_n$ of the mixed density p_0.

A MLE $\hat{p}_n$ of p_0 is a measurable function of the observations taking values in $\mathscr{P} := \{p_G : G \in \mathscr{G}\}$ such that

$$\hat{p}_n \in \arg\max_{p_G \in \mathscr{P}} \frac{1}{n} \sum_{i=1}^{n} \log p_G(X_i) = \arg\max_{p_G \in \mathscr{P}} \int (\log p_G) \, \mathrm{d}\mathbb{P}_n,$$

where $\mathbb{P}_n := n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure associated with the random sample $X_1, \ldots, X_n$, namely, the discrete uniform distribution on the sample values that puts mass $1/n$ on each one of the observations. We assume that the MLE exists, but do not require it to be unique, see Lindsay (1995), Theorem 18, page 112, for sufficient conditions ensuring its existence.

General results on rates of convergence in the Hellinger metric for the MLE of a density can be found in Birgé and Massart (1993), Wong and Shen (1995) and Van de Geer (1996): the underlying idea is that it is the metric entropy of the parameter space $\mathscr{P}$ that determines the rate of convergence, but it can be difficult to check this condition since it involves the L^1-metric entropy *with bracketing* of the square root of the densities. However, a general mixture model $\{\int_{\mathscr{Y}} K(\cdot, y) \, \mathrm{d}G(y) : G \in \mathscr{G}\}$ is the closure of the convex hull of the collection of kernels $\{K(\cdot, y) : y \in \mathscr{Y} \subseteq \mathbb{R}\}$, which is typically a much smaller class. Ball and Pajor (1990) give a bound on the metric-entropy *without bracketing* of the class of mixtures in terms of the covering number of the class of kernels. Based on this result, a relatively simple "recipe" to derive the rate of convergence in the Hellinger metric for the MLE of a density is given in Corollary 2.3 of Van de Geer (1996), page 298: it is the dimension of the class of kernels and the behaviour of p_0 near zero that determine (an upper bound on) the rate of convergence in the Hellinger metric for the MLE $\hat{p}_n$.

Proposition 4 *Suppose that the sampling density $p_0 \equiv p_{G_0} = G_0 * f$, with kernel f the Laplace density and mixing distribution $G_0 \in \mathscr{G}$. Suppose that, for a sequence of positive real numbers $\sigma_n \downarrow 0$, we have*

(*a*) $\int_{p_0 > \sigma_n} (1/p_0) \, \mathrm{d}\lambda \lesssim |\log \sigma_n|$,

(*b*) $\int_{p_0 \leq \sigma_n} p_0 \, \mathrm{d}\lambda \lesssim \sigma_n^2$.

Then, for $\sigma_n = n^{-3/8} \log^{1/8} n$,

$$h(\hat{p}_n, \, p_0) = O_{\mathbf{P}}(n^{-3/8} \log^{1/8} n).$$

Proof We begin by spelling out the remark mentioned in the introduction concerning the fact that a mixture model is the closure of the convex hull of the collection of kernels. Recall that the convex hull of a class $\mathscr{K}$ of functions, denoted by $\mathrm{conv}(\mathscr{K})$, is defined as the set of all finite convex combinations of functions in $\mathscr{K}$,

$$\mathrm{conv}(\mathscr{K}) := \left\{ \sum_{j=1}^{r} \theta_j K_j : K_j \in \mathscr{K}, \; j = 1, \ldots, r, \; r \in \mathbb{N}, \text{ and } \theta_j \geq 0, \; \sum_{j=1}^{r} \theta_j = 1 \right\}.$$

Let $\mathscr{K} := \{f(\cdot - y) : y \in \mathscr{Y} \subseteq \mathbb{R}\}$ be the collection of kernels with f the Laplace density. The class $\mathscr{P} := \{p_G : G \in \mathscr{G}\}$ of all Laplace mixture densities $p_G = G * f$ is the closure of the convex hull of $\mathscr{K}$, that is, $\mathscr{P} = \overline{\mathrm{conv}}(\mathscr{K})$. Clearly, $\mathscr{P}$ is itself a convex class. This remark enables us to apply Theorem 2.2 and Corollary 2.3 of Van de Geer (1996), pages 297-298 and 310, or, equivalently, Theorem 7.7 of Van de Geer (2000), pages 104–105, whose conditions are hereafter shown to be satisfied. Define the class

$$\tilde{\mathscr{G}} := \left\{ \frac{f(\cdot - y)}{p_0} \mathbf{1}\{p_0 > \sigma_n\} : y \in \mathscr{Y} \right\}$$

and the function

$$\bar{G}(\cdot) := \sup_{y \in \mathscr{Y}} \frac{f(\cdot - y)}{p_0(\cdot)} \mathbf{1}\{p_0 > \sigma_n\}.$$

Since, by assumption (a), we have

$$\int \bar{G}^2 \, \mathrm{d}P_0 \lesssim \int_{p_0 > \sigma_n} \frac{1}{p_0(x)} \, \mathrm{d}x \lesssim \log \frac{1}{\sigma_n} \tag{9}$$

and, by assumption (b),

$$\int_{p_0 \leq \sigma_n} \mathrm{d}P_0 = \int_{p_0 \leq \sigma_n} p_0(x) \, \mathrm{d}x \lesssim \sigma_n^2,$$

we can take, for a suitable constant $c > 0$, the sequence $\delta_n = c\sigma_n$ in condition (7.21) of Theorem 7.7, page 104. Because the Laplace kernel density f is Lipschitz,

$$\forall\, y_1,\, y_2 \in \mathscr{Y}, \quad |f(\cdot - y_1) - f(\cdot - y_2)| \leq \frac{1}{2}|y_1 - y_2|,$$

see, *e.g.*, Lemma A.1 in Scricciolo (2011), pages 299-300, over the set

$$\left\{ \int \bar{G}^2 \, \mathrm{d}\mathbb{P}_n \leq T^2 \log(1/\delta_n) \right\}, \tag{10}$$

where $T > 0$ is a suitable finite constant, we find that, for $\mathrm{d}\mathbb{Q}_n := \mathrm{d}\mathbb{P}_n/(T^2 \log(1/\delta_n))$,

$$N(\delta,\, \tilde{\mathscr{G}},\, \|\cdot\|_{2,\mathbb{Q}_n}) \lesssim \delta^{-1}, \quad \delta \in (0,\, 1),$$

where $\|\cdot\|_{2,\mathbb{Q}_n}$ denotes the $L^2(\mathbb{Q}_n)$-norm, that is, $\|g\|_{2,\mathbb{Q}_n} := (\int |g|^2 \, \mathrm{d}\mathbb{Q}_n)^{1/2}$. So, in view of the result of Ball and Pajor (1990), reported as Theorem 1.1 in Van de Geer (1996), pages 295, over the same set as in (10), we have

$$\log N(\delta,\, \overline{\mathrm{conv}}(\tilde{\mathscr{G}}),\, \|\cdot\|_{2,\mathbb{Q}_n}) \lesssim \delta^{-2/3},$$

hence

$$\log N(\delta,\, \overline{\mathrm{conv}}(\tilde{\mathscr{G}}),\, \|\cdot\|_{2,\mathbb{P}_n}) \lesssim \left(\frac{\log^{1/2}(1/\delta_n)}{\delta} \right)^{2/3}.$$

