The Annals of Statistics
2014, Vol. 42, No. 5, 2058–2091
DOI: 10.1214/14-AOS1253

ON BAYESIAN SUPREMUM NORM CONTRACTION RATES

BY ISMAËL CASTILLO[1]

CNRS—LPMA Paris

Building on ideas from Castillo and Nickl [*Ann. Statist.* **41** (2013) 1999–2028], a method is provided to study nonparametric Bayesian posterior convergence rates when "strong" measures of distances, such as the sup-norm, are considered. In particular, we show that likelihood methods can achieve optimal minimax sup-norm rates in density estimation on the unit interval. The introduced methodology is used to prove that commonly used families of prior distributions on densities, namely log-density priors and dyadic random density histograms, can indeed achieve optimal sup-norm rates of convergence. New results are also derived in the Gaussian white noise model as a further illustration of the presented techniques.

1. Introduction. In the fundamental contributions by Ghosal, Ghosh and van der Vaart [13], Shen and Wasserman [32] and Ghosal and van der Vaart [15], a general theory is developed to study the behaviour of Bayesian posterior distributions. A main tool is provided by the existence of exponentially powerful tests between a point and the complement of a ball for some distance. The use of some important distances, such as the Hellinger distance between probability measures, indeed guarantees the existence of such tests. The theory often also allows extensions to other metrics, for instance, L^2-type distances, but the question of dealing with arbitrary metrics has been left essentially open so far. Although a general theory might be harder to obtain, it is natural to consider such a problem in simple, canonical, statistical settings first, such as Gaussian white noise or density estimation. This is the starting point of the authors in Giné and Nickl [16], and this paper was the first to provide tools to get rates in strong norms, such as the L^∞-norm. Exponential inequalities for frequentist estimators are used in [16] as a way to build appropriate tests, and this enables one to obtain some rates in sup-norm in density estimation. In the case where the true density is itself supersmooth and a kernel mixture is used as a prior, the nearly parametric minimax rate is attained, at least up to a possible logarithmic term; see also the work by Scricciolo [31] for related results. In the general case where the true density belongs to a Hölder class, a sup-norm rate is obtained which differs from the minimax rate by a power of n. On the other hand, by using explicit computations, the authors in [16] show that in the Gaussian white noise model with conjugate Gaussian priors, minimax sup-norm rates are

Received April 2013; revised March 2014.
[1]Supported in part by ANR Grant "Banhdits" ANR-2010-BLAN-0113-03.
MSC2010 subject classifications. Primary 62G20; secondary 62G05, 62G07.
Key words and phrases. Bayesian nonparametrics, contraction rates, supremum norm.

attainable, which leads to the natural question to know whether this is still possible in density estimation, or in nonconjugate regression settings. This nontrivial question also arises for other likelihood methods, such as nonparametric maximum likelihood estimation; see Nickl [26].

From a general statistical perspective, density estimation in supremum-norm is a central problem both from theoretical and practical points of view. The problem was the object of much interest in the framework of minimax theory. Lower bounds in density estimation in sup-norm can be found in Hasminskii [19], upper-bounds in Ibragimov and Hasminskii [21] for density estimation and Stone [33] for regression. We refer to Goldenshluger and Lepski [17] for an overview of current work in this area. From the practical perspective, sup-norm properties are of course very desirable, since saying that two curves in a simulation picture look close is very naturally, and often implicitly, done in a sup-norm sense.

Here, we establish that minimax optimal sup-norm rates of convergence in density estimation are attainable by common and natural Bayes procedures. The methodology we introduce is in fact related to a programme initiated in [6] and continued in [7], namely nonparametric Bernstein–von Mises type results, as discussed below. In [7], we use the results of the present paper to derive nonparametric Bernstein–von Mises theorems in density estimation, as well as Donsker-type results for the posterior distribution function. The testing approach commonly used to establish posterior rates is replaced here by tools from semiparametric Bernstein–von Mises results (testing is still typically useful to establish preliminary rates); see [6] for an overview of references. We split the distance of interest in simpler pieces, each simpler piece being a semiparametric functional to study. One novelty of the paper consists in providing well-chosen uniform approximation schemes of various influence functions appearing at the semiparametric level when estimating those simple functionals.

Two natural families of nonparametric priors are considered for density estimation: priors on log-densities; see, for example, Ghosal, Ghosh and van der Vaart [13], Scricciolo [29], Tokdar and Ghosh [35], van der Vaart and van Zanten [3, 38], Rivoirard and Rousseau [27], and random (dyadic) histogram priors; see, for example, Barron [1], Barron, Schervish and Wasserman [2], Walker [39], Ghosal and van der Vaart [15], Scricciolo [30], Giné and Nickl [16] and the recent semiparametric treatment in [8]. Both classes are relevant for applications and priors of these types have been studied from the implementation perspective; see, for example, Lenk [23], Tokdar [34] and references therein for the use of logistic Gaussian process priors, and Leonard [24], Gasparini [11] for random histogram priors.

New results are also derived in the Gaussian white noise model, in the spirit of [6], for nonconjugate priors.

While working on this paper, we learned from the work by Marc Hoffmann, Judith Rousseau and Johannes Schmidt-Hieber [20], which independently obtains sup-norm properties for different priors. Their method is different from ours, and both approaches shed light on different specific aspects of the problem. In Gaussian white noise, adaptive results over Hölder classes are obtained in [20] for a class of sparse priors. In Theorem 1 below, the sup-norm minimax rate for fixed regularity is obtained for canonical priors without sparsity enforcement. The authors also give insight on the interplay between loss function and posterior rate, as well as an upper bound result for fairly abstract sieve-type priors, which are shown to attain the adaptive sup-norm rate in density estimation. This is an interesting existence result, but no method is provided to investigate sup-norm rates for general given priors. Although for simplicity we limit ourselves here to the fixed regularity case, the present paper suggests such a method and demonstrates its applicability by dealing with several commonly used classes of prior distributions. Clearly, there is still much to do in the understanding of posterior rates for strong measures of loss, and we hope that future contributions will go further in the different directions suggested by both the present paper and [20].

Let $L^2[0,1]$ and $L^\infty[0,1]$, respectively, denote the space of square integrable functions with respect to Lebesgue measure on $[0,1]$ and the space of measurable bounded functions on $[0,1]$. Theses spaces are equipped with their usual norms, respectively, denoted $\|\cdot\|_2$ (denote by $\langle\cdot,\cdot\rangle_2$ the associated inner product) and $\|\cdot\|_\infty$. Let $\mathcal{C}^\alpha := \mathcal{C}^\alpha[0,1]$ denote the class of Hölder functions on $[0,1]$ with Hölder exponent $\alpha > 0$.

For any $\alpha > 0$ and any $n \geq 1$, denote by $\bar{\varepsilon}_{n,\alpha}$ the rate

$$\bar{\varepsilon}_{n,\alpha} := n^{-\alpha/(2\alpha+1)}. \tag{1}$$

The typical minimax rate over a ball of the Hölder space $\mathcal{C}^\alpha[0,1]$, $\alpha > 0$, for the sup-norm is

$$\varepsilon^*_{n,\alpha} := \left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}. \tag{2}$$

Let us also set, omitting the dependence in α in the notation,

$$h_n = \left(\frac{n}{\log n}\right)^{-1/(2\alpha+1)}, \qquad L_n = \lfloor \log_2(1/h_n) \rfloor. \tag{3}$$

For a statistical model $\{P_f^{(n)}\}$ indexed by f in some class of functions to be specified and associated observations $X^{(n)}$, denote by f_0 the "true" function and by $E^n_{f_0}$ the expectation under $P^{(n)}_{f_0}$. Given a prior Π on a set of possible f's, denote by $\Pi[\cdot|X^{(n)}]$ the posterior distribution and by $E^\Pi[\cdot|X^{(n)}]$ the expectation operator under the law $\Pi[\cdot|X^{(n)}]$.

2. Prologue.

Let us start by a simple example in Gaussian white noise which will serve as a slightly naive yet useful illustration of the main technique of proof.

Let f be an element of $L^\infty[0,1]$. Let $n \geq 1$. Suppose one observes

$$dX^{(n)}(t) = f(t)\,dt + \frac{1}{\sqrt{n}}\,dW(t), \qquad t \in [0,1], \tag{4}$$

where W is standard Brownian motion. Let $\{\psi_{lk}, l \geq 0, 0 \leq k \leq 2^l - 1\}$ be a wavelet basis on the interval $[0,1]$. Here, we take the basis constructed in [10], see below for precise definitions. The model (4) is statistically equivalent to observing the projected observations onto the basis $\{\psi_{lk}\}$,

$$x_{lk} = f_{lk} + \frac{1}{\sqrt{n}}\varepsilon_{lk}, \qquad l \geq 0, 0 \leq k \leq 2^l - 1,$$

where $f_{lk} := \langle f, \psi_{lk}\rangle_2$ and ε_{lk} are i.i.d. standard normal. Denote $\hat{f}_{lk} := x_{lk}$, an efficient frequentist estimator of the wavelet coefficient f_{lk}.

2.1. *A first example*. Suppose the coefficients of the true function f_0 satisfy, for some $R > 0$ that we suppose to be *known* in this first example,

$$\sup_{l \geq 0, 0 \leq k \leq 2^l - 1} 2^{l(1/2+\alpha)} |f_{0,lk}| \leq R, \qquad \alpha > 0. \tag{5}$$

Define a prior Π on f via an independent product prior on its coordinates f_{lk} onto the considered basis. The component f_{lk} is assumed to be sampled from a prior with density $\sigma_l^{-1}\varphi(\cdot/\sigma_l)$ with respect to Lebesgue measure on $[0,1]$, where, for α, R as in (5), $x \in \mathbb{R}$ and a given $B > R$

$$\varphi(x) = \frac{1}{2B}\mathbb{1}_{[-B,B]}(x), \qquad \sigma_l = 2^{-l(1/2+\alpha)}. \tag{6}$$

This type of prior was considered in [16], Section 2.2, and provides a simple example of a random function with bounded α-Hölder norm.

PROPOSITION 1. *Consider observations $X^{(n)}$ from the model* (4). *Let f_0 and α satisfy* (5) *and let the prior be chosen according to* (6). *Then there exists $M > 0$ such that for $\varepsilon^*_{n,\alpha}$ defined by* (2),

$$E^n_{f_0}\int \|f - f_0\|_\infty \, d\Pi(f \mid X^{(n)}) \leq M\varepsilon^*_{n,\alpha}.$$

Uniform wavelet priors thus lead to the minimax rate of convergence in sup-norm. The result has a fairly simple proof, as we now illustrate, and is new, to the best of our knowledge.

Let L_n be defined in (3). Denote by f^{L_n} the orthogonal projection of f in $L^2[0,1]$ onto $\mathrm{Vect}\{\psi_{lk}, l \leq L_n, 0 \leq k < 2^l\}$, and $f^{L_n^c}$ the projection of f onto $\mathrm{Vect}\{\psi_{lk}, l > L_n, 0 \leq k < 2^l\}$. Then

$$f - f_0 = f^{L_n} - \hat{f}^{L_n} + \hat{f}^{L_n} - f_0^{L_n} + f^{L_n^c} - f_0^{L_n^c},$$

where $\hat{f}^{L_n}$ is the projection estimator onto the basis $\{\psi_{lk}\}$ with cut-off L_n. Note that the previous equality as such is an equality in L^2. However, if the wavelet series of f into the basis $\{\psi_{lk}\}$ is absolutely convergent Π-almost surely (which is the case for all priors considered in this paper), we also have $f(x) = f^{L_n}(x) + f^{L_n^c}(x)$ pointwise for Lebesgue-almost every x, Π-almost surely, and similarly for f_0. Now,

$$
\begin{aligned}
&E^{\Pi}[\|f - f_0\|_\infty | X^{(n)}] \\
&\quad = \int \|f - f_0\|_\infty \, d\Pi(f|X^{(n)}) \\
&\quad \le \underbrace{\int \|f^{L_n} - \hat{f}^{L_n}\|_\infty \, d\Pi(f|X^{(n)})}_{\text{(i)}} + \underbrace{\int \|f^{L_n^c}\|_\infty \, d\Pi(f|X^{(n)})}_{\text{(ii)}} + \underbrace{\|\hat{f}^{L_n} - f_0\|_\infty}_{\text{(iii)}}.
\end{aligned}
$$

We have (iii) $\le \|f_0^{L_n^c}\|_\infty + \|\hat{f}^{L_n} - f_0^{L_n}\|_\infty$. Using (5) and the localisation property of the wavelet basis $\|\sum_k |\psi_{lk}|\|_\infty \lesssim 2^{l/2}$ (see below), one obtains

$$
\|f_0^{L_n^c}\|_\infty \le \sum_{l > L_n} \left[\max_k |f_{0,lk}| \left\| \sum_k |\psi_{lk}| \right\|_\infty \right] \lesssim h_n^\alpha \lesssim \varepsilon_{n,\alpha}^*,
$$

where $\lesssim$ means less or equal to up to some universal constant. The term $\|\hat{f}^{L_n} - f_0^{L_n}\|_\infty$ depends on the randomness of the observations only,

$$
\|\hat{f}^{L_n} - f_0^{L_n}\|_\infty = \frac{1}{\sqrt{n}} \left\| \sum_{l \le L_n, k} \varepsilon_{lk} \psi_{lk}(\cdot) \right\|_\infty.
$$

This is bounded under $E_{f_0}^n$ by a constant times $\varepsilon_{n,\alpha}^*$; see Lemma 7 for a proof in the more difficult case of empirical processes.

Term (i). By definition, $\hat{f}^{L_n}$ has coordinates $\hat{f}_{lk}$ in the basis $\{\psi_{lk}\}$, so using the localisation property of the wavelet basis as above, one obtains

$$
\|f^{L_n} - \hat{f}^{L_n}\|_\infty \lesssim \frac{1}{\sqrt{n}} \sum_{l \le L_n} 2^{l/2} \Big[\max_{0 \le k < 2^l} \sqrt{n} |f_{lk} - x_{lk}| \Big].
$$

For $t > 0$, via Jensen's inequality and bounding the maximum by the sum, using Π_n as a shorthand notation for the posterior $\Pi[\cdot | X^{(n)}]$,

$$
\begin{aligned}
&t E_{f_0}^n E^{\Pi_n} \Big[\max_{0 \le k < 2^l} \sqrt{n} |f_{lk} - x_{lk}| \Big] \\
&\qquad \le \log \sum_{k=0}^{2^l - 1} E_{f_0}^n E^{\Pi_n} \big[e^{t\sqrt{n}(f_{lk} - x_{lk})} + e^{-t\sqrt{n}(f_{lk} - x_{lk})} \big],
\end{aligned}
$$

for any $l \geq 0$. Simple computations presented in Lemma 1 yield a sub-Gaussian behaviour for the Laplace transform of $\sqrt{n}(f_{lk} - x_{lk})$ under the posterior distribution, which is bounded above by $Ce^{t^2/2}$ for a constant C independent of $l \leq L_n$ and k. From this deduce, for any $t > 0$ and $l \leq L_n$,

$$E_{f_0}^n E^{\Pi}\Big[\max_{0 \leq k < 2^l} \sqrt{n}|f_{lk} - x_{lk}| \big| X^{(n)}\Big] \lesssim \frac{\log(C2^l)}{t} + \frac{t}{2}.$$

The choice $t = \sqrt{2\log(C2^l)}$ leads us to the bound

$$E_{f_0}^n(\mathrm{i}) \lesssim \frac{1}{\sqrt{n}} \sum_{l \leq L_n} \sqrt{l}2^{l/2} \lesssim \sqrt{L_n/(nh_n)} \lesssim \varepsilon_{n,\alpha}^*.$$

Term (ii). Under the considered prior, the wavelet coefficients of f are bounded by σ_l, so using again the localisation property of the wavelet basis,

$$\begin{aligned} E_{f_0}^n(\mathrm{ii}) &\lesssim \sum_{l > L_n} 2^{l/2} E_{f_0}^n E^{\Pi}\Big[\max_k |f_{lk}| \big| X^{(n)}\Big] \\ &\lesssim \sum_{l > L_n} 2^{l/2}\sigma_l \lesssim h_n^{\alpha} = \varepsilon_{n,\alpha}^*. \end{aligned}$$

This concludes the proof of Proposition 1.

