Bayesian Analysis (2014) **9**, Number 2, pp. 475–520

Adaptive Bayesian Density Estimation in L^p-metrics with Pitman-Yor or Normalized Inverse-Gaussian Process Kernel Mixtures

Catia Scricciolo *

Abstract. We consider Bayesian nonparametric density estimation using a Pitman-Yor or a normalized inverse-Gaussian process convolution kernel mixture as the prior distribution for a density. The procedure is studied from a frequentist perspective. Using the stick-breaking representation of the Pitman-Yor process and the finite-dimensional distributions of the normalized inverse-Gaussian process, we prove that, when the data are independent replicates from a density with analytic or Sobolev smoothness, the posterior distribution concentrates on shrinking L^p-norm balls around the sampling density at a minimax-optimal rate, up to a logarithmic factor. The resulting hierarchical Bayesian procedure, with a fixed prior, is adaptive to the unknown smoothness of the sampling density.

Keywords: adaptation, nonparametric density estimation, normalized inverse-Gaussian process, Pitman-Yor process, posterior contraction rate, sinc kernel

1 Introduction

Consider the problem of estimating a univariate density f_0 from independent and identically distributed (i.i.d.) observations taking a Bayesian nonparametric approach. A prior probability law is defined on a metric space of probability measures that possess Lebesgue densities and a summary of the posterior distribution, typically the posterior expected density, can be employed as a density estimator. Since the seminal articles of Ferguson (1983) and Lo (1984), the idea of constructing priors on spaces of densities by convoluting a fixed kernel with a random distribution has been successfully exploited in density estimation. A convolution kernel mixture may provide an efficient approximation scheme possibly resulting in a minimax-optimal, up to a logarithmic factor, speed of concentration for the posterior mass on shrinking balls around the sampling density. Recent literature on Bayesian density estimation has mainly focussed on posterior contraction rates relative to the Hellinger or the L^1-metric using Dirichlet process mixtures of (generalized) normal densities. Ghosal and van der Vaart (2001) found a nearly parametric rate for estimating *supersmooth* densities that are themselves mixtures of normal densities. Supersmooth cases, beyond being of interest in themselves, help developing mathematical tools to deal with the estimation of ordinary smooth densities, *i.e.*, densities that are differentiable up to a certain order, but not necessarily are kernel mixtures. In the article of Ghosal and van der Vaart (2007b), a twice continuously differentiable density f_0 is estimated using a Dirichlet process mixture of

*Bocconi University catia.scricciolo@unibocconi.it

 DOI:10.1214/14-BA863

Gaussian densities, the scale parameter, which plays the role of the smoothing window, being assigned a sample-size dependent prior obtained by re-scaling a fixed distribution with an accurately calibrated sequence converging to zero at an appropriate rate so that the *a priori* smoothness assumption on f_0 is incorporated into the prior. When the smoothness is unknown, rate-adaptive estimation over Hölder classes can be performed using *finite* Dirichlet location mixtures of Gaussian densities, cf. Kruijer *et al.* (2010). Extending this result to *infinite* mixtures, Shen *et al.* (2013) have recently proved that fully rate-adaptive multivariate density estimation over Hölder regularity scales can be performed using Dirichlet process mixtures of Gaussian densities without any bandwidth shrinkage in the prior for the scale nor any knowledge of the smoothness level.

Even if much progress has been made during the last decade in understanding frequentist asymptotic properties of kernel mixture models for Bayesian density estimation, there seems to be a lack of results concerning adaptive estimation of ordinary and infinitely smooth densities with respect to more general loss functions than the Hellinger or the L^1-distance, employing other processes, beyond the Dirichlet process, as priors for the mixing distribution. In this article, we investigate the question of how to complement and generalize existing results on posterior contraction rates by considering adaptive estimation over analytic or Sobolev density function spaces using the Pitman-Yor or the normalized inverse-Gaussian process as priors for the mixing distribution of Gaussian mixtures.

The main results describe recovery rates for smooth densities, where smoothness is measured through a scale of integrated tail bounds on the Fourier transform of the sampling density. For analytic densities, a nearly parametric rate stems under various prior laws that may only affect the power of the logarithmic term, which automatically recovers the characteristic exponent of the Fourier transform. Such a fast rate is roughly explainable from the fact that spaces of analytic functions are slightly bigger than finite-dimensional spaces in terms of metric entropy. Apart from the prior probability measures considered, the novelty of this article is in the use of stronger metrics to measure recovery rates, namely, the full scale of L^p-metrics, $1 \leq p \leq \infty$. That a large class of Bayesian procedures is capable of such a recovery is established here for the first time and is encouraging to the use of these methods. For densities in Sobolev spaces, recovery rates are found to be minimax-optimal, up to a logarithmic factor, under the Dirichlet process for L^p-metrics, with $1 \leq p \leq 2$, whereas they deteriorate by a genuine power of n as p increases beyond 2. Slower than minimax-optimal rates are found when endowing the mixing distribution with a Pitman-Yor process having a strictly positive discount parameter since small Kullback-Leibler type balls do not seem to be charged enough prior mass. We currently have no proof of the fact that posterior contraction rates are suboptimal under a Pitman-Yor process with strictly positive discount parameter, but believe that the rates cannot be substantially improved in this situation. The results of this article may be of interest for different reasons: they constitute a first step towards the study of posterior contraction rates for other process priors, beyond the Dirichlet process, recently proposed in the literature which, in many contexts, can be better suited than the Dirichlet process for the analysis of data in

a variety of applied settings, as witnessed by the burst of the use of the Pitman-Yor process in the machine learning community. Also, they provide an indication on the performance of Bayesian procedures for adaptive density estimation over function spaces extensively considered in the frequentist literature on nonparametric curve estimation.

The main challenge when proving adaptation in the infinitely smooth case lies in finding a finite mixing distribution, with a relatively small number of support points, such that the corresponding Gaussian mixture approximates the sampling density, in the Kullback-Leibler divergence, with an error of the appropriate order. Such a finitely supported mixing distribution may be found by matching the moments of an *ad hoc* constructed mixing density for which the method used by Kruijer *et al.* (2010) is not suited because of the infinite degree of smoothness of the true density. There are limitations implicitly coming from the employed kernel which can be by-passed using superkernels, whose usefulness in density estimation has been pointed out by, among others, Devroye (1992). The crux and a main contribution of this article is the development of an approximation result for analytic densities with exponentially decaying Fourier transforms, cf. Lemma 5. We believe this result can also be of autonomous interest for frequentist methods in adaptive density estimation for clustering with Gaussian mixtures along the lines of the article by Maugis-Rabusseau and Michel (2013).

When assessing posterior contraction rates, a major difficulty is the evaluation of the prior concentration rate, estimated by bounding below the probability of Kullback-Leibler type neighborhoods of the sampling density by the probability of ℓ^1-balls of appropriate dimension. For the normalized inverse-Gaussian process, likewise the Dirichlet process, the explicit expressions of the finite-dimensional distributions can be exploited to estimate the probability of ℓ^1-balls. For the Pitman-Yor process, instead, the stick-breaking representation turns out to be useful to derive lower bounds on the probabilities of ℓ^1-balls of the mixing weights and locations. We expect this technique can be applied to other stick-breaking process priors.

The present article contributes to the topic by showing that, at least, for densities in a certain scale of regularity classes, full rate adaptation can be achieved using infinite Gaussian mixtures without any bandwidth shrinkage, the use of analytic kernels being intuitively justified by the fact that, in absence of any knowledge of the smoothness level of f_0, only infinitely smooth kernels can capture the "true" regularity of f_0. Thus, whatever the smoothness of the sampling density, the asymptotic performance, in terms of posterior contraction rates, of Dirichlet process Gaussian mixture priors is optimal. The exposition is focussed on density estimation, but other statistical settings are implicitly covered, for example, fixed design linear regression with unknown error distribution as described in Ghosal and van der Vaart (2007a), pages 205–206. Extension of these results to a multivariate setting is imminent along the lines of Shen *et al.* (2013) and is not pursued here.

The organization of the article is as follows. In Section 2, we describe the model and review some preliminary definitions. In Section 3, we state results on posterior contraction rates for general convolution kernel mixtures highlighting the connection with posterior recovery rates for mixing distributions. The main results are reported in Section 4,

wherein, after investigating the achievability of the error rate $1/\sqrt{n}$, up to a logarithmic factor, for *supersmooth* densities that possess a kernel mixture representation, we focus on adaptive estimation of densities with analytic or Sobolev smoothness using Gaussian mixtures. Estimates of the probabilities of ℓ^1-balls under various priors are given in Section 5. Sections 6 and 7 report the proofs of the theorems on adaptive estimation of densities with analytic or Sobolev smoothness, respectively. Auxiliary results are deferred to the Appendix.

1.1 Notation

Calculus

For real numbers a and b, we denote by $a\wedge b$ their minimum and by $a\vee b$ their maximum. We write "$\lesssim$" and "$\gtrsim$" for inequalities valid up to a constant multiple which is universal or inessential for our purposes. For integrals where no domain of integration is indicated, integration is performed over the entire domain of variation of the variables in the integrand. For any real valued function f, we write f^+ for its non-negative part $f1_{\{f\geq 0\}}$. For real valued functions f and g, the notation $f = o(g)$ means that $f/g \to 0$ in an asymptotic regime that should be clear from the context, while $f = O(g)$ means that $|f/g|$ is (eventually) bounded. Also, $f \sim g$ means that $f/g \to 1$. 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 $n \to \infty$,
- $a_n \ll b_n$ to mean that $a_n/b_n \to 0$ as $n \to \infty$.

Probability measures

When a probability measure is clearly specified by the context, it is sometimes denoted just by P and the associated expectation operator by $\mathrm{E}[\cdot]$. Subscripts in $\mathrm{E}[\cdot]$ specify the probability measure with respect to which the expectation is taken. We use the same symbol F to denote a distribution function and the corresponding probability measure. All density functions are understood to be with respect to Lebesgue measure λ on $\mathbb{R}$ or on some subset thereof. The probability density function of a standard normal distribution is denoted by ϕ. For any pair of probability density functions f and g,

- given $1 \leq p < \infty$, $\|f-g\|_p$ stands for the L^p-metric $(\int |f-g|^p \,\mathrm{d}\lambda)^{1/p}$,
- $\|f-g\|_\infty$ stands for the supremum norm $\|f-g\|_\infty := \sup_x |f(x)-g(x)|$,
- $\mathrm{KL}(f;\, g)$ stands for the Kullback-Leibler divergence $\int f \log(f/g)\,\mathrm{d}\lambda$.

For any probability density function f, define the positive (possibly infinite) constant $S_f := \sup\{|t|\colon |\hat{f}(t)| \neq 0\}$, where $\hat{f}(t) := \int e^{itx} f(x)\,\mathrm{d}x$, $t \in \mathbb{R}$, is the Fourier transform of f. If

- $S_f < \infty$, then $\mathrm{support}(|\hat{f}|) \subseteq [-S_f,\, S_f]$,
- $S_f = \infty$, then $|\hat{f}| > 0$ everywhere.

Function spaces

- $BC(\mathbb{R})$ is the space of bounded continuous real-valued functions on $\mathbb{R}$,
- $C^\infty(\mathbb{R})$ is the space of infinitely differentiable real-valued functions on $\mathbb{R}$,
- $C^\omega(\mathbb{R})$ is the space of analytic real-valued functions on $\mathbb{R}$.

2 Model description

The model is a location mixture $f_{F,\sigma}(\cdot) := (F*K_\sigma)(\cdot) = \int \sigma^{-1}K((\cdot-\theta)/\sigma)\,\mathrm{d}F(\theta)$, where K denotes the kernel density, σ the scale parameter and F the mixing distribution. Kernels herein considered are characterized via an integrated tail bound condition on their Fourier transforms. For constants $0 < \rho, r, L < \infty$, let $\mathcal{A}^{\rho,r,L}(\mathbb{R})$ be the class of densities on $\mathbb{R}$ with Fourier transforms satisfying

$$I^{\rho,r}(\hat{f}) := \int e^{2(\rho|t|)^r}|\hat{f}(t)|^2\,\mathrm{d}t \leq 2\pi L^2. \tag{1}$$

In symbols, $\mathcal{A}^{\rho,r,L}(\mathbb{R}) := \{f : \mathbb{R} \to \mathbb{R}^+|\ \|f\|_1 = 1,\ I^{\rho,r}(\hat{f}) \leq 2\pi L^2\}$. Condition (1) implies that the behavior of $|\hat{f}(t)|$ is described by $e^{-(\rho|t|)^r}$ as $|t| \to \infty$. Densities with Fourier transforms satisfying (1) are *infinitely* differentiable on $\mathbb{R}$, see, *e.g.*, Theorem 11.6.2. in Kawata (1972), pages 438–439, and "increasingly smooth" as ρ or r increases. Also, they are bounded, $\|f\|_\infty \leq (2\pi)^{-1}\|\hat{f}\|_1 \leq L^2 + C(\rho, r)/\pi < \infty$, where $C(\rho, r) := \int_0^\infty e^{-2(\rho t)^r}\,\mathrm{d}t = (2\rho^r)^{-1/r}\Gamma(1+1/r)$, cf. Lemma 1 in Butucea and Tsybakov (2008), page 35. Densities in classes $\mathcal{A}^{\rho,r,L}(\mathbb{R})$ are called *supersmooth*. They form a larger collection than that of analytic densities, including important examples like Gaussian, Cauchy, symmetric stable laws, Student's-t, distributions with characteristic functions vanishing outside a compact set, as well as their mixtures and convolutions.

EXAMPLE 2.1. *Symmetric stable laws*, which have characteristic functions of the form $e^{-(\rho|t|)^r}$, $t \in \mathbb{R}$, for some $0 < \rho < \infty$ and $0 < r \leq 2$, are supersmooth. Cauchy laws Cauchy$(0, \sigma)$ are stable with $r = 1$ and $\rho = \sigma$. Normal laws $\mathrm{N}(0, \sigma^2)$ are stable with $r = 2$ and $\rho = \sigma/\sqrt{2}$.

EXAMPLE 2.2. *Student's-t distribution* with $\nu > 0$ degrees of freedom has characteristic function verifying (1) for $r = 1$:

$$\widehat{f}_{t_\nu}(t) \sim \sqrt{\pi}[\Gamma(\nu/2)2^{(\nu-1)/2}]^{-1}(\sqrt{\nu}|t|)^{(\nu-1)/2}e^{-\sqrt{\nu}|t|},$$

see formula (4.8) in Hurst (1995), page 5.

EXAMPLE 2.3. *Densities with characteristic functions vanishing outside a symmetric convex compact set* are supersmooth. Let Σ_Λ be the class of densities with characteristic functions equal to 0 outside a symmetric convex compact set Λ in $\mathbb{R}^k$, $k \in \mathbb{N}$. For $k = 1$, let $\Lambda = [-T, T]$, with $0 < T < \infty$. For any $f \in \Sigma_\Lambda$, it is $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$ for every $0 < \rho, r < \infty$ and $L^2 \geq \pi^{-1}Te^{2(\rho T)^r}$. The *Fejér-de la Vallée-Poussin* density $f(x) = (2\pi)^{-1}[(x/2)^{-1}\sin(x/2)]^2$, $x \in \mathbb{R}$, having characteristic function $\hat{f}(t) = (1-|t|)^+$, $t \in \mathbb{R}$, is the typical example of density in $\Sigma_{[-1,1]}$.

Classes of densities as in Example 2.3 are such that, even if infinite-dimensional, for every $2 \leq p < \infty$, $\inf_{f_n}\sup_{f\in\Sigma_\Lambda}\mathrm{E}_f^n[\|f_n - f\|_p^s] \leq c_s n^{-s/2}$, where f_n denotes any estimator for densities in Σ_Λ based on n observations and the expectation is taken over the n-fold product measure of the probability law with density f. Moreover, for $p = s = 2$, the precise asymptotic bound $\lim_{n\to\infty} n \inf_{f_n}\sup_{f\in\Sigma_\Lambda}\mathrm{E}_f^n[\|f_n - f\|_2^2] = \mathrm{meas}(\Lambda)/(2\pi)^k$ holds, see Hasminskii and Ibragimov (1990), page 1008, and the references therein. The *almost parametric* rate $(\log n)/n$ is achievable for densities with characteristic functions

decaying exponentially fast, see Watson and Leadbetter (1963). This rate was shown to be optimal in the minimax sense by Ibragimov and Hasminskii (1983). Starting from this article, function classes related to $\mathcal{A}^{\rho,r,L}(\mathbb{R})$ have been considered by many authors in frequentist nonparametric curve estimation. Just to mention a few, Belitser and Levit (2001) considered nonparametric minimax estimation of an infinitely smooth density at a given point under random censorship; Golubev *et al.* (1996) investigated nonparametric regression estimation; Guerre and Tsybakov (1998) studied estimation of the unknown signal in the Gaussian white noise model; Butucea and Tsybakov (2008) considered adaptive density estimation in deconvolution problems. Adaptive density or regression function estimation over classes $\mathcal{A}^{\rho,r,L}(\mathbb{R})$ has so far hardly been studied from a Bayesian perspective, except for the recent articles of van der Vaart and van Zanten (2009), who used a Gaussian random field with an inverse-gamma bandwidth, and of de Jonge and van Zanten (2010), who suggested the use of finite kernel mixture priors with Gaussian mixing weights for inference. The problem with the use of finite mixtures is the choice of the number of mixing components, while updating it in a fully Bayesian way is computationally intensive. Mixture models with priors on the mixing distribution admitting an infinite discrete representation, like the Dirichlet process or more general stick-breaking priors, avoid fixing a truncation level. The focus of this article is on the capability of general kernel mixture priors to adapt posterior contraction rates to the analytic or Sobolev smoothness of the sampling density, without using any knowledge of the regularity of f_0 in the construction of the prior probability measure.

Given the model $f_{F,\sigma}$, a prior probability measure is induced on the space of Lebesgue univariate densities by putting priors on the mixing distribution F and the scale parameter σ. Let Π denote the prior for F. The scale parameter is assumed to be distributed, independently of F, according to a prior G on $(0, \infty)$. The overall prior $\Pi \times G$ on $\mathscr{M}(\Theta) \times (0, \infty)$, where, unless otherwise stated, $\mathscr{M}(\Theta)$ stands for the collection of all probability measures on some set $\Theta \subseteq \mathbb{R}$, induces a prior on $\mathscr{F} := \{f_{F,\sigma} : (F, \sigma) \in \mathscr{M}(\Theta) \times (0, \infty)\}$, which is equipped with an L^p-metric, $1 \leq p \leq \infty$. The sequence of observations $(X_i)_{i \geq 1}$ is assumed to be exchangeable. Observations from a kernel mixture prior can be equivalently described as

$$X_i | (F, \sigma) \overset{\text{iid}}{\sim} f_{F,\sigma}, \quad i = 1, \ldots, n,$$
$$(F, \sigma) \sim \Pi \times G.$$

Assuming that $X^{(n)} := (X_1, \ldots, X_n)$ are i.i.d. observations from an unknown density f_0 which may or may not be itself a convolution kernel mixture, we analyze contraction properties of the posterior distribution

$$(\Pi \times G)(B | X^{(n)}) \propto \int_B \prod_{i=1}^n f_{F,\sigma}(X_i)\, \Pi(\mathrm{d}F) G(\mathrm{d}\sigma), \quad \text{for any Borel set } B,$$

under regularity conditions on the prior $\Pi \times G$ and the sampling density f_0. Given any $1 \leq p \leq \infty$, a sequence of positive numbers $\varepsilon_{n,p}$ such that $\varepsilon_{n,p} \to 0$ and $n\varepsilon_{n,p}^2 \to \infty$, as $n \to \infty$, is an *upper bound* on the posterior contraction rate, relative to the L^p-metric,

if, for a constant $0 < M < \infty$, the posterior probability $(\Pi \times G)((F, \sigma) : \|f_{F,\sigma} - f_0\|_p \geq M\varepsilon_{n,p}|X^{(n)}) \to 0$ in P_0^n-probability, where P_0^n stands for the joint law of the first n coordinate projections of the infinite product probability measure P_0^∞, with P_0 denoting the probability measure corresponding to f_0. In the following section, we present results on posterior contraction rates for general kernel mixture priors.

3 Posterior contraction rates for kernel mixture priors

In this section, we present sufficient conditions for assessing posterior contraction rates in L^p-metrics, $2 \leq p \leq \infty$, for supersmooth kernel mixture priors. Results for specific priors on the mixing distribution belonging to the class of species sampling models, which are useful in concrete applications, are later exposed in Section 4.

Posterior contraction rates in L^p-metrics, $1 \leq p \leq \infty$, have been recently investigated by Giné and Nickl (2011), who have developed a new approach to testing problems based on the concentration properties of kernel-type density estimators. This approach accounts for having sufficient control of the approximation properties of the prior support. To describe regularity properties of the sampling density, they consider a general approximation scheme in function spaces based on integrating a fixed kernel-type function $K_{2^{-j}}(\cdot, \cdot) := 2^j K(2^j \cdot, 2^j \cdot)$ against a density f, that is, $\int K_{2^{-j}}(\cdot, y) f(y)\, \mathrm{d}y$ which, in the convolution kernel case, is $\int K_{2^{-j}}(\cdot - y) f(y)\, \mathrm{d}y$. The *sinc* kernel

$$\operatorname{sinc}(x) := \begin{cases} (\sin x)/(\pi x), & \text{if } x \neq 0, \\ 1/\pi, & \text{if } x = 0, \end{cases}$$

turns out to play a key role in characterizing regular densities in terms of their approximation properties. This is an unconventional kernel, in the sense that it may take negative values, it is Riemann integrable with $\int \operatorname{sinc}\, \mathrm{d}\lambda = 1$, but not Lebesgue integrable, $\operatorname{sinc} \notin L^1(\mathbb{R})$, it has Fourier transform identically equal to 1 on $[-1, 1]$ and vanishing outside it. The key role of the sinc kernel in density estimation is known since the work of Davis (1977), who showed that, for the sinc kernel density estimator, the optimal mean integrated squared error is of the order $O(n^{-1}(\log n)^{1/r})$ for estimands satisfying (1) with characteristic exponent r. Regularity of the overall prior distribution $\Pi \times G$ is expressed through the usual Kullback-Leibler prior support condition, as discussed in Ghosal *et al.* (2000), page 504, which involves Kullback-Leibler type neighborhoods of f_0

$$B_{\mathrm{KL}}(f_0; \varepsilon^2) := \{(F, \sigma) : \mathrm{KL}(f_0; f_{F,\sigma}) \leq \varepsilon^2,\ \mathrm{E}_0[(\log(f_0/f_{F,\sigma}))^2] \leq \varepsilon^2\}, \tag{2}$$

where $\mathrm{E}_0[\cdot]$ denotes the expectation with respect to the probability measure P_0. Employing a prior distribution for σ that is fully supported on $(0, \infty)$ amounts for regularity conditions on the tails of G, the requirement on the decay rate at 0 being expectedly more restrictive than that at ∞, the most important values being those included in a neighborhood of 0. In fact, as the bandwidth tends to 0, Gaussian mixtures can approximate any density in $L^p(\mathbb{R})$, $1 \leq p < \infty$. Therefore, the Kullback-Leibler prior support condition typically accounts for some assumption on the lower tail of G, like the following one, which guarantees enough prior mass in every neighborhood of 0.

(A0) The prior distribution G for σ has a continuous and positive density g on $(0, \infty)$ such that, for constants $0 < C_1, C_2, D_1, D_2 < \infty$, $0 \leq s, t < \infty$ and $0 < \gamma \leq \infty$,

$$C_1\sigma^{-s}\exp\left(-D_1\sigma^{-\gamma}(\log(1/\sigma))^t\right) \leq g(\sigma) \leq C_2\sigma^{-s}\exp\left(-D_2\sigma^{-\gamma}(\log(1/\sigma))^t\right)$$

for all σ in a neighborhood of 0.

An inverse-gamma distribution $\mathrm{IG}(\nu, \lambda)$, with shape parameter $0 < \nu < \infty$ and scale parameter $0 < \lambda < \infty$, is an eligible prior on σ satisfying Assumption (A0) for $s = \nu+1$, $t = 0$ and $\gamma = 1$.

We are now in a position to state the main result of the section.