Defined the class

$$\mathscr{G}_{\sigma_n}^{(\mathrm{conv})} := \left\{ \frac{2p_G}{p_G + p_0} \mathbf{1}\{p_0 > \sigma_n\} :\ p_G \in \mathscr{P} \right\}$$

considered in condition (7.20) of Theorem 7.7 in Van de Geer (2000), page 104, since

$$\log N(2\delta,\, \mathscr{G}_{\sigma_n}^{(\mathrm{conv})},\, \|\cdot\|_{2,\mathbb{P}_n}) \leq \log N(\delta,\, \overline{\mathrm{conv}}(\tilde{\mathscr{G}}),\, \|\cdot\|_{2,\mathbb{P}_n}),$$

in view of (9), we have

$$\sup_{\delta > 0} \frac{\log N(\delta,\, \mathscr{G}_{\sigma_n}^{(\mathrm{conv})},\, \|\cdot\|_{2,\mathbb{P}_n})}{\delta^{-2/3} \log^{1/3}(1/\delta_n)} = O_{\mathbf{P}}(1),$$

so that the above mentioned condition is satisfied. Defined the non-increasing function of δ

$$H(\delta) := \delta^{-2/3} \log^{1/3}(1/\delta_n), \quad \delta \in (0,\, 1),$$

we take $\Psi(\delta) := \delta^{2/3} \log^{1/6}(1/\delta_n)$, so that

$$\Psi(\delta) \geq \left(\int_{\delta^2/c}^{\delta} H^{1/2}(u) \, \mathrm{d}u \right) \vee \delta$$

and, for some $\epsilon > 0$, $\Psi(\delta)/\delta^{2-\epsilon}$ is non-increasing. Then, for δ_n such that $\sqrt{n}\delta_n^2 \geq \Psi(\delta_n)$, which implies taking $\delta_n = O(n^{-3/8} \log^{1/8} n)$, cf. condition (7.22) of Theorem 7.7 in Van de Geer (2000), page 104, we have $h(\hat{p}_n,\, p_0) = O_{\mathbf{P}}(\delta_n)$ and the proof is complete. □

Remark 2 In the special case where $p_0 > 0$ and G_0 has compact support, say $[-a,\, a]$ with $a > 0$, we have $h(\hat{p}_n,\, p_0) = O_{\mathbf{P}}(n^{-3/8})$. In fact, $\sigma_n \equiv 0$, $\|\bar{G}\|_\infty \leq e^{2a}$ and $\int \bar{G}^2 \, \mathrm{d}P_0 \leq e^{2a}$ so that, over the set $\{\int \bar{G}^2 \, \mathrm{d}\mathbb{P}_n \leq c_1^2\}$ with a suitable constant $c_1 > 0$, $\log N(\delta,\, \overline{\mathrm{conv}}(\tilde{\mathscr{G}}),\, \|\cdot\|_{2,\mathbb{P}_n}) \lesssim \delta^{-2/3}$ and, reasoning as in Proposition 4, we find the rate $n^{-3/8}$.

We now derive a consequence of Proposition 4 on the rate of convergence for the MLE of G_0. A MLE $\hat{p}_n$ of the *mixed* density p_0 corresponds to a MLE $\hat{G}_n$ of the *mixing* distribution G_0, that is, $\hat{p}_n \equiv p_{\hat{G}_n}$, such that

$$\hat{G}_n \in \arg\max_{G \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \log p_G(X_i) = \arg\max_{G \in \mathscr{G}} \int (\log p_G) \, \mathrm{d}\mathbb{P}_n.$$

Note that, even if we know that $\hat{G}_n$ is a discrete distribution, we do not know the number of its components.

Corollary 2 *Under the same assumptions as in Proposition 4, if the mixing distribution G_0 possesses finite moment generating function on an interval $(-s_0, s_0)$ for some $s_0 > 1$, then*

$$W_1(\hat{G}_n, G_0) = O_{\mathbf{P}}(n^{-1/8} \log^{3/8} n).$$

Proof The assertion follows by combining Proposition 4 and Lemma 4.

4 Merging of Bayes and ML for L^1-Wasserstein deconvolution of Laplace mixtures

In this section, we show that the Bayes' estimator and the MLE of G_0 merge in the L^1-Wasserstein metric, their discrepancy vanishing, at worst, at a rate $n^{-1/8} \log^{1/2} n$ because they both consistently estimate G_0 at a speed which is within a $(\log n)$-factor of $n^{-1/8}$, cf. Corollary 2 and Proposition 3.

Proposition 5 *Under the assumptions of Proposition 3 and Corollary 2, we have*

$$W_1(\hat{G}_n^{\mathrm{B}}, \hat{G}_n) = O_{\mathbf{P}}(n^{-1/8} \log^{1/2} n). \tag{11}$$

Proof By the triangle inequality, we have $W_1(\hat{G}_n^{\mathrm{B}}, \hat{G}_n) \leq W_1(\hat{G}_n^{\mathrm{B}}, G_0) + W_1(G_0, \hat{G}_n)$, where $W_1(\hat{G}_n^{\mathrm{B}}, G_0) = O_{\mathbf{P}}(n^{-1/8} \log^{1/2} n)$ and $W_1(G_0, \hat{G}_n) = O_{\mathbf{P}}(n^{-1/8} \log^{3/8} n)$ by Proposition 3 and Corollary 2, respectively. Relationship (11) follows. □

Proposition 5 states that the Bayes' estimator and the MLE will eventually agree and the speed of convergence of their L^1-Wasserstein discrepancy is determined by the stochastic order of their errors in recovering G_0. The crucial question that remains open is whether the Bayes' and the maximum likelihood estimators are rate optimal. Concerning this issue, we note that, on the one hand, other deconvolution estimators attain the rate $n^{-1/8}$ when the error distribution is a Laplace, with the proviso, however, that the L^1-Wasserstein metric is not linked to the (integrated) risks between the c.d.f.'s used in the results we are going to mention, so that the rates are not directly comparable: for instance, the estimator $\hat{G}_n$ of G_0 based on the standard deconvolution kernel density estimator is such that $(E\|\hat{G}_n - G_0\|_2^2)^{1/2} = O(n^{-1/8})$, see Hall and Lahiri (2008); the estimator $\bar{G}_n$ proposed by Dattner *et al*. (2011) for pointwise estimation of $G(x_0)$, when Y has density belonging to a Sobolev space of order $\alpha > -1/2$, is such that, for $\alpha = 0$, $(E|\bar{G}_n(x_0) - G(x_0)|^2)^{1/2} = O(n^{-1/8})$ and is rate optimal. On the other hand, a recent lower bound result, due to Dedecker *et al*. (2015), Theorem 4.1, pages 246–248, suggests that better rates could be possible. For $M > 0$ and $q \geq 1$, let $\mathscr{D}(M, q)$ be the set of all probability measures G on $\mathbb{R}$ such that $\int |y|^q \, \mathrm{d}G(y) \leq M$. Let $r = 1/\hat{f}$, where f is the kernel density, so that $r^{(\ell)}$ denotes the ℓth derivative of r. Assume that there exist $\beta > 0$ and $c > 0$ such that, for every $\ell \in \{0, 1, 2\}$, we have $|r^{(\ell)}(t)| \leq c(1 + |t|)^{-\beta}$, $t \in \mathbb{R}$. Then, there exists a constant $C > 0$ such that, for any estimator $\hat{G}_n$,

$$\liminf_{n \to \infty} n^{p/(2\beta+1)} \sup_{G \in \mathscr{D}(M,q)} E W_p^p(\hat{G}_n, G) > C.$$

For $q = p = 1$ and a Laplace error distribution, this renders the lower bound $n^{-1/5}$, which is better than the upper bound $n^{-1/8}$ we have found, even if it is not said that either the Bayes' or the maximum likelihood estimator attains this lower bound.