Although fairly simple, the previous example is revealing of some important facts, some of which are well known from frequentist analysis of the problem, some being specific to the Bayesian approach. The previous proof shows two regimes of frequencies: $l \leq L_n$ "low frequency" and $l > L_n$ "high frequency." In the low frequency regime, the estimator x_{lk} of $f_{lk} = \langle f, \psi_{lk}\rangle_2$ is satisfactory, and the concentration of the posterior distribution around this efficient frequentist estimator is desirable. This is reminiscent of the Bernstein–von Mises (BvM) property; see van der Vaart [37], Chapter 10, which states that in regular parametric problems with unknown parameter θ, the posterior distribution is asymptotically Gaussian concentrating at rate $1/\sqrt{n}$ and centered around an efficient estimator of θ.

Here are a few words on the general philosophy of the results specifically in the Bayesian context. Such method was used as a building block in [6]. The idea is to split the distance of interest into small pieces. For the sup-norm, those pieces can, for instance, involve the wavelet coefficients $\langle f, \psi_{lk}\rangle_2$, but not necessarily, as will be seen for log-density priors. In this case, this split is obtained, for instance, from the inequality

$$\|f - f_0\|_\infty \lesssim \sum_{l \geq 0} 2^{l/2} \max_{0 \leq k \leq 2^l - 1} \big|\langle f, \psi_{lk}\rangle_2 - \langle f_0, \psi_{lk}\rangle_2\big|,$$

which holds for localised bases $\{\psi_{lk}\}$. Note that $f \to \langle f, \psi_{lk}\rangle_2$ can be seen as a semiparametric functional; see, for example, [37], Chapter 25 for an Introduction to semiparametrics and the notions of efficiency and efficient influence functions. Next, one analyses each piece separately, with different regimes of indexes l, k often arising, requiring specific techniques for each of them.

- the *BvM-regime: semiparametric bias.* For "low frequencies," what is typically needed is a concentration of the posterior distribution for the functional of interest, say $\langle f, \psi_{lk}\rangle_2$, at rate $1/\sqrt{n}$ around a semiparametrically efficient estimator of the functional. This is at the heart of the proof of semiparametric BvM results, hence the use of BvM techniques. In particular, sharp control of the bias will be essential. Regarding the BvM property, although the precise Gaussian shape will not be needed here, one needs *uniformity* in all frequencies in the considered regime. This requires nontrivial strengthenings of BvM-type results, the semiparametric efficient influence function of the functional of interest, which can be, for instance, a re-centered version of ψ_{lk}, being typically *unbounded* as l grows.
- Taking care of uniformity issues in approximation of the efficient influence functions by the prior may require various approximations regimes depending on l. For log-density priors, we will indeed see various regimes of indexes "l" arise in the obtained bounds for the bias.
- The *high-frequency bias* corresponds to frequencies where the prior should make the likelihood negligible. This part can be difficult to handle, too, especially for unbounded priors.

In the example above for uniform priors in white noise, most of the previous steps are either almost trivial or at least can be carried out by considering the explicit expression of the posterior, but for different priors or in different sampling situations some of the previous steps may become significantly harder, as we will see below.

2.2. *Wavelet basis and Besov spaces*. Central to our investigations is the tool provided by localised bases of $L^2[0, 1]$. We refer to the Lecture Notes by Härdle, Kerkyacharian, Picard and Tsybakov [18] for an Introduction to wavelets. Two bases will be used in the sequel.

The Haar basis on $[0, 1]$ is defined by $\varphi^H(x) = 1$, $\psi^H(x) := \psi^H_{0,0}(x) = -\mathbb{1}_{[0,1/2]}(x) + \mathbb{1}_{(1/2,1]}(x)$ and $\psi^H_{l,k}(x) = 2^{l/2}\psi(2^l x - k)$, for any integer l and $0 \le k \le 2^l - 1$. The supports of Haar wavelets form dyadic partitions of $[0, 1]$, corresponding to intervals $I^l_k := (k2^{-l}, (k + 1)2^{-l}]$ for $k > 0$, and where the interval is closed to the left when $k = 0$.

The boundary corrected basis of Cohen, Daubechies, Vial [10] will be referred to as CDV basis. Similar to the Haar basis, the CDV basis enables a treatment on compact intervals, but at the same time can be chosen sufficiently smooth. A few properties are lost, essentially simple explicit expressions, but most convenient localisation properties and characterisation of spaces are maintained. Below we recall some useful properties of the CDV basis. We denote this basis $\{\psi_{lk}\}$, with indexes $l \geq 0$, $0 \leq k \leq 2^l - 1$ (with respect to the original construction in [10], one starts at a sufficiently large level $l \geq J$, with J fixed large enough; for simplicity, up to renumbering, one can start the indexing at $l = 0$). Let $\alpha > 0$ be fixed.

- $\{\psi_{lk}\}$ forms an orthonormal basis of $L^2[0, 1]$.
- ψ_{lk} have support S_{lk}, with diameter at most a constant (independent of l, k) times 2^{-l}, and $\|\psi_{lk}\|_\infty \lesssim 2^{l/2}$. The ψ_{lk}'s are in the Hölder class $\mathcal{C}^S[0, 1]$, for some $S \geq \alpha$.
- At fixed level l, given a fixed ψ_{lk} with support S_{lk},
  - ⋄ the number of wavelets of the level $l' \leq l$ with support intersecting S_{lk} is bounded by a universal constant (independent of l', l, k),
  - ⋄ the number of wavelets of the level $l' > l$ with support intersecting S_{lk} is bounded by $2^{l'-l}$ times a universal constant.

  The following localisation property holds $\sum_{k=0}^{2^l-1} \|\psi_{lk}\|_\infty \lesssim 2^{l/2}$, where the inequality is up to a fixed universal constant.
- The constant function equal to 1 on $[0, 1]$ is orthogonal to high-level wavelets, in the sense that $\langle \psi_{lk}, 1\rangle_2 = \int_0^1 \psi_{lk} = 0$ whenever $l \geq M$, for a large enough constant M.
- The basis $\{\psi_{lk}\}$ characterises Besov spaces $B^s_{\infty,\infty}[0, 1]$, any $s \leq \alpha$, in terms of wavelet coefficients. That is, $g \in B^s_{\infty,\infty}[0, 1]$ if and only if

$$\|g\|_{\infty,\infty,s} := \sup_{l \geq 0, 0 \leq k \leq 2^l - 1} 2^{l(1/2+s)} \left|\langle g, \psi_{lk}\rangle_2\right| < \infty. \tag{7}$$

We note that orthonormality of the basis is not essential. Other nonorthonormal, multi-resolution dictionaries could be used instead up to some adaptation of the proofs, as long as coefficients in the expansion of f can be recovered from inner products. Also, recall that $B^s_{\infty,\infty}$ coincides with the Hölder space $\mathcal{C}^s$ when s is not an integer and that when s is an integer the inclusion $\mathcal{C}^s \subset B^s_{\infty,\infty}$ holds. If the Haar-wavelet is considered, the fact that f_0 is in $\mathcal{C}^s$, $0 < s \leq 1$, implies that the supremum in (7) with $\psi_{lk} = \psi^H_{lk}$ is finite.

3. Main results.

3.1. *Gaussian white noise.* Consider priors Π defined as coordinate-wise products of priors on coordinates specified by a density φ and scalings $\{\sigma_l\}$ as in Section 2.1. The next result allows for a much broader class of priors.

Let φ be a continuous density with respect to Lebesgue measure on $\mathbb{R}$. We assume that φ is (strictly) positive on $[-1, 1]$ and that it satisfies

$$\exists b_1, b_2, c_1, c_2, \delta > 0, \forall x : |x| \geq 1, \qquad c_1 e^{-b_1|x|^{1+\delta}} \leq \varphi(x) \leq c_2 e^{-b_2|x|^{1+\delta}}. \tag{8}$$

Consider a scaling σ_l for the prior equal to, for δ the constant in (8),

$$\sigma_l = \frac{2^{-l(1/2+\alpha)}}{(l+1)^\mu}, \qquad \mu = \frac{1}{1+\delta}. \tag{9}$$

THEOREM 1. *Let $X^{(n)}$ be observations from* (4). *Suppose f_0 belongs to $B^{\alpha}_{\infty,\infty}[0,1]$, for some $\alpha > 0$. Let the prior Π be a product prior defined through φ and σ_l satisfying* (8), (9). *Then there exists $M > 0$ such that for $\varepsilon^*_{n,\alpha}$ defined by* (2),

$$E^n_{f_0} \int \|f - f_0\|_\infty \, d\Pi(f \mid X^{(n)}) \le M\varepsilon^*_{n,\alpha}.$$

Theorem 1 can be seen as a generalisation to nonconjugate priors of Theorem 1 in [16]. Possible choices for φ cover several commonly used classes of prior distributions, such as so-called exponential power (EP) distributions; see, for example, Choy and Smith [9], Walker and Gutiérrez-Peña [40] and references therein, as well as some of the univariate Kotz-type distibutions, see, for example, Nadarajah [25]. Other choices of prior distributions are possible, up sometimes to some adaptations. For instance, Proposition 1 provides a result in the case of a uniform distribution. If one allows for some extra logarithmic term in the rate, Laplace (double-exponential) distributions can be used, as well as distributions without the control from below on the tail in (8), provided one chooses $\sigma_l = 2^{-l(1/2+\alpha)}$, as can be checked following the steps of the proof of Theorem 1. As a special case, the latter include all sub-Gaussian distributions. Also note that Theorem 1 as such applies to canonical priors, in that they do not depend on n. Results for truncated priors, which set $\langle f, \psi_{lk}\rangle = 0$ for l above a threshold, can be obtained along the same lines, with slightly simpler proofs.

Further consequences of Theorem 1 include the minimaxity in sup-norm of several Bayesian estimators. The result for the posterior mean immediately follows from a convexity argument. One can also check that the posterior coordinate-wise median is minimax. Details are omitted.

3.2. *Density estimation*. Consider independent and identically distributed observations

$$X^{(n)} = (X_1, \dots, X_n), \tag{10}$$

with unknown density function f on $[0,1]$. We use the same notation $X^{(n)}$ for observations as in the white noise model: it will always be clear from the context which model we are referring to. Let $\mathcal{F}$ be the set of densities f on $[0,1]$ which are bounded away from 0 and ∞. In other words, one can write $\mathcal{F} = \bigcup_{0<\rho\le D<\infty} \mathcal{F}(\rho, D)$, with $\mathcal{F}_{\rho,D} = \{f, 0 < \rho \le f \le D < \infty, \int_0^1 f = 1\}$. In the sequel, we assume that the "true" f_0 belongs to $\mathcal{F}_0 := \mathcal{F}(\rho_0, D_0)$, for some $0 < \rho_0 \le D_0 < \infty$. The assumption that the density is bounded away from 0 and ∞ is for simplicity. Allowing the density to tend to 0, for example, at the boundary of $[0,1]$ would be an interesting extension, but would presumably induce technicalities not related to our point here. Let h denote the Hellinger distance between densities on $[0,1]$.

3.2.1. *Log-densities priors.* Define the prior Π on densities as follows. Given a sufficiently smooth CDV-wavelet basis $\{\psi_{lk}\}$, consider the prior induced by, for any $x \in [0, 1]$ and L_n defined in (3),

$$T(x) = \sum_{l=0}^{L_n} \sum_{k=0}^{2^l-1} \sigma_{lk}\alpha_{lk}\psi_{lk}(x), \tag{11}$$

$$f(x) = \exp\{T(x) - c(T)\}, \qquad c(T) = \log \int_0^1 e^{T(x)}\,dx, \tag{12}$$

where α_{lk} are i.i.d. random variables of density φ with respect to Lebesgue measure on $\mathbb{R}$ and σ_{lk} are positive reals which for simplicity we make only depend on l, that is $\sigma_{lk} \equiv \sigma_l$. We consider the choices $\varphi(x) = \varphi_G(x) = e^{-x^2/2}/\sqrt{2\pi}$ the Gaussian density and $\varphi(x) = \varphi_H(x)$, where φ_H is any density such that its logarithm $\log \varphi_H$ is Lipschitz on $\mathbb{R}$. We refer to this as the "log-Lipschitz case." For instance, the α_{lk}'s can be Laplace-distributed or have heavier tails, such as, for a given $0 \le \tau < 1$ and $x \in \mathbb{R}$, and c_τ a normalising constant,

$$\varphi_{H,\tau}(x) = c_\tau \exp\{-(1 + |x|)^{1-\tau}\}. \tag{13}$$

Suppose the prior parameters σ_l satisfy, for some $\alpha > 1/2$ and $0 < r \le \alpha - \frac{1}{4}$,

$$\begin{aligned} &\sigma_l \ge 2^{-l(\alpha+1/2)} \qquad \text{(log-Lipschitz case)}, \\ &\sigma_l = 2^{-l(1/2+r)} \qquad \text{(Gaussian-case)}. \end{aligned} \tag{14}$$

Typically, see examples below, such priors f in (12) under $\varphi = \varphi_G$ or φ_H and (14) attain the rate $\bar{\varepsilon}_{n,\alpha}$ in (1) in terms of Hellinger loss, up to logarithmic terms. For some $\nu > 0$, suppose

$$E_{f_0}^n \Pi[f : h(f, f_0) > (\log n)^\nu \bar{\varepsilon}_{n,\alpha} | X^{(n)}] \to 0. \tag{15}$$

If (15) holds for some $\nu > 0$, we denote $\varepsilon_n := (\log n)^\nu \bar{\varepsilon}_{n,\alpha}$ and $\zeta_n := \varepsilon_n 2^{L_n/2}$, with L_n as in (3).