Proposition 1. *Let $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$. Let $\tilde{\varepsilon}_n$ be a sequence of positive numbers such that $\tilde{\varepsilon}_n \to 0$ and $n\tilde{\varepsilon}_n^2 \to \infty$, as $n \to \infty$. For every $2 \leq p \leq \infty$, let $\varepsilon_{n,p} := \tilde{\varepsilon}_n(n\tilde{\varepsilon}_n^2)^{(1-1/p)/2}$. Suppose that, for $f_0 \in L^p(\mathbb{R})$,*

$$(\Pi \times G)(B_{\mathrm{KL}}(f_0; \tilde{\varepsilon}_n^2)) \gtrsim \exp\left(-Cn\tilde{\varepsilon}_n^2\right) \quad \textit{for some constant } 0 < C < \infty, \tag{3}$$

where Π is any prior probability measure on $\mathscr{M}(\Theta)$ and G satisfies Assumption (A0) with $0 \leq s, t < \infty$ and $1 < \gamma \leq \infty$ such that $n\tilde{\varepsilon}_n^2 \gtrsim (\log n)^{1/[r(1-1/\gamma)]}$. Suppose, furthermore, that $\|f_0 - f_0 * \mathrm{sinc}_{2^{-J_n}}\|_p = O(\varepsilon_{n,p})$ for $2^{J_n} = cn\tilde{\varepsilon}_n^2$, with $(\alpha^{1/r}\rho E)^{-1} \leq c < \infty$ and $0 < E \leq \{D_2[1_{[0,\gamma-1]}(s) + \beta 1_{(\gamma-1,\infty)}(s)]/(C+4)\}^{1/\gamma} 1_{(1,\infty)}(\gamma) + 1_{\{\infty\}}(\gamma)$ for some constants $0 < \alpha, \beta < 1$. Then, for a sufficiently large constant $0 < M < \infty$,

$$(\Pi \times G)((F, \sigma) : \|f_{F,\sigma} - f_0\|_p \geq M\varepsilon_{n,p} | X^{(n)}) \to 0 \quad \textit{in } P_0^n\textit{-probability.}$$

The assertion, whose proof is reported in the Appendix, is an in-probability statement that the posterior mass outside an L^p-norm ball of radius a large multiple M of $\varepsilon_{n,p}$ is approximately 0. Condition (3) is the essential one: the prior concentration rate is the only determinant of the posterior contraction rate at regular densities f_0 having L^p-approximation error of the same order against the sinc kernel-type approximant $f_0 * \mathrm{sinc}_{2^{-J_n}}$, with $2^{J_n} = O(n\tilde{\varepsilon}_n^2)$. This is the requirement expressed by the assumption $\|f_0 - f_0 * \mathrm{sinc}_{2^{-J_n}}\|_p = O(\varepsilon_{n,p})$. For instance, densities in $\mathcal{A}^{\rho,r,L}(\mathbb{R})$ meet this condition, see Lemma 12. For concreteness, the regularity condition on f_0 has been stated in terms of the sinc kernel, but any superkernel $S \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$, with bounded p-variation for some $1 \leq p < \infty$, can be employed, see the Appendix.

Proposition 1 yields optimal rates, up to a logarithmic factor, when the prior concentration rate is nearly parametric. When f_0 is ordinary smooth, even if the prior concentration rate is minimax-optimal, up to a logarithmic factor, suboptimal posterior contraction rates are found. Nonetheless, the result has an intrinsic value. When the kernel has Fourier transform decaying at an exponential power rate and f_0 is itself a kernel mixture, Proposition 1 yields contraction rates, relative to the Wasserstein metric of order $1 \leq p < \infty$, for the posterior measure on the mixing distribution. Before stating the result, we introduce Wasserstein distances. Let $(\Theta, \mathcal{B}(\Theta))$, $\Theta \subseteq \mathbb{R}$, be a measurable metric space with the Borel σ-field. For $1 \leq p < \infty$, define the *Wasserstein distance of order p* between any two Borel probability measures ν and ν' on Θ with finite pth-moment (*i.e.*, $\int d^p(\theta, \theta_0)\nu(\mathrm{d}\theta) < \infty$ for some and hence any θ_0

in Θ) as $W_p(\nu, \nu') := (\inf_{\mu \in \Gamma(\nu, \nu')} \int d^p(\theta, \theta')\mu(\mathrm{d}\theta, \mathrm{d}\theta'))^{1/p}$, where μ runs over the set $\Gamma(\nu, \nu')$ of all joint probability measures on $\Theta \times \Theta$ with marginal distributions ν and ν'. When $p = 2$, we take d to be the Euclidean distance on $\Theta \times \Theta$. By definition, $W_p(\nu, \nu') \in [0, \mathrm{diam}(\Theta)]$, where $\mathrm{diam}(\Theta)$ denotes the diameter of Θ. If Θ is compact, then $\mathrm{diam}(\Theta) < \infty$.

Corollary 1. *Let K be a symmetric (around 0) probability density such that, for some constants $0 < \rho < \infty$ and $0 < r \leq 2$,*

$$|\hat{K}(t)| \sim e^{-(\rho|t|)^r} \quad \text{as } |t| \to \infty. \tag{4}$$

Suppose that $f_0 = f_{F_0,\sigma_0} = F_0 * K_{\sigma_0}$, with F_0 supported on a bounded set $\Theta \subset \mathbb{R}$ and $0 < \sigma_0 < \infty$ fixed. Let the model be $f_{F,\sigma_0} = F * K_{\sigma_0}$, with F distributed according to a prior measure Π on $\mathscr{M}(\Theta)$. If, for a sequence of positive numbers $\tilde{\varepsilon}_n$ such that $\tilde{\varepsilon}_n \to 0$, as $n \to \infty$, and $n\tilde{\varepsilon}_n^2 \gtrsim (\log n)^{1/r}$,

$$\Pi(B_{\mathrm{KL}}(f_0; \tilde{\varepsilon}_n^2)) \gtrsim \exp(-Cn\tilde{\varepsilon}_n^2) \quad \text{for some constant } 0 < C < \infty, \tag{5}$$

then, for every $1 \leq p < \infty$, there is a sufficiently large constant $0 < M' < \infty$ so that

$$\Pi(F : W_p(F, F_0) \geq M'(\log n)^{-1/r} | X^{(n)}) \to 0 \quad \text{in } P_0^n\text{-probability.}$$

Inspection of the proof of Corollary 1, which is reported in the Appendix, reveals that the assumption that Θ is bounded can be replaced by the following requirement:

$$\forall\, 1 \leq p < \infty,\ \exists\, p < u < \infty,\ 0 < B < \infty :\ \mathrm{E}_K[|X|^u] < \infty \ \text{ and } \ \mathrm{E}_F[|X|^u] < B \quad \text{a.s. } [\Pi].$$

If, for some $1 < r \leq 2$, we have $\hat{K}(t) = e^{-(\rho|t|)^r}$, $t \in \mathbb{R}$, then $\mathrm{E}_K[|X|^u] = 2\rho^u \Gamma(1 - u/r)\Gamma(u)\sin(\pi u/2) < \infty$ for every $0 \leq u < r$, so that the previous condition is satisfied for every $p < u < r$. In virtue of Proposition 1, condition (4), combined with (5), implies that the posterior distribution for the mixture density concentrates on L^p-norm balls centered at the "true" density f_0, which is *in* the model, with probability tending to 1. This assertion translates into a parallel statement on the contraction rate, relative to the Wasserstein metric of order p, for the posterior on the mixing distribution. The resulting rate only depends on the characteristic exponent r of the Fourier transform of the kernel density so that the greater r, the smoother the kernel, the more difficult to recover the mixing distribution and the slower the rate. The open question remains whether this rate is optimal. Dedecker and Michel (2013) have shown that, in the deconvolution problem with supersmooth errors, the rate $(\log n)^{-1/r}$ is minimax-optimal over a larger class of probability measures than the one herein considered.

Posterior contraction rates for the mixing distribution in Wasserstein metrics have been recently investigated by Nguyen (2013), who has argued how convergence in Wasserstein metrics for discrete mixing measures has a natural interpretation in terms of convergence of the single atoms providing support for the measures. He has stated sufficient entropy and remaining mass conditions in the spirit of Ghosal *et al.* (2000), expressed in terms of the Wasserstein distance on the mixing distributions as opposed to the Hellinger or the L^1-distance on the mixture densities. Corollary 1 allows to derive posterior

contraction rates in the Wasserstein metric of any order $1 \leq p < \infty$ only from the prior concentration rate, whatever the prior measure and, under this respect, is more general than Theorem 6 in the above mentioned article, whose scope of applicability is confined to Dirichlet process kernel mixtures.

4 Posterior rates for specific priors on the mixing distribution

In this section, we derive posterior contraction rates for specific priors on the mixing distribution, namely, the Pitman-Yor process and the normalized inverse-Gaussian process, which are hereafter introduced.

Stick-breaking priors and the Pitman-Yor process

Stick-breaking priors form a rich class of random probability measures, which includes the Dirichlet process, the Pitman-Yor process or two-parameter Poisson-Dirichlet process, see Pitman and Yor (1997), and beta two-parameter processes, cf. Ishwaran and Zarepour (2000), Ishwaran and James (2001). Stick-breaking priors are almost surely discrete random probability measures F that can be represented as $F(\cdot) = \sum_{j=1}^{N} W_j \delta_{Z_j}(\cdot)$, with either a finite or an infinite number of terms $1 \leq N \leq \infty$, where $\delta_{Z_j}(\cdot)$ denotes a point mass at Z_j. The $(Z_j)_{j\geq 1}$ are independent and identically distributed random elements with common distribution $\bar{\alpha}$ over a measurable Polish space $(\Theta, \mathcal{B}(\Theta))$ endowed with its Borel σ-field. It is assumed that $\bar{\alpha}$ is non-atomic (*i.e.*, $\bar{\alpha}(\{\theta\}) = 0$ for every $\theta \in \Theta$) and $\bar{\alpha} := \alpha/\alpha(\Theta)$ for some positive and finite measure α. The random variables $(W_j)_{j\geq 1}$, called random weights, are independent of the $(Z_j)_{j\geq 1}$ and such that $0 \leq W_j \leq 1$, with $\sum_{j=1}^{N} W_j = 1$ almost surely. Furthermore,

$$W_1 = V_1, \qquad W_j = V_j \prod_{h=1}^{j-1} (1 - V_h), \quad j = 2, 3, \ldots, \tag{6}$$

where $V_j \sim H_j$ independently, H_j being a probability measure on $[0, 1]$ which is typically chosen to be a Beta(a_j, b_j) distribution with parameters $0 < a_j, b_j < \infty$, $j = 1, 2, \ldots$.

When $N < \infty$, we necessarily set $V_N = 1$ to ensure that $\sum_{j=1}^{N} W_j = 1$ almost surely because $W_N = 1 - \sum_{j=1}^{N-1} W_j = \prod_{h=1}^{N-1}(1 - V_h)$. When $N = \infty$, a necessary and sufficient condition for $\sum_{j=1}^{\infty} W_j = 1$ almost surely is that $\sum_{j=1}^{\infty} \mathrm{E}_{H_j}[\log(1 - V_j)] = -\infty$, see, *e.g.*, Lemma 1 in Ishwaran and James (2001), pages 162 and 170.

A stick-breaking random probability measure F where, for $0 \leq d < 1$ and $-d < c < \infty$, the random variables V_j are independently distributed according to a Beta distribution Beta$(1 - d, c + dj)$, $j \in \mathbb{N}$, is called the *Pitman-Yor process* or *two-parameter Poisson-Dirichlet process*, with *concentration* parameter c and *discount* parameter d, denoted

by $F \sim \mathrm{PY}(c, d)$, which can be described as

$$F(\cdot) = \sum_{j=1}^{\infty} V_j \prod_{h=1}^{j-1} (1 - V_h) \delta_{Z_j}(\cdot) \quad \text{a.s.}$$
$$V_j \overset{\text{ind}}{\sim} \mathrm{Beta}(1 - d, c + dj), \quad j \in \mathbb{N}, \qquad \text{and} \qquad Z_j \overset{\text{iid}}{\sim} \bar{\alpha}, \quad j \in \mathbb{N}.$$

The case where $d = 0$ and $c = \alpha(\Theta)$, so that the $(V_j)_{j \geq 1}$ are independent and identically distributed according to a $\mathrm{Beta}(1, \alpha(\Theta))$, returns the Dirichlet process with base measure α, denoted by $\mathrm{DP}(\alpha)$. There are no known analytic expressions for the finite-dimensional distributions of the Pitman-Yor process, except when $d = 0$ and $c = \alpha(\Theta)$ or when $d = 1/2$ and $-1/2 < c < \infty$.

The Dirichlet process, the Pitman-Yor process with $d = 1/2$ and the normalized inverse-Gaussian process, which is hereafter introduced, are the only known processes for which explicit expressions of the finite-dimensional distributions are available.

Normalized inverse-Gaussian process

Let α be a finite and positive measure on a measurable Polish space $(\Theta, \mathcal{B}(\Theta))$ endowed with its Borel σ-field. Following Lijoi *et al.* (2005), we call a random probability measure F a *normalized inverse-Gaussian* (N-IG) *process* with parameter α, denoted by N-IG(α), if, for every finite measurable partition $A_1, \ldots, A_N$ of Θ, the random vector $(F(A_1), \ldots, F(A_N))$, $2 \leq N < \infty$, has a N-IG distribution with parameter $(\alpha(A_1), \ldots, \alpha(A_N))$. A random vector $(Z_1, \ldots, Z_N)$ has a N-IG distribution with parameter $(\alpha_1, \ldots, \alpha_N)$, where $0 < \alpha_j < \infty$ for every $j = 1, \ldots, N$, denoted by N-IG($\alpha_1, \ldots, \alpha_N$), if it has probability density function over the unit $(N-1)$-simplex $\Delta^{N-1} := \{(z_1, \ldots, z_N) \in \mathbb{R}^N : z_j \geq 0, 1 \leq j \leq N, \text{ and } \sum_{j=1}^N z_j = 1\}$ given by

$$\begin{aligned} f(z_1, \ldots, z_{N-1}) &= \frac{\exp\left(\sum_{j=1}^N \alpha_j\right) \prod_{j=1}^N \alpha_j}{2^{N/2-1} \pi^{N/2}} \\ &\quad \times K_{-N/2}\left(\sqrt{\mathcal{A}_N(z_1, \ldots, z_{N-1})}\right) \\ &\quad \times [\mathcal{A}_N(z_1, \ldots, z_{N-1})]^{-N/4} \\ &\quad \times [z_1 \times \ldots \times z_{N-1} \times (1 - z_1 - \ldots - z_{N-1})]^{-3/2} \\ &=: h_1 \times \prod_{r=2}^{4} h_r(z_1, \ldots, z_{N-1}), \end{aligned} \tag{7}$$

where $K_{-N/2}(\cdot)$ denotes the modified Bessel function of the third kind and

$$\mathcal{A}_N(z_1, \ldots, z_{N-1}) := \sum_{j=1}^{N-1} \frac{\alpha_j^2}{z_j} + \alpha_N^2 \left(1 - \sum_{j=1}^{N-1} z_j\right)^{-1}.$$

4.1 Estimation of densities with a kernel mixture representation

We begin the analysis from the case where the sampling density f_0 is itself a convolution kernel mixture, $f_0 = f_{F_0,\sigma_0} = F_0 * K_{\sigma_0}$, where F_0 and σ_0 denote the true values of the mixing distribution and the scale parameter, respectively. Considering this case helps developing techniques that can be used when f_0 cannot be represented as a location mixture of kernel densities. Results are obtained under the following assumptions.

Assumptions

(A1) The kernel probability density $K : \mathbb{R} \to \mathbb{R}^+$ is symmetric around 0, monotone decreasing in $|x|$ and satisfies the tail condition $K(x) \gtrsim e^{-c|x|^\kappa}$ as $|x| \to \infty$, for some constants $0 < c, \kappa < \infty$.

(A2) The true mixing distribution F_0 satisfies the tail condition $F_0(\theta : |\theta| > t) \lesssim e^{-c_0 t^\varpi}$ as $t \to \infty$, for some constants $0 < c_0 < \infty$ and $0 < \varpi \leq \infty$.

(A3) The base measure α has a continuous and positive density α' on $\mathbb{R}$ such that $\alpha'(\theta) \propto e^{-b|\theta|^\delta}$ as $|\theta| \to \infty$, for some constants $0 < b, \delta < \infty$.

Assumptions (A1)–(A3) are standard requirements on the kernel density, the true mixing distribution and the density of the base measure, respectively.

The following theorem, whose proof is deferred to the Appendix, extends results of Ghosal and van der Vaart (2001) and Scricciolo (2011) on posterior contraction rates, relative to the Hellinger or the L^1-metric, for Dirichlet process Gaussian mixtures to Pitman-Yor or normalized inverse-Gaussian kernel mixtures in L^p-metrics, $1 \leq p \leq \infty$.

Given numbers $0 < \kappa, r < \infty$, let ϖ be such that

$$0 < \max\{\kappa, [1 + 1_{(1,\infty)}(r)/(r-1)]\} \leq \varpi \leq \infty. \tag{8}$$

Also, let $\tau := 1 + \{1/r - [1 - 1_{(0,\infty)}(\varpi)/\varpi]\}1_{(0,1]}(r)/2$ and

$$\varphi(d) := \tau + (\tau - 1/2)1_{(0,1)}(d), \quad 0 \leq d < 1. \tag{9}$$

Condition (8) requires a matching between the tail decay speed of the true mixing distribution F_0 and that of the kernel density K.

Theorem 1. *Let $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$, be as in Assumption* (A1)*. Suppose that the probability density $f_0 = f_{F_0,\sigma_0} = F_0 * K_{\sigma_0}$, with*

(i) F_0 satisfying Assumption (A2) *for some constants $0 < c_0 < \infty$ and ϖ as in* (8)*.*

Let $F \sim \mathrm{PY}(c, d)$, with $0 \leq d < 1$ and $-d < c < \infty$. Alternatively, let $F \sim \text{N-IG}(\alpha)$. Assume that

(ii) α satisfies Assumption (A3) *for constants $0 < b, \delta < \infty$, with $\delta \leq \varpi$ when $\varpi < \infty$;*

(iii) G satisfies Assumption (A0) *for constants $0 \leq s, t < \infty$, $0 < \gamma \leq \infty$ if $p = 1$, $\max\{1, \{1 - [2r\varphi(d)]^{-1}\}^{-1}\} < \gamma \leq \infty$ if $2 \leq p \leq \infty$, where $\varphi(d)$ is as in* (9)*.*

When $p = 1$, assume furthermore that, for some constant $0 < \varrho \leq \infty$, the tail probability $1 - G(\sigma) \lesssim \sigma^{-\varrho}$ as $\sigma \to \infty$.

Then, the posterior contraction rate relative to the L^p-metric, $1 \leq p \leq \infty$, denoted by $\varepsilon_{n,p}$, is $n^{-1/2}(\log n)^{\mu}$, with a constant $0 < \mu < \infty$ possibly depending on p.

A caveat applies to Theorem 1: since the normalized inverse-Gaussian process behaves like the Dirichlet process, results for both these priors are obtained when $d = 0$. Theorem 1 shows that a nearly parametric rate is achievable, irrespective of the tail behavior of the kernel density (hence of the sampling density f_0), heavy-tailed distributions, like the Student's-t, which play a crucial role in modeling certain phenomena, being admitted. Estimation of heavy-tailed distributions is not covered by Theorem 2 nor by Theorem 3 on adaptation which, by requiring f_0 to have (sub-)exponential tails, rule out these distributions. Furthermore, Theorem 1 has a relevant implication for the contraction rate, relative to the Wasserstein metric of order $1 \leq p < \infty$, of the posterior distribution corresponding to a Dirichlet process prior on the mixing distribution: it yields the recovery rate for the true mixing distribution when the sampling density is a convolution kernel mixture.

Corollary 2. *Under the conditions of Theorem 1, with $F \sim \mathrm{DP}(\alpha)$ and $G = \delta_{\sigma_0}$ a point mass at σ_0, for every $1 \leq p < \infty$ there exists a sufficiently large constant $0 < M' < \infty$ so that $\Pi(F : W_p(F, F_0) \geq M'(\log n)^{-1/r}) \to 0$ in P_0^n-probability.*

4.2 Adaptive estimation of analytic densities

In this section, we consider adaptive estimation of analytic densities whose Fourier transforms have sub-exponential tails using Gaussian mixtures. We assume that the sampling density f_0 satisfies the following conditions.

(*a*) *Smoothness*: $f_0 \in \mathcal{A}^{\rho_0, r_0, L_0}(\mathbb{R})$ for some constants $0 < \rho_0$, $L_0 < \infty$ and $1 \leq r_0 \leq 2$.

(*b*) *Monotonicity*: f_0 is non-decreasing on $(-\infty, a)$, non-increasing on (b, ∞), for $-\infty < a \leq b < \infty$, with $f_0 \geq \ell > 0$ on $[a, b]$ and $0 < f_0 \leq \ell < \infty$ on $[a, b]^c$.

(*c*) *Tails*: for some constants $0 < c_0 < \infty$ and $1 < \varpi_0 < \infty$ such that $(\varpi_0 - 1)r_0 \leq \varpi_0$,

$$f_0(x) \sim e^{-c_0|x|^{\varpi_0}} \quad \text{as } |x| \to \infty.$$

In order to prove that contraction rates of posterior distributions corresponding to a Pitman-Yor or a N-IG process mixture of normal densities adapt to the "analytic smoothness" r_0 of f_0, the key step is the approximation of f_0 with a continuous mixture of normal densities, which is then discretized to have a sufficiently restricted number of support points, see Lemma 8. We suspect that this step is only possible under Condition (*c*) that f_0 asymptotically has exponential tails, which, for instance, is satisfied for a Gaussian density with $\varpi_0 = r_0 = 2$. This requirement seems to be necessary to obtain a nearly parametric rate because, when restricting the support to a compact set, it allows to take the endpoint of the order $O((\log(1/\varepsilon))^l)$, for some $0 < l < \infty$, thus entailing

a finite mixture with a relatively small number of support points. A density with polynomially decaying tails would instead incur an additional factor of order $O(\varepsilon^{-k})$, for some $0 < k < \infty$, and a power of n would be lost in the prior as well as in the posterior contraction rate.

The key step in the proof is the construction of a (not necessarily non-negative) function that uniformly approximates f_0 with an exponentially small error in terms of the inverse of the bandwidth, see Lemma 5. By suitably modifying this function, we obtain a density with the same approximation error in the Kullback-Leibler divergence. The general strategy is similar to that adopted by Kruijer *et al.* (2010), but the iterative procedure they used to construct the approximant turns out to be inefficient in the present case because of the infinite degree of smoothness of f_0. As far as we are aware, the approximation result of Lemma 5 involving the sinc kernel is novel. Once a finite mixture is constructed, we need to show that there exists a whole set of finite Gaussian mixtures close to it and contained in a Kullback-Leibler type ball around f_0 that is charged enough prior mass.

We are now in a position to state the result, to which the proviso applies that the claim for the normalized inverse-Gaussian process is obtained setting $d = 0$.

Theorem 2. *Suppose that the probability density f_0 satisfies Conditions (a)–(c). Let the model be $f_{F,\sigma} = F * \phi_\sigma$, with $F \sim \mathrm{PY}(c, d)$, $0 \le d < 1$ and $-d < c < \infty$. Alternatively, let $F \sim \text{N-IG}(\alpha)$. Assume that*

(*i*) *α satisfies Assumption* (A3) *for some constants $0 < b < \infty$ and $0 < \delta \le (\varpi_0 \wedge 2)$;*

(*ii*) *G satisfies Assumption* (A0) *for some constants $0 \le s, t < \infty$, $0 < \gamma \le \infty$ if $p = 1$, $2 \le \gamma \le \infty$ if $2 \le p \le \infty$. When $p = 1$, assume furthermore that, for some constant $0 < \varrho \le \infty$, the tail probability $1 - G(\sigma) \lesssim \sigma^{-\varrho}$ as $\sigma \to \infty$.*

Then, the posterior contraction rate relative to the L^p-metric, denoted by $\varepsilon_{n,p}$, is

$$\varepsilon_{n,p} = \begin{cases} n^{-1/2}(\log n)^{\frac{1}{2}+\{\frac{1}{2}\vee[2(\frac{1}{\delta}+\frac{1}{\gamma})\psi(r_0,d)]\}}, & \text{for} \quad p = 1, \\ n^{-1/2}(\log n)^{(2-1/p)\psi(r_0,d)}, & \text{for} \quad 2 \le p \le \infty, \end{cases}$$

where

$$\psi(r_0, d) := \max\{(\gamma/r_0 + t)/2, \{(1/2) + [1/r_0 + 1/(\varpi_0 \wedge 2)][1 + 1_{(0,\infty)}(d)]\}\}. \tag{10}$$

If conditions specific to the cases $p = 1$ and $p = 2$ are simultaneously satisfied, then $\varepsilon_{n,p} \le (\varepsilon_{n,1} \vee \varepsilon_{n,2})$ for every $1 < p < 2$.

Given that the power of n is fixed at $-1/2$, the most important element in the rate is the power of the logarithmic term, which adapts to the characteristic exponent r_0 of f_0. A main implication of Theorem 2, whose proof is deferred to Section 6, is that the choice of the kernel is not an issue in Bayesian density estimation. A well-known problem with the use of Gaussian convolutions is that the approximation error of a smooth density can only be of the order $O(\sigma^2)$, even if the density has greater smoothness. The approximation can be improved using higher-order kernels, but the resulting convolution is not guaranteed to be everywhere non-negative which, in a frequentist approach, translates

into a non-bona fide estimator. This is not an issue in a Bayesian framework because to get adaptation it suffices that the prior support contains a set of Gaussian mixtures close to f_0 receiving enough prior mass, which is the case when the mixing distribution is endowed with a Pitman-Yor or a N-IG process prior.

4.3 Adaptive estimation over Sobolev spaces

In this section, we study adaptive estimation of densities in Sobolev spaces using Gaussian mixtures. We assume that f_0 satisfies the following conditions.