Finally, a remark on the use of the term "merging". Even if this term is herein declined with a different meaning from that employed in Barron (1988), where it is understood as the convergence to one of the ratio between the marginal likelihood and the joint density of the first n observations, or from that in Diaconis and Freedman (1986), where it refers to the "intersubjective agreement", as more and more data become available, between two Bayesians having different prior opinions, the underlying idea is, in a broad sense, the same and refers to inferential procedures derived within different paradigms that will essentially be indistinguishable for large sample sizes.

5 Final remarks

In this note, we have studied rates of convergence for Bayes and maximum likelihood estimation of (convolution) Laplace mixtures as well as for their L^1-Wasserstein deconvolution. The result on the rate of convergence in the Hellinger metric for the MLE of a Laplace mixture is obtained taking a different approach from that adopted in Ghosal and van der Vaart (2001), which is based on the L^1-metric entropy with bracketing of the square root of the densities and appears difficult to apply in the present setting due to the non-analyticity of the Laplace kernel. Posterior contraction rates for Dirichlet-Laplace mixtures have been studied by Gao and van der Vaart (2016) in the case where the mixing distribution is compactly supported and have been here extended to the case of a possibly unbounded set of locations, which accounts for deriving more general entropy estimates, cf. Appendix B. An interesting extension to pursue would be that of considering general kernels with Fourier transforms having polynomially decaying tails in the sense of Definition 1. Indeed, in the proof of Proposition 2, which gives an assessement of the posterior contraction rate in the L^2-metric for Dirichlet-Laplace mixtures, all conditions, except for the Kullback-Leibler prior mass requirement, have been shown to be verified for any kernel as in Definition 1. The missing piece is an extension of Lemma 2 of Gao and van der Vaart (2016), pages 615–616, which is preliminary for checking the Kullback-Leibler prior mass condition and guarantees that a Laplace mixture, with mixing distribution that is the re-normalized restriction of G_0 to a compact interval, can be approximated in the Hellinger metric by a Laplace mixture having a carefully chosen discrete mixing distribution with a sufficiently restricted number of support points. We believe that, as for the Laplace kernel, the number of support points of the approximating mixing distribution will ultimately depend only on the decay rate of the Fourier transform of the kernel, even though the explicit expression of the kernel density cannot be exploited in the proof as for the Laplace. Extending the result on posterior contraction rates to general kernel mixtures would be of interest in itself and would allow to extend the result on their L^1-Wasserstein deconvolution, even though it would pose in general terms the question of the optimality of the rate, as for the $n^{-1/8}$-rate for Laplace deconvolution, see the remarks at the end of Section 4. We hope to report soon on this issue.

Appendix A: Auxiliary results

In this section, a sufficient condition on a convolution kernel $K \in L^1(\mathbb{R})$ is stated in terms of its Fourier transform so that the exact order of the L^2-norm error for approximating any probability density f with polynomially decaying characteristic function of order $\beta > 1/2$, see Definition 1 below, by its convolution with $K_h := h^{-1}K(\cdot/h)$, that is, by $f * K_h$, is assessed in terms of the bandwidth h. The result is instrumental to the proof of Proposition 2 to show that any mixture density $p_G = G * f$, irrespective of the mixing distribution $G \in \mathscr{G}$, verifies the *bias* condition $\|p_G * K_h - p_G\|_2 = O(h^{\beta-1/2})$ involved in the definition of the set in (15) of Theorem 2 in Giné and Nickl (2011), page 2891. We have referred to the difference $(p_G * K_h - p_G)$ as the *bias* because this is indeed the bias of the kernel density estimator $\mathbb{P}_n * K_h$, when the observations are sampled from p_G. The condition in (13) below, which traces back to Watson and Leadbetter (1963), Theorem 3, pages 486–487, is verified by any kernel K of order r greater than or equal to β, as later on commented in Remark 3.

Definition 1 *Let f be a probability density on $\mathbb{R}$. The Fourier transform of f or the characteristic function of the probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ corresponding to f, denoted by $\hat{f}$, is said to decrease algebraically of degree $\beta > 0$ if*

$$\lim_{|t|\to+\infty} |t|^{\beta}|\hat{f}(t)| = B_f, \quad 0 < B_f < +\infty. \tag{12}$$

Relationship (12) describes the tail behaviour of $\hat{f}$ which decays polynomially as $|t|^{-\beta}$. The class of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ that have characteristic functions satisfying condition (12) includes

- any gamma distribution with shape and scale parameters $\nu > 0$ and $\lambda > 0$, respectively, whose characteristic function has expression $(1 + it/\lambda)^{-\nu}$, the role of β in (12) being played by ν;
- any distribution with characteristic function $(1 + |t|^{\alpha})^{-1}$, $t \in \mathbb{R}$, for $0 < \alpha \leq 2$, which is called an *α-Laplace distribution* or *Linnik's distribution*, cf. Devroye (1990); the case $\alpha = 2$ renders the characteristic function of a Laplace distribution with density $e^{-|x|}/2$, $x \in \mathbb{R}$. The role of β in (12) is played by α;
- any distribution with characteristic function $(1 + |t|^{\alpha})^{-1/\beta}$, which, for $\beta = 1$, reduces to that of an α-Laplace distribution. The exponent α/β plays the role of the polynomial's degree β in (12). Devroye (1990) observes that, if S_{α} is any symmetric stable r.v. with characteristic function $e^{-|t|^{\alpha}}$, $0 < \alpha \leq 2$, and V_{β} is an independent r.v. with density $\Gamma(1 + 1/\beta)e^{-v^{\beta}}$, $v > 0$, then the r.v. $S_{\alpha}V_{\beta}^{\beta/\alpha}$ has characteristic function $(1 + |t|^{\alpha})^{-1/\beta}$.