THEOREM 2. *Consider observations $X^{(n)}$ from model* (10). *Suppose* $\log f_0$ *belongs to $\mathcal{C}^\alpha[0, 1]$, with $\alpha \ge 1$. Let Π be the prior on $\mathcal{F}$ defined by* (12), *with $\varphi = \varphi_G$ or φ_H. Suppose that σ_l satisfy* (14) *and that* (15) *holds. Then, for $\alpha > 1$ and $\varepsilon_{n,\alpha}^*$ defined by* (2), *any $M_n \to \infty$, it holds, as $n \to \infty$,*

$$E_{f_0}^n \Pi[f : \|f - f_0\|_\infty > M_n \varepsilon_{n,\alpha}^* | X^{(n)}] \to 0.$$

In the case $\alpha = 1$, the same holds with $\varepsilon_{n,\alpha}^*$ replaced by $(\log n)^\eta \varepsilon_{n,\alpha}^*$, *for some $\eta > 0$.*

Theorem 2 implies that log-density priors for many natural priors on the coefficients achieve the precise optimal minimax rate of estimation over Hölder spaces under sup-norm loss, as soon as the regularity is at least 1.

In the case $1/2 < \alpha < 1$, examination of the proof reveals that the presented techniques provide the sup-norm rate $\rho_n = n^{(1/2-3\alpha/2)/(1+2\alpha)}$ up to logarithmic terms. For $1/2 < \alpha < 1$, we have $\varepsilon^*_{n,\alpha} \ll \rho_n \ll \zeta_n$. So, although the minimax rate is not exactly attained for those low regularities, the obtained rate improves on the intermediate rate ζ_n, which was obtained in [16] for slightly different priors. In the next subsection, a prior is proposed which attains the minimax rate for the sup-norm in the case $1/2 < \alpha < 1$.

Let us give some examples of prior distributions satisfying the assumptions of Theorem 2. In the Gaussian case, any sequence of the type $\sigma_l = 2^{-l(1/2+\gamma)}$ with $0 < \gamma \le \alpha - 1/4$ satisfies both (14) and (15). In the log-Lipschitz case, the choice $\varphi = \varphi_{H,\tau}$ in (13) with any $0 \le \tau < 1$ combined with $\sigma_l = 2^{-l\alpha}$ satisfies (14)–(15). Both claims follow from minor adaptations of Theorem 4.5 in [38] and Theorem 2.1 in [27], respectively; see Lemma 8. In both Gaussian and log-Lipschitz cases, we in fact expect (15) to hold true for many other choices of σ_l under (14) and $\log \varphi_H$ Lipschitz, or under $\sigma_l \ge 2^{-l(1/4+\alpha)}$ in the Gaussian case, although such a general statement in Hellinger distance is not yet available in the literature, to the best of our knowledge.

3.2.2. *Random dyadic histograms*. Associated to the regular dyadic partition of $[0, 1]$ at level $L \in \mathbb{N}^*$, given by $I_0^L = [0, 2^{-L}]$ and $I_k^L = (k2^{-L}, (k+1)2^{-L}]$ for $k = 1, \dots, 2^L - 1$, is a natural notion of histogram

$$\mathcal{H}_L = \left\{ h \in L^\infty[0, 1], h(x) = \sum_{k=0}^{2^L-1} h_k \mathbb{1}_{I_k^L}(x), h_k \in \mathbb{R}, k = 0, \dots, 2^L - 1 \right\}$$

the set of all histograms with 2^L regular bins on $[0, 1]$. Let $\mathcal{S}_L = \{\omega \in [0, 1]^{2^L}; \sum_{k=0}^{2^L-1} \omega_k = 1\}$ be the unit simplex in $\mathbb{R}^{2^L}$. Further denote

$$\mathcal{H}_L^1 = \left\{ f \in L^\infty[0, 1], f(x) = 2^L \sum_{k=0}^{2^L-1} \omega_k \mathbb{1}_{I_k^L}(x), (\omega_0, \dots, \omega_{2^L-1}) \in \mathcal{S}_L \right\}.$$

The set $\mathcal{H}_L^1$ is the subset of $\mathcal{H}_L$ consisting of histograms which are densities on $[0, 1]$. Let $\mathcal{H}^1$ be the set of all histograms which are densities on $[0, 1]$.

A simple way to specify a prior on $\mathcal{H}_L^1$ is to set $L = L_n$ deterministic and to fix a distribution for $\omega_L := (\omega_0, \dots, \omega_{2^L-1})$. Set $L = L_n$ as defined in (3). Choose some fixed constants $a, c_1, c_2 > 0$ and let

$$(16) \qquad L = L_n, \qquad \omega_L \sim \mathcal{D}(\alpha_0, \dots, \alpha_{2^L-1}), \qquad c_1 2^{-La} \le \alpha_k \le c_2,$$

for any admissible index k, where $\mathcal{D}$ denotes the Dirichlet distribution on S_L. Unlike suggested by the notation, the coefficients α of the Dirichlet distribution are allowed to depend on L_n, so that $\alpha_k = \alpha_{k,L_n}$.

THEOREM 3. *Let* $f_0 \in \mathcal{F}_0$ *and suppose* f_0 *belongs to* $\mathcal{C}^\alpha[0,1]$, *where* $1/2 < \alpha \leq 1$. *Let* Π *be the prior on* $\mathcal{H}^1 \subset \mathcal{F}$ *defined by* (16). *Then, for* $\varepsilon_{n,\alpha}^*$ *defined by* (2) *and any* $M_n \to \infty$ *it holds, as* $n \to \infty$,

$$E_{f_0}^n \Pi\big[f : \|f - f_0\|_\infty > M_n \varepsilon_{n,\alpha}^* | X^{(n)}\big] \to 0.$$

According to Theorem 3, random dyadic histograms achieve the precise minimax rate in sup-norm over Hölder balls. Condition (16) is quite mild. For instance, the uniform choice $\alpha_0 = \cdots = \alpha_{2^L-1} = 1$ is allowed, as well as a variety of others, for instance, one can take $\alpha_k = \alpha_{k,L_n}$ to originate from a measure $A = A_{L_n}$ on the interval $[0,1]$, of finite total mass $\bar{A}_{L_n} := A([0,1])$. By this we mean, $\alpha_k = A(I_k^{L_n})$. If $A/\bar{A}_{L_n}$ has say a fixed continuous and positive density a with respect to Lebesgue measure on $[0,1]$, then (16) is satisfied as soon as there exists a $\delta > 0$ with $2^{-\delta L_n} \lesssim \bar{A}_{L_n} \lesssim 2^{L_n}$.

3.2.3. *Further examples.* A referee of the paper, whom we thank for the suggestion, has asked whether the proposed technique would work for other priors, more specifically for non-n-dependent priors in density estimation. Although not considered here for lack of space, we would like to mention the important class of Pólya tree priors; see, for example, Lavine [22]. For well-chosen parameters, it can be shown that these priors achieve supremum-norm consistency in density estimation (consistency in the, weaker, Hellinger sense was studied, e.g., in [2]) and minimax rates of convergence in the sup-norm can be obtained. In particular, this class contain canonical (i.e., non-n-dependent) priors that achieve such optimal rates in density estimation. This will be studied elsewhere.

3.3. *Discussion.* We have introduced new tools which allow to obtain optimal minimax rates of contraction in strong distances for posterior distributions. The essence of the technique is to view the problem semiparametrically as the uniform study of a collection of semiparametric Bayes concentration results, very much in the spirit of nonparametric Bernstein–von Mises results as studied in [6]. For the sake of clarity, we refrain of carrying out further extensions in the present paper but briefly mention a few applications. From the sup-norm rates, optimal results—up to logarithmic terms- in L^q-metrics, $q \geq 2$, can be immediately obtained by interpolation. Adaptation to the unknown α could also be considered. This will be the object of future work. However, note that "fixed α" nonparametric results as such are already very desirable in strong norms. They can, for instance, be used in the study of remainder terms of semiparametric functional expansions or of LAN-expansions as, for example, to check the conditions of application of semiparametric Bernstein–von Mises theorems as in [4]. In this semiparametric perspective, adaptation to f is in fact not always desirable, since posteriors for functionals may behave pathologically when an adaptive prior on the nuisance is chosen; see [27] and [8], where it is shown that too large discrepancies in smoothness between the semiparametric functional and the unknown f can lead to undesirable bias. Also, we expect the present methodology to give results in a broad variety of statistical models and/or for different classes of priors. Indeed, it reduces the problem of the strong-distance rate to two parts: (1) uniform semiparametric study of functionals and (2) high-frequency bias. The first part is very much related to obtaining (uniform) semiparametric Bernstein–von Mises (BvM) results. So, any advance in BvM theory for classes of priors will automatically lead to advances in (1). As for (2), the studied examples suggest that for frequencies above the cut-off the posterior behaves essentially as the prior itself. So, contrary to the BvM-regime (1) where the prior washes out asymptotically, one does not expect a universal behaviour for this part. However, showing that the posterior is close to the prior provides a possible method of proof.

4. Proofs.

4.1. *Gaussian white noise.*

LEMMA 1. *Let $X^{(n)}$ follow model* (4). *Let f_0 satisfy* (5) *and let the prior Π be chosen according to* (6). *There exists $C > 0$ such that for any real t, any $n \geq 2$ and $l \leq L_n$, with L_n defined in* (3),

$$E^n_{f_0} E^\Pi\big[e^{t\sqrt{n}(f_{lk}-x_{lk})}|X^{(n)}\big] \leq C e^{t^2/2}.$$

PROOF. The proof is similar to the first lines of the proof of Theorem 5 in [6]: one uses Bayes' formula to express the posterior expectation in the lemma. Next, using (3) and (5), one checks that for any $v \in [-L_0, L_0]$ with $L_0 := \sqrt{2}(B-R)/2$, the ratio $|f_{0,lk}+v/\sqrt{n}|/\sigma_l$ is at most $R+(B-R)/2 < B$, for any $l \leq L_n$ and k. For such v's, since φ is the uniform density on $[-B, B]$, the expression involving φ in the next line is constant, and thus can be removed from the expression, leading to

$$\begin{aligned}
&E^\Pi\big[e^{t\sqrt{n}(f_{lk}-x_{lk})}|X^{(n)}\big] \\
&\quad = e^{-t\varepsilon_{lk}} \frac{\int e^{-v^2/2+(t+\varepsilon_{lk})v}\varphi((f_{0,lk}+v/\sqrt{n})/\sigma_l)\,dv}{\int e^{-v^2/2+\varepsilon_{lk}v}\varphi((f_{0,lk}+v/\sqrt{n})/\sigma_l)\,dv} \\
&\quad \lesssim e^{-t\varepsilon_{lk}} \frac{\int e^{tv-(v-\varepsilon_{lk})^2/2}\,dv}{\int_{-L_0}^{L_0} e^{-(v-\varepsilon_{lk})^2/2}\,dv} \lesssim \frac{\int e^{tu-u^2/2}\,du}{\int_{-L_0}^{L_0} e^{-(v-\varepsilon_{lk})^2/2}\,dv} \\
&\quad \lesssim e^{t^2/2}\bigg[\int_{-L_0}^{L_0} e^{-(v-\varepsilon_{lk})^2/2}\,dv\bigg]^{-1}.
\end{aligned}$$

Since ε_{lk} are standard normal, simple calculations show that the expectation of the inverse of the quantity under brackets is bounded by a universal constant, as in [6], pages 2015–2016. □

PROOF OF THEOREM 1. *Small* l. Let us first consider indexes l with $l \le L_n$. For any real t, set $Q_{lk}(t) := E^n_{f_0} E^{\Pi}[e^{t\sqrt{n}(f_{lk}-X_{lk})} | X^{(n)}]$. Using the fact that φ is bounded,

$$Q_{lk}(t) \lesssim E^n_{f_0} \frac{\int e^{t(v-\varepsilon_{lk})-((v-\varepsilon_{lk})^2/2)}\,dv}{\int e^{-(v-\varepsilon_{lk})^2/2}\varphi((f_{0,lk}+v/\sqrt{n})/\sigma_l)\,dv}.$$

Introduce the set, for any possibly l-dependent sequence M_l,

$$\mathcal{A}(M_l) := \left\{ v : \left| \frac{f_{0,lk}+v/\sqrt{n}}{\sigma_l} \right| \le M_l \right\}. \tag{17}$$

Choose $M_l = C(l+1)^{\mu}$ with $\mu = (1+\delta)^{-1}$. This implies, with our choices of M_l, σ_l and taking C large enough, that $\mathcal{A}(M_l)$ contains the interval $(-1,1)$. First restricting the integral on the denominator to $(-1,1)$ and next using the tail condition on φ and the fact that $\varphi \ge c_\varphi > 0$ on $(-1,1)$, one gets

$$Q_{lk}(t) \lesssim E^n_{f_0} \frac{e^{t^2/2}}{e^{-l}\int_{-1}^{1} e^{-(v-\varepsilon_{lk})^2/2}\,dv} \lesssim e^{t^2/2+l}.$$

The maximal inequality argument from Section 2.1 directly yields (i) $\le \varepsilon^*_{n,\alpha}$.

