Pólya tree posterior distributions on densities

Ismaël Castillo

Laboratoire de Probabilités et Modèles Aléatoires (LPMA) UMR 7599, Universités Paris 6 et 7, France

ismael.castillo@upmc.fr

Abstract

Pólya trees form a popular class of prior distributions used in Bayesian nonparametrics. For some choice of parameters, Pólya trees are prior distributions on density functions. In this paper we carry out a frequentist analysis of the induced posterior distributions. We investigate the contraction rate of Pólya tree posterior densities in terms of the supremum loss and study the limiting shape distribution. A nonparametric Bernstein–von Mises theorem is established, as well as a Bayesian Donsker theorem for the posterior cumulative distribution function.

1 Introduction

Pólya trees are a class of random probability distributions that are commonly used as prior distributions in the Bayesian nonparametrics study of infinite-dimensional statistical models. The name 'Pólya tree' appears in works by Mauldin et al. [34] and Lavine [30] and it has been used since then, although the object and related ones such as tail-free processes already appear in works by Freedman [13]-[14], Kraft [29], and Ferguson [12]. It should be noted that the name 'Pólya tree' is also used for a *different* object, not considered in the present work, in the literature on trees, where it refers to a rooted unordered tree. The origin of the name in the statistical literature comes from a beautiful connection with Pólya urns, themselves named after the 1930 article [37] by George Pólya in *Annales de l'Institut Henri Poincaré*. It was indeed shown in [34] that Pólya trees are de Finetti measures of certain exchangeable sampling schemes defined from a tree of Pólya urns.

In Bayesian nonparametric statistics, the starting point is the construction of a prior distribution, a probability measure that 'samples at random' the parameter to be estimated. If the parameter is itself a distribution, one needs to build a 'distribution on distributions'. A popular distribution on probability measures is the Dirichlet process introduced by Ferguson [11], see [15] for a review of its use in the statistics literature. The Dirichlet process, as we recall below, is actually a special case of Pólya tree for certain choices of parameters. However, the resulting measure is not directly suited for modelling a smooth

object such as a density function, as draws from the Dirichlet process are discrete almost surely. On the contrary, different choices of parameters of the Pólya tree lead to a probability measure that is absolutely continuous with respect to Lebesgue measure, and hence admits a density. In this paper we focus on this type of 'density Pólya trees'.

Let $X = X^{(n)} = (X_1, \ldots, X_n)$ be a sample from an unknown distribution P on the interval $[0,1]$, and suppose that P admits a density f with respect to Lebesgue measure, denoting $P = P_f$. Following a Bayesian approach, we view P as random and put an a priori distribution Π on P, equal to a Pólya tree distribution. The data $X_1, \ldots, X_n$ are viewed as, given P, i.i.d. from P. From this one forms the posterior distribution, the conditional law $P \,|\, X_1, \ldots, X_n$, that we denote $\Pi[\cdot \,|\, X_1, \ldots, X_n] = \Pi[\cdot \,|\, X]$. To study the convergence of this random (it depends on the data) distribution as $n \to \infty$, we undertake a so-called frequentist analysis of the Bayesian procedure: we assume that the data has actually been generated i.i.d. from a fixed distribution $P_0 = P_{f_0}$, for some density f_0 on $[0,1]$, and are interested in the convergence of the posterior $\Pi[\cdot \,|\, X^{(n)}]$ in probability under P_{f_0}, as $n \to \infty$. A natural question is: does the posterior $\Pi[\cdot \,|\, X^{(n)}]$ converge to δ_{P_0}, a Dirac mass at the 'true' distribution? If so, this is the so-called consistency property of the posterior at P_0. Further, what can be said about the rate of convergence, and, perhaps, the form of the limit after rescaling?

General conditions for consistency of posterior distributions were given in Schwartz [40], and the theory was further developed among others in [1]. For Pólya trees, posterior consistency in density estimation in the weak topology follows from results in [31], see also [16], and consistency in the Hellinger topology was obtained in [1], whose conditions where further refined in [43]. A next natural step once consistency is obtained is to investigate the convergence rate of the posterior distribution. This has been the object of much attention in the last 15 years, with fundamental contributions such as [17], [41], [18], where general sufficient conditions on model and prior are given ensuring posterior convergence towards the true distribution at some rate.