Lemma 1 *Let $f \in L^2(\mathbb{R})$ be a probability density with characteristic function $\hat{f}$ satisfying condition (12) for some $\beta > 1/2$ and some constant $0 < B_f < +\infty$. If $K \in L^1(\mathbb{R})$ has Fourier transform $\hat{K}$ such that*

$$I_\beta^2[\hat{K}] := \int_{-\infty}^{+\infty} \frac{|1 - \hat{K}(t)|^2}{|t|^{2\beta}} \, \mathrm{d}t < +\infty, \tag{13}$$

then

$$h^{-2(\beta-1/2)} \|f - f * K_h\|_2^2 \to B_f^2 \times I_\beta^2[\hat{K}] \quad \text{as } h \to 0.$$

Proof Since $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and $K \in L^1(\mathbb{R})$, the norm $\|f * K_h\|_p \leq \|f\|_p \|K_h\|_1 < +\infty$ for $p = 1, 2$. Thus, $(f - f * K_h) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and, by Plancherel's Theorem, $\|f - f * K_h\|_2 = \|\hat{f} - \hat{f} \times \hat{K}_h\|_2$. By the change of variable $z = ht$,

$$\begin{aligned} \|f - f * K_h\|_2^2 &= \int_{-\infty}^{+\infty} |\hat{f}(t)|^2 |1 - \hat{K}(ht)|^2 \, \mathrm{d}t \\ &= h^{2(\beta-1/2)} \left\{ B_f^2 \times I_\beta^2[\hat{K}] + \int_{-\infty}^{+\infty} \frac{|1 - \hat{K}(z)|^2}{|z|^{2\beta}} \left[|z/h|^{2\beta} |\hat{f}(z/h)|^2 - B_f^2 \right] \mathrm{d}z \right\}, \end{aligned}$$

where, for every sequence of positive real numbers $h_n \to 0$, the integral on the right-hand side of the last display tends to zero by the dominated convergence theorem due to assumption (13). The assertion follows. □

In the following remark, which is essentially due to Davis (1977), cf. Section 3, pages 532–533, a sufficient condition is given for a kernel $K \in L^1(\mathbb{R})$ to satisfy the requirement in (13).

Remark 3 Let $K \in L^1(\mathbb{R})$. If $\beta > 1/2$, then $\int_1^{+\infty} |t|^{-2\beta} |1 - \hat{K}(t)|^2 \, \mathrm{d}t < +\infty$. The problem with the condition in (13) is the integrability of the function $t \mapsto |t|^{-2\beta} |1 - \hat{K}(t)|^2$ over the interval $(0, 1)$. Suppose that

$$\exists \, r \in \mathbb{N}\colon \int_{-\infty}^{+\infty} x^m K(x) \, \mathrm{d}x = \mathbf{1}_{\{0\}}(m) \text{ for } m = 0, 1, \ldots, r-1 \text{ and } \int_{-\infty}^{+\infty} x^r K(x) \, \mathrm{d}x \neq 0, \tag{14}$$

the value r being called the *characteristic exponent* of the Fourier transform $\hat{K}$ of K, cf. Parzen (1962), page 1072. The assumption in (14) requires that K is a *kernel of order r*, the order of a kernel being the first non-zero "moment" of the kernel, cf. Definition 1.3 in Tsybakov (2004), page 5. Then, for every real $t \neq 0$,

$$\begin{aligned} \frac{1 - \hat{K}(t)}{t^r} = -\frac{\hat{K}(t) - 1}{t^r} &= -\frac{1}{t^r} \int_{-\infty}^{+\infty} \left[e^{itx} - \sum_{j=0}^{r-1} \frac{(itx)^j}{j!} \right] K(x) \, \mathrm{d}x \\ &= -\frac{i^r}{(r-1)!} \int_{-\infty}^{+\infty} x^r K(x) \int_0^1 (1-u)^{r-1} e^{itux} \, \mathrm{d}u \, \mathrm{d}x. \end{aligned}$$

Defined

$$\kappa_r := -\frac{i^r}{r!} \int_{-\infty}^{+\infty} x^r K(x) \, \mathrm{d}x,$$

which is called the *characteristic coefficient* of the Fourier transform $\hat{K}$ of K, cf. Parzen (1962), page 1072–1073, under the assumption on K stated in (14), we have $\kappa_r \neq 0$. Then,

$$\frac{1 - \hat{K}(t)}{t^r} \to \kappa_r \text{ as } t \to 0.$$

For $r \geq \beta$, the integral $\int_0^1 |t|^{-2\beta} |1 - \hat{K}(t)|^2 \, \mathrm{d}t < +\infty$. Conversely, for $r < \beta$, the integral diverges. If, for instance, $1/2 < \beta \leq 2$, then any symmetric probability density K on $\mathbb{R}$, with finite second moment $\mu_2 := \int_{-\infty}^{+\infty} x^2 K(x) \, \mathrm{d}x = 2\kappa_2 \neq 0$, is such that $I_\beta^2[\hat{K}] < +\infty$ and condition (13) is satisfied.

Appendix B: Entropy estimates

In this section, Hellinger and L^1-metric entropy estimates for a class of Laplace mixtures with mixing distributions having tails dominated by a given decreasing function are provided. The result extends, along the lines of Theorem 7 of Ghosal and van der Vaart (2007), page 708-709, Proposition 2 of Gao and van der Vaart (2016), page 617, which deals with Laplace mixtures having compactly supported mixing distributions. The lemma is invoked in the proof of Proposition 1, reported in Appendix C, to verify that the entropy condition is satisfied.

Lemma 2 *For a given decreasing function $A : (0, +\infty) \to [0, 1]$, with inverse A^{-1}, define the class of Laplace mixture densities*

$$\mathscr{P}_A := \{p_G : G([-a, a]^c) \leq A(a) \text{ for all } a > 0\}.$$

Then, for every $0 < \varepsilon < 1$,

– *taking $a \equiv a_\varepsilon := A^{-1}(\varepsilon)$ in the definition of $\mathscr{P}_A$, we have*

$$\log N(3\varepsilon, \mathscr{P}_A, \|\cdot\|_1) \lesssim \varepsilon^{-2/3} \log \frac{A^{-1}(\varepsilon)}{\varepsilon^2}, \tag{15}$$

– *taking $a \equiv a_{\varepsilon^2} := A^{-1}(\varepsilon^2)$ in the definition of $\mathscr{P}_A$, we have*

$$\log N((\sqrt{2}+1)\varepsilon, \mathscr{P}_A, h) \lesssim \varepsilon^{-2/3} \log \frac{A^{-1}(\varepsilon^2)}{\varepsilon^2}. \tag{16}$$

Proof Concerning the L^1-metric entropy in (15), since $a \equiv a_\varepsilon := A^{-1}(\varepsilon)$ satisfies $G([-a_\varepsilon, a_\varepsilon]^c) \leq A(a_\varepsilon) = \varepsilon$ for all G as in the definition of $\mathscr{P}_A$, Lemma A.3 of Ghosal and van der Vaart (2001), page 1261, implies that the L^1-distance between any density $p_G \in \mathscr{P}_A$ and the corresponding density p_{G^*}, with mixing distribution G^* defined as the re-normalized restriction of G to $[-a_\varepsilon, a_\varepsilon]$, is bounded above by 2ε. Then, in virtue of the inequality $\|f - g\|_1 \leq 2h(f, g)$ in (2), a Hellinger $(\varepsilon/2)$-net over the class of densities $\mathscr{P}_{a_\varepsilon} := \{p_G : G([-a_\varepsilon, a_\varepsilon]) = 1\}$ is an L^1-metric 3ε-net over $\mathscr{P}_A$, where

$$\log N(\varepsilon/2, \mathscr{P}_{a_\varepsilon}, h) \lesssim \varepsilon^{-2/3} \log \frac{a_\varepsilon}{\varepsilon^2}$$

because of Proposition 2 of Gao and van der Vaart (2016), pages 617-618. The inequality in (15) follows.