(a') *Smoothness*: for $k_0 \in \mathbb{N}$, the probability density $f_0 \in L^2(\mathbb{R})$ has Fourier transform $\widehat{f}_0$ satisfying

$$\int (1+|t|^2)^{k_0} |\widehat{f}_0(t)|^2 \, \mathrm{d}t < \infty. \tag{11}$$

In addition,

$$f_0^{(k_0)} \in L^\infty(\mathbb{R}) \quad \text{with} \quad \mathrm{E}_0[|(f_0^{(k_0)}/f_0)(X)|^2] < \infty \tag{12}$$

and, for $k_0 \in \{2, 3, \dots\}$,

$$\mathrm{E}_0[|(f_0^{(j)}/f_0)(X)|^{2k_0/j}] < \infty, \quad j = 1, \dots, k_0 - 1. \tag{13}$$

(b') *Tails*: for some constants $0 < c_0$, $M_0 < \infty$ and $2 \leq \varpi_0 < \infty$,

$$f_0(x) \leq M_0 e^{-c_0 |x|^{\varpi_0}} \quad \text{for large } |x|.$$

Condition (11), which is the essential one, implies that f_0 is k_0-times continuously differentiable on $\mathbb{R}$, with derivatives that vanish at ∞, *i.e.*, $f_0^{(j)}(x) \to 0$, as $|x| \to \infty$, for every integer $1 \leq j \leq k_0$. Conditions (12) and (13) are technical requirements. Condition (b') postulates that the sampling density f_0 has (sub-)Gaussian tails, which when restricting to a compact set the support of the mixing density of a suitably constructed Gaussian convolution mixture that uniformly approximates f_0 allows for keeping the number of support points relatively small.

The following theorem, whose proof is deferred to Section 7, asserts that, whatever the "Sobolev smoothness" k_0 of f_0, the posterior distribution corresponding to a Dirichlet process mixture of normal densities contracts at a rate at least as fast as $n^{-k_0/(2k_0+1)}(\log n)^\kappa$, with $0 < \kappa < \infty$, in all L^p-metrics, $1 \leq p \leq 2$, where the rate $n^{-k_0/(2k_0+1)}$ is known to be optimal in the minimax sense for the L^2-metric, see Theorem 2 in Donoho *et al.* (1996), page 515.

Theorem 3. *Suppose that the probability density f_0 satisfies Conditions (a') and (b'). Let the model be $f_{F,\sigma} = F * \phi_\sigma$, with $F \sim \mathrm{DP}(\alpha)$. Assume that*

(*i*) *α satisfies Assumption* (A3) *for some constants $0 < b < \infty$ and $0 < \delta \leq 2$;*

(*ii*) *G satisfies Assumption* (A0) *for some constants $0 \leq s, t < \infty$ and $\gamma = 1$. When $p = 1$, assume furthermore that, for some constant $0 < \varrho \leq \infty$, the tail probability $1 - G(\sigma) \lesssim \sigma^{-\varrho}$ as $\sigma \to \infty$.*

Then, the posterior contraction rate relative to the L^p-metric, denoted by $\varepsilon_{n,p}$, is

$$\varepsilon_{n,p} = \begin{cases} n^{-k_0/(2k_0+1)}(\log n)^{\tau+1+(2\delta)^{-1}}, & \text{for } p=1, \\ n^{-k_0/(2k_0+1)}(\log n)^{\tau}, & \text{for } p=2, \end{cases}$$

with a suitable constant $0 < \tau < \infty$. If conditions specific to the cases $p = 1$ and $p = 2$ are simultaneously satisfied, then $\varepsilon_{n,p} \leq n^{-k_0/(2k_0+1)}(\log n)^{\tau+1+(2\delta)^{-1}}$ for every $1 < p < 2$.

A few comments are in order. Slower rates are found when endowing the mixing distribution with a Pitman-Yor process having a strictly positive discount parameter since small Kullback-Leibler type neighborhoods of f_0 do not seem to be charged enough prior mass. The open question remains whether posterior contraction rates are indeed suboptimal for a non-degenerate Pitman-Yor process prior on the mixing distribution. Also, rates in L^p-metrics, $2 < p \leq \infty$, are found to deteriorate by a genuine power of n.

The result of Theorem 3 differs from the one of Theorem 1 in Shen *et al.* (2013), page 627, in so far that they obtained adaptation for multivariate Hölder densities using a Dirichlet process mixture of normal densities with a Gaussian base measure and an inverse-Wishart prior on the covariance matrix, while we treat the univariate case and allow for a more general base measure. In addition, the smoothness assumption is expressed through an integrated tail bound on the Fourier transform of f_0 and the integrability conditions in (12) and (13) are weaker. Besides, while Shen *et al.* (2013) made use of the stick-breaking representation of the Dirichlet process to show that the remaining mass and entropy conditions are satisfied, we verified the remaining mass condition using the fact that the tails of almost every sample distribution function from a Dirichlet process are much smaller than the tails of the base measure. We could thus exploit the same sieve set employed for estimating analytic densities because, in the univariate case, the entropy correctly scales to the smoothness level.

5 Estimates of the probabilities of ℓ^1-balls

Estimates, under different priors, of the probabilities of ℓ^1-balls are essential to evaluate the prior probability mass of Kullback-Leibler type neighborhoods of the sampling density f_0 as in (2). While for the N-IG process, the expressions of the finite-dimensional distributions can be exploited as in Lemma A.1 of Ghosal *et al.* (2000), pages 518–519, which deals with the Dirichlet process, for the Pitman-Yor process, the stick-breaking representation turns out to be useful to obtain lower bounds on the probabilities of ℓ^1-balls of the mixing weights and locations.

5.1 Pitman-Yor process

Lemma 1. *Let $F \sim \mathrm{PY}(c, d)$, with $0 \leq d < 1$ and $-d < c < \infty$. For $0 < \varepsilon < 1$, let $F' = \sum_{j=1}^{N} p_j \delta_{\theta_j}$, $1 \leq N < \infty$, be a probability measure on $\mathbb{R}$ (possibly depending on ε) with $p_1 \geq p_2 \geq \ldots \geq p_N > 0$. If $N = 1$, define $v_1 := 1$. If $2 \leq N < \infty$,*

define $v_1 := p_1$, $v_j := p_j[\prod_{h=1}^{j-1}(1-v_h)]^{-1}$, $j = 2, \ldots, N-1$, *and* $v_N := 1$. *Assume that* $3\varepsilon/N^2 \leq \min_{1\leq j\leq N-1} v_j \leq \max_{1\leq j\leq N-1} v_j \leq 1 - 2\varepsilon(1+\xi)/N^2$ *for some constant* $0 < \xi < 1$. *Let* $U := (\sum_{j=1}^{N}\sum_{h=1}^{j}|V_h - v_h| \leq 2\varepsilon,\ \min_{1\leq j\leq N} V_j \geq \varepsilon/N^2)$, *where the random variables* $V_1, \ldots, V_N$ *are those stemming from the representation in* (6). *Then, there exist constants* $0 < c_1, C_1 < \infty$ *(depending only on* c *and* d*) such that*

$$\mathrm{P}(U) \geq C_1 \times \begin{cases} \exp\left(-(c+d)\log(1/\varepsilon)\right), & \text{for} \quad N = 1, \\ \exp\left(-c_1(1 \vee dN)N\log(N/\varepsilon)\right), & \text{for} \quad 2 \leq N < \infty. \end{cases}$$

Proof. If $N = 1$, then $v_1 = 1$, $U = (|V_1 - 1| \leq 2\varepsilon,\ V_1 \geq \varepsilon)$, with $V_1 \sim \mathrm{Beta}(1-d, c+d)$, and

$$\mathrm{P}(U) \geq \frac{\Gamma(1+c)}{\Gamma(1-d)\Gamma(c+d)} \int_0^{2\varepsilon} z^{c+d-1}\,\mathrm{d}z \gtrsim \exp\left(-(c+d)\log(1/\varepsilon)\right).$$

If $N \geq 2$, then $v_N = 1$. If $|V_j - v_j| \leq 2\varepsilon/N^2$ for $j = 1, \ldots, N$, then $\sum_{j=1}^{N}\sum_{h=1}^{j}|V_h - v_h| < 2\varepsilon$. Thus, the event U is implied by $V := (|V_j - v_j| \leq 2\varepsilon/N^2,\ V_j \geq \varepsilon/N^2,\ j = 1, \ldots, N)$. Let $l_j := [(v_j - 2\varepsilon/N^2) \vee (\varepsilon/N^2)]$ and $u_j := [(v_j + 2\varepsilon/N^2) \wedge 1]$, $j = 1, \ldots, N$. We have $l_N = 1 - 2\varepsilon/N^2$ and $u_N = 1$. By assumption, the V_j are independent $\mathrm{Beta}(1-d, c+dj)$, $j \in \mathbb{N}$, thus, using the identity $\Gamma(z+1) = z\Gamma(z)$, $z > 0$,

$$\mathrm{P}(V) \geq \frac{[\Gamma(1-d)]^{-N}\Gamma(c)c^N}{(c+dN)\Gamma(c+dN)}(2\varepsilon/N^2)^{c+dN}\prod_{j=1}^{N-1}\int_{l_j}^{u_j}(1-v)^{c+dj}\,\mathrm{d}v.$$

Using the constraints $3\varepsilon/N^2 \leq \min_{1\leq j\leq N-1} v_j \leq \max_{1\leq j\leq N-1} v_j \leq 1 - 2\varepsilon(1+\xi)/N^2$, with $0 < \xi < 1$,

$$\mathrm{P}(V) \geq \frac{[\Gamma(1-d)]^{-N}\Gamma(c)c^N}{\Gamma(c+dN+1)}(2\varepsilon/N^2)^{c+dN}(4\varepsilon/N^2)^{N-1} \times \left(1 - 2\varepsilon/N^2 - \max_{1\leq j\leq N-1} v_j\right)^{c(N-1)+dN(N-1)/2}.$$

For $d = 0$, $\mathrm{P}(V) \gtrsim \exp\left(-c_1 N\log(N/\varepsilon)\right)$. For $d > 0$, if N is fixed or, in the case where $N \to \infty$ as $\varepsilon \to 0$, using $\Gamma(c+dN+1) \sim (2\pi)^{1/2}\exp\left(-dN\right)(dN)^{dN+c+1/2}$, we have $\mathrm{P}(V) \gtrsim \exp\left(-c_1(1 \vee dN)N\log(N/\varepsilon)\right)$. Conclude by noting that $\mathrm{P}(U) \geq \mathrm{P}(V)$. □

Lemma 2. *Let* $F \sim \mathrm{PY}(c, d)$, *with* $0 \leq d < 1$, $-d < c < \infty$ *and the base measure* α *satisfying Assumption* (A3) *for some constants* $0 < b, \delta < \infty$. *For* $0 < \varepsilon < 1$, *let* $F' = \sum_{j=1}^{N} p_j\delta_{\theta_j}$, $1 \leq N < \infty$, *be a probability measure (possibly depending on* ε*) with* $\mathrm{support}(F') \subset [-a, a]$ *for sufficiently large* $0 < a < \infty$. *Then,* $\mathrm{P}(\sum_{j=1}^{N}|Z_j - \theta_j| \leq \varepsilon) \gtrsim \exp\left(-N[\log(N\alpha(\mathbb{R})/(2\varepsilon)) + ba^{\delta}]\right)$, *where the* Z_j *are the random locations of* F.

Proof. If $|Z_j - \theta_j| \leq \varepsilon/N$ for every $j = 1, \ldots, N$, then $\sum_{j=1}^{N}|Z_j - \theta_j| \leq \varepsilon$. Since $Z_1, \ldots, Z_N$ are independent and identically distributed according to $\bar{\alpha}$,

$$\mathrm{P}\left(\sum_{j=1}^{N}|Z_j - \theta_j| \leq \varepsilon\right) \geq \prod_{j=1}^{N}\int_{\theta_j - \varepsilon/N}^{\theta_j + \varepsilon/N}\frac{\alpha'(z)}{\alpha(\mathbb{R})}\,\mathrm{d}z \gtrsim \exp\left(-N[\log(N\alpha(\mathbb{R})/(2\varepsilon)) + ba^{\delta}]\right),$$

where the second inequality follows from Assumption (A3) and the fact that a is large enough. □

Remark 1. For $d = 0$, if $N = O((1/\varepsilon)^{\varrho})$ for some $0 < \varrho < \infty$, by Lemma 1, $\mathrm{P}(U) \gtrsim \exp(-cN\log(1/\varepsilon))$ which, combined with Lemma 2, yields an estimate of the probability of an ℓ^1-ball under the $\mathrm{PY}(c, 0)$, for $0 < c < \infty$, that is consistent with the one known for the Dirichlet process $\mathrm{DP}(\alpha) = \mathrm{PY}(\alpha(\mathbb{R}), 0)$ from Lemma 6.1 in Ghosal *et al.* (2000), pages 518–519, or Lemma A.1 in Ghosal (2001), pages 1278–1279.

For the Pitman-Yor process, the stick-breaking representation has turned out to be useful to obtain lower bounds on the probabilities of ℓ^1-balls of the mixing weights and locations. However, when $d = 1/2$ and $-1/2 < c < \infty$, the expressions of the finite-dimensional distributions of the Pitman-Yor process $\mathrm{PY}(c, 1/2)$ are known and could be exploited to estimate the probability of an ℓ^1-ball around a given mixing distribution as in Lemma 6.1 in Ghosal *et al.* (2000), pages 518–519, or as in Lemma A.1 in Ghosal (2001), pages 1278–1279.

Let $F \sim \mathrm{PY}(c, 1/2)$, with $-1/2 < c < \infty$, be a random probability measure on a measurable Polish space $(\Theta, \mathcal{B}(\Theta))$ endowed with its Borel σ-field. For every finite measurable partition $A_1, \ldots, A_N$ of Θ, the random vector $(F(A_1), \ldots, F(A_N))$, $2 \leq N < \infty$, has probability density function over the unit $(N-1)$-simplex Δ^{N-1}

$$f(p_1, \ldots, p_N) := \frac{\Gamma(c + N/2)}{\pi^{(N-1)/2}\Gamma(c + 1/2)} \times \frac{\prod_{j=1}^{N}(\bar{\alpha}_j/p_j^{3/2})}{(\sum_{j=1}^{N}\bar{\alpha}_j^2/p_j)^{c+N/2}} 1_{\Delta^{N-1}}(p_1, \ldots, p_N), \quad (14)$$

where $\bar{\alpha}_j := \bar{\alpha}(A_j)$ for $j = 1, \ldots, N$, see Theorem 3.1 in Carlton (2002), pages 768–769.

Lemma 3. *Let $(P_1, \ldots, P_N)$, $2 \leq N < \infty$, have distribution with density as in (14). For $0 < \varepsilon < 1$, let $U := (\sum_{j=1}^{N}|P_j - p_{j0}| \leq 2\varepsilon)$, with $(p_{10}, \ldots, p_{N0}) \in \Delta^{N-1}$ such that $\varepsilon^2 < \min_{1\leq j\leq N} p_{j0} \leq \max_{1\leq j\leq N} p_{j0} < 1 - \varepsilon^2$. Assume that $A\varepsilon^b \leq \bar{\alpha}_j < 1$, $j = 1, \ldots, N$, for some constants $0 < A, b < \infty$. If $0 < \varepsilon \leq 1/N$, then*

$$\mathrm{P}(U) > C\frac{2^{(N-1)(N/2+c-1/2)}}{(N/2 + c - 1/2)^{N-1}}\varepsilon^{N+2c-3}(A\varepsilon^b)^N \times \exp(-(N-1)(N + 2c - 1)\log(1/\varepsilon)),$$

where $C := \Gamma(c + N/2)/[\pi^{(N-1)/2}\Gamma(c + 1/2)]$.

Proof. Note that for $0 < p_j < 1$, $j = 1, \ldots, N$,

$$\left(\sum_{j=1}^{N}\bar{\alpha}_j^2/p_j\right)^{c+N/2} = \left(\prod_{j=1}^{N}p_j\right)^{-(c+N/2)}\left(\sum_{j=1}^{N}\bar{\alpha}_j^2\prod_{j'\neq j}p_{j'}\right)^{c+N/2} < \prod_{j=1}^{N}p_j^{-(c+N/2)}$$

because $\bar{\alpha}_j < 1$ for every $j = 1, \ldots, N$. Reasoning as in Lemma 6.1 in Ghosal *et al.* (2000), pages 518–519, we can assume that $p_{N0} \geq 1/N$. For $l_j := [(p_{j0} - \varepsilon^2) \vee 0]$ and

$u_j := [(p_{j0} + \varepsilon^2) \wedge 1]$, $j = 1, \ldots, N-1$,

$$\begin{aligned}
\mathrm{P}(U) &\geq \mathrm{P}(|P_j - p_{j0}| \leq \varepsilon^2,\ \ j = 1, \ldots, N-1) \\
&> C \frac{\varepsilon^{N+2c-3}(A\varepsilon^b)^N}{(N/2+c-1/2)^{N-1}} \prod_{j=1}^{N-1} (u_j^{N/2+c-1/2} - l_j^{N/2+c-1/2}) \\
&\geq C \frac{\varepsilon^{N+2c-3}(A\varepsilon^b)^N}{(N/2+c-1/2)^{N-1}} \prod_{j=1}^{N-1} (u_j - l_j)^{N/2+c-1/2} \\
&\geq C \frac{2^{(N-1)(N/2+c-1/2)}}{(N/2+c-1/2)^{N-1}} \varepsilon^{N+2c-3}(A\varepsilon^b)^N \exp\left(-(N-1)(N+2c-1)\log(1/\varepsilon)\right)
\end{aligned}$$

because $x^p - y^p \geq (x-y)^p$ for $0 \leq x, y \leq 1$ and $1 \leq p < \infty$. □

The approach used to estimate the probability of a small ℓ^1-ball does not impact the rate: the bound in Lemma 3 agrees with the one in Lemma 1 in showing that, when $d = 1/2$, the prior probability of an ℓ^1-ball around a given mixing distribution is bounded below by a term of the order $O(\exp(-cN^2 \log(1/\varepsilon)))$ which, differently from the one for the Dirichlet process, involves the squared number N^2 of support points. This does not have remarkable consequences on posterior contraction rates in the supersmooth case because N is of logarithmic order and a possible loss in the rate would only incur an additional logarithmic factor, but may have serious consequences in the ordinary smooth case, where N is of polynomial order, thus possibly leading to suboptimal posterior contraction rates because small Kullback-Leibler type balls around the sampling density f_0 are not charged enough prior mass. Admittedly, these are only lower bounds, but we believe they cannot be substantially improved.

5.2 Normalized inverse-Gaussian process

We prove an analogue of Lemma 6.1 in Ghosal *et al.* (2000), pages 518–519, or Lemma A.1 in Ghosal (2001), pages 1278–1279, which provides an estimate of the probability of an ℓ^1-ball in $\mathbb{R}^N$ under the N-IG distribution.

Lemma 4. *Let* $(Z_1, \ldots, Z_N) \sim \text{N-IG}(\alpha_1, \ldots, \alpha_N)$, $2 \leq N < \infty$. *For* $0 < \varepsilon < 1$, *let* $U := (\sum_{j=1}^N |Z_j - z_{j0}| \leq 2\varepsilon,\ \min_{1\leq j\leq N} Z_j > \varepsilon^2/2)$, *with* $(z_{10}, \ldots, z_{N0}) \in \Delta^{N-1}$. *Assume that* $A\varepsilon^b \leq \alpha_j \leq 1$, $j = 1, \ldots, N$, *for some constants* $0 < A, b < \infty$. *If* $2/N \ll \min_{1\leq j\leq N} z_{j0} - \varepsilon$ *and* $\max_{1\leq j\leq N} z_{j0} < 1 - \varepsilon^2$, *then there exist constants* $0 < c, C < \infty$ *(depending only on* A, b *and* $m := \sum_{j=1}^N \alpha_j$*) such that, for* $0 < \varepsilon < 1/N$ *and* $N \to \infty$ *as* $\varepsilon \to 0$, *we have* $\mathrm{P}(U) \geq C \exp\left(-cN \log(1/\varepsilon)\right)$.

Proof. As in the proof of Lemma 6.1 of Ghosal *et al.* (2000), pages 518–519, we can assume that $z_{N0} \geq 1/N$. If $|Z_j - z_{j0}| \leq \varepsilon^2$ for every $j = 1, \ldots, N-1$, then $\sum_{j=1}^N |Z_j - z_{j0}| \leq 2\varepsilon$ and $Z_N \geq \varepsilon^2 > \varepsilon^2/2$. Therefore, the event U is implied by $V := (|Z_j - z_{j0}| \leq \varepsilon^2,\ \ Z_j > \varepsilon^2/2,\ \ j = 1, \ldots, N-1)$. For $l_j := [(z_{j0} - \varepsilon^2) \vee (\varepsilon^2/2)]$ and $u_j := [(z_{j0} + \varepsilon^2) \wedge 1]$,

$j = 1, \ldots, N-1$, the probability $\mathrm{P}(V) = \int_{l_1}^{u_1} \cdots \int_{l_{N-1}}^{u_{N-1}} f(z_1, \ldots, z_{N-1})\, \mathrm{d}z_1 \cdots \mathrm{d}z_{N-1}$, where $f(\cdot) = h_1 \prod_{r=2}^{4} h_r(\cdot)$, with h_1, $h_2(\cdot)$, $\ldots$, $h_4(\cdot)$ as in (7). Then,

$$\begin{aligned}\mathrm{P}(V) &\geq \frac{e^m (A\varepsilon^b)^N}{2^{N/2-1}\pi^{N/2}} \times (em)^{-N/2} \left(\min_{1 \leq j \leq N} z_{j0} - \varepsilon \right)^{N/2} \times (2\varepsilon^2)^{N-1} \\ &\gtrsim \exp\left(-cN \max\left\{ \log(1/\varepsilon),\, -\log\left(\min_{1 \leq j \leq N} z_{j0} - \varepsilon \right) \right\} \right),\end{aligned}$$

where h_1 is bounded below using the constraint $\alpha_j \geq A\varepsilon^b$, while $h_4(z_1, \ldots, z_{N-1}) \geq 1$ because $z_j \leq 1$ for every $j = 1, \ldots, N-1$. To bound below $h_2(\cdot)$, note that $K_{-N/2}(\cdot) = K_{N/2}(\cdot)$, see 9.6.6 in Abramowitz and Stegun (1964), page 375. Since, for $\varepsilon > 0$ small enough,

$$[\mathcal{A}_N(z_1, \ldots, z_{N-1})]^{1/2} < m^{1/2} \left(\min_{1 \leq j \leq N} z_{j0} - \varepsilon \right)^{-1/2} \ll (N/2+1)^{1/2},$$

we have $h_2(z_1, \ldots, z_{N-1}) \sim 2^{N/2-1}\Gamma(N/2)[\mathcal{A}_N(z_1, \ldots, z_{N-1})]^{-N/4}$ for $N \to \infty$ as $\varepsilon \to 0$, *ibidem*, formula 9.6.9. By Stirling's formula,

$$h_2(z_1, \ldots, z_{N-1}) \gtrsim e^{-N/2} m^{-N/4} \left(\min_{1 \leq j \leq N} z_{j0} - \varepsilon \right)^{N/4}.$$

Consequently, $h_2(z_1, \ldots, z_{N-1}) \times h_3(z_1, \ldots, z_{N-1}) \gtrsim (em)^{-N/2} (\min_{1 \leq j \leq N} z_{j0} - \varepsilon)^{N/2}$. □

In principle, one could employ the stick-breaking representation for the N-IG process discovered by Favaro *et al.* (2012), which represents the first case of a tractable prior with explicit stick-breaking representation based on dependent weights. However, the approach based on bounding below the prior probability mass of a Kullback-Leibler type ball around the "truth" by the product of the prior masses of ℓ^1-balls of the mixing weights and locations (see the events in a) and b) in the proof of Theorem 1), would yield too small prior estimates using the arguments of Lemma 1. So, in this case, even if the technique developed in the article to deal with stick-breaking priors could still be applied to the N-IG process, the approach based on the finite-dimensional distributions seems to yield a more accurate estimate.

6 Approximation results and proof of Theorem 2

The main difficulty when proving the result on adaptive estimation of analytic densities using Gaussian convolution mixtures lies in finding a finite mixing distribution with a number $N = O((\log(1/\varepsilon))^k)$, for some $0 < k < \infty$, of support points such that the corresponding Gaussian mixture is within ε Kullback-Leibler distance from f_0. Such a finite mixing distribution can be found by matching a certain number of its moments with those of an *ad hoc* constructed mixing density. The crux is the approximation of an

analytic density having sub-exponentially decaying Fourier transform by convoluting the Gaussian kernel with an operator whose expression is a series with suitably calibrated coefficients and derivatives convoluted with the sinc kernel or, more generally, with a superkernel. Inspection of the proof of Lemma 5 below reveals that the operation of convoluting the above described transformation of f_0 with the Gaussian density allows to reproduce the tail behavior of the Fourier transform of f_0. Once this (not necessarily non-negative) function is modified to be a density with the same tail behavior as f_0 and with the same approximation properties in the supremum norm as well as in the Kullback-Leibler divergence, the re-normalized restriction to a compact set of the corresponding continuous mixture is discretized.