Large l. Let us now consider the case $l > L_n$. For any real t set,

$$\begin{aligned}
&E^n_{f_0} E^{\Pi}[e^{tf_{lk}} | X^{(n)}] \\
&\quad = E^n_{f_0} \frac{\int e^{t(f_{0,lk}+v/\sqrt{n})} e^{-v^2/2+\varepsilon_{lk}v} 1/\sqrt{n}\sigma_l \varphi((f_{0,lk}+v/\sqrt{n})/\sigma_l)\,dv}{\int e^{-v^2/2+\varepsilon_{lk}v} 1/\sqrt{n}\sigma_l \varphi((f_{0,lk}+v/\sqrt{n})/\sigma_l)\,dv} \\
&\quad =: E^n_{f_0} \frac{N_{lk}(t)}{D_{lk}}.
\end{aligned}$$

To bound the denominator, first restrict the integral to the set $\mathcal{A} := \mathcal{A}(1)$ as defined in (17). Set

$$\zeta_l = \int_{\mathcal{A}} v \frac{1}{\sqrt{n}\sigma_l} \varphi\left(\frac{f_{0,lk}+v/\sqrt{n}}{\sigma_l}\right) dv,$$

next apply Jensen's inequality with the logarithm function to get, with $|\mathcal{A}|$ the diameter of $\mathcal{A}$ and some constant $C > 0$,

$$\begin{aligned}
\log D_{lk} &\ge -\frac{|\mathcal{A}|\|\varphi\|_\infty}{2\sqrt{n}\sigma_l} \sup_{v\in\mathcal{A}} v^2 + \varepsilon_{lk}\zeta_l \\
&\ge -Cn(\sigma_l^2 + f_{0,lk}^2) + \varepsilon_{lk}\zeta_l,
\end{aligned}$$

where we have used that $M_l = 1$ in (17). Below we shall also use that

$$|\zeta_l| \lesssim \frac{|\mathcal{A}|\|\varphi\|_\infty}{\sqrt{n}\sigma_l} \sup_{v\in\mathcal{A}} |v| \lesssim \sqrt{n}(|f_{0,lk}| + \sigma_l).$$

To bound the numerator from above, split the integrating set into $\mathcal{A} := \mathcal{A}(1)$ and $\mathcal{A}^c$ and write $N_{lk}(t) =: N_{lk}^{(1)}(t) + N_{lk}^{(2)}(t)$ for the integrals over each respective set. Using the previous bound on D_{lk}, that $|t(f_{0,lk} + \frac{v}{\sqrt{n}})| \le |t|\sigma_l$ by definition of $\mathcal{A}$, and Fubini's theorem,

$$
\begin{aligned}
E_{f_0}^n \frac{N_{lk}^{(1)}(t)}{D_{lk}}
&\le e^{|t|\sigma_l + Cn(\sigma_l^2 + f_{0,lk}^2)} \int_{\mathcal{A}} E_{f_0}^n\big[e^{(v-\zeta_l)\varepsilon_{lk}}\big] \frac{e^{-v^2/2}}{\sqrt{n}\sigma_l} \varphi\left(\frac{f_{0,lk} + v/\sqrt{n}}{\sigma_l}\right) dv \\
&\le e^{|t|\sigma_l + Cn(\sigma_l^2 + f_{0,lk}^2) + \zeta_l^2/2} \|\varphi\|_\infty \sup_{v \in \mathcal{A}} e^{|v\zeta_l|} \\
&\lesssim e^{|t|\sigma_l + C'n(\sigma_l^2 + f_{0,lk}^2)}.
\end{aligned}
$$

On the other hand, the term over $\mathcal{A}^c$ can be bounded as follows:

$$
\begin{aligned}
E_{f_0}^n \frac{N_{lk}^{(2)}(t)}{D_{lk}}
&\lesssim e^{Cn(\sigma_l^2 + f_{0,lk}^2)} \\
&\quad \times \int_{(-1,1)^c} e^{t\sigma_l w} E_{f_0}^n\big[e^{-n/2(w\sigma_l - f_{0,lk})^2 + \varepsilon_{lk}(\sqrt{n}(w\sigma_l - f_{0,lk}) - \zeta_l)}\big] \varphi(w)\, dw \\
&\lesssim e^{Cn(\sigma_l^2 + f_{0,lk}^2) + \zeta_l^2/2} \int_{(-1,1)^c} e^{t\sigma_l w - \sqrt{n}\zeta_l(\sigma_l w - f_{0,lk})} \varphi(w)\, dw \\
&\lesssim e^{C'n(\sigma_l^2 + f_{0,lk}^2)} \int_{(-1,1)^c} e^{(t\sigma_l + \sqrt{n}\sigma_l\zeta_l)w} \varphi(w)\, dw.
\end{aligned}
$$

Using the tail behaviour of φ leads to

$$
E_{f_0}^n \frac{N_{lk}^{(2)}(t)}{D_{lk}} \lesssim e^{C'n(\sigma_l^2 + f_{0,lk}^2)} e^{C\{\sigma_l(|t| + \sqrt{n}|\zeta_l|)\}^{(\delta+1)/\delta}}.
$$

One deduces, using that for $l > L_n$, one has $n(\sigma_l^2 + f_{0,lk}^2) \le n2^{-l(1+2\alpha)} \lesssim \log n \lesssim l$, that for $t > 0$,

$$
\begin{aligned}
R_{lk}(t) := E_{f_0}^n E^{\Pi}\Big[\max_k |f_{lk}| \big| X^{(n)}\Big]
&\lesssim \frac{1}{t}\big(l + \log\big(e^{t\sigma_l} + e^{C\{\sigma_l(t + \sqrt{n}|\zeta_l|)\}^{(\delta+1)/\delta}}\big)\big) \\
&\lesssim \frac{l}{t} + \sigma_l + \frac{1}{t}\big\{\sigma_l\big(t + \sqrt{n}|\zeta_l|\big)\big\}^{(\delta+1)/\delta}.
\end{aligned}
$$

Set $t = \sigma_l^{-1} l^{\delta/(\delta+1)}$ to deduce, using $\sigma_l\sqrt{n}|\zeta_l| \lesssim l^{\delta/(\delta+1)}$ for $l > L_n$,

$$
R_{lk}(t) \lesssim l^{1/(\delta+1)}\sigma_l \lesssim 2^{-l(1/2+\alpha)}
$$

and further obtain (ii) $\le \sum_{l>L_n} 2^{l/2}2^{-l(1/2+\alpha)} \lesssim h_n^\alpha = \varepsilon_{n,\alpha}^*$. Therefore, for any $\delta > 0$, the rate is precisely $\varepsilon_{n,\alpha}^*$. $\square$

4.2. *Density estimation*, *notation*. Given observations $X^{(n)}$ from (10), denote by $\ell_n(f)$ the log-likelihood $\ell_n(f) = \sum_{i=1}^n \log f(X_i)$. For any u, v in $L^2(P_{f_0}) =: L^2(f_0)$, define the inner-product $\langle \cdot, \cdot \rangle_L$ with associated norm $\| \cdot \|_L$, together with a stochastic term $W_n(u)$, as follows:

$$\langle u, v \rangle_L = \int_0^1 (u - P_{f_0}u)(v - P_{f_0}v) f_0,$$

$$W_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \bigl[u(X_i) - P_{f_0}u\bigr].$$

In particular, in empirical process notation $W_n(u) = \mathbb{G}_n(u)$. For any f in $\mathcal{F}$, set $R_n(f, f_0) = \sqrt{n} P_{f_0} \log(f/f_0) + n\| \log(f/f_0)\|_L^2/2$. For any $f \in \mathcal{F}$, it holds

$$\ell_n(f) - \ell_n(f_0) = -\frac{n}{2}\|\log(f/f_0)\|_L^2 + \sqrt{n} W_n\bigl(\log(f/f_0)\bigr) + R_n(f, f_0).$$

Denote, for any density f in $\mathcal{F}$ and any given u in $L^2(f_0)$,

$$\mathcal{B}(u, f, f_0) = \left\langle \frac{f - f_0}{f_0}, u \right\rangle_L - \bigl\langle \log(f/f_0), u \bigr\rangle_L.$$

Let D_n be a measurable set. Denote by Π^{D_n} the restriction of Π to D_n. Suppose, as $n \to \infty$,

$$E_{f_0}^n \Pi\bigl(D_n | X^{(n)}\bigr) = 1 + o(1). \tag{18}$$

Combining (18) and Markov's inequality leads to, for any $M_n \to \infty$,

$$\begin{aligned} &E_{f_0}^n \Pi\bigl[f : \|f - f_0\|_\infty > M_n \varepsilon_{n,\alpha}^* | X^{(n)}\bigr] \\ &\qquad \le \bigl(M_n \varepsilon_{n,\alpha}^*\bigr)^{-1} E_{f_0}^n \bigl[E^{\Pi^{D_n}}\bigl[\|f - f_0\|_\infty | X^{(n)}\bigr] \Pi\bigl(D_n | X^{(n)}\bigr)\bigr] + o(1). \end{aligned}$$

In the sequel, we focus on bounding $E^{\Pi^{D_n}}[\|f - f_0\|_\infty | X^{(n)}]$ from above.

4.3. *Density estimation*, *log-density priors*. Let us define the set D_n by, for $\varepsilon_n = (\log n)^\nu \bar{\varepsilon}_{n,\alpha}$ the rate in (15), L_n as in (3) and $\zeta_n = \varepsilon_n 2^{L_n/2}$,

$$D_n = \bigl\{f, \|f - f_0\|_2 \le \varepsilon_n, \|f - f_0\|_\infty \le \zeta_n\bigr\}. \tag{19}$$

It follows from Lemma 4 below that $\Pi(D_n | X^{(n)})$ goes to 1 in probability, up to replacing ε_n by $M\varepsilon_n$ for a large enough constant M, and similarly for ζ_n. Indeed, since a ε_n-Hellinger-contraction rate for the posterior is assumed, see (15), the conditions of Lemma 4 are satisfied.

4.3.1. *First step*, *reduction to the logarithmic scale*. Let us set $g = \log f$ and $g_0 = \log f_0$. With T defined in (11), one has $g = T - c(T)$. First, one notes that obtaining a rate going to 0 for $\|g - g_0\|_\infty$ implies the same rate up to constants for $\|f - f_0\|_\infty$. Indeed, $\|f - f_0\|_\infty = \|e^{g_0}(e^{g-g_0} - 1)\|_\infty \lesssim \|g - g_0\|_\infty$ using the bound $|e^x - 1| \lesssim |x|$ for small x and that $\|f_0\|_\infty$ is bounded. So, instead of writing Markov's inequality as above with f, we write it with g, the set D_n still being the one defined in (19) with the dependence on $f - f_0$.

That is, we focus on bounding $E^{\Pi^{D_n}}[\|g - g_0\|_\infty | X^{(n)}]$ from above. Now write, with the notation g^{L_n} denoting the L^2-projection up to level L_n as in Section 2.1, and L_n as in (3),

$$
\begin{aligned}
&E^{\Pi^{D_n}}[\|g - g_0\|_\infty | X^{(n)}] \\
&\quad \le \underbrace{\int \|g^{L_n} - g_0^{L_n}\|_\infty \, d\Pi^{D_n}(f | X^{(n)})}_{\text{(i)}} + \underbrace{\int \|g^{L_n^c}\|_\infty \, d\Pi^{D_n}(f | X^{(n)})}_{\text{(ii)}} + \underbrace{\|g_0^{L_n^c}\|_\infty}_{\text{(iii)}}.
\end{aligned}
$$

The term (ii) is 0 because the sum defining T goes up to level $l \le L_n$ under the prior distribution, and the constant function 1 is orthogonal to higher levels. Since $g_0 = \log f_0$ belongs to $B^\alpha_{\infty,\infty}$ by assumption, the term (iii) is bounded by a constant times $\varepsilon^*_{n,\alpha}$.

We now start analysing the term (i). First, let us introduce, for $\{\mathcal{A}_{l,k}\}_{l,k}$ a collection of elements of $L^2(f_0)$ to be chosen later, and L_n as in (3),

$$
\Gamma^{L_n}(\cdot) := g_0^{L_n}(\cdot) + \frac{1}{\sqrt{n}} \sum_{l=0}^{L_n} \sum_{k=0}^{2^l - 1} W_n(\mathcal{A}_{l,k}) \psi_{lk}(\cdot). \tag{20}
$$

Next, let us write

$$
\text{(i)} \le \int \|g^{L_n} - \Gamma^{L_n}\|_\infty \, d\Pi^{D_n}(f | X^{(n)}) + \|\Gamma^{L_n} - g_0^{L_n}\|_\infty.
$$

The second term is bounded with the help of Lemma 7. For the first term, following the scheme of proof of the maximal inequality in Section 2.1 via the moment generating function, one sees that it is enough to bound for $t > 0$ the following quantity, uniformly in l, k and $l \le L_n$

$$
\mathcal{M}_{lk}(t) := e^{-t W_n(\mathcal{A}_{l,k})} E^{\Pi^{D_n}}\big[e^{t\sqrt{n}\langle g - g_0, \psi_{lk}\rangle_2} | X^{(n)}\big]. \tag{21}
$$

Denote $\rho(x) := \log(1 + x) - x$. It holds

$$
\begin{aligned}
\int_0^1 (g - g_0)\psi_{lk} &= \int_0^1 \log\left[\frac{f - f_0}{f_0} + 1\right] \psi_{lk} \\
&= \int_0^1 \frac{f - f_0}{f_0} \frac{\psi_{lk}}{f_0} f_0 + \int_0^1 \rho\left(\frac{f - f_0}{f_0}\right) \psi_{lk}.
\end{aligned}
$$

On D_n we have an intermediate sup-norm rate $\zeta_n = o(1)$ when $\alpha > 1/2$. In this case the argument of ρ in the previous display tends to 0. Using the bound $|\rho(u)| \le u^2$ for small u, one gets

$$\left|\int_0^1 \rho\left(\frac{f-f_0}{f_0}\right)\psi_{lk}\right| \le \|\psi_{lk}\|_\infty \int_0^1 \left(\frac{f-f_0}{f_0}\right)^2 \lesssim 2^{l/2}\|f-f_0\|_2^2. \tag{22}$$

This bound is a $O(1/\sqrt{n})$ on D_n as soon as $2^{L_n/2}\varepsilon_n^2 = O(1/\sqrt{n})$, which is satisfied if $\alpha > 1$. This implies that the inner-product $\langle g - g_0, \psi_{lk}\rangle_2$ in (21) can be replaced by $\langle f - f_0, \zeta_{l,k}\rangle$, where

$$\zeta_{l,k} = \frac{\psi_{lk}}{f_0}. \tag{23}$$

That is, we can reason as if one would be considering the semiparametric problem of estimating the linear functional of the density $f \to \langle \zeta_{l,k}, f\rangle_2$. The corresponding efficient influence function is $\tilde{\zeta}_{l,k} = \zeta_{l,k} - P_{f_0}\zeta_{l,k}$, with respect to the tangent set $\mathcal{H}_{f_0} := \{h : [0,1] \to \mathbb{R}, h \text{ bounded}, \int_0^1 hf_0 = 0\}$; see [37], Chapter 25 for definitions.

There is one difficulty with $\zeta_{l,k}$. It is not an element of the basis of expansion of the prior Π, so it needs to be properly approximated by the prior in some sense. In fact, there is a fundamental difference with what has been done so far in proving BvM-type results; see, for example, [4, 8, 27]. Here, we need to study approximating sequences *uniformly* in the indexes l, k and a sharp control on this dependence is essential; see the key Lemma 2, where two regimes of indexes "l" arise, depending on whether l is small or close to L_n.