Concerning the Hellinger-metric entropy in (16), by taking $a \equiv a_{\varepsilon^2} := A^{-1}(\varepsilon^2)$, for every $p_G \in \mathscr{P}_A$ and the corresponding p_{G^*}, with mixing distribution G^* defined as the re-normalized restriction of G to $[-a_{\varepsilon^2}, a_{\varepsilon^2}]$, by the inequality $h^2(f, g) \leq \|f - g\|_1$ in (1), we have $h^2(p_G, p_{G^*}) \leq \|p_G - p_{G^*}\|_1 \leq 2G([-a_{\varepsilon^2}, a_{\varepsilon^2}]^c) \leq 2\varepsilon^2$, which implies that $h(p_G, p_{G^*}) \leq \sqrt{2}\varepsilon$. Thus, a Hellinger ε-net over $\mathscr{P}_{a_{\varepsilon^2}}$ is a $(\sqrt{2}+1)\varepsilon$-net over $\mathscr{P}_A$, where

$$\log N(\varepsilon, \mathscr{P}_{a_{\varepsilon^2}}, h) \lesssim \varepsilon^{-2/3} \log \frac{a_{\varepsilon^2}}{\varepsilon^2}$$

again by Proposition 2 of Gao and van der Vaart (2016), pages 617-618. The inequality in (16) follows. □

Appendix C: Posterior contraction rates in L^r-metrics, $1 \leq r \leq 2$, for Dirichlet-Laplace mixtures

In this section, we prove Propositions 1 and 2 of Section 2 on contraction rates in the Hellinger and the L^2-metric, respectively, for the posterior distribution corresponding to a Dirichlet process mixture of Laplace densities.

Proof of Proposition 1 In order to derive the Hellinger or the L^1-metric posterior contraction rate, we appeal to Theorem 2.1 of Ghosal *et al*. (2000), page 503, or Theorem 2.1 of Ghosal and van der Vaart (2001), page 1239. We define a sieve set for which conditions (2.2) or (2.8) and (2.3) or (2.9), postulated in the abovementioned theorems, are satisfied. To the aim, let $\bar{\alpha} := \alpha/\alpha(\mathbb{R})$ be the probability measure corresponding to the baseline measure of the Dirichlet process. Consistently with the notation adopted throughout the manuscript, $\bar{\alpha}$ is also used to denote the corresponding cumulative distribution function. By a result of Doss and Sellke (1982), page

1304, which concerns the tails of probability measures chosen from a Dirichlet prior, we have that, for almost every sample distribution $G \in \mathscr{G}$, if $a > 0$ is large enough so that $\bar{\alpha}(-a) = 1 - \bar{\alpha}(a)$ is sufficiently small, then

$$\begin{aligned} G([-a, a]^c) &\leq G(-a) + 1 - G(a) \\ &\leq \exp\left\{ - \frac{1}{\bar{\alpha}(-a)|\log \bar{\alpha}(-a)|^2} \right\} + \exp\left\{ - \frac{1}{[1 - \bar{\alpha}(a)]|\log[1 - \bar{\alpha}(a)]|^2} \right\} \\ &\leq 2 \exp\left\{ - \frac{1}{\bar{\alpha}(-a)\,|\log \bar{\alpha}(-a)|^2} \right\} \\ &\leq A_\eta(a), \end{aligned}$$

having set the position $A_\eta(a) := 2\exp\{-[\bar{\alpha}(-a)]^{-\eta}\}$ for some fixed $0 < \eta < 1$. The inverse function $A_\eta^{-1} : (0, 1) \to (0, +\infty)$ is defined as $A_\eta^{-1} : y \mapsto -\bar{\alpha}^{-1}((y^{-1}\log 2)^{-1/\eta})$, where $\bar{\alpha}^{-1}(u) := \inf\{x \in \mathbb{R} : \bar{\alpha}(x) \geq u\}$, $u \in (0, 1)$. Considered the class of densities $\mathscr{P}_{A_\eta} := \{p_G : G([-a, a]^c) \leq A_\eta(a) \text{ for all } a > 0\}$, we have $\Pi(\mathscr{P}_{A_\eta}) = 1$. For any sequence of positive real numbers $\bar{\varepsilon}_n \downarrow 0$, set the position $a \equiv a_{\bar{\varepsilon}_n} := A_\eta^{-1}(\bar{\varepsilon}_n)$ and defined the sieve set $\mathscr{P}_n := \{p_G : G([-a_{\bar{\varepsilon}_n}, a_{\bar{\varepsilon}_n}]^c) \leq A_\eta(a_{\bar{\varepsilon}_n}) = \bar{\varepsilon}_n\}$, we have

$$\Pi(\mathscr{P} \setminus \mathscr{P}_n) = 0$$

and condition (2.3) or (2.9) is satisfied. As for condition (2.2) or (2.8), taking $\bar{\varepsilon}_n = n^{-3/8}\log^{3/8} n$, by Lemma 2, we have

$$\log D(\bar{\varepsilon}_n, \mathscr{P}_n, \|\cdot\|_1) \leq \log N(\bar{\varepsilon}_n/2, \mathscr{P}_n, \|\cdot\|_1) \lesssim (\bar{\varepsilon}_n)^{-2/3} \log \frac{A_\eta^{-1}(\bar{\varepsilon}_n/6)}{\bar{\varepsilon}_n^2} \lesssim n\bar{\varepsilon}_n^2. \tag{17}$$

The same bound as in (17) also holds for the Hellinger metric entropy. The Kullback-Leibler prior mass condition (2.4) of Theorem 2.1 of Ghosal *et al*. (2000), page 503, or, equivalently, condition (2.10) of Theorem 2.1 of Ghosal and van der Vaart (2001), page 1239, is satisfied for $\tilde{\varepsilon}_n = n^{-3/8}\log^{1/2} n$. For the verification of condition (2), we refer the reader to Proposition 2 below, whose requirements (5) and (6) are satisfied under assumptions (3) and (4), respectively, of Proposition 1. The proof is completed by taking $\varepsilon_n := \max\{\bar{\varepsilon}_n, \tilde{\varepsilon}_n\} = n^{-3/8}\log^{1/2} n$. □

We now prove Proposition 2 on the posterior contraction rate in the L^2-metric. The result relies on Theorem 3 of Giné and Nickl (2011), page 2892, which gives sufficient conditions for deriving posterior contraction rates in L^r-metrics, $1 < r < +\infty$. All assumptions of Theorem 3, except for condition (2), are shown to be satisfied for any kernel density f as in Definition 1 with $\beta \geq 1$. This includes the Laplace kernel density as a special case when $\beta = 2$. Condition (2), which requires the prior mass in Kullback-Leibler type neighborhoods of the sampling density $p_0 \equiv p_{G_0} = G_0 * f$ to be not exponentially small, relies on a preliminary approximation result of the mixture density, having mixing distribution obtained as the re-normalized restriction of the true mixing distribution G_0 to a compact interval, by the mixture having as a mixing distribution a carefully chosen discrete distribution with a sufficiently restricted number of support points. This result is known to hold for the Laplace kernel density in virtue of Lemma 2 of Gao and van der Vaart (2016), pages 615–616.