Yet, to the best of our knowledge, there has been no study so far of posterior convergence rates when the prior distribution is a Pólya tree. One reason may be that density Pólya trees are often perceived as relatively 'rough' objects: it can be shown for instance that the corresponding density has jumps at a countable number of points almost surely. From this it could seem as if Pólya trees are just 'smooth enough' for consistency, not for rates. One first result in the paper implies that for well-chosen parameters, Pólya trees are able to model smooth functions and to induce posterior distributions with optimal convergence rates in the minimax sense for a range of Hölder regularities. Here we will follow a multiscale approach to obtaining rates and limiting shape results, introduced in [6], [7], [5], with connections to semiparametric functionals [8].

In the Bayesian nonparametrics literature, there has been a recent interest in Pólya trees and related constructions. Wong and Ma [44] introduce optional Pólya trees, where the tree is cut using stopping times in a data-driven way. The work [36] studies Rubbery Pólya trees, an extension of Pólya trees that enables some dependence in the tree while keeping its essential properties unchanged. Quantile pyramids [24] reverse the construction of the measure by fixing the probabilities but making the interval lengths random. As a way to 'smooth' Pólya trees, one can consider mixtures, as in [2], [22]. We also note that

Pólya trees are particular cases of the more general class of tail-free processes, introduced in [13], [10]; mixtures of such processes were recently considered in [27].

An in-depth introduction to Pólya trees, including their construction as well as the proofs of many useful properties, can be found in the forthcoming book by Ghosal and van der Vaart [19]. The present work directly benefited from their exposition on the subject.

1.1 Definition

First let us introduce some notation relative to dyadic partitions. For any fixed indexes (k, l), $0 \leq k < 2^l$, $l \geq 0$, the rational number $r = k2^{-l}$ can be written in a unique way as $\varepsilon(r) := \varepsilon_1(r) \dots \varepsilon_l(r)$, its finite expression of length l in base $1/2$ (note that it can end with one or more '0'). That is, $\varepsilon_i \in \{0, 1\}$ and $k2^{-l} = \sum_{i=1}^{l} \varepsilon_i(r) 2^{-i}$. Let $\mathcal{E} := \cup_{l \geq 0} \{0, 1\}^l \cup \{\varnothing\}$ be the set of finite binary sequences. We write $|\varepsilon| = l$ if $\varepsilon \in \{0, 1\}^l$ and $|\varnothing| = 0$.

Let us introduce a sequence of partitions $\mathcal{I} = \{(I_\varepsilon)_{\varepsilon:\, |\varepsilon| = l},\ l \geq 0\}$ of the unit interval. Here we will consider regular partitions, as defined below. This is mostly for simplicity of presentation, and other partitions, based for instance on quantiles of a given distribution, could be considered as well. Set $I_\varnothing = [0, 1)$ and, for any $\varepsilon \in \mathcal{E}$ such that $\varepsilon = \varepsilon(l, k)$ is the expression in base $1/2$ of $k2^{-l}$, set

$$I_\varepsilon := \left[\frac{k}{2^l}, \frac{k+1}{2^l} \right) =: I_k^l.$$

For any $l \geq 0$, the collection of all such dyadic intervals is a partition of $[0, 1)$.