So, instead of working with $\zeta_{l,k}$ directly, one replaces it by an approximation $\mathcal{A}_{l,k}$ defined in (27) below. This induces a *bias* term for any l, k, familiar in the context of semiparametric BvM results; see, for example, [4, 5, 8], equal to

$$\sqrt{n}\langle f - f_0, \zeta_{l,k} - \mathcal{A}_{l,k}\rangle_2 = \sqrt{n}\int_0^1 (f - f_0)(\zeta_{l,k} - \mathcal{A}_{l,k}). \tag{24}$$

This term is controlled using Lemma 2 below. Indeed, on D_n the bounds of (24) of Lemma 2 are at most $\sqrt{nh_n}\varepsilon_n = o(1)$ if $\alpha > 1$. Next, apply Lemma 3 with $\gamma_n = \mathcal{A}_{l,k}$. The estimates of L^2 and sup-norm of $\mathcal{A}_{l,k}$ imply that the conditions of application of Lemma 3 are satisfied. Thus,

$$\mathcal{M}_{lk}(t) \le e^{Ct^2}\frac{\int e^{\ell_n(f_t)-\ell_n(f_0)}\,d\Pi^{D_n}(f)}{\int e^{\ell_n(f)-\ell_n(f_0)}\,d\Pi^{D_n}(f)}, \tag{25}$$

where we have set $f_t = e^{g_t}$ with $g_t = g_{t,l,k}$ defined as in Lemma 3 by (the expression is invariant under adding a constant to g, so one can write it either with $\mathcal{A}_{l,k}$ or $\tilde{\mathcal{A}}_{l,k} = \mathcal{A}_{l,k} - P_{f_0}\mathcal{A}_{l,k}$)

$$g_t = g - \frac{t}{\sqrt{n}}\mathcal{A}_{l,k} - \log\int e^{g - t/\sqrt{n}\mathcal{A}_{l,k}}. \tag{26}$$

In the case $1/2 < \alpha \le 1$, the cost of replacing f by $\log f$ is controlled by $\|g - g_0 - (f - f_0)\|_\infty = \|f_0\rho(f/f_0 - 1)\|_\infty \lesssim \|f/f_0 - 1\|_\infty^2$, which is bounded on D_n by a constant times ζ_n^2 via Lemma 4. Using Remark 1 below, the bias (24) leads to an extra term $\exp\{t\sqrt{n}n^{(-3\alpha/2)/(1+2\alpha)}\}$ in (25).

4.3.2. *"Uniform" approximations of efficient influence functions* $\tilde{\zeta}_{l,k}$. For any $l \le L_n$ and k between 0 and $2^l - 1$, define $\mathcal{A}_{l,k}$ to be the L^2-projection of $\zeta_{l,k}$ on the space spanned by the first L_n levels of wavelet coefficients,

$$\mathcal{A}_{l,k} = \sum_{1\le\lambda\le L_n} \sum_{0\le\mu\le 2^\lambda-1} \langle \zeta_{l,k}, \psi_{\lambda\mu}\rangle_2 \psi_{\lambda\mu}. \tag{27}$$

For any l, k in the previous ranges, we also set

$$\tilde{\mathcal{A}}_{l,k} = \mathcal{A}_{l,k} - P_{f_0}\mathcal{A}_{l,k}.$$

LEMMA 2. *Let f_0 belong to $\mathcal{F}_0 \cap \mathcal{C}^\alpha[0,1]$, with $\alpha \ge 1$. For any l such that $1 \le 2^l \le 2^{L_n}$ and $0 \le k \le 2^l - 1$, any density f in $\mathcal{F}$, and $\mathcal{A}_{l,k}$ as in* (27),

$$\|\mathcal{A}_{l,k} - \zeta_{l,k}\|_\infty \lesssim 2^{l(1/2+\alpha)} 2^{-\alpha L_n},$$

$$\left|\int_0^1 (\mathcal{A}_{l,k} - \zeta_{l,k})(f - f_0)\right| \lesssim \left(2^{(l-L_n)\alpha} \wedge 2^{-l}\right)\|f - f_0\|_2.$$

PROOF. For any admissible indexes λ, μ, let $S_{\lambda\mu}$ denote the support of the wavelet $\psi_{\lambda\mu}$ in $[0,1]$ and $|S_{\lambda\mu}|$ its Lebesgue measure. The following identity holds both in $L^2[0,1]$ (definition of the L^2-projection) and in $L^\infty[0,1]$ (because $\zeta_{l,k} \in B^S_{\infty,\infty}$ with $S > 0$)

$$\zeta_{l,k} - \mathcal{A}_{l,k} = \sum_{\lambda > L_n} \sum_{\mu=0}^{2^\lambda-1} \langle \zeta_{l,k}, \psi_{\lambda\mu}\rangle_2 \psi_{\lambda\mu}. \tag{28}$$

Since ψ_{lk} belongs to $B^\alpha_{\infty,\infty}$ and f_0 to $\mathcal{F}_0 \cap \mathcal{C}^\alpha$, Lemma 5 implies that $\zeta_{l,k} = \psi_{lk} \cdot f_0^{-1}$ belongs to $B^\alpha_{\infty,\infty}$, with $\|\cdot\|_{\infty,\infty,\alpha}$-norm bounded above by a constant times $\|\psi_{lk}\|_{\infty,\infty,\alpha}\|f_0^{-1}\|_{\infty,\infty,\alpha} \lesssim 2^{l(1/2+\alpha)}$, again by Lemma 5, using $f_0^{-1} \in \mathcal{C}^\alpha \subset B^\alpha_{\infty,\infty}$. Now using the localisation property of the wavelet basis,

$$\begin{aligned}
\|\mathcal{A}_{l,k} - \zeta_{l,k}\|_\infty &\le \left\| \sum_{\lambda > L_n} \sum_{\mu=0}^{2^\lambda-1} \langle \zeta_{l,k}, \psi_{\lambda\mu}\rangle_2 \psi_{\lambda\mu} \right\|_\infty \\
&\le \sum_{\lambda > L_n} 2^{\lambda/2} 2^{-\lambda(\alpha+1/2)} \max_{0\le\mu\le 2^\lambda-1} \left[2^{\lambda(\alpha+1/2)} |\langle \zeta_{l,k}, \psi_{\lambda\mu}\rangle_2|\right] \\
&\le \|\zeta_{l,k}\|_{\infty,\infty,\alpha} \sum_{\lambda > L_n} 2^{-\lambda\alpha} \lesssim 2^{l(1/2+\alpha)} 2^{-\alpha L_n}.
\end{aligned}$$

Now let us prove that the support of $\zeta_{l,k} - \mathcal{A}_{l,k}$ has diameter at most a constant times $|S_{lk}|$. Indeed, $\zeta_{l,k} - \mathcal{A}_{l,k}$ written above is a linear combination of "high"-frequency wavelets ($\lambda > L_n$), with support diameter thus at most of the order of $|S_{\lambda\mu}| \leq R|S_{lk}|$, for a fixed constant R, for any $\lambda > L_n$, any admissible μ, since $\lambda > l$. But in the sum (28), one may keep only those $\psi_{\lambda\mu}$ whose support intersects the one of $\zeta_{l,k}$, otherwise the coefficient $\langle \zeta_{l,k}, \psi_{\lambda\mu}\rangle_2$ is 0. So, all supports of the $\psi_{\lambda\mu}$'s which have a nonzero contribution to (28) are contained in an interval of [0, 1] of diameter at most $(2R+1)|S_{lk}|$. Thus, the diameter of the support of $\zeta_{l,k} - \mathcal{A}_{l,k}$ is at most $(2R+1)|S_{lk}|$.

Now we focus on $\int_0^1 (\mathcal{A}_{l,k} - \zeta_{l,k})(f - f_0) = \int_0^1 (\mathcal{A}_{l,k} - \zeta_{l,k})(f - f_0)\mathbb{1}_{\Delta_{l,k}}$, where $\Delta_{l,k}$ denotes the support of $\mathcal{A}_{l,k} - \zeta_{l,k}$. Bounding $\mathcal{A}_{l,k} - \zeta_{l,k}$ by its supremum and next applying Cauchy–Schwarz inequality,

$$\begin{aligned}\left|\int_0^1 (\mathcal{A}_{l,k} - \zeta_{l,k})(f - f_0)\right| &\lesssim \|\mathcal{A}_{l,k} - \zeta_{l,k}\|_\infty \sqrt{|\Delta_{lk}|}\|f - f_0\|_2 \\ &\lesssim 2^{(l-L_n)\alpha}\|f - f_0\|_2.\end{aligned}$$

To obtain the other part of the bound, the idea is to use a different approximating sequence $\mathcal{D}_{l,k}$ for which the comparison to $\zeta_{l,k}$ is easier for large l's. Define $\mathcal{D}_{l,k}$ to be the function obtained by replacing f_0 in (23) by its average on the support of ψ_{lk},

$$\mathcal{D}_{l,k} = \frac{\psi_{lk}}{[\bar{f}_0]_{lk}}, \tag{29}$$

where we have set

$$[\bar{f}_0]_{lk} = \frac{1}{|S_{lk}|}\int_{S_{lk}} f_0.$$

Note that since $l \leq L_n$, by definition $\mathcal{D}_{l,k}$ belongs to the vector space generated by the first L_n levels of wavelet coefficients. In particular, it holds $\|\mathcal{A}_{l,k} - \zeta_{l,k}\|_2 \leq \|\mathcal{D}_{l,k} - \zeta_{l,k}\|_2$ by definition of the L^2-projection. Since by definition again $\mathcal{D}_{l,k} - \zeta_{l,k}$ has support included in S_{lk}, one gets

$$\|\mathcal{A}_{l,k} - \zeta_{l,k}\|_2^2 \leq \int \mathbb{1}_{S_{lk}}(\mathcal{D}_{l,k} - \zeta_{l,k})^2 \leq \|\mathcal{D}_{l,k} - \zeta_{l,k}\|_\infty^2 |S_{lk}|.$$

Next, one bounds the last sup-norm. Denoting $\rho_0 := \inf_{x\in[0,1]} f_0(x)$,

$$\begin{aligned}\|\mathcal{D}_{l,k} - \zeta_{l,k}\|_\infty &\leq \rho_0^{-1}\|\psi_{lk}\|_\infty \sup_{x\in S_{lk}} |f_0(x) - [\bar{f}_0]_{lk}| \\ &\leq \rho_0^{-1}\|\psi_{lk}\|_\infty \sup_{x\in S_{lk}} |S_{lk}|^{-1}\left|\int_{S_{lk}} \big(f_0(x) - f_0(u)\big)\,du\right| \\ &\leq \rho_0^{-1}\|\psi_{lk}\|_\infty \sup_{x\in S_{lk}} |S_{lk}|^{-1}\int_{S_{lk}} |x - u|\,du \\ &\leq \rho_0^{-1}\|\psi_{lk}\|_\infty |S_{lk}|^{-1}|S_{lk}|^2/2 \lesssim 2^{-l/2},\end{aligned}$$

where for the third inequality we have used the fact that f_0 is at least Hölder 1 and for the last inequality that $|S_{lk}|$ is of the order 2^{-l} and $\|\psi_{lk}\|_\infty$ of the order $2^{l/2}$ up to constants. Thus,

$$\begin{aligned}\left|\int_0^1 (\mathcal{A}_{l,k} - \zeta_{l,k})(f - f_0)\right| &\le \|\mathcal{A}_{l,k} - \zeta_{l,k}\|_2 \|f - f_0\|_2 \\ &\le \sqrt{|S_{lk}|}\|\mathcal{D}_{l,k} - \zeta_{l,k}\|_\infty \|f - f_0\|_2 \\ &\lesssim 2^{-l}\|f - f_0\|_2. \qquad \square\end{aligned}$$

REMARK 1. In the case $1/2 < \alpha < 1$, similarly one gets the bound $(2^{-l\alpha} \wedge (2^{l-L_n})^\alpha)\|f - f_0\|_2$. The minimum of the bounds is attained for $2^l = 2^{L_n/2}$. This leads to a bound for the integral $\int_0^1 (\mathcal{A}_{l,k} - \zeta_{l,k})(f - f_0)$ equal to $n^{(-3\alpha/2)/(1+2\alpha)}$ for all considered indexes l, k.

4.3.3. *Change of variables.* Now everything is in place to start exploiting (25)–(26). First, rewrite (25) as

$$\frac{\int e^{\ell_n(f_t) - \ell_n(f_0)}\, d\Pi^{D_n}(f)}{\int e^{\ell_n(f) - \ell_n(f_0)}\, d\Pi^{D_n}(f)} = \frac{\int \mathbb{1}_{D_n}(f) e^{\ell_n(f_t) - \ell_n(f_0)}\, d\Pi(f)}{\Pi(D_n|X^{(n)}) \int e^{\ell_n(f) - \ell_n(f_0)}\, d\Pi(f)}. \tag{30}$$

The expression $\log f_t =$ (26), due to its invariance by adding a constant and recalling that $g = T - c(T)$ from (12), can be seen as a function of $T - t\mathcal{A}_{l,k}/\sqrt{n}$ [the constant $c(T)$ vanishes]. More precisely, we are now ready to change variables in the prior by setting

$$\tilde{T} = T - \frac{t}{\sqrt{n}}\mathcal{A}_{l,k}. \tag{31}$$

Essentially, if the "complexity" of $\mathcal{A}_{l,k}$ is not too large in view of the chosen prior Π, the fact of having f_t instead of f in (25) will not matter much and the corresponding ratio of integrals will be close to 1. We treat the case of log-Lipschitz priors on coefficients first. In fact, as can intuitively be guessed, a prior with heavy tails is less influenced under shift transformations than a more concentrated prior.

Denote by $\mathcal{C}_n = \{(\lambda, \mu) \in \mathbb{N}^2, 0 \le \mu \le 2^\lambda - 1, 1 \le 2^\lambda \le 2^{L_n}\}$. By the definition (27) of $\mathcal{A}_{l,k}$,

$$\langle \mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2 = \langle \zeta_{l,k}, \psi_{\lambda\mu}\rangle_2, \qquad (\lambda, \mu) \in \mathcal{C}_n, (l, k) \in \mathcal{C}_n.$$

Log-Lipschitz prior. With the chosen prior on f, the numerator in (30) is in fact an integral over the law of the coefficients of T in (11), that is, over (a subset of) $\mathbb{R}^{2^{L_n}}$. The change of variables (31) is thus a shift in $\mathbb{R}^{2^{L_n}}$, and its Jacobian is 1. The coordinates of T in the wavelet basis $\{\psi_{\lambda\mu}\}$ have densities $\sigma_\lambda^{-1}\varphi(\theta_{\lambda,\mu}/\sigma_\lambda)$

with respect to $d\theta_{\lambda,\mu}$ (we denote by $\theta_{\lambda,\mu}$ the integrating variable). The transformation in density can be controlled by, since $\varphi = \varphi_H$ has a Lipschitz logarithm,

$$\begin{aligned}
&\log \prod_{(\lambda,\mu)\in\mathcal{C}_n} \frac{\varphi(\theta_{\lambda,\mu}/\sigma_\lambda)}{\varphi(\{\theta_{\lambda,\mu} - t/\sqrt{n}\langle\mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2\}/\sigma_\lambda)} \\
&\quad = \sum_{(\lambda,\mu)\in\mathcal{C}_n} (\log\varphi)(\theta_{\lambda,\mu}/\sigma_\lambda) - (\log\varphi)\left(\left\{\theta_{\lambda,\mu} - \frac{t}{\sqrt{n}}\langle\mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2\right\}\Big/\sigma_\lambda\right) \\
&\quad \le \sum_{(\lambda,\mu)\in\mathcal{C}_n} \frac{|t|}{\sqrt{n}\sigma_\lambda} |\langle\zeta_{l,k}, \psi_{\lambda\mu}\rangle_2| \qquad \forall (l,k) \in \mathcal{C}_n.
\end{aligned}$$

We now study conditions on the σ_λ's under which the last display is bounded above by $C|t|$. Let us split the sum over $\mathcal{C}_n$ in the two cases $\lambda \le l$ and $\lambda > l$. When $\lambda \le l$, for any fixed level λ there is a constant number of wavelets $\psi_{\lambda\mu}$ intersecting the support of $\zeta_{l,k}$. Combined with $|\langle\zeta_{l,k}, \psi_{\lambda\mu}\rangle_2| \le \rho_0^{-1}$, this leads to the condition $\sum_{\lambda\le l} \sigma_\lambda^{-1} \lesssim \sqrt{n}$. When $\lambda > l$, for any fixed level λ there is a constant times $2^{\lambda-l}$ wavelets $\psi_{\lambda\mu}$ intersecting the support of $\zeta_{l,k}$, leading to

$$\begin{aligned}
\sum_{(\lambda,\mu)\in\mathcal{C}_n} \sigma_\lambda^{-1} |\langle\zeta_{l,k}, \psi_{\lambda\mu}\rangle_2| &\le \sum_{l<\lambda\le L_n} \sigma_\lambda^{-1} 2^{\lambda-l} 2^{-\lambda(1/2+\alpha)} \|\zeta_{l,k}\|_{\infty,\infty,\alpha} \\
&\le \sum_{l<\lambda\le L_n} \sigma_\lambda^{-1} 2^{(l-\lambda)(\alpha-1/2)},
\end{aligned}$$

where we have used that $\zeta_{l,k} = \psi_{lk}/f_0$ is α-smooth and applied Lemma 5. These conditions are quite mild. In particular, for $\alpha > 1/2$ they are implied by $\sum_{\lambda\le L_n} \sigma_\lambda^{-1} \lesssim \sqrt{n}$.