Proof of Proposition 2 We apply Theorem 3 of Giné and Nickl (2011), page 2892, with $r = 2$. We refer to the conditions of this theorem using the same letters/numbers as in the original article. Let $\gamma_n \equiv 1$ and $\delta_n \equiv \varepsilon_n := n^{-3/8}\log^{1/2} n$, $n \in \mathbb{N}$.

– *Verification of condition* (b):
Condition (b), which requires that $\varepsilon_n^2 = O(n^{-1/2})$, is satisfied in the general case for $\varepsilon_n = n^{-(\beta-1/2)/2\beta}\log^\kappa n$, with some $\kappa > 0$ and $\beta \geq 1$.

– *Verification of condition* (1):
Condition (1) requires that the prior probability of the complement of a sieve set $\mathscr{P}_n$ is exponentially small. We show that, in the present setting, the prior probability of a sieve set $\mathscr{P}_n$, chosen as prescribed by (15) in Theorem 2 of Giné and Nickl (2011), page 2891, is equal to zero. Let J_n be any sequence of positive real numbers satisfying $2^{J_n} \leq cn\varepsilon_n^2$ for some fixed constant $0 < c < +\infty$. Let K be a convolution kernel such that it is of bounded p-variation for some finite real number $p \geq 1$, right (or left) continuous and satisfies $\|K\|_\infty < +\infty$, $\int_{-\infty}^{+\infty}(1 + |z|)^w K(z)\,\mathrm{d}z < +\infty$ for some $w > 2$ and $I_\beta^2[\hat{K}] < +\infty$, cf. condition (13) in Lemma 1. Defined the sieve set

$$\mathscr{P}_n := \{p_G, G \in \mathscr{G} : \|p_G - p_G * K_{2^{-J_n}}\|_2 \leq C\delta_n\},$$

where $K_{2^{-J_n}}(\cdot) := 2^{J_n}K(\cdot 2^{J_n})$ and $C > 0$ is a finite constant depending only on K and f, we have

$$\Pi(\mathscr{P} \setminus \mathscr{P}_n) = 0 \quad \text{for all } n \in \mathbb{N}.$$

In fact, for every $G \in \mathscr{G}$, by Plancherel's Theorem, $\|p_G - p_G * K_{2^{-J_n}}\|_2^2 = \|\hat{p}_G - \hat{p}_G \times \hat{K}_{2^{-J_n}}\|_2^2 \leq \|\hat{f} - \hat{f} \times \hat{K}_{2^{-J_n}}\|_2^2$ and, by Lemma 1, $\|\hat{f} - \hat{f} \times \hat{K}_{2^{-J_n}}\|_2^2 \sim (2^{-J_n})^{2\beta-1} B_f^2 \times I_\beta^2[\hat{K}]$, where, for $\beta = 2$, we have $(2^{-J_n})^{2\beta-1} = (2^{-J_n})^3 = O(\delta_n^2)$. Thus,

$$\forall\, G \in \mathscr{G}, \quad \|p_G - p_G * K_{2^{-J_n}}\|_2 = O(\delta_n) \tag{18}$$

and condition (1) is verified. Relationship (18) also holds for $p_0 \equiv p_{G_0}$. Furthermore, $p_0 \in L^2(\mathbb{R})$ if $f \in L^2(\mathbb{R})$, which is the case for the Laplace kernel density, because $\|p_0\|_2 = \|\hat{p}_0\|_2 = \|\hat{G}_0 \times \hat{f}\|_2 \leq \|\hat{f}\|_2 = \|f\|_2$.

– *Verification of condition* (2):
Condition (2) requires that, for some finite constant $C > 0$, the prior probability of Kullback-Leibler type neighborhoods of p_0 of radius ε_n^2 is at least $e^{-Cn\varepsilon_n^2}$, that is, $\Pi(B_{\mathrm{KL}}(p_0; \varepsilon_n^2)) \gtrsim \exp(-Cn\varepsilon_n^2)$. Fix $0 < \varepsilon \leq (1 - e^{-1})/\sqrt{2}$ and let $a_\varepsilon := A_0^{-1}(\varepsilon^2)$, where A_0^{-1} is the inverse of the function A_0 in (5). Define G_0^* as the re-normalized restriction of G_0 to $[-a_\varepsilon, a_\varepsilon]$. By Lemma A.3 of Ghosal and van der Vaart (2001), page 1261, and assumption (5), we have $\|p_{G_0} - p_{G_0^*}\|_1 \lesssim G_0([-a_\varepsilon, a_\varepsilon]^c) \lesssim \varepsilon^2$. From the inequality $h^2(f, g) \leq \|f - g\|_1$ in (1), we have $h^2(p_{G_0}, p_{G_0^*}) \leq \|p_{G_0} - p_{G_0^*}\|_1 \lesssim \varepsilon^2$, whence $h(p_{G_0}, p_{G_0^*}) \lesssim \varepsilon$. It is known from Lemma 2 of Gao and van der Vaart (2016), pages 615–616, that there exists a discrete distribution G_0' such that $h(p_{G_0'}, p_{G_0^*}) \lesssim \varepsilon$. The distribution G_0' matches the moments of G_0^* up to the order $N \lesssim \varepsilon^{-2/3}$ and has at most N support points $\theta_1, \ldots, \theta_N$ in $[-a_\varepsilon, a_\varepsilon]$, which we may assume to be at least $2\varepsilon^2$-separated. If not, we can take a maximal $2\varepsilon^2$-separated set in the support points of G_0' and replace G_0' with the discrete measure G_0'' obtained by relocating the masses of G_0' to the nearest points of the $2\varepsilon^2$-net. Then, as shown in Proposition 2 of Gao and van der Vaart (2016), page 617, we have $h(p_{G_0'}, p_{G_0''}) \lesssim (\max_{1\leq j\leq N} |\theta_j' - \theta_j''|)^2 \lesssim \varepsilon^4 \lesssim \varepsilon^2$. Let $G_0' = \sum_{j=1}^N p_j \delta_{\theta_j}$, with $|\theta_j - \theta_k| \geq 2\varepsilon^2$ for all $1 \leq j \neq k \leq N$. For any distribution G on $\mathbb{R}$ such that

$$\sum_{j=1}^N |G([\theta_j - \varepsilon^2, \theta_j + \varepsilon^2]) - p_j| \leq \varepsilon^2, \tag{19}$$

we have $\|p_G - p_{G_0'}\|_1 \lesssim \varepsilon^2$ by Lemma 5 of Gao and van der Vaart (2016), page 620. Thus,

$$\begin{aligned} h^2(p_G, p_{G_0}) &\leq h^2(p_G, p_{G_0'}) + h^2(p_{G_0'}, p_{G_0^*}) + h^2(p_{G_0^*}, p_{G_0}) \\ &\lesssim \|p_G - p_{G_0'}\|_1 + \varepsilon^2 + \|p_{G_0^*} - p_{G_0}\|_1 \lesssim \varepsilon^2. \end{aligned}$$