We begin by presenting the result on the approximation of analytic densities having Fourier transforms with sub-exponentially decaying tails by convolutions with the Gaussian kernel. For every $j \in \mathbb{N}$, let $m_j := \int y^j \phi(y)\,\mathrm{d}y$ be the moment of order j of a standard normal distribution. Odd moments are null and even moments $m_{2j} = (2j)!\,/(2^j j!\,)$, $j = 1, 2, \ldots$. We define two collections of numbers $(c_{2j})_{j\geq 1}$ and $(d_{2j})_{j\geq 1}$ that only depend on the moments of ϕ. Set $c_2 := 0$ and $d_2 := m_2/2!$, let

$$c_{2j} := \sum_{\substack{k,l\geq 1\\ k+l=j}} (-1)^l \frac{m_{2k}}{(2k)!}\frac{m_{2l}}{(2l)!} \qquad \text{and} \qquad d_{2j} := \frac{m_{2j}}{(2j)!} + c_{2j}, \quad j = 2, 3, \ldots .$$

For any $0 < \sigma < \infty$ and any function $f \in C^\infty(\mathbb{R})$, define the transform

$$T_\sigma(f) := f - \sum_{j=1}^{\infty} d_{2j}\sigma^{2j}(f^{(2j)} * \mathrm{sinc}_\sigma).$$

The following lemma asserts that any analytic density f_0, whose Fourier transform has sub-exponentially decaying tails, can be uniformly approximated, with exponentially small error in terms of the inverse of the bandwidth σ, by convoluting its transform $T_\sigma(f_0)$ with the Gaussian kernel.

Lemma 5. *Let $f_0 \in \mathcal{A}^{\rho_0, r_0, L_0}(\mathbb{R})$, with $0 < \rho_0$, $L_0 < \infty$ and $1 \leq r_0 < \infty$. For $\sigma > 0$ small enough,*

$$\|T_\sigma(f_0) * \phi_\sigma - f_0\|_\infty \lesssim e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}) \tag{15}$$

and

$$T_\sigma(f_0) = 2f_0 - f_0 * \phi_\sigma - \sum_{j=1}^{\infty} c_{2j}\sigma^{2j} f_0^{(2j)} + O(e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0})). \tag{16}$$

Proof. For $1 \leq r_0 < \infty$, the density f_0 is analytic on $\mathbb{R}$, $f_0 \in C^\omega(\mathbb{R})$, see, *e.g.*, Theorem 11.7.1 in Kawata (1972), pages 439–440. For any $0 < \sigma < \infty$, by definition of

$T_\sigma(f_0)$, the fact that $f_0 \in C^\omega(\mathbb{R})$ and Taylor's formula, for every $x \in \mathbb{R}$,

$$
\begin{aligned}
&[T_\sigma(f_0) * \phi_\sigma - f_0](x) \\
&\quad = \int \left[f_0(x-y) - f_0(x) - \sum_{j=1}^{\infty} d_{2j}\sigma^{2j}(f_0^{(2j)} * \mathrm{sinc}_\sigma)(x-y) \right] \phi_\sigma(y)\,\mathrm{d}y \\
&\quad = \sum_{j=1}^{\infty} \left[\frac{m_{2j}}{(2j)!}\sigma^{2j} f_0^{(2j)}(x) - d_{2j}\sigma^{2j}(f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \right] \\
&\quad = \sum_{j=1}^{\infty} \left[\frac{m_{2j}}{(2j)!}\sigma^{2j}(f_0^{(2j)} - f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) - c_{2j}\sigma^{2j}(f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \right],
\end{aligned}
$$

where, in the third line, the order of integration and summation has been interchanged, which is licit, and, in the last one, the definition of the d_{2j} has been used. For every $j \in \mathbb{N}$,

$$
\begin{aligned}
&(f_0^{(2j)} - f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \\
&\quad = (f_0^{(2j)} - f_0^{(2j)} * \mathrm{sinc}_\sigma)(x) + (f_0^{(2j)} * \mathrm{sinc}_\sigma - f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \\
&\quad = \frac{1}{2\pi} \int_{|t|>1/\sigma} e^{-itx}(-it)^{2j}\widehat{f}_0(t)\,\mathrm{d}t \\
&\qquad\qquad + \frac{1}{2\pi}\int e^{-itx}(-it)^{2j}\widehat{f}_0(t)1_{[-1,\,1]}(\sigma t)\,\mathrm{d}t - (f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \\
&\quad =: T_1(2j,\,\sigma,\,x) + T_2(2j,\,\sigma,\,x), \quad x \in \mathbb{R}.
\end{aligned}
$$

By the Cauchy-Schwarz inequality and the assumption that $\widehat{f}_0$ satisfies (1), for $\sigma > 0$ small enough, $T_1(2j,\,\sigma,\,x) \lesssim \sigma^{-2j} e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0})$, $x \in \mathbb{R}$. Thus,

$$
\sum_{j=1}^{\infty} \frac{m_{2j}}{(2j)!}\sigma^{2j} T_1(2j,\,\sigma,\,x) \lesssim e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}), \quad x \in \mathbb{R},
$$

because $\sum_{j=1}^{\infty} m_{2j}/(2j)! < \infty$. We show that

$$
\sum_{j=1}^{\infty} \left[\frac{m_{2j}}{(2j)!}\sigma^{2j} T_2(2j,\,\sigma,\,x) - c_{2j}\sigma^{2j}(f_0^{(2j)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \right] = 0, \quad x \in \mathbb{R}.
$$

Algebra leads to $T_2(2j,\,\sigma,\,x) = \sum_{k=1}^{\infty}[(-1)^k m_{2k}\sigma^{2k}(f_0^{(2j+2k)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x)/(2k)!\,]$, $x \in \mathbb{R}$. Then, by definition of the c_{2s},

$$
\begin{aligned}
\sum_{j=1}^{\infty} \frac{m_{2j}}{(2j)!}\sigma^{2j} T_2(2j,\,\sigma,\,x) &= \sum_{j=1}^{\infty} \frac{m_{2j}}{(2j)!} \sum_{k=1}^{\infty} (-1)^k \frac{m_{2k}}{(2k)!}\sigma^{2(j+k)}(f_0^{(2j+2k)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x) \\
&= \sum_{s=2}^{\infty} c_{2s}\sigma^{2s}(f_0^{(2s)} * \mathrm{sinc}_\sigma * \phi_\sigma)(x), \quad x \in \mathbb{R},
\end{aligned}
$$

which completes the proof of (15).

We now prove (16). We know that, for $\sigma > 0$ small enough,
$T_1(2j,\, \sigma,\, x) \lesssim \sigma^{-2j} e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0})$ for every $x \in \mathbb{R}$,
therefore $T_\sigma(f_0) = f_0 - \sum_{j=1}^{\infty} d_{2j}\sigma^{2j} f_0^{(2j)} + O(e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0})) = 2f_0 - f_0 * \phi_\sigma - \sum_{j=1}^{\infty} c_{2j}\sigma^{2j} f_0^{(2j)} + O(e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}))$ by definition of the d_{2j} and the fact that $\sum_{j=1}^{\infty}[m_{2j}\sigma^{2j} f_0^{(2j)}/(2j)!] = f_0 * \phi_\sigma - f_0$. □

Let $S \in L^2(\mathbb{R})$ be a superkernel as defined in the Appendix. For any $0 < \sigma < \infty$ and any function $f \in C^\infty(\mathbb{R})$, let

$$\widetilde{T}_\sigma(f) := f - \sum_{j=1}^{\infty} d_{2j}\sigma^{2j}(f^{(2j)} * S_\sigma).$$

If $f_0 \in \mathcal{A}^{\rho_0, r_0, L_0}(\mathbb{R})$, for $0 < \rho_0,\, L_0 < \infty$ and $1 \le r_0 < \infty$, then Lemma 5 holds with $\widetilde{T}_\sigma(f_0)$ replacing $T_\sigma(f_0)$. Furthermore, as shown in the next lemma, $\widetilde{T}_\sigma(f_0)$ integrates to 1 because of the absolute integrability of superkernels.

Lemma 6. *Let $f_0 \in \mathcal{A}^{\rho_0, r_0, L_0}(\mathbb{R})$, $0 < \rho_0, r_0, L_0 < \infty$. Suppose that $f_0^{(j)} \in L^1(\mathbb{R})$, $j \in \mathbb{N}$. Then, for every $0 < \sigma < \infty$, we have $\int \widetilde{T}_\sigma(f_0)\, \mathrm{d}\lambda = 1$.*

Proof. The assumption $f_0 \in \mathcal{A}^{\rho_0, r_0, L_0}(\mathbb{R})$ implies that, for every $j \in \mathbb{N} \cup \{0\}$, $f_0^{(j)} \in BC(\mathbb{R})$. Also, $f_0^{(j)} \in L^1(\mathbb{R})$, $j \in \mathbb{N}$, implies that, for every $j \in \mathbb{N}$, $f_0^{(j)}(x) \to 0$ as $|x| \to \infty$. Interchanging the order of integration and summation, which is licit,

$$\int \widetilde{T}_\sigma(f_0)\, \mathrm{d}\lambda = 1 - \sum_{j=1}^{\infty} d_{2j}\sigma^{2j} \int (f_0^{(2j)} * S_\sigma)\, \mathrm{d}\lambda = 1.$$

In fact, for every $j \in \mathbb{N}$, $\|f_0^{(2j)} * S_\sigma\|_1 \le \|f_0^{(2j)}\|_1 \|S_\sigma\|_1 < \infty$ and, by Fubini's theorem, $\int (f_0^{(2j)} * S_\sigma)\, \mathrm{d}\lambda = \int [\int f_0^{(2j)}(x - y) S_\sigma(y)\, \mathrm{d}y]\, \mathrm{d}x = \int [\int f_0^{(2j)}(x - y)\, \mathrm{d}x] S_\sigma(y)\, \mathrm{d}y = 0$. □

Suppose that f_0 satisfies Conditions (a) and (c). Let $(a_j)_{j \ge 1}$ be a sequence of positive numbers such that

$$D := \sum_{j=1}^{\infty} a_{2j} |d_{2j}| < \infty.$$

Given $0 < \beta < 1/2$, $0 < c_1 < [\rho_0^{r_0} \wedge (c_0/2^{\varpi_0 \wedge 2})]$, $0 < B,\, \sigma < \infty$ and $1 < M < \infty$, let

$$\begin{aligned}
B_\sigma &:= \{x \in \mathbb{R} :\, f_0(x) \ge B\sigma^{-M} e^{-c_1(1/\sigma)^{r_0}}\},\\
G_\sigma &:= \{x \in \mathbb{R} :\, \widetilde{T}_\sigma(f_0)(x) \ge \beta f_0(x)\},\\
U_\sigma &:= \{x \in \mathbb{R} :\, |f_0^{(2j)}(x)| \le [(1 - 2\beta)/D] a_{2j} \sigma^{-2j} f_0(x),\ \ j \in \mathbb{N}\}.
\end{aligned}$$

The function $\widetilde{T}_\sigma(f_0)$ is modified to be everywhere non-negative by setting it equal to a multiple of f_0 when it is below it. Let

$$g_\sigma := \widetilde{T}_\sigma(f_0) 1_{G_\sigma} + \beta f_0 1_{G_\sigma^c}$$

be the modified function.

Lemma 7. *Suppose that the probability density f_0 satisfies Conditions (a) and (c). Then, for $\sigma > 0$ small enough, $\int g_\sigma \, \mathrm{d}\lambda \geq \beta$ and $\int g_\sigma \, \mathrm{d}\lambda = 1 + O(e^{-c_3(1/\sigma)^{r_0}})$ for a suitable constant $0 < c_3 < \infty$.*

Proof. By definition, $g_\sigma \geq \beta f_0(1_{G_\sigma} + 1_{G_\sigma^c}) = \beta f_0$. Thus, $\int g_\sigma \, \mathrm{d}\lambda \geq \beta$. Rewritten g_σ as $g_\sigma = \widetilde{T}_\sigma(f_0) + [\beta f_0 - \widetilde{T}_\sigma(f_0)]1_{G_\sigma^c}$, by Lemma (6), for $\sigma > 0$ small enough, $\int g_\sigma \, \mathrm{d}\lambda = 1 + \int[\beta f_0 - \widetilde{T}_\sigma(f_0)]1_{G_\sigma^c} \, \mathrm{d}\lambda$ because, for every $j \in \mathbb{N}$, $\|f_0^{(j)}\|_1 \leq \{\mathrm{E}_0[|(f_0^{(j)}/f_0)(X)|^2]\}^{1/2} < \infty$. To show that $\int[\beta f_0 - \widetilde{T}_\sigma(f_0)]1_{G_\sigma^c} \, \mathrm{d}\lambda = O(e^{-c_3(1/\sigma)^{r_0}})$, we first show that, for $\sigma > 0$ small enough, $B_\sigma \cap U_\sigma \subseteq G_\sigma$. In fact, over $B_\sigma \cap U_\sigma$,

$$
\begin{aligned}
|\widetilde{T}_\sigma(f_0) - f_0| &\leq f_0 \sum_{j=1}^{\infty} |d_{2j}| \sigma^{2j} (|f_0^{(2j)} - f_0^{(2j)} * S_\sigma|/f_0) + f_0 \sum_{j=1}^{\infty} |d_{2j}| \sigma^{2j} (|f_0^{(2j)}|/f_0) \\
&\lesssim f_0 [e^{-(\rho_0/\sigma)^{r_0}}/f_0 + (1 - 2\beta) \sum_{j=1}^{\infty} a_{2j} |d_{2j}|/D] \\
&= [O(\sigma^M) + 1 - 2\beta] f_0 < (1 - \beta) f_0,
\end{aligned}
$$

because, over B_σ, we have $e^{-(\rho_0/\sigma)^{r_0}}/f_0 = O(\sigma^M)$. Hence, $\widetilde{T}_\sigma(f_0) > \beta f_0$ and $B_\sigma \cap U_\sigma \subseteq G_\sigma$. Furthermore, U_σ^c has exponentially small probability. By Markov's inequality, for some constant $1 < \xi < \infty$ and, for instance, $a_{2j} = 2^{j(\log_2 j)/2}$,

$$
\begin{aligned}
P_0(U_\sigma^c) &\leq \sum_{j=1}^{\infty} P_0(|(f_0^{(2j)}/f_0)(X)| > (1 - 2\beta) a_{2j} \sigma^{-2j}/D) \\
&< \sup_{j \geq 1} \mathrm{E}_0[\exp(c_0 |(f_0^{(2j)}/f_0)(X)|^{r_0/2j}/\xi)] \\
&\qquad \times \sum_{j=1}^{\infty} \exp\left(-c_0 [(1 - 2\beta) a_{2j}/D]^{r_0/2j} (1/\sigma)^{r_0}/\xi\right) \\
&\lesssim e^{-k_3(1/\sigma)^{r_0}},
\end{aligned}
$$

where the expected value is finite under Condition (c). In fact, for every $j \in \mathbb{N}$,

$$
f_0(x) \exp(c_0 |(f_0^{(2j)}/f_0)(x)|^{r_0/2j}/\xi) \sim \exp\left(-c_0 |x|^{\varpi_0} + c_0 |x|^{(\varpi_0 - 1) r_0}/\xi\right),
$$

with $\exp\left(-c_0 |x|^{\varpi_0} + c_0 |x|^{(\varpi_0 - 1) r_0}/\xi\right) \to 0$ as $|x| \to \infty$ provided that $(\varpi_0 - 1) r_0 \leq \varpi_0$. Using the bounds $P_0(U_\sigma^c) \lesssim e^{-k_3(1/\sigma)^{r_0}}$, $P_0(B_\sigma^c) \lesssim (\sigma^{-M} e^{-c_1(1/\sigma)^{r_0}})^\nu$ valid for every $0 < \nu < 1$, and the fact that, by (16), up to $O(e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}))$, the transform $\widetilde{T}_\sigma(f_0)$ is a linear combination of f_0, $f_0 * \phi_\sigma$ and the $f_0^{(2j)}$, we prove that $\int[\beta f_0 - \widetilde{T}_\sigma(f_0)]1_{G_\sigma^c} \, \mathrm{d}\lambda \lesssim e^{-c_3(1/\sigma)^{r_0}}$. We begin by showing that $\int_{U_\sigma^c} (f_0 * \phi_\sigma) \, \mathrm{d}\lambda \lesssim e^{-(k_3 \wedge 4^{-1})(1/\sigma)^{r_0}}$. For random variables $Y \sim f_0$ and $Z \sim \mathrm{N}(0, 1)$, $\int_{U_\sigma^c} (f_0 * \phi_\sigma) \, \mathrm{d}\lambda \leq \mathrm{P}(Y + \sigma Z \in U_\sigma^c, |Z| \leq \sigma^{-r_0/2}) + \mathrm{P}(|Z| > \sigma^{-r_0/2}) =: T_1 + T_2$, where $T_2 \lesssim e^{-(1/\sigma)^{r_0}/4}$ and $T_1 \leq P_0(U_\sigma^c) \lesssim e^{-k_3(1/\sigma)^{r_0}}$ for

$\sigma > 0$ small enough. By the result just shown, $\int_{B_\sigma^c}(f_0 * \phi_\sigma)\,\mathrm{d}\lambda \lesssim \int_{B_\sigma^c \cap U_\sigma}(f_0 * \phi_\sigma)\,\mathrm{d}\lambda + e^{-(k_3 \wedge 4^{-1})(1/\sigma)^{r_0}}$, where, for $1 < \xi' < \infty$,

$$\int_{B_\sigma^c \cap U_\sigma}(f_0 * \phi_\sigma)\,\mathrm{d}\lambda \leq \mathrm{P}(Y + \sigma Z \in B_\sigma^c \cap U_\sigma, |Z| \leq \sigma^{-r_0/2}, Y \in B_{\xi'\sigma} \cap U_\sigma) + \mathrm{P}(Y \in U_\sigma^c) + \mathrm{P}(Y \in B_{\xi'\sigma}^c) + \mathrm{P}(|Z| > \sigma^{-r_0/2}) \lesssim e^{-k_4(1/\sigma)^{r_0}}.$$

Since $\widetilde{T}_\sigma(f_0) \to f_0$ as $\sigma \to 0$, we have $\lambda(G_\sigma^c) = o(1)$. For $c_3 := \min\{\nu c_1/2,\, k_3/2,\, k_4\}$,

$$\begin{aligned}\int [\beta f_0 - \widetilde{T}_\sigma(f_0)]1_{G_\sigma^c}\,\mathrm{d}\lambda &\leq \int |\beta f_0 - \widetilde{T}_\sigma(f_0)|1_{G_\sigma^c}\,\mathrm{d}\lambda \\ &\leq (\beta+2)[P_0(B_\sigma^c) + P_0(U_\sigma^c)] + \int_{G_\sigma^c}(f_0 * \phi_\sigma)\,\mathrm{d}\lambda \\ &\quad + \sum_{j=1}^{\infty}|c_{2j}|\sigma^{2j}\int_{G_\sigma^c}|f_0^{(2j)}|\,\mathrm{d}\lambda + O(e^{-(\rho_0/\sigma)^{r_0}}1_{\{\infty\}}(S_{f_0}))\lambda(G_\sigma^c) \\ &\lesssim [(\sigma^{-M}e^{-c_1(1/\sigma)^{r_0}})^\nu + e^{-k_3(1/\sigma)^{r_0}}]^{1/2} + e^{-k_4(1/\sigma)^{r_0}} \\ &\quad + e^{-(k_3 \wedge 4^{-1})(1/\sigma)^{r_0}} + o(e^{-(\rho_0/\sigma)^{r_0}}1_{\{\infty\}}(S_{f_0})) \\ &\lesssim e^{-c_3(1/\sigma)^{r_0}},\end{aligned}$$

where, for every $j \in \mathbb{N}$, $\int_{G_\sigma^c}|f_0^{(2j)}|\,\mathrm{d}\lambda \leq \{\mathrm{E}_0[|(f_0^{(2j)}/f_0)(X)|^2]P_0(G_\sigma^c)\}^{1/2}$ and

$$\sum_{j=1}^{\infty}|c_{2j}|\sigma^{2j}\{\mathrm{E}_0[|(f_0^{(2j)}/f_0)(X)|^2]\}^{1/2} < \infty.$$

□

Next, a finite mixture of normal densities is constructed from the re-normalized restriction to a compact set of the density derived from g_σ such that it still approximates f_0, in the Kullback-Leibler divergence, with an error of the order $O(e^{-c(1/\sigma)^{r_0}})$ for some constant $0 < c < \infty$.

Lemma 8. *Suppose that the probability density f_0 satisfies Conditions (a)–(c). For $\sigma > 0$ small enough, there exists a finite Gaussian mixture m_σ, having at most $N_\sigma = O((a_\sigma/\sigma)^2)$ support points in $[-a_\sigma, a_\sigma]$, with $a_\sigma = O((1/\sigma)^{r_0/(\varpi_0 \wedge 2)})$, such that, for suitable constants $0 < c_5$, $S < \infty$,*

$$\max\{\mathrm{KL}(f_0;\, m_\sigma),\, \mathrm{E}_0[(\log(f_0/m_\sigma))^2]\} \lesssim \sigma^{-S}e^{-c_5(1/\sigma)^{r_0}}. \tag{17}$$

Proof. We present a detailed proof only for the upper bound on the Kullback-Leibler divergence, which is decomposed into the sum of three integrals, see (18) below. We bound the first integral. Fix $0 < \zeta < 1$ and let C_ζ be the same constant appearing in Lemma 21. Set $C_{g_\sigma} := \int g_\sigma\,\mathrm{d}\lambda$, by Lemma 7, for $\sigma > 0$ small enough, $C_{g_\sigma} =$

$1 + Ae^{-c_3(1/\sigma)^{r_0}}$ for a suitable constant $0 < A < \infty$. Define the density $h_\sigma := g_\sigma / C_{g_\sigma}$. For τ_ζ as in Lemma 21,

$$\forall\, 0 < \sigma \le \tau_\zeta, \quad h_\sigma * \phi_\sigma \ge \frac{\beta(f_0 * \phi_\sigma)}{1 + Ae^{-c_3(1/\sigma)^{r_0}}} \ge \frac{\beta C_\zeta}{1 + Ae^{-c_3(1/\sigma)^{r_0}}} f_0$$

because $g_\sigma \ge \beta f_0$ and, by virtue of Condition (b), Lemma 21 can be invoked. Furthermore, $|h_\sigma * \phi_\sigma - f_0| \le C_{g_\sigma}^{-1} |g_\sigma * \phi_\sigma - f_0| + |C_{g_\sigma}^{-1} - 1| f_0 \lesssim |g_\sigma * \phi_\sigma - f_0| + e^{-c_3(1/\sigma)^{r_0}} f_0$. Then, Lemma 5 implies that

$$\begin{aligned} |g_\sigma * \phi_\sigma - f_0| &\le |\widetilde{T}_\sigma(f_0) * \phi_\sigma - f_0| + |\{[\beta f_0 - \widetilde{T}_\sigma(f_0)] 1_{G_\sigma^c}\} * \phi_\sigma| \\ &\lesssim e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}) + |\{[\beta f_0 - \widetilde{T}_\sigma(f_0)] 1_{G_\sigma^c}\} * \phi_\sigma|. \end{aligned}$$

Now, $\mathrm{KL}(f_0;\, h_\sigma * \phi_\sigma) = (\int_{B_\sigma \cap U_\sigma} + \int_{B_\sigma^c \cup U_\sigma^c})[f_0 \log(f_0/(h_\sigma * \phi_\sigma))]\, \mathrm{d}\lambda =: I_1 + I_2$. Recalling that $c_3 := \min\{\nu c_1/2,\, k_3/2,\, k_4\}$,

$$\begin{aligned} I_1 \le \int_{B_\sigma \cap U_\sigma} f_0 \frac{|h_\sigma * \phi_\sigma - f_0|}{h_\sigma * \phi_\sigma}\, \mathrm{d}\lambda &\le \int_{B_\sigma \cap U_\sigma} f_0 \frac{|g_\sigma * \phi_\sigma - f_0| + e^{-c_3(1/\sigma)^{r_0}} f_0}{h_\sigma * \phi_\sigma}\, \mathrm{d}\lambda \\ &\lesssim e^{-[c_3 \wedge (\rho_0^{r_0} - c_1)](1/\sigma)^{r_0}}, \end{aligned}$$

where $\|\{[\beta f_0 - \widetilde{T}_\sigma(f_0)] 1_{G_\sigma^c}\} * \phi_\sigma\|_1 \le \|[\beta f_0 - \widetilde{T}_\sigma(f_0)] 1_{G_\sigma^c}\|_1 \|\phi_\sigma\|_1 \lesssim e^{-c_3(1/\sigma)^{r_0}}$. By Lemma 7, $\int_{B_\sigma^c \cup U_\sigma^c} f_0\, \mathrm{d}\lambda \lesssim (\sigma^{-M} e^{-c_1(1/\sigma)^{r_0}})^\nu + e^{-k_3(1/\sigma)^{r_0}}$. Therefore,

$$I_2 \lesssim (\sigma^{-\nu M} e^{-\nu c_1 (1/\sigma)^{r_0}} + e^{-k_3(1/\sigma)^{r_0}}) \log((1 + Ae^{-c_3(1/\sigma)^{r_0}})/(\beta C_\zeta)),$$

where the logarithmic term is positive because $0 < \beta C_\zeta < 1$. Thus,

$$\mathrm{KL}(f_0;\, h_\sigma * \phi_\sigma) \lesssim \sigma^{-\nu M} e^{-[c_3 \wedge (\rho_0^{r_0} - c_1)](1/\sigma)^{r_0}}.$$

Let $C_{h_\sigma} := \int_{-a_\sigma}^{a_\sigma} h_\sigma\, \mathrm{d}\lambda$ and define $\tilde{h}_\sigma := h_\sigma 1_{[-a_\sigma,\, a_\sigma]} / C_{h_\sigma}$ as the re-normalized restriction of h_σ to $[-a_\sigma,\, a_\sigma]$. By Lemma 17, there exists a discrete distribution $\tilde{F}$ on $[-a_\sigma,\, a_\sigma]$, with at most $N_\sigma = O((a_\sigma/\sigma)^2)$ support points, such that $\|\tilde{h}_\sigma * \phi_\sigma - \tilde{F} * \phi_\sigma\|_\infty \lesssim \sigma^{-1} e^{-N_\sigma}$. Set $\widetilde{m}_\sigma := C_{h_\sigma}(\tilde{F} * \phi_\sigma)$, we have $|h_\sigma * \phi_\sigma - \widetilde{m}_\sigma| \le \sigma^{-1} e^{-N_\sigma} + (h_\sigma 1_{[-a_\sigma,\, a_\sigma]^c}) * \phi_\sigma$. For $\sigma > 0$ small enough, $(h_\sigma 1_{[-a_\sigma,\, a_\sigma]^c}) * \phi_\sigma \lesssim e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}) + e^{-c_0 (a_\sigma/2)^{\varpi_0 \wedge 2}}$ by virtue of Lemma 5 and Condition (c) on f_0. Thus, for a constant $c_1 < c'' \le [\rho_0^{r_0} \wedge (c_0/2^{\varpi_0 \wedge 2})]$ and $a_\sigma \propto (1/\sigma)^{r_0/(\varpi_0 \wedge 2)}$,

$$\|h_\sigma * \phi_\sigma - \widetilde{m}_\sigma\|_\infty \lesssim \sigma^{-1} e^{-N_\sigma} + e^{-(\rho_0/\sigma)^{r_0}} 1_{\{\infty\}}(S_{f_0}) + e^{-c_0 (a_\sigma/2)^{\varpi_0 \wedge 2}} \lesssim e^{-c''(1/\sigma)^{r_0}}.$$

Let $t_\sigma := \widetilde{m}_\sigma + D_\sigma \phi_\sigma$, with $D_\sigma := \sigma^{-(R-1)} e^{-\tilde{c}(1/\sigma)^{r_0}}$ for $1 < R < M$ and $c_1 < \tilde{c} < \infty$. Define the finite Gaussian mixture $m_\sigma := (\int t_\sigma\, \mathrm{d}\lambda)^{-1} t_\sigma = (\widetilde{m}_\sigma + D_\sigma \phi_\sigma)/(C_{h_\sigma} + D_\sigma)$. Write

$$\begin{aligned} \mathrm{KL}(f_0;\, m_\sigma) &= \int f_0 \log \frac{f_0}{h_\sigma * \phi_\sigma}\, \mathrm{d}\lambda + \int f_0 \log \frac{h_\sigma * \phi_\sigma}{t_\sigma}\, \mathrm{d}\lambda + \int f_0 \log \frac{t_\sigma}{m_\sigma}\, \mathrm{d}\lambda \qquad (18) \\ &=: J_1 + J_2 + J_3, \end{aligned}$$

where $J_1 = \mathrm{KL}(f_0; h_\sigma * \phi_\sigma)$.