Gaussian prior. Let us write explicitly the log-ratio of densities

$$\begin{aligned}
&\sum_{(\lambda,\mu)\in\mathcal{C}_n} (\log\varphi)(\theta_{\lambda,\mu}/\sigma_\lambda) - (\log\varphi)\left(\left\{\theta_{\lambda,\mu} - \frac{t}{\sqrt{n}}\langle\mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2\right\}\Big/\sigma_\lambda\right) \\
&\quad = \sum_{(\lambda,\mu)\in\mathcal{C}_n} \frac{t^2}{2n\sigma_\lambda^2} \langle\mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2^2 - \frac{t}{\sqrt{n}\sigma_\lambda^2} \theta_{\lambda,\mu} \langle\mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2.
\end{aligned} \tag{32}$$

The obtained quantity still depends on the integrating variables $\theta_{\lambda,\mu}$. The idea is to exploit the fact that on D_n, it holds $\|g - g_0\|_2 \lesssim \varepsilon_n$, which is obtained along the way in the proof of Lemma 4. But $g = T - c(T) = T - c(T)1$, where 1 denotes the constant function equal to 1. Since 1 is orthogonal to high levels of wavelet coefficients, it means that for large enough λ, say $\lambda > K$, and any μ, it holds $|\theta_{\lambda,\mu} - g_{0,\lambda,\mu}| \le \|g - g_0\|_2 \lesssim \varepsilon_n$. So, for such λ, μ, we decompose $\theta_{\lambda,\mu} = \theta_{\lambda,\mu} - g_{0,\lambda,\mu} + g_{0,\lambda,\mu}$. For coefficients $\theta_{\lambda,\mu}$ such that $\lambda \le K$, we use a different argument.

Let us first deal with the term containing $\theta_{\lambda,\mu}$ in (32) when $\lambda \le K$. From the beginning of the proof, using Lemma 9, one can restrict slightly the set D_n by intersecting it with the set $\{T : \max_{\lambda \le K,\mu} |\langle T, \psi_{lk}\rangle| \le C\sqrt{n}\varepsilon_n\}$. Hence, using that $|\langle \mathcal{A}_{l,k}, \psi_{\lambda\mu}\rangle_2| \lesssim 1$ and the assumed specific form of σ_λ, one gets that the term at stake is at most a fixed constant times $|t|\varepsilon_n$.

Now we bound (32) and only have to deal with $\lambda > K$ for the part depending on $\theta_{\lambda,\mu}$. For any $(l,k) \in \mathcal{C}_n$, the first term in (32) is

$$\frac{t^2}{2n} \sum_{(\lambda,\mu)\in\mathcal{C}_n} \sigma_\lambda^{-2} \left\langle \frac{\psi_{lk}}{f_0}, \psi_{\lambda\mu}\right\rangle_2^2 \lesssim \frac{t^2}{n} \sum_{\lambda \le l} \sigma_\lambda^{-2} + \frac{t^2}{n} \sum_{l<\lambda\le L_n} \sigma_\lambda^{-2} 2^{(l-\lambda)2\alpha},$$

where the bound is obtained in a similar way as in the log-Lipschitz case by distinguishing the cases $\lambda \le l$ and $l > \lambda$. For the second term, we decompose $\theta = \theta - g_0 + g_0$ and use Cauchy–Schwarz inequality on the $\theta - g_0$ part,

$$\begin{aligned} &\frac{|t|}{\sqrt{n}} \|g - g_0\|_2 \left\{ \sum_{(\lambda,\mu)\in\mathcal{C}_n} \sigma_\lambda^{-4} \left\langle \frac{\psi_{lk}}{f_0}, \psi_{\lambda\mu}\right\rangle_2^2 \right\}^{1/2} \\ &\qquad \le |t| \frac{\varepsilon_n}{\sqrt{n}} \left\{ \sum_{\lambda\le l} \sigma_\lambda^{-4} + \sum_{l<\lambda\le L_n} \sigma_\lambda^{-4} 2^{(l-\lambda)2\alpha} \right\}^{1/2}. \end{aligned}$$

Finally, the remaining term with g_0 is bounded by $|t|/\sqrt{n}$ times

$$\begin{aligned} &\sum_{(\lambda,\mu)\in\mathcal{C}_n} \sigma_\lambda^{-2} \left| g_{0,\lambda\mu} \left\langle \frac{\psi_{lk}}{f_0}, \psi_{\lambda\mu}\right\rangle_2 \right| \\ &\qquad \le \sum_{\lambda\le l} \sigma_\lambda^{-2} 2^{-\lambda(1/2+\alpha)} + \sum_{l<\lambda\le L_n} \sigma_\lambda^{-2} 2^{(l-\lambda)2\alpha} 2^{-\lambda(1/2+\alpha)}. \end{aligned}$$

Under the condition of the theorem $\sigma_\lambda \ge 2^{-\lambda(1/4+\alpha)}$ for any $0 \le \lambda \le L_n$, all bounds obtained above in the Gaussian case are less than $C(|t| + t^2)$.

4.3.4. *End of proof.* To conclude for both considered classes of priors, note that the indicator $\mathbb{1}_{D_n}(f)$ in (30) becomes $\mathbb{1}_{D_n'}(f_t)$ under the change of variables—for a set D_n' that one can write explicitly, although this will not be needed here—and one simply further bounds the indicator $\mathbb{1}_{D_n'}$ by 1 on the numerator. Once the change of variable is done, the assumed conditions on $\{\sigma_l\}$ ensure that the ratio of densities is bounded by $e^{C(|t|+t^2)}$ for some constant C and one gets

$$\frac{\int e^{\ell_n(f_t)-\ell_n(f_0)} \mathbb{1}_{D_n}(f)\, d\Pi(f)}{\int e^{\ell_n(f)-\ell_n(f_0)}\, d\Pi(f)} \le e^{C(|t|+t^2)} \frac{\int e^{\ell_n(f)-\ell_n(f_0)}\, d\Pi(f)}{\int e^{\ell_n(f)-\ell_n(f_0)}\, d\Pi(f)} = e^{C(|t|+t^2)}.$$

Inequality (25) can thus be further written, again for a fixed constant C,

$$E^{\Pi^{D_n}}\left[e^{t\sqrt{n}\langle g - \Gamma^{L_n}, \psi_{lk}\rangle_2} | X^{(n)}\right] \le e^{C(|t|+t^2)} \Pi\left(D_n | X^{(n)}\right)^{-1}.$$

For any $s > 0$, similar to the Gaussian white noise case,

$$
\begin{aligned}
&\int \|g^{L_n} - \Gamma^{L_n}\|_\infty \, d\Pi^{D_n}(f \mid X^{(n)}) \\
&\quad \le \sum_{l \le L_n} \frac{2^{l/2}}{\sqrt{n}} E^{\Pi^{D_n}}\Big[\max_{0 \le k \le 2^l - 1} \big|\sqrt{n}\langle g - \Gamma^{L_n}, \psi_{lk}\rangle_2\big| \big| X^{(n)}\Big] \\
&\quad \le \sum_{l \le L_n} 2^{l/2} \frac{1}{\sqrt{n}s} \log\Bigg[\sum_{k=0}^{2^l-1} E^{\Pi^{D_n}}\big[e^{s\sqrt{n}\langle g - \Gamma^{L_n}, \psi_{lk}\rangle_2} + e^{-s\sqrt{n}\langle g - \Gamma^{L_n}, \psi_{lk}\rangle_2} \big| X^{(n)}\big]\Bigg] \\
&\quad \lesssim \sum_{l \le L_n} 2^{l/2} \frac{1}{\sqrt{n}s} \log\big[2^l e^{C(s+s^2)}\big] + \sum_{l \le L_n} 2^{l/2} \frac{1}{\sqrt{n}s} \log \frac{1}{\Pi(D_n \mid X^{(n)})}.
\end{aligned}
$$

Set $s = \sqrt{l}$. The first term in the last display is bounded by a constant times $\frac{1}{\sqrt{n}} \sum_{l \le L_n} \sqrt{l} 2^{l/2} \lesssim \varepsilon^*_{n,\alpha}$. Now coming back to the application of Markov's inequality one gets, with Ξ the function $\Xi : u \to u \log u^{-1}$,

$$
E^n_{f_0} \Pi\big[\|g - g_0\|_\infty > M_n \varepsilon^*_{n,\alpha} \mid X^{(n)}\big] \lesssim M_n^{-1} + M_n^{-1} E^n_{f_0} \Xi\big(\Pi(D_n \mid X^{(n)})\big) + o(1).
$$

With $M_n \to \infty$, the fact that Ξ is bounded on [0, 1] completes the proof.

In the case $1/2 < \alpha \le 1$, the only difference is that one gets two extra terms: one from going at the logarithmic level, which eventually leads to a rate ζ_n^2; another one from the semiparametric bias in (24), which leads to a rate $\rho_n = n^{(1/2 - 3\alpha/2)/(1+2\alpha)}$. This leads to a sup-norm rate of $\zeta_n^2 \vee \rho_n = \zeta_n^2$.

Once the rate ζ_n^2 has been obtained, one can restart the proof once again, but this time knowing that one can use a better intermediate rate of ζ_n^2 in sup-norm. One can then write

$$
\left|\int \rho\left(\frac{f - f_0}{f_0}\right)\psi_{lk}\right| \le \left\|\rho\left(\frac{f - f_0}{f_0}\right)\right\|_\infty \|\psi_{lk}\|_1 \lesssim \left\|\frac{f - f_0}{f_0}\right\|_\infty^2 2^{-l/2} \lesssim (\zeta_n^2)^2 2^{-l/2}.
$$

This eventually leads to the accelerated rate $(\zeta_n^2)^2 \vee \rho_n$. Iterating this procedure leads to $(\zeta_n^2)^{2^p} \vee \rho_n$, any $p \ge 1$ and, for any given $\alpha > 1/2$, to ρ_n as final rate, up to logarithmic terms.

4.4. *Density estimation, dyadic histogram priors.* We follow the scheme of proof used for uniform priors in white noise, this time in density estimation, with the wavelet basis of expansion being the Haar system. The very specific properties of the Haar basis, particularly its close links to approximation by dyadic histograms, enable a simplified argument. In particular, as we demonstrate below, the semiparametric bias is always negligible, provided the parameters of the prior are reasonably chosen.

Let us set

$$
\hat{f}_{lk} = \mathbb{P}_n \psi^H_{lk} = \frac{1}{n} \sum_{i=1}^n \psi^H_{lk}(X_i).
$$

Set $D_n = \{f, h(f, f_0) \le \varepsilon_n\}$, where here $\varepsilon_n = \varepsilon^*_{n,\alpha}$ up to multiplication by a large enough constant. Lemma 10 implies that $E^n_{f_0}\Pi[D_n|X^{(n)}] \to 1$. Let h_n, L_n be defined as in (3), and denote by f^{L_n} the projection of f onto the subspace $\mathcal{V}_{L_n} := \mathrm{Vect}\{\varphi^H, \psi^H_{lk}, l < L_n, 0 \le k < 2^l\}$ and $f^{L_n^c}$ the projection of f onto $\mathrm{Vect}\{\psi^H_{lk}, l \ge L_n, 0 \le k < 2^l\}$ (i.e., for simplicity we keep the notation f^{L_n} from Section 2.1, although the basis of projection is now the Haar basis *and* $l < L_n$ replaces $l \le L_n$). If $\hat{f}^{L_n}$ denotes the element of $\mathcal{V}_{L_n}$ of coordinates $\{\hat{f}_{lk}\}$ in the basis $\{\psi^H_{lk}\}$, one can bound $E^{\Pi^{D_n}}[\|f - f_0\|_\infty | X^{(n)}]$ from above by

$$\underbrace{\int \|f^{L_n} - \hat{f}^{L_n}\|_\infty \, d\Pi^{D_n}(f|X^{(n)})}_{\text{(i)}} + \underbrace{\int \|f^{L_n^c}\|_\infty \, d\Pi(f|X^{(n)})}_{\text{(ii)}} + \underbrace{\|\hat{f}^{L_n} - f_0\|_\infty}_{\text{(iii)}}.$$

The term (iii) can be bounded by $\|\hat{f}^{L_n} - f_0^{L_n}\|_\infty + \|f_0^{L_n^c}\|_\infty$. The second term in this sum is pure bias and the first term is

$$\left\| (\mathbb{P}_n\varphi^H - P_{f_0}\varphi^H)\varphi^H(\cdot) + \sum_{l=0}^{L_n-1}\sum_{k=0}^{2^l-1} (\mathbb{P}_n\psi^H_{lk} - P_{f_0}\psi^H_{lk})\psi^H_{lk}(\cdot) \right\|_\infty.$$

This term is bounded in expectation by $\varepsilon^*_{n,\alpha}$, exactly as in Lemma 7.

Next, the high-frequency bias term (ii) is zero. Indeed, for any draw f from the prior, in the inner-product $\langle f, \psi^H_{lk}\rangle_2$, the first element is a dyadic histogram at resolution level L_n, so is constant over the support of the Haar basis element ψ^H_{lk} if $l \ge L_n$. Hence, the previous inner-product is zero Π-almost surely, and thus also $\Pi[\cdot|X^{(n)}]$ almost surely.