We can now invoke Lemma A.10 in Scricciolo (2011), page 305, taking into account Remark A.3 of the same article. To this aim, note that, if G satisfies (19), then $G([-(a_\varepsilon + 1), (a_\varepsilon + 1)]) > 1/2$. The inclusion

$$\left\{G : \sum_{j=1}^N |G([\theta_j - \varepsilon^2, \theta_j + \varepsilon^2]) - p_j| \leq \varepsilon^2\right\} \subseteq B_{\mathrm{KL}}(p_0; \varepsilon^2 |\log \varepsilon|)$$

holds. To apply Lemma A.2 of Ghosal and van der Vaart (2001), page 1260, note that, for every θ_j, we have $\alpha([\theta_j - \varepsilon^2, \theta_j + \varepsilon^2]) \gtrsim \varepsilon^2 A_0(a_\varepsilon) \gtrsim \varepsilon^4$, $j = 1, \ldots, N$. Thus, for suitable constants $0 < c, c_1 < +\infty$,

$$\Pi(B_{\mathrm{KL}}(p_0; \varepsilon^2 |\log \varepsilon|)) \gtrsim \exp(-cN |\log \varepsilon|) \gtrsim \exp(-c_1 \varepsilon^{-2/3} |\log \varepsilon|).$$

Taking $\varepsilon_n := \varepsilon |\log \varepsilon|^{1/2}$, we have $\Pi(B_{\mathrm{KL}}(p_0; \varepsilon_n^2)) \gtrsim \exp(-c_1 n \varepsilon_n^2)$ and condition (2) is satisfied.

– *Verification of condition* (3):
Condition (3) requires that there exists a finite constant $B > 0$ such that $\Pi(G : \|p_G\|_\infty > B \mid X^{(n)}) = o_{\mathbf{P}}(1)$. If $\|f\|_\infty < +\infty$, then $\|p_G\|_\infty \leq \|f\|_\infty < +\infty$ for all $G \in \mathscr{G}$. Taking $B = \|f\|_\infty$, we have

$$\forall\, n \in \mathbb{N}, \;\; \Pi(G : \|p_G\|_\infty > B \mid X^{(n)}) = 0 \quad P_0^n\text{-almost surely},$$

and condition (3) is satisfied. For the Laplace kernel density, we have $\|f\|_\infty = 1/2$.

The proof is thus complete and assertion (7) follows. □

Appendix D: Inversion inequalities

In this section, we state a result relating, for every $p \geq 1$, the L^p-Wasserstein distance between any pair of mixing distributions G, $G' \in \mathscr{G}$ to the L^2-distance between the corresponding mixed densities $p_G = G * f$ and $p_{G'} = G' * f$ with an ordinary smooth kernel f, see condition (21) below. The result extends Lemma 7 of Gao and van der Vaart (2016), pages 621-622, beyond the case of compactly supported mixing distributions to mixing distributions with finite moment generating functions on an interval $(-s_0, s_0)$, with $s_0 > 1$. If, furthermore, the kernel density f is bounded, $\|f\|_\infty < +\infty$, then the inversion inequality in (22) below also holds for the Hellinger distance because of the following known result, which is reported for the reader's convenience.

Lemma 3 *For given kernel density f, let $p_G = G * f$, with $G \in \mathscr{G}$. If $\|f\|_\infty < +\infty$, then*

$$\forall\, G \in \mathscr{G}, \quad p_G(x) \leq \|f\|_\infty \ \text{ for all } x \in \mathbb{R},$$

and

$$\forall\, G,\, G' \in \mathscr{G}, \quad \|p_G - p_{G'}\|_2^2 \leq 4\|f\|_\infty h^2(p_G,\, p_{G'}). \tag{20}$$

We now state and prove an inequality translating the Hellinger or the L^2-distance between mixed densities into any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions.

Lemma 4 *Let G and G' be probability measures on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, $\mathscr{Y} \subseteq \mathbb{R}$, such that the associated moment generating functions $M_G(s)$ and $M_{G'}(s)$ are finite for all $|s| < s_0$, with $s_0 > 1$. Let f be a probability density on $\mathbb{R}$ satisfying, for some real number $\beta > 0$, the condition*

$$\inf_{t\in\mathbb{R}}(1 + |t|^\beta)|\hat{f}(t)| > 0. \tag{21}$$

Let d stand for either the Hellinger or the L^2-distance between the mixed densities $p_G = G * f$ and $p_{G'} = G' * f$. Then, for any real number $p \geq 1$,

$$W_p(G,\, G') \lesssim \left(d \log \frac{1}{d}\right)^{1/(p+\beta)} \qquad \text{with } d = \|p_G - p_{G'}\|_2 \text{ small enough.} \tag{22}$$

If, furthermore, $\|f\|_\infty < +\infty$, then the bound in (22) *also holds for d being the Hellinger distance, $d = h(p_G,\, p_{G'})$.*

Proof For any real number $\delta > 0$, by the triangle inequality, we have

$$W_p^p(G,\, G') \leq W_p^p(G,\, G * \Phi_\delta) + W_p^p(G * \Phi_\delta,\, G' * \Phi_\delta) + W_p^p(G' * \Phi_\delta,\, G'), \tag{23}$$

where Φ_δ stands for a zero-mean Gaussian probability measure with variance δ^2, whose density is denoted by $\phi_\delta(\cdot) := \delta^{-1}\phi(\cdot/\delta)$, for ϕ the density of a standard normal r.v. Z. The first and third terms on the right-hand side of (23) can be bounded above as follows. By standard arguments, see, for instance, the proof of Theorem 2 in Nguyen (2013), pages 389–391,

$$\max\{W_p^p(G,\, G * \Phi_\delta),\, W_p^p(G' * \Phi_\delta,\, G')\} \leq E[|\delta Z|^p] \lesssim \delta^p,$$

where $E[|Z|^p] < +\infty$ for every real number $p > 0$, hence, for $p \geq 1$. Concerning the second term on the right-hand side of (23), reasoning as in Lemma 7 of Gao and van der Vaart (2016), pages 621–622, for any real number $M > 0$,

$$W_p^p(G * \Phi_\delta,\, G' * \Phi_\delta) \lesssim \left(\int_{|x|\leq M} + \int_{|x|>M}\right)|x|^p|(G - G') * \phi_\delta(x)|\,\mathrm{d}x =: T_1 + T_2,$$

where

$$T_1 \lesssim M^{p+1/2}\|(G - G') * \phi_\delta\|_2 \lesssim M^{p+1/2}\delta^{-\beta}\|p_G - p_{G'}\|_2, \tag{24}$$