• *Control of* J_1. It has been shown that $J_1 \lesssim \sigma^{-\nu M} e^{-[c_3 \wedge (\rho_0^{r_0} - c_1)](1/\sigma)^{r_0}}$.

• *Control of* J_2. Write $J_2 = (\int_{B_\sigma} + \int_{B_\sigma^c})[f_0 \log((h_\sigma * \phi_\sigma)/t_\sigma)]\,\mathrm{d}\lambda =: J_{21} + J_{22}$. Since $0 < c_1 < (c'' \wedge \tilde{c})$,

$$\begin{aligned} J_{21} &\leq \int_{B_\sigma} f_0 \frac{|h_\sigma * \phi_\sigma - t_\sigma|}{t_\sigma}\,\mathrm{d}\lambda \\ &\lesssim \frac{\sigma^{-R} e^{-(c'' \wedge \tilde{c})(1/\sigma)^{r_0}}}{B\sigma^{-M} e^{-c_1(1/\sigma)^{r_0}} - e^{-c''(1/\sigma)^{r_0}}} \int_{B_\sigma} f_0\,\mathrm{d}\lambda \lesssim \sigma^{M-R} e^{-[(c'' \wedge \tilde{c}) - c_1](1/\sigma)^{r_0}}, \end{aligned}$$

because $|h_\sigma * \phi_\sigma - t_\sigma| \leq |h_\sigma * \phi_\sigma - \widetilde{m}_\sigma| + D_\sigma \phi_\sigma \lesssim \sigma^{-R} e^{-(c'' \wedge \tilde{c})(1/\sigma)^{r_0}}$ and, over B_σ, $h_\sigma * \phi_\sigma \gtrsim f_0 \geq B\sigma^{-M} e^{-c_1(1/\sigma)^{r_0}}$ so that $t_\sigma > \widetilde{m}_\sigma \geq h_\sigma * \phi_\sigma - |h_\sigma * \phi_\sigma - \widetilde{m}_\sigma| \gtrsim B\sigma^{-M} e^{-c_1(1/\sigma)^{r_0}} - e^{-c''(1/\sigma)^{r_0}}$. Because $\|h_\sigma * \phi_\sigma\|_\infty \leq C_0 < \infty$ for a constant C_0 (possibly depending on f_0) and $t_\sigma \geq D_\sigma \phi_\sigma$,

$$\begin{aligned} J_{22} &\lesssim \log(\sigma/D_\sigma) \int_{B_\sigma^c} f_0\,\mathrm{d}\lambda + \frac{1}{2\sigma^2} \int_{B_\sigma^c} x^2 f_0(x)\,\mathrm{d}x \\ &\lesssim \sigma^{-(\nu M + r_0)} e^{-\nu c_1 (1/\sigma)^{r_0}} + \sigma^{-(\nu M + 2)} e^{-\nu c_1 (1/\sigma)^{r_0}} \lesssim \sigma^{-[\nu M + (r_0 \vee 2)]} e^{-\nu c_1 (1/\sigma)^{r_0}}. \end{aligned}$$

• *Control of* J_3. Noting that $(t_\sigma/m_\sigma) = C_{h_\sigma} + D_\sigma \leq 1 + D_\sigma$, we get that $J_3 \leq \log(1 + D_\sigma) \leq D_\sigma = \sigma^{-(R-1)} e^{-\tilde{c}(1/\sigma)^{r_0}}$.

Combining partial results, $\mathrm{KL}(f_0; m_\sigma) \lesssim \sigma^{-S} e^{-c_5(1/\sigma)^{r_0}}$, where $S \geq \max\{R - 1, \nu M + 2\}$ and $c_5 := \min\{[c_3 \wedge (\rho_0^{r_0} - c_1)], [(c'' \wedge \tilde{c}) - c_1]\}$ are finite constants.

The same reasoning applies to $\mathrm{E}_0[(\log(f_0/m_\sigma))^2]$ and (17) follows. Constants stemming from the upper bound on this term may possibly be just multiplied by 2. We still denote the overall constants by S and c_5. □

Proof of Theorem 2. As in the proof of Theorem 1, we first prove the result for the L^1-metric. We then deal with L^p-metrics, $2 \leq p \leq \infty$. The case of L^p-metrics, with $1 < p < 2$, is covered by interpolation.

• L^1*-metric.* As in the proof of Theorem 1, for $\tilde{\varepsilon}_n = n^{-1/2}(\log n)^{\psi(r_0, d)}$, with $\psi(r_0, d)$ as in (10), since $2\psi(r_0, d) > 1$ for every $0 \leq d < 1$, we have $\varepsilon_{n,1} := (\bar{\varepsilon}_n \vee \tilde{\varepsilon}_n) = \bar{\varepsilon}_n = n^{-1/2}(\log n)^{\frac{1}{2} + \{\frac{1}{2} \vee [2(\frac{1}{\delta} + \frac{1}{\gamma})\psi(r_0, d)]\}}$.

• L^p*-metrics,* $2 \leq p \leq \infty$. We appeal to Proposition 1. Let $\varepsilon_{n,p} := \tilde{\varepsilon}_n (n\tilde{\varepsilon}_n^2)^{(1-1/p)/2}$. By the assumption that $f_0 \in \mathcal{A}^{\rho_0, r_0, L_0}(\mathbb{R})$, we have $f_0 \in L^p(\mathbb{R})$, $2 \leq p \leq \infty$, and, by virtue of Lemma 12, for $2^{J_n} = cn\tilde{\varepsilon}_n^2$, with c as in Proposition 1 ($r = 2$ and $\rho = 2^{-1/2}$ for the Gaussian kernel), $\|f_0 - f_0 * \mathrm{sinc}_{2^{-J_n}}\|_p \lesssim \exp(-(\rho_0 cn\tilde{\varepsilon}_n^2)^{r_0}) \lesssim n^{-1} \lesssim \varepsilon_{n,p}$ for n large enough because $2r_0\psi(r_0, d) > 1$ for every $0 \leq d < 1$. The requirements of Proposition 1 that $1 < \gamma \leq \infty$ and $(\log n)^{2\psi(r_0, d)} \gtrsim (\log n)^{1/[2(1-1/\gamma)]}$ are both satisfied for $2 \leq \gamma < \infty$.

• *Small ball probability estimate.* We show that, for a suitable constant $0 < c_2 < \infty$, $(\Pi \times G)(B_{\mathrm{KL}}(f_0; \tilde{\varepsilon}_n^2)) \gtrsim \exp(-c_2 n \tilde{\varepsilon}_n^2)$ for every sufficiently large n. By Lemma 8, for $\sigma > 0$ small enough, there exists a finite mixture of Gaussian densities m_σ, with $N_\sigma = O((a_\sigma/\sigma)^2)$ support points $\theta_1, \ldots, \theta_{N_\sigma}$ in $[-a_\sigma, a_\sigma]$, where $a_\sigma = O((1/\sigma)^{r_0/(\varpi_0 \wedge 2)})$, such that (17) holds. Let $p_1, \ldots, p_{N_\sigma}$ denote the mixing weights of m_σ. The inequality in (17) holds for every mixture of Gaussian densities $m_{\sigma'}$, with $\sigma' \in [\sigma, \sigma + e^{-d_1(1/\sigma)^{r_0}})$, having support points $\theta'_1, \ldots, \theta'_{N_\sigma}$ such that $\sum_{j=1}^{N_\sigma} |\theta'_j - \theta_j| \leq e^{-d_2(1/\sigma)^{r_0}}$ and mixing weights $p'_1, \ldots, p'_{N_\sigma}$ such that $\sum_{j=1}^{N_\sigma} |p'_j - p_j| \leq e^{-d_3(1/\sigma)^{r_0}}$ for suitable constants $0 < d_1, d_2, d_3 < \infty$. Let $\tilde{B}_\sigma := \{x : f_0(x) \geq \zeta_\sigma\}$, with $\zeta_\sigma := B' \sigma^{-S'} e^{-c^*(1/\sigma)^{r_0}}$, where $S' := 2(S-2)$ and $0 < c^* < (3c_5 \wedge c'')$, the constants $S > 2$, c_5 and c'' being those appearing in Lemma 8. For every $F \in \mathscr{M}(\mathbb{R})$ and $\sigma' \in [\sigma, \sigma + e^{-d_1(1/\sigma)^{r_0}})$, we have $\mathrm{KL}(f_0; f_{F,\sigma'}) \lesssim \sigma^{-S} e^{-c_5(1/\sigma)^{r_0}} + (\int_{\tilde{B}_\sigma} + \int_{\tilde{B}_\sigma^c})[f_0 \log(m_{\sigma'}/f_{F,\sigma'})] \, \mathrm{d}\lambda$. We provide an upper bound on the second integral. For every F such that $F([-a_{\sigma'}, a_{\sigma'}]) \geq 1/2$, we have $f_{F,\sigma'}(x) \gtrsim (\sigma')^{-1} \exp(-(x^2 + a_{\sigma'}^2)/(\sigma')^2)$ for all $x \in \mathbb{R}$. From Lemma 8, $\|m_{\sigma'}\|_\infty \lesssim (\sigma')^{-1}$. For any $0 < \omega < 1$, we have $\int_{\tilde{B}_\sigma^c} (x/\sigma')^2 f_0(x) \, \mathrm{d}x \lesssim \sigma^{-2} \zeta_\sigma^\omega$ and $\int_{\tilde{B}_\sigma^c} f_0 \, \mathrm{d}\lambda \lesssim \zeta_\sigma^\omega$. Therefore, for a suitable constant $0 < c' < \omega c^*$, we get $\int_{\tilde{B}_\sigma^c} f_0 \log(m_{\sigma'}/f_{F,\sigma'}) \, \mathrm{d}\lambda \lesssim \int_{\tilde{B}_\sigma^c} (x/\sigma')^2 f_0(x) \, \mathrm{d}x + (a_{\sigma'}/\sigma')^2 \int_{\tilde{B}_\sigma^c} f_0 \, \mathrm{d}\lambda \lesssim e^{-c'(1/\sigma)^{r_0}}$.

As in the proof of Theorem 1, we distinguish the case where the prior for F is a Dirichlet or a N-IG process, from the case where the prior for F is a general Pitman-Yor process with $0 \leq d < 1$ and $-d < c < \infty$.

– *Dirichlet or N-IG process.* Clearly,
$\int_{\tilde{B}_\sigma} f_0 \log(m_{\sigma'}/f_{F,\sigma'}) \, \mathrm{d}\lambda \leq \int_{\tilde{B}_\sigma} f_0 (\|m_{\sigma'} - f_{F,\sigma'}\|_\infty / f_{F,\sigma'}) \, \mathrm{d}\lambda$. Using Lemma 5 of Ghosal and van der Vaart (2007b), page 711,
$\|m_{\sigma'} - f_{F,\sigma'}\|_\infty \lesssim \sigma^{-2} \max_{1 \leq j \leq N_\sigma} \lambda(U_j) + \sigma^{-1} \sum_{j=1}^{N_\sigma} |F(U_j) - p'_j|$, where $U_0, \ldots, U_{N_\sigma}$ is a partition of $\mathbb{R}$, with $U_0 := (\bigcup_{j=1}^{N_\sigma} U_j)^c$ and $U_j \ni \theta'_j$ for $j = 1, \ldots, N_\sigma$. The support points of $m_{\sigma'}$ can be taken to be at least $\sigma^{-3(S-2)} e^{-3c_5(1/\sigma)^{r_0}}$-separated. If not, $m_{\sigma'}$ can be projected onto a mixture $m'_{\sigma'}$, with $\sigma^{-3(S-2)} e^{-3c_5(1/\sigma)^{r_0}}$-separated points, such that $\|m_{\sigma'} - m'_{\sigma'}\|_\infty \lesssim \sigma^{-(3S-4)} e^{-3c_5(1/\sigma)^{r_0}}$. There thus exist disjoint intervals $U_1, \ldots, U_{N_\sigma}$ such that $U_j \ni \theta'_j$ and $\sigma^{-3(S-2)} e^{-3c_5(1/\sigma)^{r_0}} \leq \lambda(U_j) \leq 2\sigma^{-3(S-2)} e^{-3c_5(1/\sigma)^{r_0}}$, $j = 1, \ldots, N_\sigma$. Let F be such that

$$\sum_{j=1}^{N_\sigma} |F(U_j) - p'_j| \leq \sigma^{-(3S-5)} e^{-3c_5(1/\sigma)^{r_0}}. \tag{19}$$

Then, $\|m_{\sigma'} - f_{F,\sigma'}\|_\infty \lesssim \sigma^{-(3S-4)} e^{-3c_5(1/\sigma)^{r_0}}$ and, over $\tilde{B}_\sigma$, we have $f_{F,\sigma'} \gtrsim m_{\sigma'} - \sigma^{-(3S-4)} e^{-3c_5(1/\sigma)^{r_0}} \gtrsim \zeta_\sigma$. Therefore, $\int_{\tilde{B}_\sigma} f_0 \log(m_{\sigma'}/f_{F,\sigma'}) \, \mathrm{d}\lambda \lesssim \sigma^{-S} e^{-(3c_5 - c^*)(1/\sigma)^{r_0}}$. Note that, for F satisfying (19), $F([-a_{\sigma'}, a_{\sigma'}]) \geq 1/2$. Combining partial results, $\max\{\mathrm{KL}(f_0; f_{F,\sigma'}), \mathrm{E}_0[(\log(f_0/f_{F,\sigma'}))^2]\} \lesssim \sigma^{-S} e^{-c_6(1/\sigma)^{r_0}}$ for $0 < c_6 \leq \min\{c_5, c', 3c_5 - c^*\}$. To apply Lemma A.2 of Ghosal and van der Vaart (2001), pages 1260–1261, in order to estimate the prior probability of $\{F : \sum_{j=1}^{N_\sigma} |F(U_j) - p'_j| \leq \sigma^{-(3S-5)} e^{-3c_5(1/\sigma)^{r_0}}\}$, note that $\alpha(U_j) \geq \lambda(U_j) \inf_{|\theta| \leq a_{\sigma'}} \alpha'(\theta) \gtrsim \sigma^{-3(S-2)} e^{-(3c_5 + b)(1/\sigma)^{r_0}}$ because $0 < \delta \leq (\varpi_0 \wedge 2)$.

Also, $N_\sigma \sigma^{-(3S-5)} e^{-3c_5(1/\sigma)^{r_0}} \lesssim 1$. Since $r_0 \geq 1$,

$$
\begin{aligned}
&(\Pi \times G)(B_{\mathrm{KL}}(f_0; \sigma^{-S} e^{-c_6(1/\sigma)^{r_0}})) \\
&\quad \gtrsim \Pi \left(F : \sum_{j=1}^{N_\sigma} |F(U_j) - p'_j| \leq \sigma^{-(3S-5)} e^{-3c_5(1/\sigma)^{r_0}} \right) \times G([\sigma, \sigma + e^{-d_1(1/\sigma)^{r_0}})) \\
&\quad \gtrsim \exp\left(-(c_7 N_\sigma + d_1)(1/\sigma)^{r_0} - D_1(1/\sigma)^\gamma (\log(1/\sigma))^t\right).
\end{aligned}
$$

Taking $\sigma \equiv \sigma_n = O((\log n)^{-1/r_0})$, we have $\sigma_n^{-S} e^{-c_6(1/\sigma_n)^{r_0}} \propto n^{-1}(\log n)^{S/r_0} = \tilde{\varepsilon}_n^2$, provided that $(S/r_0) = 2\psi(r_0, 0)$, and $(c_7 N_{\sigma_n} + d_1)(1/\sigma_n)^{r_0} + D_1(1/\sigma_n)^\gamma (\log(1/\sigma_n))^t \lesssim (\log n)^{2\psi(r_0, 0)}$. We need to take $S = 2r_0\psi(r_0, 0)$ while having $S \geq \max\{R-1, \nu M + 2\}$ as prescribed by Lemma 8. Since $2r_0\psi(r_0, 0) > 2$, the latter constraint is satisfied by suitably choosing M and R.

– *Pitman-Yor process with* $0 \leq d < 1$ *and* $-d < c < \infty$. It is enough to note that $\|m_{\sigma'} - f_{F,\sigma'}\|_\infty \lesssim \sigma^{-1} \sum_{j=1}^{M_\sigma} |W_j - p'_j| + \sigma^{-2} \sum_{j=1}^{M_\sigma} p'_j |Z_j - \theta'_j|$ and estimate the probabilities in a) and b) of Theorem 1. We have $\mathrm{P}(\sum_{j=1}^{M_\sigma} |Z_j - \theta'_j| \leq \sigma^{-(3S-5)} e^{-3c_5(1/\sigma)^{r_0}}) \gtrsim \exp(-N_\sigma (1/\sigma)^{r_0})$. Thus,

$$
\begin{aligned}
&(\Pi \times G)(B_{\mathrm{KL}}(f_0; \sigma^{-S} e^{-c_6(1/\sigma)^{r_0}})) \\
&\qquad \gtrsim \exp\left(-[c_8(1 \vee dN_\sigma)N_\sigma + d_1](1/\sigma)^{r_0} - D_1(1/\sigma)^\gamma (\log(1/\sigma))^t\right).
\end{aligned}
$$

Taking $\sigma \equiv \sigma_n = O((\log n)^{-1/r_0})$, we have $\sigma_n^{-S} e^{-c_6(1/\sigma_n)^{r_0}} \propto \tilde{\varepsilon}_n^2$ and

$$
[c_8(1 \vee dN_{\sigma_n})N_{\sigma_n} + d_1](1/\sigma_n)^{r_0} + D_1(1/\sigma_n)^\gamma (\log(1/\sigma_n))^t \lesssim (\log n)^{2\psi(r_0, d)}
$$

as long as $S = 2r_0\psi(r_0, d)$ and $S \geq \max\{R-1, \nu M + 2\}$. □

7 Proof of Theorem 3

Some instrumental results are preliminarily presented. For any $0 < \sigma < \infty$ and any k-times differentiable function f on $\mathbb{R}$, define the transform

$$
T_{k,\sigma}(f) := \begin{cases} f, & \text{for } k = 1, \\ f - \displaystyle\sum_{j=1}^{k-1} d_j \sigma^j f^{(j)}, & \text{for } k = 2, 3, \ldots, \end{cases}
$$

where, abusing the notation introduced in Section 6, $c_1 := 0$, $d_1 := -m_1 = 0$,

$$
c_j := -\sum_{\substack{k,l \geq 1 \\ k+l=j}} (-1)^k \frac{m_k}{k!} d_l \qquad \text{and} \qquad d_j := (-1)^j \frac{m_j}{j!} + c_j, \quad j = 2, 3, \ldots,
$$

m_j being the moment of order j of a standard normal distribution. The following approximation result is an adaptation of Lemma 3.4 in de Jonge and van Zanten (2010), pages 3311 and 3317–3318, which deals with the approximation in the supremum norm of multivariate Hölder functions, see Kruijer *et al.* (2010) for the univariate case.