A random probability measure P follows a Pólya tree distribution $PT(\mathcal{A})$ with parameters $\mathcal{A} = \{\alpha_\varepsilon,\ \varepsilon \in \mathcal{E}\}$ on the sequence of partitions $\mathcal{I}$ if there exist random variables $0 \leq Y_\varepsilon \leq 1$ such that,

1. the variables $Y_{\varepsilon 0}$ for $\varepsilon \in \mathcal{E}$ are mutually independent and $Y_{\varepsilon 0}$ follows a $\text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$ distribution.
2. for any $\varepsilon \in \mathcal{E}$, we have $Y_{\varepsilon 1} = 1 - Y_{\varepsilon 0}$
3. for any $l \geq 0$ and $\varepsilon = \varepsilon_1 \dots \varepsilon_l \in \{0, 1\}^l$, we have

$$P(I_\varepsilon) = \prod_{j=1}^{l} Y_{\varepsilon_1 \dots \varepsilon_j}. \tag{1}$$

This construction can be visualised using a tree representation, see Figure 1: to compute the random mass that P assigns to the subset I_ε of $[0, 1]$, one follows a dyadic tree along the expression of ε: $\varepsilon_1, \varepsilon_1\varepsilon_2, \dots, \varepsilon_1\varepsilon_2 \cdots \varepsilon_l = \varepsilon$. The mass $P(I_\varepsilon)$ is a product of Beta variables whose parameters depend on whether one goes 'left' ($\varepsilon_j = 0$) or 'right' ($\varepsilon_j = 1$) along the tree:

$$P(I_\varepsilon) = \prod_{j=1;\ \varepsilon_j = 0}^{l} Y_{\varepsilon_1 \dots \varepsilon_{j-1} 0} \times \prod_{j=1;\ \varepsilon_j = 1}^{l} (1 - Y_{\varepsilon_1 \dots \varepsilon_{j-1} 0}). \tag{2}$$

This construction uniquely defines a random probability distribution on distributions on $[0, 1]$. For details we refer to Ferguson [12] and Lavine [30].

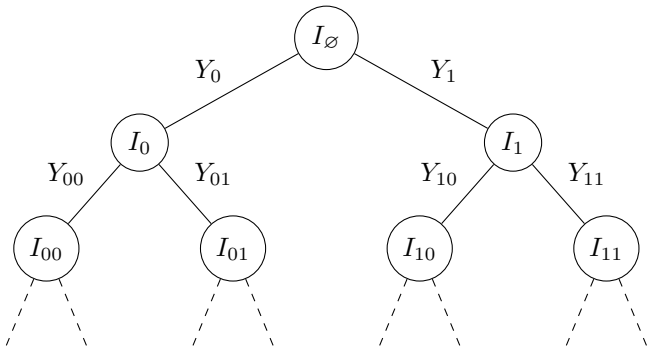

Figure 1: Indexed binary tree with levels $l \leq 2$ represented. The nodes index the intervals I_ε. Edges are labelled with random variables Y_ε.

The corresponding object, the class of Pólya tree distributions, is quite flexible: as will be seen in the results below, different behaviours of the sequence of parameters $\{\alpha_\varepsilon\}$ give a Pólya tree with different properties. A standard assumption is that the parameters $\{\alpha_\varepsilon\}$ only depend on the depth $|\varepsilon|$, so that

$$\alpha_\varepsilon = a_l, \quad \forall\, \varepsilon : |\varepsilon| = l, \tag{3}$$

for any $l \geq 1$ and a sequence $(a_l)_{l\geq 1}$ of positive numbers, which will be assumed henceforth.

The class of Pólya trees contains as special cases several important distributions used in Bayesian nonparametrics. A distinguished special case is the Dirichlet process [11], which corresponds to the choice $\alpha_\varepsilon = 2^{-|\varepsilon|}$, or more generally $MG(I_\varepsilon)$ for some $M > 0$ and a given distribution function G. It can be shown that the corresponding random probability measure is discrete almost surely and in particular does not provide a prior on densities. On the other hand, if a_l goes to ∞ fast enough with l, more precisely if

$$\sum_{l\geq 1} a_l^{-1} < \infty \tag{4}$$

one can show, see [29] or [35], that a Pólya tree distribution on the canonical dyadic partition has a.s. a density with respect to Lebesgue measure. As we are interested in random density priors, we shall work under that assumption.