Second, one studies $\langle f^{L_n} - \hat{f}^{L_n}, \psi^H_{lk}\rangle_2$ in the BvM-regime $l < L_n$. Following the maximal inequality approach from Section 2.1, it is enough to bound the posterior expectation of $\exp(t\sqrt{n}\langle f^{L_n} - \hat{f}^{L_n}, \psi^H_{lk}\rangle_2)$, for any possible k and $l < L_n$ and say $|t| \lesssim \log n$ (also, to simplify the notation below we omit mentioning the scaling function φ^H, but the same Laplace transform control is obtained for it in a similar way). To do so, apply Lemma 3 with $\gamma_n = \psi^H_{lk}$, for any given l, k with $l < L_n$. The conditions of the lemma are satisfied with $a_n = \varepsilon_n$, since ψ^H_{lk} is bounded in $L^2[0,1]$ and has a sup-norm bounded by a constant times $2^{l/2} \lesssim 2^{L_n/2} = o(1/\varepsilon_n)$ if $\alpha > 1/2$. Noting that, again for l, k with $l < L_n$,

$$\begin{aligned} \langle f^{L_n} - \hat{f}^{L_n}, \psi^H_{lk}\rangle_2 &= \langle f^{L_n} - f_0^{L_n}, \psi^H_{lk}\rangle_2 + \langle f_0^{L_n} - \hat{f}^{L_n}, \psi^H_{lk}\rangle_2 \\ &= \langle f - f_0, \psi^H_{lk}\rangle_2 - W_n(\psi^H_{lk}), \end{aligned}$$

an application of Lemma 3 leads to, with Π^{D_n} the restriction of Π to D_n,

$$E^{\Pi^{D_n}}\left[e^{t\sqrt{n}\langle f^{L_n} - \hat{f}^{L_n}, \psi^H_{lk}\rangle_2} | X^{(n)}\right] \lesssim e^{Ct^2} \frac{\int e^{\ell_n(f_t) - \ell_n(f_0)}\, d\Pi^{D_n}(f)}{\int e^{\ell_n(f) - \ell_n(f_0)}\, d\Pi^{D_n}(f)},$$

where f_t is defined by $\log f_t = \log f - t\psi_{lk}^H/\sqrt{n} - c(fe^{-t\psi_{lk}^H/\sqrt{n}})$ (again, having $\tilde{\psi}_{lk}^H$ or ψ_{lk}^H at both places in the last equality does not matter since the constant simplifies). Below to simplify the notation, we denote $\gamma_n = \psi_{lk}^H$.

In the last display, once coming back to Π via $d\Pi^{D_n}(f) = \mathbb{1}_{D_n} d\Pi(f)/\Pi(D_n)$, the variable f is a random dyadic histogram over the subdivision with intervals $I_\mu^{L_n} = (\mu 2^{-L_n}, (\mu+1)2^{-L_n})$ and $0 \le \mu \le 2^{L_n} - 1$. Denote by $\gamma_{n,\mu}$ the value of γ_n over the interval $I_\mu^{L_n}$. Next, observe that both f_t and f are, writing the histogram prior in terms of its coefficients over the subdivision, functions of $\omega = (\omega_0, \dots, \omega_{2^{L_n}-1})$ and the integral over f is nothing but an integral over $\omega \in \mathcal{S}_{L_n}$. On the other hand, from the expression of f_t, using the fact that ψ_{lk}^H is constant over each individual interval I_μ (since $l < L_n$), one sees that f_t is a dyadic histogram over I_μ with weights given by the vector ζ

$$\zeta := (\zeta_\mu)_{0 \le \mu \le 2^{L_n}-1} = \left(\frac{\omega_\mu e^{-t\gamma_{n,\mu}/\sqrt{n}}}{\sum_\mu \omega_\mu e^{-t\gamma_{n,\mu}/\sqrt{n}}} \right)_{0 \le \mu \le 2^{L_n}-1}. \tag{33}$$

Now we are in position to change variables in the above expression by taking ζ as the new variable (this technique is developed in [8] for general, *fixed* influence functions). The change of variables introduces a multiplicative factor $M(\zeta)$ in front of $d\Pi(\zeta)$, factor which is the product of the variation in density of the Dirichlet law under the change of variable and the Jacobian of the change of variable,

$$\frac{\int e^{\ell_n(f_t)-\ell_n(f_0)}\, d\Pi^{D_n}(f)}{\int e^{\ell_n(f)-\ell_n(f_0)}\, d\Pi^{D_n}(f)} = \Pi[D_n | X^{(n)}]^{-1} \frac{\int_{\tilde{D}_n} e^{\ell_n(f(\zeta))-\ell_n(f_0)} M(\zeta)\, d\mathcal{D}_\alpha(\zeta)}{\int e^{\ell_n(f(\omega))-\ell_n(f_0)}\, d\mathcal{D}_\alpha(\omega)},$$

where $\tilde{D}_n$ is the new integrating set after change of variables and the notation $f(\zeta)$ is used for $f(\zeta)(\cdot) = 2^L \sum_{\mu=0}^{2^L-1} \zeta_\mu \mathbb{1}_{I_k^L}(\cdot)$ and similarly for $f(\omega)$. Computation of the Jacobian say $\Delta(\zeta)$ is done in Lemma 11. Calculating $M(\zeta) = d\mathcal{D}_\alpha(\omega)/d\mathcal{D}_\alpha(\zeta)\Delta(\zeta)$ gives that $M(\zeta)$ satisfies

$$\begin{aligned} M(\zeta) e^{-t/\sqrt{n} \sum_\mu \alpha_\mu \gamma_{n,\mu}} &= \left[\int_0^1 e^{t\gamma_n(x)/\sqrt{n}} f(\zeta)(x)\, dx \right]^{-\sum_\mu \alpha_\mu} \\ &= \left[\int_0^1 e^{-t\gamma_n(x)/\sqrt{n}} f(\omega)(x)\, dx \right]^{\sum_\mu \alpha_\mu}. \end{aligned}$$

For the term on the right-hand side, since $\alpha > 1/2$ it holds $|t| \|\gamma_n\|_\infty/\sqrt{n} = o(1)$ so one can expand the exponential function by writing $e^u = 1 + O(u)$ as $u = o(1)$. Next, write $f = f_0 + (f - f_0)$, so that the expression under brackets writes $1 + O((t/\sqrt{n})[\langle f_0, \gamma_n \rangle_2 + \langle f - f_0, \gamma_n \rangle_2])$. The last term is a $O(|t|/\sqrt{n})$, since the Haar-coefficients of f_0 are certainly bounded and those of $f - f_0$ are bounded above by a constant times $\|\gamma_n\|_\infty h(f, f_0)$ which is bounded because $2^{L_n/2}\varepsilon_n = O(1)$. So if $\sum_\mu \alpha_\mu/\sqrt{n} = O(1)$, the term at stake is $O(1)$. But this condition follows from (16), because the number of terms in the previous sum is $2^{L_n} = O(\sqrt{n})$ when $\alpha > 1/2$.

Now we deal with the exponential term on the left-hand side of the last display. A term in the sum $\sum_\mu \alpha_\mu \gamma_{n,\mu}$ is nonzero only if the support of the Haar basis element $\gamma_n = \psi_{lk}^H$ (or φ^H) intersects $I_\mu^{L_n}$. This is the case for $2^{L_n - l}$ terms for ψ_{lk}^H and 2^{L_n} terms for φ^H. Using $\|\psi_{lk}^H\|_\infty \lesssim 2^{l/2}$, this shows that the considered sum is at most a constant times $2^{L_n - l/2} \lesssim 2^{L_n}$. In particular, for $\alpha > 1/2$ the considered term is a $O(1)$.

The previous reasoning shows that the change of variable part generates a multiplicative factor $O(1)$, times the term e^{Ct^2} coming from the application of Lemma 3. To conclude, it now suffices to apply the maximal inequality technique in the same way as for log-densities prior.

4.5. *Tools for density estimation.* The notation here follows the one introduced in Section 4.2, in particular $\|\cdot\|_L$, W_n, R_n and $\mathcal{B}$.

LEMMA 3. *Let f_0 belong to $\mathcal{F}_0$. Let $\{a_n\}$ be a sequence of real numbers with $na_n^2 \geq 1$, any $n \geq 1$. Let $\{\Pi_n\}$ be a collection of priors on densities restricted to the set $\{f, h(f, f_0) \leq a_n\}$. Let $\{\gamma_n\}$ be an arbitrary sequence in $L^\infty[0,1]$. Set $\tilde{\gamma}_n := \gamma_n - P_{f_0}\gamma_n$. Suppose, for some $m > 0$ and all $n \geq 1$,*

$$\|\tilde{\gamma}_n\|_L \leq m, \qquad \|\tilde{\gamma}_n\|_\infty \leq \left(4a_n \log(n+1)\right)^{-1}.$$

Then there exists $C > 0$ depending on m, $\|f_0\|_\infty$ only such that for any $n \geq 1$ and any $|t| \leq \log n$,

$$E^{\Pi_n}\left[e^{t\sqrt{n}\langle f - f_0, \gamma_n\rangle_2} | X^{(n)}\right] \leq e^{Ct^2 + tW_n(\gamma_n)} \frac{\int e^{\ell_n(f_t) - \ell_n(f_0)}\, d\Pi_n(f)}{\int e^{\ell_n(f) - \ell_n(f_0)}\, d\Pi_n(f)},$$

where f_t is defined by $\log f_t = \log f - t\tilde{\gamma}_n/\sqrt{n} - c(fe^{-t\tilde{\gamma}_n/\sqrt{n}})$.

PROOF. Denote $g = \log f$, $g_0 = \log f_0$. From elementary algebra, it follows that

$$\begin{aligned}
&t\sqrt{n}\langle f - f_0, \gamma_n\rangle_2 + \ell_n(f) - \ell_n(f_0) \\
&\quad = -\frac{n}{2}\left\|g - g_0 - \frac{t}{\sqrt{n}}\tilde{\gamma}_n\right\|_L^2 + \sqrt{n}W_n\left(g - g_0 - \frac{t}{\sqrt{n}}\tilde{\gamma}_n\right) \\
&\qquad + t\sqrt{n}\mathcal{B}(\tilde{\gamma}_n, f, f_0) + R_n(f, f_0) + \frac{t^2}{2}\|\tilde{\gamma}_n\|_L^2 + tW_n(\tilde{\gamma}_n) \\
&\quad = \ell_n(f_t) - \ell_n(f_0) + \left[t\sqrt{n}\mathcal{B}(\tilde{\gamma}_n, f, f_0) + R_n(f, f_0) - R_n(f_t, f_0)\right] \\
&\qquad + \frac{t^2}{2}\|\tilde{\gamma}_n\|_L^2 + tW_n(\tilde{\gamma}_n).
\end{aligned}$$

Let us show that the bracketed term is small. From the definition of R_n, f_t,

$$R_n(f, f_0) - R_n(f_t, f_0)$$
$$= -\frac{t^2}{2}\|\tilde{\gamma}_n\|_L^2 + t\sqrt{n}\langle g - g_0, \tilde{\gamma}_n\rangle_L + n\log F\big[e^{-t/\sqrt{n}\tilde{\gamma}_n}\big].$$

Next, expand the last logarithmic term. Using the assumption on $\|\tilde{\gamma}_n\|_\infty$ and $|t| \le \log n$, the absolute value of the exponent in this term is at most $1/4$. This enables us to expand successively the logarithm and exponential functions, using the inequalities (the first is valid for $|x| \le 1/4$),

$$e^{-x} \le 1 - x + x^2, \qquad \log(1+x) \le x,$$

leading to

$$\begin{aligned}\log F\big[e^{-t/\sqrt{n}\tilde{\gamma}_n}\big] &\le \log F\left[1 - \frac{t}{\sqrt{n}}\tilde{\gamma}_n + \frac{t^2}{n}\tilde{\gamma}_n\right]\\ &\le \log\left[1 - \frac{t}{\sqrt{n}}F\tilde{\gamma}_n + \frac{t^2}{n}F\tilde{\gamma}_n^2\right]\\ &\le -\frac{t}{\sqrt{n}}F\tilde{\gamma}_n + \frac{t^2}{n}F_0\tilde{\gamma}_n^2 + \frac{t^2}{n}(F - F_0)\tilde{\gamma}_n^2.\end{aligned}$$

The last term, using Lemma 6 and $h(f, f_0)\|\tilde{\gamma}_n\|_\infty \le 1$ together with $\|\tilde{\gamma}_n\|_2^2 \lesssim \|\tilde{\gamma}_n\|_L^2 \lesssim m$, is a $O(t^2/n)$. On the other hand,

$$F\tilde{\gamma}_n = (F - F_0)\tilde{\gamma}_n = \left\langle \frac{f - f_0}{f_0}, \tilde{\gamma}_n\right\rangle_L = \langle g - g_0, \tilde{\gamma}_n\rangle_L + \mathcal{B}(\tilde{\gamma}_n, f, f_0).$$

Combine the previous results to obtain the desired bound. $\square$

LEMMA 4. *Consider the log-density prior* (12). *Suppose* $g_0 = \log f_0$ *belongs to* $\mathcal{C}^\alpha$, *with* $\alpha > 1/2$. *Suppose* (15) *holds and let* ε_n, ζ_n *be defined as below* (15). *Then* M *large enough*,

$$E_{f_0}^n \Pi\big[f : \|f - f_0\|_2 \le M\varepsilon_n, \|f - f_0\|_\infty \le M\zeta_n | X^{(n)}\big] \to 1.$$

PROOF. Obtaining this result could be done following the arguments in [16]. Here we use instead an approach from [27], extending their argument on the second moment of $\log(f/f_0)$ to further get a rate ζ_n for $\|\log(f/f_0)\|_\infty$.