Lemma 9. *Suppose that, for $k \in \mathbb{N}$, f is a k-times continuously differentiable function on $\mathbb{R}$ with $f^{(k)} \in L^\infty(\mathbb{R})$. For every $x \in \mathbb{R}$,*

$$|[T_{k,\sigma}(f) * \phi_\sigma - f](x)| \sim \frac{m_k}{2(k-1)!}\sigma^k |f^{(k)}(x)| \quad \text{as } \sigma \to 0.$$

Proof. For $x, y \in \mathbb{R}$, let

$$R_k(x, y) := \frac{(-y)^k}{(k-1)!}\int_0^1 f^{(k)}(x - sy)(1-s)^{k-1}\,\mathrm{d}s.$$

Since $f^{(k)} \in BC(\mathbb{R})$, by the dominated convergence theorem, for every $x \in \mathbb{R}$,

$$\int R_k(x, y)\phi_\sigma(y)\,\mathrm{d}y \sim (-1)^k \frac{m_k}{k!}\sigma^k f^{(k)}(x). \tag{20}$$

The proof is by induction on k. For $k = 1$, by definition of $T_{1,\sigma}(f)$, Taylor's formula and (20), for every $x \in \mathbb{R}$,

$$|[T_{1,\sigma}(f) * \phi_\sigma - f](x)| = \left|\int R_1(x, y)\phi_\sigma(y)\,\mathrm{d}y\right| \sim |-m_1\sigma f^{(1)}(x)| = 0.$$

For $k = 2$, by definition, $T_{2,\sigma}(f) = f - d_1\sigma f^{(1)} = f$ because $d_1 = 0$. By Taylor's formula and (20), for every $x \in \mathbb{R}$,

$$|[T_{2,\sigma}(f) * \phi_\sigma - f](x)| = \left|\int R_2(x, y)\phi_\sigma(y)\,\mathrm{d}y\right| \sim \frac{m_2}{2}\sigma^2 |f^{(2)}(x)|.$$

For $k \geq 3$, assume that, for every $1 \leq j \leq k-1$,

$$[f^{(j)} - T_{k-j,\sigma}(f^{(j)}) * \phi_\sigma](x) = C_{k-j}(-1)^{k-j}\frac{m_{k-j}}{(k-j)!}\sigma^{k-j} f^{(k)}(x), \quad x \in \mathbb{R},$$

where, for $(k-j)$ even, $C_{k-j} := \binom{k/2}{j/2}^{-1}$. Then, by Taylor's formula and (20),

$$\begin{aligned}|[T_{k,\sigma}(f) * \phi_\sigma - f](x)| = \Bigg|&\sum_{j=1}^{k-1}(-1)^j\frac{m_j}{j!}\sigma^j f^{(j)}(x) - \sum_{j=1}^{k-1} d_j\sigma^j (f^{(j)} * \phi_\sigma)(x) \\ &+ \int R_k(x, y)\phi_\sigma(y)\,\mathrm{d}y\Bigg| \\ \sim \left|\frac{k}{2}(-1)^k\frac{m_k}{k!}\sigma^k f^{(k)}(x)\right|,&\end{aligned}$$

because

$$\begin{aligned}
&\sum_{j=1}^{k-1}(-1)^j\frac{m_j}{j!}\sigma^j f^{(j)} - \sum_{j=1}^{k-1} d_j\sigma^j f^{(j)} * \phi_\sigma \\
&\quad= \sum_{j=1}^{k-1}(-1)^j\frac{m_j}{j!}\sigma^j (f^{(j)} - f^{(j)} * \phi_\sigma) - \sum_{j=1}^{k-1} c_j\sigma^j f^{(j)} * \phi_\sigma \\
&\quad= \sum_{j=1}^{k-1}(-1)^j\frac{m_j}{j!}\sigma^j [f^{(j)} - T_{k-j,\sigma}(f^{(j)}) * \phi_\sigma] \\
&\qquad\qquad + \sum_{j=1}^{k-1}(-1)^j\frac{m_j}{j!}\sigma^j [T_{k-j,\sigma}(f^{(j)}) * \phi_\sigma - f^{(j)} * \phi_\sigma] - \sum_{j=1}^{k-1} c_j\sigma^j f^{(j)} * \phi_\sigma \\
&\quad= \sum_{j=1}^{k-1}(-1)^j\frac{m_j}{j!}\sigma^j [f^{(j)} - T_{k-j,\sigma}(f^{(j)}) * \phi_\sigma] \\
&\quad= \left(\frac{k}{2} - 1\right)(-1)^k\frac{m_k}{k!}\sigma^k f^{(k)},
\end{aligned}$$

where the term in the fourth line is identically null by definition of the coefficients c_j. □

Let $k_0 \in \{2, 3, \dots\}$. Suppose that f_0 is a k_0-times differentiable density. Let $0 < \vartheta < 1$ and $S_0 := \sum_{j=1}^{k_0-1}|d_j|$. Given $0 < \sigma < \infty$, define

$$\begin{aligned}
G_\sigma &:= \{x \in \mathbb{R} : T_{k_0,\sigma}(f_0)(x) \geq \vartheta f_0(x)\}, \\
U_\sigma &:= \{x \in \mathbb{R} : |f_0^{(j)}(x)| \leq (1-\vartheta)\sigma^{-j} f_0(x)/S_0,\ \ j = 1, \dots, k_0 - 1\}.
\end{aligned}$$

Lemma 10. *Suppose that, for some $k_0 \in \{2, 3, \dots\}$, the probability density f_0 is k_0-times differentiable, with $f_0^{(j)}(x) \to 0$ as $|x| \to \infty$, $j = 1, \dots, k_0 - 1$, and satisfies the integrability conditions in (13). Define $g_{k_0,\sigma} := T_{k_0,\sigma}(f_0)1_{G_\sigma} + \vartheta f_0 1_{G_\sigma^c}$. Then, for $\sigma > 0$ small enough, $\vartheta \leq \int g_{k_0,\sigma}\,\mathrm{d}\lambda = 1 + O(\sigma^{2k_0})$.*

Proof. By definition, $g_{k_0,\sigma} \geq \vartheta f_0$. Hence, $\int g_{k_0,\sigma}\,\mathrm{d}\lambda \geq \vartheta$. Write $g_{k_0,\sigma}$ as $T_{k_0,\sigma}(f_0) + [\vartheta f_0 - T_{k_0,\sigma}(f_0)]1_{G_\sigma^c}$. By the assumption that $f_0^{(j)}(x) \to 0$ as $|x| \to \infty$, $j = 1, \dots, k_0 - 1$, we have $\int g_{k_0,\sigma}\,\mathrm{d}\lambda = 1 + \int[\vartheta f_0 - T_{k_0,\sigma}(f_0)]1_{G_\sigma^c}\,\mathrm{d}\lambda$. We now prove that $\int[\vartheta f_0 - T_{k_0,\sigma}(f_0)]1_{G_\sigma^c}\,\mathrm{d}\lambda = O(\sigma^{2k_0})$. We begin by showing that $U_\sigma \subseteq G_\sigma$. Over U_σ, we have $|T_{k_0,\sigma}(f_0) - f_0| \leq f_0(1-\vartheta)\sum_{j=1}^{k_0-1}|d_j|/S_0 \leq (1-\vartheta)f_0$. Hence, $T_{k_0,\sigma}(f_0) \geq \vartheta f_0$. Consequently, $U_\sigma \subseteq G_\sigma$. The set U_σ^c has exponentially small probability. In fact, by Markov's inequality and the integrability conditions in (13),

$$P_0(U_\sigma^c) \lesssim \sigma^{2k_0}\sum_{j=1}^{k_0-1}\mathrm{E}_0[|(f_0^{(j)}/f_0)(X)|^{2k_0/j}] \lesssim \sigma^{2k_0}$$

and $\int[\vartheta f_0 - T_{k_0,\sigma}(f_0)]1_{U_\sigma^c}\,\mathrm{d}\lambda \lesssim P_0(U_\sigma^c) \lesssim \sigma^{2k_0}$. □

The following lemma can be proved similarly to Proposition 1 in Shen *et al.* (2013), pages 629 and 635, invoking Lemma 9 and Lemma 10.

Lemma 11. *Suppose that the probability density f_0 satisfies Conditions (a') and (b'). For $\sigma > 0$ small enough, there exists a finite Gaussian mixture m_σ, having at most $N_\sigma = O(a_\sigma/\sigma)$ support points in $[-a_\sigma, a_\sigma]$, with $a_\sigma = O((\log(1/\sigma))^{1/2})$, such that* $\max\{\mathrm{KL}(f_0; m_\sigma), \mathrm{E}_0[(\log(f_0/m_\sigma))^2]\} \lesssim \sigma^{2k_0}$.

Proof of Theorem 3. We prove the result for the L^1- and L^2-metrics. The case of L^p-metrics, $1 < p < 2$, is covered by interpolation.

• *L^1-metric.* The entropy condition (2.8) and the small ball probability estimate condition (2.10) of Theorem 2.1 in Ghosal and van der Vaart (2001), page 1239, are shown to be satisfied for $\bar\varepsilon_n = n^{-k_0/(2k_0+1)}(\log n)^{\tau+1+(2\delta)^{-1}}$ and $\tilde\varepsilon_n = n^{-k_0/(2k_0+1)}(\log n)^{\tau}$, respectively, with an appropriate constant $0 < \tau < \infty$. The posterior rate is then $\varepsilon_{n,1} := (\bar\varepsilon_n \vee \tilde\varepsilon_n) = \bar\varepsilon_n$. We begin by considering the entropy condition. For $0 < a, \underline\sigma, \bar\sigma < \infty$ and $0 < \eta < 1$, let $\mathscr{F}_{a,\eta,\underline\sigma,\bar\sigma} := \{f_{F,\sigma} : F([-a, a]) \geq 1 - \eta,\ \underline\sigma \leq \sigma \leq \bar\sigma\}$ and $\mathscr{F}_{a,\underline\sigma,\bar\sigma} := \{f_{F,\sigma} : F([-a, a]) = 1,\ \underline\sigma \leq \sigma \leq \bar\sigma\}$. Combining Lemma A.3 in Ghosal and van der Vaart (2001), page 1261, with Lemma 3 in Ghosal and van der Vaart (2007b), pages 705–707,

$$\begin{aligned}\log D(\eta, \mathscr{F}_{a,\eta/4,\underline\sigma,\bar\sigma}, \|\cdot\|_1) &\leq \log N(\eta/2, \mathscr{F}_{a,\underline\sigma,\bar\sigma}, \|\cdot\|_1)\\ &\lesssim \log\left(\frac{2\bar\sigma}{\eta\underline\sigma}\right) + \left(\frac{a}{\underline\sigma} \vee 1\right)\left(\log\frac{2}{\eta}\right)\left[\log\left(\frac{2a}{\eta\underline\sigma} + 1\right) + \log\frac{2}{\eta}\right].\end{aligned}$$

Choosing $a_n = L(\log n)^{1/\delta}$, $\eta_n = \bar\varepsilon_n$, $\underline\sigma_n = E(n\tilde\varepsilon_n^2)^{-1}$ and $\bar\sigma_n = e^{Fn\tilde\varepsilon_n^2}$ with suitable constants $0 < E, F, L < \infty$, for $\mathscr{F}_n := \mathscr{F}_{a_n,\eta_n/4,\underline\sigma_n,\bar\sigma_n}$, we have $\log D(\bar\varepsilon_n, \mathscr{F}_n, \|\cdot\|_1) \lesssim n\bar\varepsilon_n^2$.

Using results of Doss and Sellke (1982), page 1304, the prior probability of $\mathscr{F}_n^c$ can be bounded above as follows:

$$\begin{aligned}(\Pi \times G)(\mathscr{F}_n^c) &\leq G(\underline\sigma_n) + [1 - G(\bar\sigma_n)] + \frac{4}{\eta_n}\mathrm{E}_\Pi[F([-a_n, a_n]^c)]\\ &\lesssim \underline\sigma_n^{-s}\exp\left(-[D_2\underline\sigma_n^{-1}(\log(1/\underline\sigma_n))^t]\right) + \bar\sigma_n^{\varrho}\\ &\qquad + \frac{4}{\eta_n}\exp\left(-\frac{1}{\bar\alpha(-a_n)[\log\bar\alpha(-a_n)]^2}\right)\\ &\qquad + \frac{4}{\eta_n}\exp\left(-\frac{1}{[1-\bar\alpha(a_n)][\log(1-\bar\alpha(a_n))]^2}\right)\\ &\lesssim \exp\left(-(c_2+4)n\tilde\varepsilon_n^2\right),\end{aligned}$$

where $0 < c_2 < \infty$ is the constant stemming from the small ball probability estimate.

• *L^2-metric.* We appeal to Theorem 3 of Giné and Nickl (2011), page 2892. Choosing their $\gamma_n = 1$, $n \in \mathbb{N}$, the sequence $\varepsilon_{n,2} := \tilde\varepsilon_n$ plays the same role as δ_n. Condition (b) that $\tilde\varepsilon_n^2 = O(n^{-1/2})$ is satisfied for every $k_0 \in \mathbb{N}$. Condition (1) of Theorem 2, *ibidem*, page 2891, is now checked. Let $\underline\sigma_n := E(n\tilde\varepsilon_n^2)^{-1}(\log n)^{\psi}$, for $1/2 < \psi < t$

and a constant $0 < E < \infty$ to be suitably chosen later on, and let $2^{J_n} = cn\tilde{\varepsilon}_n^2$, with any $0 < c < \infty$. Define $\mathscr{F}_n := \{f_{F,\sigma} : F \in \mathscr{M}(\mathbb{R}), \sigma \geq \underline{\sigma}_n\}$. We show that $\mathscr{F}_n \subseteq \{f_{F,\sigma} : \|f_{F,\sigma} - \mathrm{sinc}_{2^{-J_n}}(f_{F,\sigma})\|_2 \leq C(\mathrm{sinc})\varepsilon_{n,2}\}$ for every sufficiently large n. For every $f_{F,\sigma} \in \mathscr{F}_n$ such that $S_F < \infty$, we have $\|f_{F,\sigma} - \mathrm{sinc}_{2^{-J_n}}(f_{F,\sigma})\|_2 = 0$ because $2^{J_n} > S_F$ for all n large enough. For every $f_{F,\sigma} \in \mathscr{F}_n$ such that $S_F = \infty$, we have $\|f_{F,\sigma} - \mathrm{sinc}_{2^{-J_n}}(f_{F,\sigma})\|_2 \lesssim \underline{\sigma}_n^{-1} \exp(-(\rho\underline{\sigma}_n 2^{J_n})^2) \lesssim n^{-1} < \varepsilon_{n,2}$ because $(\underline{\sigma}_n 2^{J_n})^2 \propto (\log n)^{2\psi} \gtrsim (\log n)$ as $\psi > 1/2$. Now, $(\Pi \times G)(\mathscr{F}_n^c) \lesssim \underline{\sigma}_n^{-s} \exp(-[D_2\underline{\sigma}_n^{-1}(\log(1/\underline{\sigma}_n))^t]) \lesssim \exp(-(c_2+4)n\tilde{\varepsilon}_n^2)$ because $\psi < t$.

By the assumption that f_0 has Fourier transform satisfying the integrability condition in (11), we have $f_0 \in L^\infty(\mathbb{R})$ and $\|f_0 - f_0 * \mathrm{sinc}_{2^{-J_n}}\|_2 = O(\varepsilon_{n,2})$. Concerning Condition (3), we first apply Theorem 2, *ibidem*, page 2891, for the supremum norm (Condition (1) for the supremum norm can be seen to be satisfied as for the L^2-norm and $\|f_0 - f_0 * \mathrm{sinc}_{2^{-J_n}}\|_\infty = O(n^{1/2}\tilde{\varepsilon}_n^2)$) and then use the conclusion that the posterior concentrates on a shrinking supremum norm neighborhood of f_0 to see that the posterior accumulates on a fixed supremum norm ball of radius $B := 1 + \|f_0\|_\infty$ with probability tending to 1.

• *Small ball probability estimate.* By routine computations, see, e.g., Theorem 4 in Shen *et al.* (2013), pages 629–630, it can be seen that, for the Dirichlet process, there exists a constant $0 < c_2 < \infty$ so that $(\Pi \times G)(B_{\mathrm{KL}}(f_0; \tilde{\varepsilon}_n^2)) \gtrsim \exp(-c_2 n\tilde{\varepsilon}_n^2)$ for all n large enough. □

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, Applied Mathematics Series, 55. U.S. Government Printing Office, Washington, D.C. Available online at http://www.math.sfu.ca/~cbm/aands/ 494

Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*. New York: Springer. 511

Belitser, E. and Levit, B. (2001). "Asymptotically local minimax estimation of infinitely smooth density with censored data." *Annals of the Institute of Statistical Mathematics*, 53: 289–306. 480

Billingsley, P. (1995). *Probability and Measure*. New York: John Wiley & Sons, 3rd edition. 518

Butucea, C. and Tsybakov, A. B. (2008). "Sharp optimality in density deconvolution with dominating bias. I." *Theory of Probability & Its Applications*, 52: 24–39. 479, 480

Carlton, M. A. (2002). "A family of densities derived from the three-parameter Dirichlet process." *Journal of Applied Probability*, 39: 764–774. 492

Davis, K. B. (1977). "Mean integrated square error properties of density estimates." *The Annals of Statistics*, 5: 530–535. 481

Dedecker, J. and Michel, B. (2013). "Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension." *Journal of Multivariate Analysis*, 122: 278–291. 483

de Jonge, R. and van Zanten, J. H. (2010). "Adaptive nonparametric Bayesian inference using location-scale mixture priors." *The Annals of Statistics*, 38: 3300–3320. 480, 503

Devroye, L. (1992). "A note on the usefulness of superkernels in density estimation." *The Annals of Statistics*, 20: 2037–2056. 477, 511, 517, 518

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). "Density estimation by wavelet thresholding." *The Annals of Statistics*, 24: 508–539. 489

Doss, H. and Sellke, T. (1982). "The tails of probabilities chosen from a Dirichlet prior." *The Annals of Statistics*, 10: 1302–1305. 506, 516

Favaro, S., Lijoi, A. and Prünster, I. (2012). "On the stick-breaking representation of normalized inverse Gaussian priors." *Biometrika*, 99: 663–674. 494

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions." In *Recent Advances in Statistics*, eds. Rizvi, M. H., Rustagi, J. S. and Siegmund, D., New York: Academic Press, pp. 287–302. 475

Ghosal, S. (2001). "Convergence rates for density estimation with Bernstein polynomials." *The Annals of Statistics*, 29: 1264–1280. 492, 493

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27: 143–158. 519

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28: 500–531. 481, 483, 490, 492, 493

Ghosal, S. and van der Vaart, A. W. (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities." *The Annals of Statistics*, 29: 1233–1263. 475, 486, 502, 506, 513, 514, 515, 518

Ghosal, S. and van der Vaart, A. (2007a). "Convergence rates of posterior distributions for noniid observations." *The Annals of Statistics*, 35: 192–223. 477

— (2007b). "Posterior convergence rates of Dirichlet mixtures at smooth densities." *The Annals of Statistics*, 35: 697–723. 475, 502, 506, 514, 520

Giné, E. and Nickl, R. (2011). "Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$." *The Annals of Statistics*, 39: 2883–2911. 481, 506, 512

Golubev, Y. K., Levit, B. Y. and Tsybakov, A. B. (1996). "Asymptotically efficient estimation of analytic functions in Gaussian noise." *Bernoulli*, 2: 167–181. 480

Guerre, E. and Tsybakov, A. B. (1998). "Exact asymptotic minimax constants for the estimation of analytical functions in L_p." *Probability Theory and Related Fields*, 112: 33–51. 480

Hasminskii, R. and Ibragimov, I. (1990). "On density estimation in the view of Kolmogorov's ideas in approximation theory." *The Annals of Statistics*, 18: 999–1010. 479

Hurst, S. (1995). "The characteristic function of the Student t distribution." *Financial Mathematics Research Report* No. FMRR 006-95, *Statistics Research Report* No. SRR044-95. 479

Ibragimov, I. A. and Hasminskii, R. Z. (1983). "Estimation of distribution density." *Journal of Soviet Mathematics*, 21: 40–57. 480

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. 484

Ishwaran, H. and Zarepour, M. (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." *Biometrika*, 87: 371–390. 484

Kawata, T. (1972). *Fourier Analysis in Probability Theory*. *Probability and Mathematical Statistics*, No. 15. New York-London: Academic Press. 479, 495

Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). "Adaptive Bayesian density estimation with location-scale mixtures." *Electronic Journal of Statistics*, 4: 1225–1257. 476, 477, 488, 503

Lijoi, A., Mena, R. H. and Prünster, I. (2005). "Hierarchical mixture modeling with normalized inverse-Gaussian priors." *Journal of the American Statistical Association*, 100: 1278–1291. 485

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12: 351–357. 475

Maugis-Rabusseau, C. and Michel, B. (2013). "Adaptive density estimation for clustering with Gaussian mixtures." *ESAIM: Probability and Statistics*, 17: 698–724. 477

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models." *The Annals of Statistics*, 41: 370–400. 483, 512, 517

Norets, A. and Pelenis, J. (2014). "Posterior consistency in conditional density estimation by covariate dependent mixtures." Forthcoming in *Econometric Theory*. 520

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25: 855–900. 484

Scricciolo, C. (2011). "Posterior rates of convergence for Dirichlet mixtures of exponential power densities." *Electronic Journal of Statistics*, 5: 270–308. 486

Shen, W., Tokdar, S. T. and Ghosal, S. (2013). "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures." *Biometrika*, 100: 623–640. 476, 477, 490, 506, 507

Titchmarsh, E. C. (1937). *Introduction to the Theory of Fourier Integrals*. Oxford: Clarendon Press. 511

van der Vaart, A. W. and van Zanten, J. H. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth." *The Annals of Statistics*, 37: 2655–2675. 480, 512

Villani, C. (2008). *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg. 512

Watson, G. S. and Leadbetter, M. R. (1963). "On the estimation of the probability density, I." *The Annals of Mathematical Statistics*, 34: 480–491. 480

Wong, W. H. and Shen, X. (1995). "Probability inequalities for likelihood ratios and convergence rates of sieve MLEs." *The Annals of Statistics*, 23: 339–362. 515, 516

Acknowledgments

The author would like to thank the Editor, an Associate Editor and an anonymous referee for constructive and valuable remarks that helped improving the original manuscript. She is grateful to Prof. A. W. van der Vaart for his availability, insightful comments and suggestions while visiting the Department of Mathematics at VU University Amsterdam, whose kind hospitality is acknowledged. This research was mainly supported by grants from Bocconi University.

Appendix

The Appendix is split into three parts: the first one presents the proofs of the results in Section 3, preceded by some lemmas; the second one contains the proofs of Theorem 1 and Corollary 2; the third one reports some auxiliary results.

Proofs of the results in Section 3

The following lemma provides, for every $2 \leq p \leq \infty$, an upper bound on the L^p-norm approximation error of a density, whose Fourier transform either vanishes outside a compact set or decays exponentially fast, by its convolution with the sinc kernel. Recall that if $\hat{f} \in L^1(\mathbb{R})$, then f can be recovered from $\hat{f}$ using the inversion formula for Fourier transforms, $f(x) = (2\pi)^{-1} \int e^{-itx} \hat{f}(t)\, \mathrm{d}t$, $x \in \mathbb{R}$. Furthermore, f is continuous and bounded, $f \in BC(\mathbb{R})$. In what follows, S_f is as defined in Section 1.1.

Lemma 12. *Let $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$. Let $0 < \sigma < \infty$ be fixed. If $S_f \leq 1/\sigma$, then $\|f - f * \mathrm{sinc}_\sigma\|_p = 0$ for every $1 \leq p \leq \infty$. If $S_f = \infty$, then $\|f - f * \mathrm{sinc}_\sigma\|_p \lesssim \sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r}$ for every $2 \leq p \leq \infty$.*

Proof. Since $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, we have $\hat{f} \in L^1(\mathbb{R})$. Then, $\hat{f}(1 - \widehat{\mathrm{sinc}_\sigma}) \in L^1(\mathbb{R})$. By the inversion formula for Fourier transforms and the fact that $\widehat{\mathrm{sinc}}(t) = 1_{[-1,1]}(t)$, $t \in \mathbb{R}$, we have $(f - f * \mathrm{sinc}_\sigma)(x) = (2\pi)^{-1} \int_{|t|>1/\sigma} e^{-itx} \hat{f}(t)\, \mathrm{d}t$, $x \in \mathbb{R}$. If $S_f \leq 1/\sigma$, then $\int_{|t|>1/\sigma} e^{-itx} \hat{f}(t)\, \mathrm{d}t = 0$ identically on $\mathbb{R}$ and $\|f - f * \mathrm{sinc}_\sigma\|_p = 0$ for every $1 \leq p \leq \infty$. Now, suppose $S_f = \infty$. For any function $g \in L^p(\mathbb{R})$, $2 \leq p < \infty$, we have $\|g\|_p^p \leq C_p \|\hat{g}\|_q^q$, where $q^{-1} := (1 - p^{-1}) \in [1/2, 1)$ and $0 < C_p < \infty$ is a

constant depending only on p, see, *e.g.*, Theorem 74 in Titchmarsh (1937), page 96. By the assumption that $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, we have $f \in L^p(\mathbb{R})$ for every $2 \leq p \leq \infty$. Then, $\|f - f * \mathrm{sinc}_\sigma\|_p \leq \|f\|_p + \|f\|_1 \|\mathrm{sinc}_\sigma\|_p < \infty$. For every $2 \leq p < \infty$, we have $\|f - f * \mathrm{sinc}_\sigma\|_p^p \leq C_p \|\hat{f}(1 - \widehat{\mathrm{sinc}_\sigma})\|_q^q = C_p \int_{|t|>1/\sigma} |\hat{f}(t)|^q \, \mathrm{d}t$. By the Cauchy-Schwarz inequality and the assumption that $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$,

$$\int_{|t|>1/\sigma} |\hat{f}(t)|^q \, \mathrm{d}t < \int_{|t|>1/\sigma} |\hat{f}(t)| \, \mathrm{d}t \lesssim \sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r}, \tag{21}$$

where $\int_{1/\sigma}^\infty e^{-2(\rho t)^r} \, \mathrm{d}t = r^{-1}(2\rho^r)^{-1/r} \Gamma(r^{-1}, 2(\rho/\sigma)^r)$, with $\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} \, \mathrm{d}t$, for $0 < a, z < \infty$, the upper incomplete gamma function. It is known that $\Gamma(a, z) \sim z^{a-1} e^{-z}$ as $z \to \infty$. The case where $p = \infty$ is treated implicitly in (21). □

When $S_f = \infty$, the result of Lemma 12 can be extended to all L^p-metrics, $1 \leq p \leq \infty$, replacing the sinc kernel with a superkernel, which, unlike the sinc kernel, is an absolutely integrable function. A *superkernel* S is a symmetric (around 0), absolutely integrable function, with $\int S \, \mathrm{d}\lambda = 1$, that has an absolutely integrable Fourier transform $\hat{S}$ (hence S is continuous and bounded) with the properties that $\hat{S} = 1$ identically on $[-1, 1]$ and $|\hat{S}| < 1$ outside $[-1, 1]$. The interval $[-1, 1]$ is chosen for convenience only: $\hat{S}$ is required to be equal to 1 in a neighborhood of 0. Superkernels necessarily have infinite support. They can be obtained as iterated convolutions of re-scaled versions of the sinc kernel, cf. Example 1 in Devroye (1992), page 2039.