because $\sup_{t\in\mathbb{R}}|\hat{\phi}(\delta t)|/|\hat{f}(t)| \lesssim \delta^{-\beta}$ in virtue of assumption (21), which implies the existence of a finite constant $B_f > 0$ such that $(1 + |t|^\beta)|\hat{f}(t)| \geq B_f$ for all $t \in \mathbb{R}$. If, furthermore, $\|f\|_\infty < +\infty$, then the L^2-distance between p_G and $p_{G'}$ in (24) can be replaced with the Hellinger distance, see Lemma 3, so that

$$T_1 \lesssim M^{p+1/2}\delta^{-\beta}h(p_G,\, p_{G'}).$$

We now deal with the term T_2. By applying the inequality, valid for every real number $p > 0$,

$$|x|^p \leq e^{|x|} < (e^x + e^{-x}), \quad x \in \mathbb{R},$$

and taking into account the expression of the moment generating function of a standard Gaussian distribution $M_\Phi(s) = e^{s^2/2}$, $s \in \mathbb{R}$, we have

$$\int_{-\infty}^{+\infty} \max\{1,\ |x|^p\} e^{|x|} \phi_\delta(x)\,\mathrm{d}x \leq E e^{2\delta|Z|} < E[e^{2\delta Z} + e^{-2\delta Z}] = 2e^{(2\delta)^2/2},$$

so that

$$\begin{aligned} T_2 &\lesssim e^{-M} \int_{|x|>M} |x|^p e^{|x|} [(G + G') * \phi_\delta(x)]\,\mathrm{d}x \\ &\lesssim e^{-M} \int_{\mathscr{Y}} (1 + |y|^p) e^{|y|} \left(\int_{-\infty}^{+\infty} \max\{1,\ |x|^p\} e^{|x|} \phi_\delta(x)\,\mathrm{d}x \right) \mathrm{d}(G + G')(y) \\ &\lesssim e^{-M} e^{2\delta^2} \int_{\mathscr{Y}} (1 + |y|^p) e^{|y|}\,\mathrm{d}(G + G')(y) \lesssim e^{-M}, \end{aligned}$$

because, for $\delta > 0$ small enough, the term $e^{2\delta^2}$ is bounded above by a constant, the integral

$$\int_{\mathscr{Y}} e^{|y|}\,\mathrm{d}(G + G')(y) < \int_{\mathscr{Y}} (e^y + e^{-y})\,\mathrm{d}(G + G')(y) = (M_G + M_{G'})(-1) + (M_G + M_{G'})(1) < +\infty$$

and, for $0 < \epsilon < 1$ such that $1 + \epsilon < s_0$,

$$\begin{aligned} \int_{\mathscr{Y}} |y|^p e^{|y|}\,\mathrm{d}(G + G')(y) &< \int_{\mathscr{Y}} e^{(1+\epsilon)|y|}\,\mathrm{d}(G + G')(y) \\ &< \int_{\mathscr{Y}} (e^{(1+\epsilon)y} + e^{-(1+\epsilon)y})\,\mathrm{d}(G + G')(y) \\ &= (M_G + M_{G'})(-(1+\epsilon)) + (M_G + M_{G'})(1+\epsilon) < +\infty \end{aligned}$$

by the assumption that both G and G' have finite moment generating functions on $(-s_0,\ s_0)$, for some $s_0 > 1$. Combining partial results, we have

$$W_p^p(G,\ G') \lesssim \delta^p + M^{p+1/2} \delta^{-\beta} d + e^{-M} \tag{25}$$

and the conclusion follows, for sufficiently small d, by minimizing the expression in (25) with respect to δ and M, which implies taking $M = O(\log(1/d))$ and $\delta^{p+\beta} = O(d \log^{p+\beta}(1/d))$. □

Remark 4 The Laplace kernel density is bounded, with $\|f\|_\infty = 1/2$, and satisfies condition (21) with $\beta = 2$.

References

Ball K, Pajor A (1990) The entropy of convex bodies with "few" extreme points. Geometry of Banach spaces (Strobl, 1989), 25–32, London Math. Soc. Lecture Note Ser., 158, Cambridge Univ. Press, Cambridge

Barron, AR (1988) The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report #7, University of Illinois at Urbana-Champaign

Birgé L, Massart P (1993) Rates of convergence for minimum contrast estimators. Probability Theory and Related Fields 97(1): 113–150

Dattner I, Goldenshluger A, Juditsky A (2011) On deconvolution of distribution functions. The Annals of Statistics 39(5): 2477–2501

Davis KB (1977) Mean integrated square error properties of density estimates. The Annals of Statistics 5(3): 530–535

Dedecker J, Fischer A, Michel B (2015) Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. Electronic Journal of Statistics 9(1): 234–265

Devroye L (1990) A note on linnik's distribution. Statistics & Probability Letters 9(4): 305–306

Diaconis P, Freedman D (1986) On the consistency of Bayes estimates. The Annals of Statistics 14(1): 1–26

Doss H, Sellke T (1982) The tails of probabilities chosen from a Dirichlet prior. The Annals of Statistics 10(4): 1302–1305

Donnet S, Rivoirard V, Rousseau J, Scricciolo C (2014) Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Forthcoming in* Bernoulli

Fan J (1991) On the optimal rates of convergence for nonparametric deconvolution problems. The Annals of Statistics 19(3): 1257–1272
Gao F, van der Vaart A (2016) Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. Electronic Journal of Statistics 10(1): 608–627
Ghosal S, Ghosh JK, van der Vaart AW (2000) Convergence rates of posterior distributions. The Annals of Statistics 28(2): 500–531
Ghosal S, van der Vaart A (2007) Posterior convergence rates of Dirichlet mixtures at smooth densities. The Annals of Statistics 35(2): 697–723
Ghosal S, van der Vaart AW (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. The Annals of Statistics 29(5): 1233–1263
Giné E, Nickl R (2011) Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. The Annals of Statistics 39(6): 2883–2911
Hall P, Lahiri SN (2008) Estimation of distributions, moments and quantiles in deconvolution problems. The Annals of Statistics 36(5): 2110–2134
Lindsay BG (1995) Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics, Hayward, CA
Meister A (2009) Deconvolution Problems in Nonparametric Statistics. Lecture Notes in Statistics 193. Springer-Verlag Berlin Heidelberg
Nguyen X (2013) Convergence of latent mixing measures in finite and infinite mixture models. The Annals of Statistics 41(1): 370–400
Parzen, E (1962) On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3): 1065–1076
Scricciolo C (2011) Posterior rates of convergence for Dirichlet mixtures of exponential power densities. Electronic Journal of Statistics 5: 270–308
Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley, New York
Tsybakov AB (2004) Introduction à l'estimation non-paramétrique. Springer-Verlag, Berlin
Van de Geer S (1993) Hellinger-consistency of certain nonparametric maximum likelihood estimators. The Annals of Statistics 21(1): 14–44
Van de Geer S (1996) Rates of convergence for the maximum likelihood estimator in mixture models. Journal of Nonparametric Statistics 6(4): 293–310
Van de Geer SA (2000) Empirical processes in M-estimation. Cambridge University Press
Watson GS, Leadbetter MR (1963) On the estimation of the probability density, I. The Annals of Mathematical Statistics 34(2): 480–491
Wong WH, Shen X (1995) Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. The Annals of Statistics 23(2): 339–362