Since (15) is assumed, one can restrict to the event $\{f : h(f, f_0) \le \varepsilon_n\} \subset \{f : \|f - f_0\|_2 \lesssim \varepsilon_n\}$. We have (see, e.g., Lemma 8 in [14]),

$$\|\log(f/f_0)\|_2^2 \lesssim h^2(f, f_0)\big(1 + \log\|f/f_0\|_\infty\big).$$

The last term is bounded by a constant times $\varepsilon_n^2(1+\log\|T-c(T)-g_0\|_\infty)$ by assumption, where $g_0=\log f_0$. Next, one writes

$$\begin{aligned}\|T-c(T)-g_0\|_\infty &= \Big\|\sum_{l,k}\langle T-c(T)-g_0,\psi_{lk}\rangle_2\psi_{lk}\Big\|_\infty\\ &\le \sum_{l\le L_n}2^{l/2}\max_k|\langle T-c(T)-g_0,\psi_{lk}\rangle_2| + \sum_{l>L_n}2^{l/2}\max_k|\langle g_0,\psi_{lk}\rangle_2|.\end{aligned}$$

Since g_0 is α-Hölder, the last term is of the order $\varepsilon_{n,\alpha}^*=o(1)$. For the middle term, Cauchy–Schwarz inequality yields the bound $2^{L_n/2}\|T-c(T)-g_0\|_2$, using $\sum_{l\le L_n}2^l\lesssim 2^{L_n}$ and bounding the maximum of squares by the sum. Deduce that

$$\|\log(f/f_0)\|_2^2\lesssim\varepsilon_n^2+\varepsilon_n^2 2^{L_n/2}\|\log(f/f_0)\|_2\lesssim\varepsilon_n^2+\varepsilon_n^2 2^{L_n}\|\log(f/f_0)\|_2^2.$$

Since $\alpha>1/2$, one has $\varepsilon_n^2 2^{L_n}=o(1)$. So gathering the L^2-norm terms on the same side of the inequality one obtains $\|\log(f/f_0)\|_2^2\lesssim\varepsilon_n^2$. Also, along the way we have obtained the bound

$$\|T-c(T)-g_0\|_\infty\lesssim 2^{L_n/2}\|T-c(T)-g_0\|_2+\varepsilon_{n,\alpha}^*\lesssim 2^{L_n/2}\varepsilon_n=\zeta_n.$$

Now the squared L^2-norm of $f-f_0$ can be expressed as

$$\int_0^1(f-f_0)^2=\int_0^1 f_0^2\big(e^{T-c(T)-g_0}-1\big)^2.$$

The inequality $|e^x-1|\le C|x|$, valid for x in a compact subset of $\mathbb{R}$ and C a large enough constant, implies

$$\int_0^1(f-f_0)^2\le C^2\int_0^1 f_0^2\big(T-c(T)-g_0\big)^2\lesssim\|T-c(T)-g_0\|_2^2\lesssim\varepsilon_n^2.$$

Similarly, since $\|f_0\|_\infty<\infty$, one obtains $\|f-f_0\|_\infty\lesssim\zeta_n$. □

4.6. *Other lemmas.* Given $R>0$, let $B_{\infty,\infty}^\alpha(R)$ denote the centered ball of $B_{\infty,\infty}^\alpha[0,1]$ of radius R for the norm $\|\cdot\|_{\infty,\infty,\alpha}$ given in Section 2.2.

LEMMA 5. *There exists a constant $C>0$ such that for any f,g in $B_{\infty,\infty}^\alpha(R_f)$ and $B_{\infty,\infty}^\alpha(R_g)$, respectively, the product fg belongs to $B_{\infty,\infty}^\alpha(CR_fR_g)$. If f belongs to $\mathcal{C}^\alpha[0,1]$ and is bounded away from 0 then f^{-1} belongs to $\mathcal{C}^\alpha[0,1]$. Moreover, for any indexes l,k,*

$$\|\psi_{lk}\|_{\infty,\infty,\alpha}=2^{l(1/2+\alpha)}.$$

PROOF. The first claim follows from the main result of Section 2.8.3 in [36] (strictly speaking the last result is for functions on $\mathbb{R}$, but the latter functions can be shown to be restrictions to $[0,1]$ of elements of $B^{\alpha}_{\infty,\infty}$ whose norm is equivalent to the one of the restriction; see [26] Proposition 2 for a similar argument for Sobolev spaces). The second claim is a simple computation using the definition of Hölder spaces. For the last claim, one uses the characterisation of $B^{\alpha}_{\infty,\infty}$ in terms of wavelet coefficients from Section 2.2, which yields $\|\psi_{lk}\|_{\infty,\infty,\alpha} = \max_{l',k'} 2^{l'(1/2+\alpha)}\langle\psi_{lk},\psi_{l'k'}\rangle_2 = 2^{l(1/2+\alpha)}$. □

LEMMA 6. *Let f, f_0 be two densities on $[0,1]$ such that f_0 is bounded. For any $g \in L^2[0,1]$ such that $h(f,f_0)\|g\|_\infty \le C_1$ and $\|g\|_2 \le C_2$, for some constants $C_1, C_2 > 0$,*

$$\left|(F-F_0)g^2\right| \le C_1^2 + C_1\sqrt{4C_2\|f_0\|_\infty + C_1^2}.$$

PROOF. Denote $\Sigma := \int_0^1 |f-f_0|g^2$. Then by Cauchy–Schwarz inequality,

$$\begin{aligned}\Sigma^2 &\le 2h(f,f_0)^2\int_0^1 (f+f_0)g^4\\ &\le 2h(f,f_0)^2\left[\|g\|_\infty^2\Sigma + 2\|f_0\|_\infty\|g\|_\infty^2\|g\|_2^2\right]\\ &\le 2C_1^2\left[\Sigma + 2C_2\|f_0\|_\infty\right].\end{aligned}$$

This implies that Σ is less than the largest root of the polynomial $X^2 - 2C_1X + 4C_1^2\|f_0\|_\infty$, which can be expressed in terms of C_1, $\|f_0\|_\infty$. □

LEMMA 7. *Let $f_0 \in \mathcal{F}_0$ and $g_0 = \log f_0$, and let Γ^{L_n} be defined by (20), with $\mathcal{A}_{l,k}$ any elements of $L^\infty[0,1]$ such that there exists constants c_1, c_2 with, for any l, k with $k < 2^l \le 2^{L_n}$, any $n \ge 2$,*

$$\|\mathcal{A}_{l,k}\|_\infty \le c_1\sqrt{n/\log n}, \qquad \|\mathcal{A}_{l,k}\|_2 \le c_2.$$

Then for any $n \ge 2$ and L_n defined by (3), it holds

$$E^n_{f_0}\|\Gamma^{L_n} - g_0^{L_n}\|_\infty \lesssim \varepsilon^*_{n,\alpha}.$$

PROOF. We proceed exactly as for the proof of the maximal inequality in Section 2.1. For any $t > 0$,

$$E^n_{f_0}\|\Gamma^{L_n} - g_0^{L_n}\|_\infty \le \frac{1}{\sqrt{n}}\sum_{l\le L_n}\frac{2^{l/2}}{t}\log\sum_{k=0}^{2^l-1}E^n_{f_0}\left[e^{tW_n(\mathcal{A}_{l,k})} + e^{-tW_n(\mathcal{A}_{l,k})}\right].$$

We have $W_n(\mathcal{A}_{l,k}) = \mathbb{G}_n(\mathcal{A}_{l,k})$ and bounds on exponential moments of the last empirical quantity are well known. From Laplace transform controls, one gets, for any real s,

$$E^n_{f_0}\left[e^{sW_n(\mathcal{A}_{l,k})}\right] \le e^{(s^2/2)[\int \mathcal{A}_{l,k}^2 f_0]e^{|s|\|\mathcal{A}_{l,k}\|_\infty/\sqrt{n}}}.$$

Let us choose $t = \sqrt{l} \lesssim \sqrt{L_n}$. Under the conditions of the lemma the last display with $s = t$ or $s = -t$ is bounded above by e^{Ct^2}. This leads to the bound $E^n_{f_0} \|\Gamma^{L_n} - g_0^{L_n}\|_\infty \lesssim \varepsilon^*_{n,\alpha}$. $\square$

LEMMA 8. *Let $\varphi = \varphi_G$ and $\sigma_l = 2^{-l(1/2+\gamma)}$ for all $l \le L_n$, and any given $0 < \gamma \le \alpha - 1/4$. Then (14)–(15) hold. The same applies for $\varphi = \varphi_{H,\tau}$ and $\sigma_l = 2^{-l\alpha}$ for all $l \le L_n$, for any given value of the parameter $0 \le \tau < 1$.*

PROOF. The first result is a minor adaptation of Theorem 4.5 in [38], where the authors consider a cut-off at $n^{1/(2\alpha+1)}$ instead of $2^{L_n} = h_n^{-1} = (n/\log n)^{1/(2\alpha+1)}$ (equality up to fixed multiplicative constants). Taking the cut-off at 2^{L_n} only changes logarithmic factors in their argument. In the log-Lipschitz case, one adapts Theorem 2.1 in [27]. There are two points to note. First, taking 2^{L_n} instead of $n^{1/(2\alpha+1)}$ induces only, again, an extra logarithmic power in the rate. Second, strictly speaking the authors in [28] consider wavelets on the interval via periodisation, which imposes conditions at the boundary (periodic Besov spaces), conditions which can be dropped when using the CDV wavelet basis. Explicit (re)derivation of the previous two results is omitted. $\square$

LEMMA 9. *Let $\varphi = \varphi_G$ and σ_l satisfy (14). Then the prior Π defined by (12), with L_n as in (3) and ε_n as below (15), satisfies, for $C > 0$ large enough and any fixed given integer K,*

$$E^n_{f_0} \Pi\Big[\max_{\lambda \le K,\mu} |\langle T, \psi_{\lambda\mu}\rangle| \le C\sqrt{n}\varepsilon_n \big| X^{(n)}\Big] \to 1.$$

PROOF. The maximum in the display of the lemma only involves a finite number of terms λ, μ with $\lambda \le K$, $\mu \le 2^K - 1$, and these terms are Gaussian with variances σ_l^2 bounded above by positive constants. Thus, by Gaussian concentration one gets

$$\Pi\Big[\max_{\lambda \le K,\mu} |\langle T, \psi_{\lambda\mu}\rangle| > C\sqrt{n}\varepsilon_n\Big] \le e^{-cn\varepsilon_n^2},$$

where c can be made arbitrarily large by taking C large enough. Next, one applies Lemma 1 in [15]. To do so, one needs to bound from below the prior probability of a Kullback–Leibler neighborhood of f_0 of size ε_n by $e^{-dn\varepsilon_n^2}$ for some $d > 0$. This follows from the conclusion of Theorem 5 in [38], which (modulo the fact that our ε_n is within a logarithmic factor of theirs, as noted in Lemma 8 to accommodate our slightly different choice of cut-off 2^{L_n}), provides the bound $\Pi[\|g - g_0\|_\infty < 4\varepsilon_n) \ge e^{-n\varepsilon_n^2}$. Switching from the sup-norm on $g - g_0$ to the Kullback–Leibler divergence between g and g_0 follows from Lemma 3.1 in [38]. $\square$

LEMMA 10. *Suppose f_0 belongs to $\mathcal{F}_0 \cap \mathcal{C}^\alpha[0,1]$, for $0 < \alpha \le 1$. Let Π be a prior on histogram densities defined by* (16). *Then, for M large enough*

$$E_{f_0}^n \Pi[f, h(f, f_0) \le M(\log n/n)^{-\alpha/(2\alpha+1)} | X^{(n)}] \to 1.$$

PROOF. One can proceed as in the proof of Theorem 2 in [30]. It suffices to use as sieve the set of dyadic density histograms $\mathcal{H}_{L_n}$, so that $\Pi[\mathcal{H}_{L_n}^c] = 0$. Entropy and prior mass conditions can then be verified with the same rate $(\log n/n)^{\alpha/(2\alpha+1)}$; see also the remark before Theorem 2.4 in [12]. $\square$

LEMMA 11. *Let $\Delta(\zeta)$ be the Jacobian of the change of variables $\omega \to \zeta$ given by* (33) *over the unit simplex $\mathcal{S}_{2^{L_n}}$. It holds, with $f(\zeta) = 2^L \sum_{\mu=0}^{2^L-1} \zeta_\mu \mathbb{1}_{I_k^L}$,*

$$\Delta(\zeta) = \prod_{\mu=0}^{2^{L_n}-1} \frac{e^{t\gamma_{n,\mu}/\sqrt{n}}}{\int_0^1 e^{t\gamma_n(x)/\sqrt{n}} f(\zeta)(x)\, dx}.$$

PROOF. This follows from elementary calulations; see [8]. $\square$

Acknowledgements. This work started from ideas developed in [6] in common with Richard Nickl. His numerous helpful comments are warmly acknowledged. I am also indebted to Judith Rousseau for early explanations on semiparametric analysis of histograms priors. Although the treatment for histograms here is a bit different, it owes to her ideas, and we refer to [8] for a systematic treatment of histograms for fixed functionals. I would like to thank Catia Scricciolo for early discussions and Gérard Kerkyacharian and Dominique Picard for helpful comments, as well as the referees and the Associate Editor for insightful and constructive comments.

REFERENCES

[1] BARRON, A. (1988). Convergence of Bayes estimators. Technical Report 7, Univ. Illinois at Urbana–Champaign.

[2] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[3] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287

[4] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753

[5] CASTILLO, I. (2012). Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya* **74** 194–221. MR3021557

[6] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856

[7] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969.

[8] CASTILLO, I. and ROUSSEAU, J. (2013). A general Bernstein–von Mises theorem in semiparametric models. Preprint. Available at arXiv:1305.4482.

[9] CHOY, S. T. B. and SMITH, A. F. M. (1997). On robust analysis of a normal location parameter. *J. Roy. Statist. Soc. Ser. B* **59** 463–474. MR1440591

[10] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. MR1256527

[11] GASPARINI, M. (1996). Bayesian density estimation via Dirichlet density processes. *J. Nonparametr. Stat.* **6** 355–366. MR1386345

[12] GHOSAL, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29** 1264–1280. MR1873330

[13] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[14] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[15] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[16] GINÉ, E. and NICKL, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \le r \le \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395

[17] GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields* **159** 479–543. MR3230001

[18] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. Springer, New York. MR1618204

[19] HAS'MINSKIĬ, R. Z. (1978). A lower bound for risks of nonparametric density estimates in the uniform metric. *Teor. Veroyatn. Primen.* **23** 824–828. MR0516279

[20] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2013). On adaptive posterior concentration rates. Preprint. Available at arXiv:1305.5270.

[21] IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **98** 61–85, 161–162, 166. MR0591862

[22] LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235. MR1186248

[23] LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83** 509–516. MR0971380

[24] LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60** 297–308. MR0326902

[25] NADARAJAH, S. (2003). The Kotz-type distribution with applications. *Statistics* **37** 341–358. MR1997185

[26] NICKL, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields* **138** 411–449. MR2299714

[27] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033

[28] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. MR2934953

[29] SCRICCIOLO, C. (2006). Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.* **34** 2897–2920. MR2329472

[30] SCRICCIOLO, C. (2007). On rates of convergence for Bayesian density estimation. *Scand. J. Stat.* **34** 626–642. MR2368802

[31] SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.* **9** 475–520.

[32] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[33] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[34] TOKDAR, S. T. (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *J. Comput. Graph. Statist.* **16** 633–655. MR2351083

[35] TOKDAR, S. T. and GHOSH, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *J. Statist. Plann. Inference* **137** 34–42. MR2292838

[36] TRIEBEL, H. (1983). *Theory of Function Spaces. Monographs in Mathematics* **78**. Birkhäuser, Basel. MR0781540

[37] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[38] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[39] WALKER, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043. MR2102501

[40] WALKER, S. G. and GUTIÉRREZ-PEÑA, E. (1999). Robustifying Bayesian procedures. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) 685–710. Oxford Univ. Press, New York. MR1724876

CNRS—LABORATOIRE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UNIVERSITÉS PARIS VI & VII
BÂTIMENT SOPHIE GERMAIN
75205 PARIS CEDEX 13
FRANCE
E-MAIL: ismael.castillo@upmc.fr