Lemma 13. *Let $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$. Let S be a superkernel and let $0 < \sigma < \infty$ be fixed. If $S_f \leq 1/\sigma$, then $\|f - f * S_\sigma\|_p = 0$ for every $1 \leq p \leq \infty$. If $S_f = \infty$, then $\|f - f * S_\sigma\|_p \lesssim \sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r}$ for every $2 \leq p \leq \infty$. If, furthermore, when $S_f = \infty$, we have $0 < \sigma < 1$ and $\int f^\upsilon \, \mathrm{d}\lambda < \infty$ for some $0 < \upsilon < 1$, then $\|f - f * S_\sigma\|_p \lesssim (\sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r})^{1-\upsilon}$ for every $1 \leq p < 2$.*

Proof. We have $(f - f * S_\sigma)(x) = (2\pi)^{-1} \int_{|t|>1/\sigma} e^{-itx} \hat{f}(t)[1 - \hat{S}(\sigma t)] \, \mathrm{d}t$, $x \in \mathbb{R}$. If $S_f \leq 1/\sigma$, then $\|f - f * S_\sigma\|_p = 0$ for every $1 \leq p \leq \infty$. If $S_f = \infty$, since, for every $2 \leq p < \infty$, $\|f - f * S_\sigma\|_p \leq (1 + \|S_\sigma\|_1)\|f\|_p < \infty$ for all $0 < \sigma < \infty$, by repeating the same reasoning as for the sinc kernel in Lemma 12, we conclude that $\|f - f * S_\sigma\|_p^p \leq C_p \|\hat{f}(1 - \widehat{S_\sigma})\|_q^q = C_p \int_{|t|>1/\sigma} (|\hat{f}(t)||1 - \hat{S}(\sigma t)|)^q \, \mathrm{d}t < 2^q C_p \int_{|t|>1/\sigma} |\hat{f}(t)|^q \, \mathrm{d}t \lesssim \sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r}$ because $|\hat{S}| < 1$ on $[-1, 1]^c$. The case where $p = \infty$ follows from the bound on $\int_{|t|>1/\sigma} |\hat{f}(t)| \, \mathrm{d}t$ in (21). Now, let $1 \leq p < 2$. From Lemma 1 in Devroye (1992), page 2040, and the assumption that $f \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, we have $\|f - f * S_\sigma\|_1 < 2(\int f^\upsilon \, \mathrm{d}\lambda)(\int_{|t|>1/\sigma} |\hat{f}(t)| \, \mathrm{d}t/\pi)^{1-\upsilon} \lesssim (\int f^\upsilon \, \mathrm{d}\lambda)(\sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r})^{1-\upsilon}$ for every $0 < \upsilon < 1$ as in the statement. For any L^p-metric, $1 < p < 2$, using the inequality $\|f - f * S_\sigma\|_p \leq \max\{\|f - f * S_\sigma\|_1, \|f - f * S_\sigma\|_2\}$ (see, *e.g.*, Athreya and Lahiri (2006), page 104), we get that $\|f - f * S_\sigma\|_p \lesssim (\sigma^{-(1-r)/2} e^{-(\rho/\sigma)^r})^{1-\upsilon}$. □

Before proving Proposition 1, a remark is in order. If $\hat{K} \in L^1(\mathbb{R})$, then $\|\widehat{f_{F,\sigma}}\|_1 \leq \int |\hat{K}(\sigma t)| \, \mathrm{d}t < \infty$ for every $0 < \sigma < \infty$. So, if $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$, then, not

only is $\hat{K} \in L^1(\mathbb{R})$, but $f_{F,\sigma} \in \mathcal{A}^{\rho\sigma,r,L/\sqrt{\sigma}}(\mathbb{R})$. Therefore, the absolute integrability of $\hat{K}$ allows to recover any convolution mixture $f_{F,\sigma}$ by just inverting its Fourier transform.

Proof of Proposition 1. We appeal to Theorem 2 of Giné and Nickl (2011), page 2891. Choosing their $\gamma_n = 1$, $n \in \mathbb{N}$, the sequence $\varepsilon_{n,p} := \tilde{\varepsilon}_n(n\tilde{\varepsilon}_n^2)^{(1-1/p)/2}$ plays the same role as δ_n in the cited theorem. For $\underline{\sigma}_n := E(n\tilde{\varepsilon}_n^2)^{-1/\gamma}$, $0 < E < \infty$ being a constant as in the statement, let $\mathscr{F}_n := \{f_{F,\sigma} : F \in \mathscr{M}(\Theta),\ \sigma \geq \underline{\sigma}_n\}$. For every $f_{F,\sigma} \in \mathscr{F}_n$, we have $I^{\rho_n,r}(\widehat{f_{F,\sigma}}) \leq 2\pi L_n^2$, with $\rho_n := \rho\underline{\sigma}_n$ and $L_n^2 := L^2/\underline{\sigma}_n$. Condition 1(a), *ibidem*, page 2890, for the convolution kernel case is satisfied for the sinc kernel. In fact, $\mathrm{sinc} \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$ because $\int \mathrm{sinc}^2 \,\mathrm{d}\lambda = \|\mathrm{sinc}\|_\infty = (1/\pi) < \infty$. Besides, the sinc kernel is continuous and, as shown in Lemma 14, is of bounded quadratic variation on $\mathbb{R}$. For every $f_{F,\sigma} \in \mathscr{F}_n$ for which $S_{f_{F,\sigma}} < \infty$, since $\underline{\sigma}_n 2^{J_n} \propto (n\tilde{\varepsilon}_n^2)^{1-1/\gamma} \to \infty$ as $n \to \infty$, we have $\|f_{F,\sigma} - \mathrm{sinc}_{2^{-J_n}}(f_{F,\sigma})\|_p = 0$ for every $2 \leq p \leq \infty$. For every $f_{F,\sigma} \in \mathscr{F}_n$ for which $S_{f_{F,\sigma}} = \infty$, taking into account that $2^{J_n} = cn\tilde{\varepsilon}_n^2$, with $(\alpha^{1/r}\rho E)^{-1} \leq c < \infty$, and using the constraint on γ, we have $\|f_{F,\sigma} - \mathrm{sinc}_{2^{-J_n}}(f_{F,\sigma})\|_p \lesssim \exp\left(-\alpha(\rho\underline{\sigma}_n 2^{J_n})^r\right) \lesssim \exp\left(-\alpha(\rho Ec)^r(n\tilde{\varepsilon}_n^2)^{r(1-1/\gamma)}\right) \lesssim n^{-1} \lesssim \varepsilon_{n,p}$ for every $2 \leq p \leq \infty$. Therefore, for every sufficiently large n, $\mathscr{F}_n \subseteq \{f_{F,\sigma} : \|f_{F,\sigma} - \mathrm{sinc}_{2^{-J_n}}(f_{F,\sigma})\|_p \leq C(\mathrm{sinc})\varepsilon_{n,p}\}$, where $0 < C(\mathrm{sinc}) < \infty$ is an appropriate constant depending only on the operator sinc kernel. If $1 < \gamma < \infty$, for $0 < E \leq \{D_2[1_{[0,\,\gamma-1]}(s) + \beta 1_{(\gamma-1,\,\infty)}(s)]/(C+4)\}^{1/\gamma}$, where $0 < C < \infty$ is the constant stemming from the small ball probability estimate, by Lemma 4.9 of van der Vaart and van Zanten (2009), page 2669, and Assumption (A0),

$$(\Pi \times G)(\mathscr{F}_n^c) = \int_0^{\underline{\sigma}_n} g(\sigma)\,\mathrm{d}\sigma \lesssim \underline{\sigma}_n^{-(s-\gamma+1)}(\log(1/\underline{\sigma}_n))^{-t}\exp\left(-D_2\underline{\sigma}_n^{-\gamma}(\log(1/\underline{\sigma}_n))^t\right) \lesssim \exp\left(-(C+4)n\tilde{\varepsilon}_n^2\right)$$

and Condition (1), *ibidem*, page 2891, is fulfilled. □

Proof of Corollary 1. Under the stated conditions, Proposition 1 holds with the prior distribution G for the scale parameter being a point mass at σ_0 and with t playing the role of p. It is verified that $K \in \mathcal{A}^{\rho',r,L}(\mathbb{R})$ for every $0 < \rho' < \rho$ and a suitable constant $0 < L < \infty$. Consequently, $f_{F_0,\sigma_0} \in L^t(\mathbb{R})$, $2 \leq t \leq \infty$, and, defined $\varepsilon_{n,t} := \tilde{\varepsilon}_n(n\tilde{\varepsilon}_n^2)^{(1-1/t)/2}$, we have $\|f_{F_0,\sigma_0} - \mathrm{sinc}_{2^{-J_n}}(f_{F_0,\sigma_0})\|_t = O(\varepsilon_{n,t})$ for $2^{J_n} = cn\tilde{\varepsilon}_n^2$, with $[\alpha^{1/r}\rho(\sigma_0 \wedge 1)]^{-1} \leq c < \infty$ for any fixed $0 < \alpha < 1$. Thus, for every $2 \leq t \leq \infty$, there exists a sufficiently large constant $0 < M < \infty$ so that $\Pi(F : \|f_{F,\sigma_0} - f_{F_0,\sigma_0}\|_t \geq M\varepsilon_{n,t}|X^{(n)}) \to 0$ in P_0^n-probability. By Lemma 15, for every $2 \leq t < \infty$ and any $0 < u < \infty$ such that $\mathrm{E}_K[|X|^u] < \infty$, we have $\|f_{F,\sigma_0} - f_{F_0,\sigma_0}\|_1 \lesssim (\|f_{F,\sigma_0} - f_{F_0,\sigma_0}\|_t)^{1/[1+(su)^{-1}]}$, where $f_{F,\sigma_0}, f_{F_0,\sigma_0} \in L^\infty(\mathbb{R})$ because $K \in \mathcal{A}^{\rho',r,L}(\mathbb{R})$ and $\sup_{F' \in \mathscr{M}(\Theta)} \mathrm{E}_{f_{F',\sigma_0}}[|X|^u] \leq (1 \vee 2^{u-1})\{\sigma_0^u\,\mathrm{E}_K[|X|^u] + \int|\theta|^u\,\mathrm{d}F'(\theta)\} < \infty$ since Θ is bounded. By modifying Theorem 2 in Nguyen (2013), page 377, for every $2 \leq p, t < \infty$, we get that $W_p(F, F_0) \lesssim [-\log d_{\mathrm{TV}}(f_{F,\sigma_0}, f_{F_0,\sigma_0})]^{-1/r} \lesssim (-\log\|f_{F,\sigma_0} - f_{F_0,\sigma_0}\|_t)^{-1/r}$, where $d_{\mathrm{TV}}(\cdot,\cdot)$ denotes the total variation distance. Thus, for every pair p, p' such that $1 \leq p' \leq 2 \leq p < \infty$, since $W_{p'}(\cdot,\cdot) \leq W_p(\cdot,\cdot)$, see, *e.g.*, Remark 6.6 in Villani (2008), page 95, we have $\{F : \|f_{F,\sigma_0} - f_{F_0,\sigma_0}\|_t \lesssim \varepsilon_{n,t}\} \subseteq \{F : W_p(F, F_0) \lesssim (\log n)^{-1/r}\} \subseteq \{F : W_{p'}(F, F_0) \lesssim (\log n)^{-1/r}\}$. Hence, for every $1 \leq p < \infty$, there exists a sufficiently

large constant $0 < M' < \infty$ so that $\Pi(F : W_p(F, F_0) \geq M'(\log n)^{-1/r}|X^{(n)}) \to 0$ in P_0^n-probability. □

Proofs of the results in Section 4.1

Recall that, given a metric space (S, d) and a totally bounded subset C of S, for fixed $0 < \varepsilon < \infty$, the ε-packing number of C, denoted by $D(\varepsilon, C, d)$, is defined as the largest integer m such that there is a set $\{s_1, \ldots, s_m\} \subseteq C$ with $d(s_k, s_l) > \varepsilon$ for all pairs of integers $1 \leq k \neq l \leq m$. The ε-capacity of (C, d) is defined as $\log D(\varepsilon, C, d)$.

Proof of Theorem 1. We prove the result for the L^1-metric invoking Theorem 2.1 of Ghosal and van der Vaart (2001), page 1239. We deal with L^p-metrics, $2 \leq p \leq \infty$, appealing to Proposition 1. For $1 < p < 2$, the result follows from the interpolation inequality $\|f_{F,\sigma} - f_0\|_p \leq \max\{\|f_{F,\sigma} - f_0\|_1, \|f_{F,\sigma} - f_0\|_2\} \lesssim n^{-1/2}(\log n)^{\iota}$ for a suitable constant $0 < \iota < \infty$.

• *L^1-metric.* We show that conditions (2.8) and (2.9) in Theorem 2.1 of Ghosal and van der Vaart (2001), page 1239, are satisfied for sequences $\bar{\varepsilon}_n = n^{-1/2}(\log n)^{\chi}$, with a suitable constant $0 < \chi < \infty$, and $\tilde{\varepsilon}_n = n^{-1/2}(\log n)^{\varphi(d)}$, with $\varphi(d) = \tau + (\tau - 1/2)1_{(0,\infty)}(d)$ as defined in (9), the latter sequence stemming from the small ball probability estimate below. Then, the posterior rate is $\varepsilon_{n,1} := (\bar{\varepsilon}_n \vee \tilde{\varepsilon}_n)$. Given $0 < \eta_n < 1/5$, for constants $0 < E, F, L < \infty$ to be suitably chosen, let $a_n := L(\log(1/\eta_n))^{2\varphi(d)/\delta}$, $\underline{\sigma}_n := E(\log(1/\eta_n))^{-2\varphi(d)/\gamma}$ and $\bar{\sigma}_n := \exp(F(\log(1/\eta_n))^{2\varphi(d)})$. For $\mathscr{F}_n := \{f_{F,\sigma} : F([-a_n, a_n]) \geq 1 - \eta_n, \ \underline{\sigma}_n \leq \sigma \leq \bar{\sigma}_n\}$, by Lemma A.3 of Ghosal and van der Vaart (2001), page 1261, and Lemma 20,

$$\log D(\eta_n, \mathscr{F}_n, \|\cdot\|_1) \lesssim \log\left(\frac{\bar{\sigma}_n}{\underline{\sigma}_n \eta_n}\right) + \left(\frac{a_n}{\underline{\sigma}_n}\right)^{1_{(0,1]}(r)} \times \left(\log\frac{1}{\eta_n}\right)^{1+1_{(0,1]}(r)/r} \times \max\left\{\left(\frac{a_n}{\underline{\sigma}_n}\right)^{r/(r-1)}, \left(\log\frac{1}{\eta_n}\right)\right\}^{1_{(1,\infty)}(r)}.$$

Taking $\eta_n = \bar{\varepsilon}_n$, we have $\log D(\bar{\varepsilon}_n, \mathscr{F}_n, \|\cdot\|_1) \lesssim n\bar{\varepsilon}_n^2$. Concerning condition (2.9), by assumptions (ii)–(iii) and the fact that $2\tau > 1$, for appropriate choices of E, F, L as functions of the constant c_2 stemming from the small ball probability estimate, the prior probability of $\mathscr{F}_n^c$ is bounded above by $\exp(-D_2[\beta 1_{(0,s+1)}(\gamma) + 1_{[s+1,\infty]}(\gamma)]\underline{\sigma}_n^{-\gamma}) + \bar{\sigma}_n^{-\varrho} + e^{-ba_n^{\delta}}/\eta_n \lesssim \exp(-(c_2 + 4)n\bar{\varepsilon}_n^2)$ because, by Markov's inequality and the independence of $(W_j)_{j\geq 1}$ and $(Z_j)_{j\geq 1}$,

$$\Pi(F : F([-a_n, a_n]^c) > \eta_n) < \frac{1}{\eta_n} \mathrm{E}\left[\sum_{j=1}^{\infty} W_j 1_{[-a_n, a_n]^c}(Z_j)\right] \lesssim \frac{\alpha([-a_n, a_n]^c)}{\eta_n} \lesssim \frac{e^{-ba_n^{\delta}}}{\eta_n}.$$

• *L^p-metrics, $2 \leq p \leq \infty$.* The conditions of Proposition 1 are satisfied. Let $\varepsilon_{n,p} := \tilde{\varepsilon}_n(n\tilde{\varepsilon}_n^2)^{(1-1/p)/2}$. By the assumption that $f_0 = f_{F_0,\sigma_0} = F_0 * K_{\sigma_0}$, we have $f_0 \in$

$\mathcal{A}^{\rho\sigma_0,r,L/\sqrt{\sigma_0}}(\mathbb{R}) \cap L^p(\mathbb{R})$, $2 \le p \le \infty$. By Lemma 12, letting $2^{J_n} = cn\tilde{\varepsilon}_n^2$, with c defined as in Proposition 1, $\|f_0 - f_0 * \mathrm{sinc}_{2^{-J_n}}\|_p = O(\varepsilon_{n,p})$ for n large enough.

• *Small ball probability estimate.* We show that, for $0 < \varepsilon \le [(1/4) \wedge (\sigma_0/2)]$, there exists a constant $0 < c_2 < \infty$ so that $(\Pi \times G)(B_{\mathrm{KL}}(f_0; \varepsilon^2)) \gtrsim \exp(-c_2(\log(1/\varepsilon))^{2\varphi(d)})$. A remark is in order. The case where $\varpi = \infty$ corresponds to F_0 having compact support, *i.e.*, $F_0([-a_0, a_0]) = 1$ for some $0 < a_0 < \infty$. Let $a_\varepsilon := a_0^{1_{\{\infty\}}(\varpi)}(c_0^{-1}\log(1/\varepsilon))^{1/\varpi}$ and let F_0^* be the re-normalized restriction of F_0 to $[-a_\varepsilon, a_\varepsilon]$. By Lemma A.3 of Ghosal and van der Vaart (2001), page 1261, and Assumption (A2), $\|f_{F_0^*,\sigma_0} - f_0\|_1 \lesssim \varepsilon$. We show that there exists a discrete probability measure F_0' on $[-a_\varepsilon, a_\varepsilon]$, with at most

$$N \lesssim \left(\log \frac{1}{\varepsilon}\right)^{2\tau-1} \tag{22}$$

support points, such that $\|f_{F_0^*,\sigma_0} - f_{F_0',\sigma_0}\|_\infty \lesssim \varepsilon$. The support points of F_0' can be taken to be at least 2ε-separated. We distinguish the case where $0 < r \le 1$ from the case where $r > 1$. In the latter case, the assertion follows immediately from Lemma 17: in fact, a_ε can be taken to be large enough so that $a_\varepsilon/(\rho\sigma_0) \ge e^{-1}$. If $0 < r \le 1$, Lemma 17 cannot be directly applied because the requirement on $a_\varepsilon/(\rho\sigma_0)$ may not be met. Yet, an argument similar to the one used in Lemma 2 of Ghosal and van der Vaart (2007b), page 705, can be adopted. Consider a partition of $[-a_\varepsilon, a_\varepsilon]$ into $k = \lceil a_0^{1_{\{\infty\}}(\varpi)}(c_0^{[1-1_{\{\infty\}}(\varpi)]/\varpi}\sigma_0)^{-1}(\log(1/\varepsilon))^{1/r-1+1_{(0,\infty)}(\varpi)/\varpi}\rceil$ subintervals $I_1, \ldots, I_k$ of equal length $0 < l \le 2\sigma_0(\log(1/\varepsilon))^{-(1-r)/r}$ and, possibly, a final interval I_{k+1} of length $0 \le l_{k+1} < l$. Let J be the total number of intervals in the partition, which can be either k or $k+1$. Write $F_0^* = \sum_{j=1}^J F_0^*(I_j)F_{0,j}^*$, where $F_{0,j}^*$ denotes the re-normalized restriction of F_0^* to I_j. Then, $f_{F_0^*,\sigma_0}(\cdot) = \sum_{j=1}^J F_0^*(I_j) f_{F_{0,j}^*,\sigma_0}(\cdot) = \sum_{j=1}^J F_0^*(I_j)(F_{0,j}^* * K_{\sigma_0})(\cdot)$. For every $j = 1, \ldots, J$, by Lemma 17 (and Remark 2) applied to every $f_{F_{0,j}^*,\sigma_0}$, with $a/\sigma = (l/2)/\sigma_0 \propto (\log(1/\varepsilon))^{-(1-r)/r}$, there exists a discrete distribution $F_{0,j}'$, with at most $N_j \lesssim \log(1/\varepsilon)$ support points, such that $\|f_{F_{0,j}^*,\sigma_0} - f_{F_{0,j}',\sigma_0}\|_\infty \lesssim \varepsilon$. Defined $F_0' := \sum_{j=1}^J F_0^*(I_j)F_{0,j}'$, we have $\|f_{F_0^*,\sigma_0} - f_{F_0',\sigma_0}\|_\infty \le \sum_{j=1}^J F_0^*(I_j)\|f_{F_{0,j}^*,\sigma_0} - f_{F_{0,j}',\sigma_0}\|_\infty \lesssim \varepsilon$, where F_0' has at most $N \lesssim \sum_{j=1}^J N_j \lesssim k \times \log(1/\varepsilon) \lesssim (\log(1/\varepsilon))^{1/r+1_{(0,\infty)}(\varpi)/\varpi}$ support points. Combining the result on the total number N of support points of F_0' in the case where $0 < r \le 1$ with the one for the case where $r > 1$, we obtain the bound in (22). Let $0 < q < \infty$ be such that $\mathrm{E}_K[|X|^q] < \infty$. For any v such that $(1+q)^{-1} < v < 1$, by Hölder's inequality, $\int f_{F_0^*,\sigma_0}^v \,\mathrm{d}\lambda \lesssim (1 + \int |x|^q f_{F_0^*,\sigma_0}(x)\,\mathrm{d}x)^v \lesssim \{(1 \vee 2^{q-1})[\sigma_0^q \,\mathrm{E}_K[|X|^q] + \int_{|\theta|\le a_\varepsilon} |\theta|^q \,\mathrm{d}F_0^*(\theta)]\}^v \lesssim a_\varepsilon^{vq}$, this implying that $\|f_{F_0^*,\sigma_0} - f_{F_0',\sigma_0}\|_1 \lesssim \varepsilon^{1-v} a_\varepsilon^{vq}$ by virtue of Lemma 16.

Next, we distinguish the case where the prior for F is a Dirichlet process, *i.e.*, a Pitman-Yor process with $d = 0$ and $c = \alpha(\mathbb{R})$, from the case where the prior for F is a general Pitman-Yor process with $0 \le d < 1$ and $-d < c < \infty$. The proof for the Dirichlet process is paradigmatic to deal with other process priors, like the N-IG process, whose finite-dimensional distributions are known.

– *Dirichlet process.* Represented F_0' as $\sum_{j=1}^N p_j \delta_{\theta_j}$, with $|\theta_j - \theta_k| \ge 2\varepsilon$ for all $1 \le j \ne$

$k \leq N$, and set $U_j := [\theta_j - \varepsilon, \theta_j + \varepsilon]$, $j = 1, \ldots, N$, for every $F \in \mathscr{M}(\mathbb{R})$ such that

$$\sum_{j=1}^{N} |F(U_j) - p_j| \leq \varepsilon \tag{23}$$

and every $0 < \sigma < \infty$ such that $|\sigma - \sigma_0| \leq \varepsilon$, we have $\|f_{F,\sigma} - f_{F_0',\sigma_0}\|_1 \lesssim \|K_\sigma - K_{\sigma_0}\|_1 + \varepsilon/(\sigma \wedge \sigma_0) + \sum_{j=1}^{N} |F(U_j) - p_j| \lesssim \varepsilon$ by virtue of Lemma 18, Lemma 19 and condition (23). Thus, $\|f_{F,\sigma} - f_{F_0',\sigma_0}\|_1 \lesssim \varepsilon$ and the squared Hellinger distance $\|f_{F,\sigma}^{1/2} - f_0^{1/2}\|_2^2 = \int (f_{F,\sigma}^{1/2} - f_0^{1/2})^2 \, \mathrm{d}\lambda \leq \|f_{F,\sigma} - f_{F_0',\sigma_0}\|_1 + \|f_{F_0',\sigma_0} - f_{F_0^*,\sigma_0}\|_1 + \|f_{F_0^*,\sigma_0} - f_0\|_1 \lesssim \varepsilon^{1-\upsilon} a_\varepsilon^{\upsilon q}$. In order to appeal to Theorem 5 of Wong and Shen (1995), pages 357–358, we show that, for densities in the set $S_\varepsilon := \{f_{F,\sigma} : \sum_{j=1}^{N} |F(U_j) - p_j| \leq \varepsilon, \ |\sigma - \sigma_0| \leq \varepsilon\}$ and a suitable constant $0 < \varrho \leq 1$, $M_\varrho^2 := \int_{\{(f_0/f_{F,\sigma}) \geq e^{1/\varrho}\}} f_0 (f_0/f_{F,\sigma})^\varrho \, \mathrm{d}\lambda = O((1/\varepsilon)^\xi)$ for some $0 \leq \xi \leq \kappa/\varpi$. For every F satisfying (23), $F([-a_\varepsilon, a_\varepsilon]) > 1/2$, thus, by symmetry and monotonicity of K, $f_{F,\sigma}(x) \geq \int_{|\theta| \leq a_\varepsilon} K_\sigma(x - \theta) \, \mathrm{d}F(\theta) > K_\sigma(|x| + a_\varepsilon)/2$, $x \in \mathbb{R}$. By Assumption (A1), $K(a_\varepsilon) \gtrsim \exp(-c a_\varepsilon^\kappa)$ for a_ε large enough. Hence, $\int_{|x| \leq a_\varepsilon} f_0^{1+\varrho}(x) K_\sigma^{-\varrho}(|x| + a_\varepsilon) \, \mathrm{d}x \lesssim \exp(\varrho c (4a_\varepsilon/\sigma_0)^\kappa)$ because $|\sigma - \sigma_0| \leq \varepsilon \leq \sigma_0/2$ and $\|f_0\|_\infty < \infty$. Also,

$$\int_{|x| > a_\varepsilon} \frac{f_0^{1+\varrho}(x)}{K_\sigma^\varrho(|x| + a_\varepsilon)} \, \mathrm{d}x \lesssim \int_{|x| > a_\varepsilon} K_{\sigma_0}^{-\varrho}(4|x|) [K_{\sigma_0}(|x|/2) + F_0(\theta : |\theta| > |x|/2)] \, \mathrm{d}x < \infty,$$

where the last integral is finite for a suitable choice of ϱ by virtue of Assumption (A2). Thus, $S_\varepsilon \subseteq B_{\mathrm{KL}}(f_0; c_1 \varepsilon^{1-\upsilon} a_\varepsilon^{\upsilon q} (\log(1/\varepsilon))^2)$. To apply Lemma A.2 of Ghosal and van der Vaart (2001), pages 1260–1261, note that, for each $|\theta_j| \leq a_\varepsilon$, by Assumption (A3), $\alpha(U_j) \gtrsim \varepsilon e^{-b a_\varepsilon^\delta} \gtrsim \varepsilon^{b'}$ for some constant $b' > 0$ because, when $\varpi < \infty$, we have $0 < \delta \leq \varpi$ by hypothesis. Thus, $\tilde{\varepsilon}_n = n^{-1/2} (\log n)^\tau$.

– *Pitman-Yor process with* $0 \leq d < 1$ *and* $-d < c < \infty$. We need to modify the arguments to control $\|f_{F,\sigma} - f_{F_0',\sigma_0}\|_1$. To the aim, the stick-breaking construction for the random weights of F is exploited. Let $F_0' = \sum_{j=1}^{N} p_j \delta_{\theta_j}$ be the finite distribution that approximates F_0^* in the supremum norm. By relabelling, we can assume that $p_1 \geq p_2 \geq \ldots \geq p_N \geq 0$. Let $1 \leq M \leq N$ be the number of strictly positive mixing weights. For every $0 < \sigma < \infty$, by Lemma 18 and the inequality $\sum_{j=M+1}^{\infty} W_j \leq \sum_{j=1}^{M} |W_j - p_j|$,

$$\|f_{F,\sigma} - f_{F_0',\sigma}\|_1 \leq 2 \sum_{j=1}^{M} |W_j - p_j| + \frac{2\|K\|_\infty}{\sigma} \sum_{j=1}^{M} p_j |Z_j - \theta_j|. \tag{24}$$

Let $v_1 := p_1$ and $v_j := p_j [\prod_{h=1}^{j-1} (1 - v_h)]^{-1}$ for $j = 2, \ldots, M$. Note that $0 < v_j < 1$ for every $j = 1, \ldots, M-1$ and $v_M = 1$ because $p_M = 1 - \sum_{j=1}^{M-1} p_j = \prod_{h=1}^{M-1} (1 - v_h)$. We have $|W_j - p_j| \leq |V_j - v_j| \prod_{h=1}^{j-1} (1 - V_h) + v_j |\prod_{h=1}^{j-1} (1 - V_h) - \prod_{h=1}^{j-1} (1 - v_h)| \leq \sum_{h=1}^{j} |V_h - v_h|$, where the inequality $|\prod_{h=1}^{j-1} y_h - \prod_{h=1}^{j-1} z_h| \leq \sum_{h=1}^{j-1} |y_h - z_h|$, valid for complex numbers $y_1, \ldots, y_{j-1}$ and $z_1, \ldots, z_{j-1}$ of modulus at most 1, has been used. If, for $0 < \varepsilon \leq \sigma_0/2$,

a) $\sum_{j=1}^{M}\sum_{h=1}^{j}|V_h - v_h| \leq \varepsilon$, b) $\sum_{j=1}^{M}|Z_j - \theta_j| \leq \varepsilon$, c) $|\sigma - \sigma_0| \leq \varepsilon$,

then, by Lemma 19 and inequality (24), we have $\|f_{F,\sigma} - f_{F_0',\sigma_0}\|_1 \lesssim \|K_\sigma - K_{\sigma_0}\|_1 + \sum_{j=1}^{M}\sum_{h=1}^{j}|V_h - v_h| + \sum_{j=1}^{M} p_j|Z_j - \theta_j| \lesssim \varepsilon$. Next, we show that, for $B_\varepsilon = a_\varepsilon$ (or $B_\varepsilon = a_\varepsilon + 1$, the latter case being considered if any support point θ_j of F_0' is equal to $-a_\varepsilon$ and/or a_ε), the events in a) and b) together imply that, for $0 < \varepsilon \leq [(1/4) \wedge (\sigma_0/2)]$, we have $F([-B_\varepsilon, B_\varepsilon]) > 1/2$. This inequality is used when checking that, for a suitable constant $0 < \varrho \leq 1$, $M_\varrho^2 = O((1/\varepsilon)^\xi)$, with $0 \leq \xi \leq \kappa/\varpi$, so that Theorem 5 of Wong and Shen (1995), pages 357–358, can be invoked. By the event in b), for $\varepsilon > 0$ small enough, all the Z_j are in $[-B_\varepsilon, B_\varepsilon]$. Using this fact and the inequality $\sum_{j=1}^{M}|W_j - p_j| \leq \sum_{j=1}^{M}\sum_{h=1}^{j}|V_h - v_h|$, the event in a) implies that $F([-B_\varepsilon, B_\varepsilon]^c) \leq \sum_{j=1}^{M}\sum_{h=1}^{j}|V_h - v_h| \leq \varepsilon < 1/2$.

Concerning the probability in c), by Assumption (A0), $\int_{\sigma_0-\varepsilon}^{\sigma_0+\varepsilon} g(\sigma)\,\mathrm{d}\sigma \gtrsim \varepsilon$, therefore the prior concentration rate is driven by the probabilities of the events in a) and b). By the independence of $(W_j)_{j\geq 1}$ and $(Z_j)_{j\geq 1}$, Lemma 1 and Lemma 2, when $0 < d < 1$, for $(3\varepsilon/M^2) \leq \min_{1\leq j\leq M-1} v_j \leq \max_{1\leq j\leq M-1} v_j \leq 1 - (4\varepsilon/M^2)$ (if the v_j do not satisfy the condition, f_{F_0',σ_0} can be projected into a new density f_{F_0'',σ_0} which is, at most, within ε L^1-distance from f_{F_0',σ_0} so that the new v_j' satisfy the constraints),

$$\mathrm{P}\left(\sum_{j=1}^{M}|W_j - p_j| \leq \varepsilon\right) \times \mathrm{P}\left(\sum_{j=1}^{M}|Z_j - \theta_j| \leq \varepsilon\right) \times \mathrm{P}(|\sigma - \sigma_0| \leq \varepsilon) \gtrsim \exp\left(-c_2 M^2 \log(1/\varepsilon)\right),$$

because, by (22), $1 \leq M \leq N \lesssim (\log(1/\varepsilon))^{2\tau-1}$, where $\tau \geq 1$, and, for $\varpi < \infty$, we have $0 < \delta \leq \varpi$ by assumption, so that $a_\varepsilon^\delta \lesssim \log(1/\varepsilon)$. Thus, $\tilde{\varepsilon}_n = n^{-1/2}(\log n)^{2\tau-1/2}$. For $d = 0$, the same lower bound as for the Dirichlet process is obtained. □

Proof of Corollary 2. We appeal to Corollary 1. Since Θ may be unbounded, we check that, given $1 \leq p < \infty$, for every $p < u < \infty$, $\int|\theta|^u\,\mathrm{d}F(\theta) < \infty$ with DP(α)-probability 1. For $\theta \in \mathbb{R}$, let $\bar{\alpha}(\theta) := \int_{-\infty}^{\theta}[\alpha'(t)/\alpha(\mathbb{R})]\,\mathrm{d}t$. Using results in Doss and Sellke (1982), page 1304, for any fixed $0 < T < \infty$ large enough, by Assumption (A3),

$$\int|\theta|^u\,\mathrm{d}F(\theta) \leq 2uT^{u-1} + \int_T^\infty u\theta^{u-1}\exp\left(-\frac{1}{\bar{\alpha}(-\theta)[\log\bar{\alpha}(-\theta)]^2}\right)\mathrm{d}\theta + \int_T^\infty u\theta^{u-1}\exp\left(-\frac{1}{[1-\bar{\alpha}(\theta)][\log(1-\bar{\alpha}(\theta))]^2}\right)\mathrm{d}\theta < \infty$$

for almost every sample distribution function F when sampling from DP(α). □

Auxiliary results

This section reports some auxiliary results used throughout the article. Proofs that are an adaptation of those of results known in the literature are omitted.

In the following lemma, the sinc kernel is shown to be of bounded quadratic variation on $\mathbb{R}$. By definition, for $1 \leq p < \infty$, a real-valued function h is of bounded p-variation on $\mathbb{R}$ if its p-variation $V_p(h, \mathbb{R}) := \sup\{\sum_{k=1}^n |h(x_k) - h(x_{k-1})|^p\colon\ -\infty < x_0 < \ldots < x_n < \infty,\ n \in \mathbb{N}\}$ is finite.

Lemma 14. *The function $x \mapsto \operatorname{sinc}(x)$ is of bounded quadratic variation on $\mathbb{R}$.*

Proof. For every $n \in \mathbb{N}$, the sum $\sum_{k=1}^n [\operatorname{sinc}(x_k) - \operatorname{sinc}(x_{k-1})]^2$ is maximum at $x_k := (2k+1)\pi/2$, $k = 1, \ldots, n$. Split the sum into two terms,

$$\sum_{2 \leq k = 2j \leq n} [\operatorname{sinc}(x_k) - \operatorname{sinc}(x_{k-1})]^2 = \frac{4}{\pi^4} \sum_{2 \leq 2j \leq n} \left[\frac{4(2j)}{(4j+1)(4j-1)}\right]^2$$

and

$$\sum_{1 \leq k = 2j+1 \leq n} [\operatorname{sinc}(x_k) - \operatorname{sinc}(x_{k-1})]^2 = \frac{4}{\pi^4} \sum_{1 \leq 2j+1 \leq n} \left[\frac{4(2j+1)}{(4j+3)(4j+1)}\right]^2.$$

Then, $V_2(\operatorname{sinc}, \mathbb{R}) < \infty$ as a consequence of the fact that $\sum_{j=1}^\infty j^{-2} < \infty$. □

The next lemma provides an upper bound on the L^p-distance, $1 \leq p < \infty$, between densities on $\mathbb{R}$ with finite absolute moment of (some) order $0 < u < \infty$, in terms of the product of their L^∞-distance and any L^q-distance, $1 < q < \infty$. The proof is similar to that of statement (b) in Lemma 6 of Nguyen (2013), pages 389 and 397–398.

Lemma 15. *Let f, $g \in L^\infty(\mathbb{R})$ be probability densities with $\max\{\mathrm{E}_f[|X|^u], \mathrm{E}_g[|X|^u]\} < \infty$ for some $0 < u < \infty$. For every $1 \leq p < \infty$ and $1 < t < \infty$,*

$$\|f - g\|_p^p < (s^{-1} + u) \times [s^{1/s}(2^{1/s}/u)^u \|f - g\|_{pt}^{pu} \|f - g\|_\infty^{(p-1)/s} (\mathrm{E}_f[|X|^u] + \mathrm{E}_g[|X|^u])^{1/s}]^{s/(1+su)},$$

where $s^{-1} := 1 - t^{-1}$.

Proof. For any $0 < R < \infty$, by Hölder's inequality, $\int_{|x| \leq R} |f(x) - g(x)|^p \, \mathrm{d}x \leq (2R)^{1/s} \|f - g\|_{pt}^p$. Also, $\int_{|x|>R} |f(x) - g(x)|^p \, \mathrm{d}x < R^{-u} \|f - g\|_\infty^{p-1} (\mathrm{E}_f[|X|^u] + \mathrm{E}_g[|X|^u])$. Therefore, $\|f - g\|_p^p < \min_{R>0}\{(2R)^{1/s} \|f - g\|_{pt}^p + R^{-u} \|f - g\|_\infty^{p-1} (\mathrm{E}_f[|X|^u] + \mathrm{E}_g[|X|^u])\}$. The inequality in the assertion follows from

$$\min_{x>0}(Ax^\alpha + Bx^{-\beta}) = (\alpha + \beta)[(A/\beta)^\beta (B/\alpha)^\alpha]^{1/(\alpha+\beta)}$$

for every A, α, B, $\beta > 0$. □

The next lemma provides an upper bound on the L^1-distance between densities on $\mathbb{R}$ in terms of their L^∞-distance. It is implicit in the proof of Lemma 1 in Devroye (1992), page 2040.

Lemma 16. *Let f, $g \in L^\infty(\mathbb{R})$ be probability densities. For any $0 < \upsilon \leq 1$ such that $\int f^\upsilon \, \mathrm{d}\lambda < \infty$, we have $\|f - g\|_1 \leq 2\|f - g\|_\infty^{1-\upsilon} \int f^\upsilon \, \mathrm{d}\lambda$.*

Proof. Write $\|f-g\|_1 = 2\int (f-g)^+ \,\mathrm{d}\lambda \le 2\int \min\{f, \|f-g\|_\infty\}\,\mathrm{d}\lambda \le 2\|f-g\|_\infty^{1-\upsilon}\int f^\upsilon \,\mathrm{d}\lambda$. The assertion follows. □

As noted in Remark 3 by Devroye (1992), page 2042, if

$$\text{for some } 0 < q < \infty, \quad \mathrm{E}_f[|X|^q] < \infty, \tag{25}$$

then $\int f^\upsilon \,\mathrm{d}\lambda < \infty$ for every υ such that $(1+q)^{-1} < \upsilon < 1$. For example, condition (25) is satisfied for a Student's-t distribution with $0 < \nu < \infty$ degrees of freedom when $0 < q < \nu$. For the special case of a Cauchy distribution, we have $0 < q < 1$.

The following lemma provides an upper bound on the number of mixing components of a convolution mixture, with kernel density K belonging to some class $\mathcal{A}^{\rho,r,L}(\mathbb{R})$, which uniformly approximates a given mixture with the same kernel and a compactly supported mixing distribution. The definition of S_K is in accordance with that in Section 1.1.

Lemma 17. *Let $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$. Let $0 < \varepsilon < 1$ and $0 < a, \sigma < \infty$ be given. For any probability measure F on $[-a, a]$, there exists a discrete probability measure F' on $[-a, a]$, with at most N support points, such that*

$$\|F * K_\sigma - F' * K_\sigma\|_\infty \lesssim \varepsilon/\sigma,$$

where

$$N \lesssim \max\left\{\log(1/\varepsilon),\, (a/\sigma)\right\}, \quad \text{if} \quad S_K < \infty,$$

and

$$N \lesssim \begin{cases} \log(1/\varepsilon), & \text{for } 0 < r < 1 \text{ and } \rho\sigma/a = O((\log(1/\varepsilon))^{(1-r)/r}), \\ \log(1/\varepsilon), & \text{for } r = 1 \text{ and } a/(\rho\sigma) \le e^{-1}, \\ \max\left\{\log(1/\varepsilon),\, (a/\sigma)^{r/(r-1)}\right\}, & \text{for } r > 1 \text{ and } a/(\rho\sigma) \ge e^{-1}, \end{cases}$$

if $S_K = \infty$.

Proof. By Lemma A.1 of Ghosal and van der Vaart (2001), page 1260, there exists a discrete probability measure F' on $[-a, a]$, with at most $N+1$ support points, N being suitably chosen later on, such that it matches the moments of F up to the order N,

$$\int_{-a}^{a} \theta^j \,\mathrm{d}F'(\theta) = \int_{-a}^{a} \theta^j \,\mathrm{d}F(\theta), \quad j = 1, \ldots, N. \tag{26}$$

By the moment matching condition in (26),

$$|\hat{F}(t) - \widehat{F'}(t)| \le \int_{-a}^{a} \frac{|t\theta|^N}{N!} \min\left\{\frac{|t\theta|}{N+1}, 2\right\} \mathrm{d}(F+F')(\theta), \quad t \in \mathbb{R}, \tag{27}$$

where the inequality holds because F and F' have finite absolute moments of any order, see, *e.g.*, inequality (26.5) in Billingsley (1995), page 343. By the assumption that $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, we have $\int |\hat{K}(\sigma t)|\,\mathrm{d}t < \infty$ for every $0 < \sigma < \infty$, hence $F * K_\sigma$ and

$F' * K_\sigma$ can be recovered using the inversion formula for Fourier transforms. By (27), $\|F*K_\sigma - F'*K_\sigma\|_\infty \leq 2a^N/(\pi N!)\int |t|^N |\hat{K}(\sigma t)|\,\mathrm{d}t$. Next, we distinguish the case where $S_K < \infty$ from that where $S_K = \infty$. If $S_K < \infty$, by the assumption that $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$,

$$\|F*K_\sigma - F'*K_\sigma\|_\infty \leq \frac{2a^N}{\pi N!}\int_{|t|\leq S_K/\sigma} |t|^N |\hat{K}(\sigma t)|\,\mathrm{d}t \leq \frac{4}{\sigma}[L^2 + C(\rho, r)/\pi]\left(\frac{aeS_K}{\sigma N}\right)^N \lesssim \frac{\varepsilon}{\sigma}$$

for $N = \max\{\log(1/\varepsilon), (ae^2 S_K/\sigma)\}$. If $S_K = \infty$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\|F * K_\sigma - F' * K_\sigma\|_\infty &\leq \frac{2a^N}{\pi N!}\left(\frac{2\pi L^2}{\sigma}\right)^{1/2}\left(\int |t|^{2N} e^{-2(\rho\sigma|t|)^r}\,\mathrm{d}t\right)^{1/2} \\ &\lesssim \frac{1}{\sigma}\left(\frac{a}{2^{1/r}\rho\sigma}\right)^N \frac{[\Gamma((2N+1)/r)]^{1/2}}{\Gamma(N+1)}.\end{aligned}$$

For N large enough, using $\Gamma(az+b) \sim (2\pi)^{1/2} e^{-az}(az)^{az+b-1/2}$ $(z \to \infty)$ with $a > 0$,

$$\|F * K_\sigma - F' * K_\sigma\|_\infty \lesssim \frac{1}{\sigma}\left(\frac{a}{\rho\sigma}\right)^N e^{N(1-1/r)} r^{-N/r} N^{-N(1-1/r)+(1/r-3/2)/2}.$$

If $0 < r < 1$ and $(\rho\sigma/a)^{r/(1-r)} = O(\log(1/\varepsilon))$, for N such that $\log(1/\varepsilon) \lesssim N \lesssim (\sigma/a)^{r/(1-r)}$,

$$\begin{aligned}\|F * K_\sigma - F' * K_\sigma\|_\infty &\lesssim \frac{1}{\sigma} N^{(1/r-3/2)/2} \exp\left(-N\left[\log\frac{\rho\sigma/a}{N^{1/r-1}} - \left(1 - \frac{1}{r} + \frac{1}{r}\log\frac{1}{r}\right)\right]\right) \\ &\lesssim \frac{\varepsilon}{\sigma}.\end{aligned}$$

If $r = 1$ and $a/(\rho\sigma) \leq e^{-1}$, for $N = \log(1/\varepsilon)$,

$$\|F * K_\sigma - F' * K_\sigma\|_\infty \lesssim \frac{1}{\sigma}\left(\frac{a}{\rho\sigma}\right)^N \lesssim \frac{\varepsilon}{\sigma}.$$

If $r > 1$ and $a/(\rho\sigma) \geq e^{-1}$, for $N = O(\max\{\log(1/\varepsilon), (a/\sigma)^{r/(r-1)}\})$,

$$\|F * K_\sigma - F' * K_\sigma\|_\infty \lesssim \frac{1}{\sigma}\exp\left(-N\left[\log\frac{N^{1-1/r}}{a/(\rho\sigma)} - \frac{1}{r}(r - 1 - \log r)\right]\right) \lesssim \frac{\varepsilon}{\sigma}$$

and the proof is complete. □

Remark 2. Although Lemma 17 is stated for a probability measure F supported on a symmetric interval around 0, it holds for every F with support(F) being any bounded set in $\mathbb{R}$.

The inequality in the next lemma can be proved similarly to the one for the Gaussian kernel, see, *e.g.*, the first part of Lemma 1 in Ghosal *et al.* (1999), pages 156–157.

Lemma 18. *Let K be a probability density on $\mathbb{R}$, bounded and symmetric around 0. For every $0 < \sigma < \infty$ and every pair $\theta_j, \theta_k \in \mathbb{R}$,*

$$\|K_\sigma(\cdot - \theta_j) - K_\sigma(\cdot - \theta_k)\|_1 \leq 2\|K\|_\infty \frac{|\theta_j - \theta_k|}{\sigma} \lesssim \frac{|\theta_j - \theta_k|}{\sigma}.$$

In the following lemma, a sufficient condition is provided for the L^1-distance between convolution mixtures with different scales to be bounded above by the distance between the scales.

Lemma 19. *Let K be a probability density on $\mathbb{R}$, bounded, symmetric around 0 and monotone decreasing in $|x|$. For every probability measure F on $\mathbb{R}$ and every pair $0 < \sigma, \sigma' < \infty$, we have $\|F * K_\sigma - F * K_{\sigma'}\|_1 \leq \|K_\sigma - K_{\sigma'}\|_1 \leq 2|\sigma - \sigma'|/(\sigma \wedge \sigma')$.*

Proof. Note that $\|F * K_\sigma - F * K_{\sigma'}\|_1 \leq \int \|K_\sigma(\cdot - \theta) - K_{\sigma'}(\cdot - \theta)\|_1 \, \mathrm{d}F(\theta) = \|K_\sigma - K_{\sigma'}\|_1$. The second inequality in the statement can be proved as in Norets and Pelenis (2014). □

The next lemma provides an upper bound on the L^1-metric entropy of a class of convolution mixtures with a supersmooth kernel. For $0 < \varepsilon < \infty$, the metric entropy of a set B in a metric space with metric d is defined as $\log N(\varepsilon, B, d)$, where $N(\varepsilon, B, d)$ is the minimum number of balls of radius ε needed to cover B. The result is based on Lemma 17, Lemma 18, Lemma 19 and can be proved similarly to Lemma 3 of Ghosal and van der Vaart (2007b), pages 705–707, which deals with mixtures of normal densities.

Lemma 20. *Let $K \in \mathcal{A}^{\rho,r,L}(\mathbb{R})$, $0 < \rho, r, L < \infty$, be symmetric around 0 and monotone decreasing in $|x|$. Let $0 < \varepsilon < 1/5$, $0 < a < \infty$ and $0 < \underline{\sigma} \leq \bar{\sigma} < \infty$ be such that $(a/\underline{\sigma}) \lesssim (\log(1/\varepsilon))^\nu$ for some constant $0 < \nu < \infty$. Define $\mathscr{F}_{a,\underline{\sigma},\bar{\sigma}} := \{F * K_\sigma : F([-a, a]) = 1, \ \underline{\sigma} \leq \sigma \leq \bar{\sigma}\}$. Then,*

$$\log N(\varepsilon, \mathscr{F}_{a,\underline{\sigma},\bar{\sigma}}, \|\cdot\|_1) \lesssim \log\left(\frac{\bar{\sigma}}{\underline{\sigma}\varepsilon}\right) + N \times \left[\log\left(\frac{2a}{\underline{\sigma}\varepsilon} + 1\right) + \log\frac{1}{\varepsilon}\right],$$

where

$$N \lesssim \begin{cases} (a/\underline{\sigma}) \times (\log(1/\varepsilon))^{1/r}, & \text{for } \ 0 < r \leq 1, \\ \max\{\log(1/\varepsilon), (a/\underline{\sigma})^{r/(r-1)}\}, & \text{for } \ r > 1. \end{cases}$$

The following lemma is a variant of Lemma 6 in Ghosal and van der Vaart (2007b), page 711.

Lemma 21. *Let K be a probability density on $\mathbb{R}$ symmetric around 0. Let f be a bounded probability density on $\mathbb{R}$, non-decreasing on $(-\infty, a)$, non-increasing on (b, ∞), for $-\infty < a \leq b < \infty$, with $f \geq \ell > 0$ on $[a, b]$ and $0 < f \leq \ell < \infty$ on $[a, b]^c$. For any $0 < \zeta < 1$, let $0 < \tau_\zeta < \infty$ be such that $\int_0^{b-a} K_{\tau_\zeta}(x) \, \mathrm{d}x \geq \zeta$. Then, $f * K_\sigma \geq C_\zeta f$ for every $0 < \sigma \leq \tau_\zeta$, with $C_\zeta := (\zeta\ell/\|f\|_\infty) \in (0, 1)$.*