The reader familiar with random multifractal structures will certainly have noticed the similarity of the previous object with random multiplicative *cascades*, as introduced by Mandelbrot and further studied in [28] and many others since then. In random cascades, the variables Y_ε are assumed to be i.i.d., and *conservative* cascades are those for which $Y_{\varepsilon 1} = 1 - Y_{\varepsilon 0}$, as we assumed above. An important difference between (density) Pólya trees random measures and standard cascades is that for Pólya trees the variables $Y_{\varepsilon 0}$ along the tree are not assumed i.i.d.: they are still independent, but their distribution may depend on the level $|\varepsilon| = l$ in the tree: this is in fact necessary under the assumed (4). Not only $a_l \to \infty$ fast enough guarantees the existence of a density, but the faster

a_l increases with l, the more 'regular' the corresponding density. This is particularly important for the approximation ability of the prior and the statistical properties considered in the present paper.

Despite this differences, and although we do not use properties of cascades in the present paper, one may expect that some properties or techniques for cascades could be of interest in the study of Pólya trees. We note that, for instance, the authors in [38] carry out a wavelet analysis of *conservative* cascades. In the present paper, such a multiscale analysis of the random measure at stake will also be central, but with two important differences: the variables $Y_{\varepsilon 0}$ are not i.i.d. and, more importantly, we do not study the Pólya tree above in itself (we shall use it as prior distribution), but rather the posterior distribution, so the random measure we analyse also depends on the data $X_1, \ldots, X_n$ and this dependence is crucial for the statistical properties considered here.

1.2 Function spaces and wavelets

We briefly introduce some standard notation appearing in the statements below.

Haar basis. The Haar wavelet basis is $\{\varphi,\ \psi_{lk}, 0 \le k < 2^l, l \ge 0\}$, where $\varphi = 1\!\!1_{[0,1]}$ and, for $\psi = -1\!\!1_{(0,1/2]} + 1\!\!1_{(1/2,1]}$,

$$\psi_{lk}(\cdot) = 2^{l/2}\psi(2^l \cdot -k), \qquad 0 \le k < 2^l, l \ge 0.$$

In this paper our interest is in density functions, that is nonnegative functions g with $\int_0^1 g\varphi = \int_0^1 g = 1$, so that their first Haar-coefficient is always 1. So, we will only need to consider the basis functions ψ_{lk} and will simply write slightly informally $\{\psi_{lk}\}$ for the Haar basis.

Function classes. Let $L^2 = L^2[0,1]$ denote the space of square-integrable functions on $[0,1]$ relative to Lebesgue measure equipped with the $\|\cdot\|_2$-norm. For $f, g \in L^2$, denote $\langle f, g\rangle_2 = \int_0^1 fg$. Let $L^\infty = L^\infty[0,1]$ denote the space of all measurable functions on $[0,1]$ that are bounded up to a set of Lebesgue-measure 0, equipped with the (essential) supremum norm $\|\cdot\|_\infty$.

The class $\mathcal{C}^\alpha[0,1]$, $\alpha \in (0,1]$, of Hölder functions on the interval $[0,1]$ is the set of functions g on $[0,1]$ such that $\sup_{x\neq y\in[0,1]} |g(x) - g(y)|/|x-y|^\alpha$ is finite. Let us recall that if a function g belongs to $\mathcal{C}^\alpha$, $\alpha \in (0,1]$, then the sequence of its Haar-wavelet coefficients $\langle g, \psi_{lk}\rangle_2$ satisfies

$$\sup_{0\le k<2^l, l\ge 0} 2^{l(\frac{1}{2}+\alpha)} |\langle g, \psi_{lk}\rangle_2| < \infty. \tag{5}$$

For a given $\alpha > 0$, and $n \ge 1$, define

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

This is the minimax rate for estimating a density function in a ball of α-Hölder functions, when the supremum-norm is considered as a loss, see [23] and [26].

1.3 Outline

In Section 2, we state our main results. Posterior rates of convergence for the density f are considered first. Next, a Donsker-type theorem is established for

the cumulative distribution function, as well as a more general nonparametric Bernstein-von Mises theorem. Section 3 gathers the proofs of Theorems 1 and 3. The second proof uses some intermediate results obtained in the first one. Section 4 gives some technical results used in the proofs, including two lemmas on Beta variables that are of independent interest.

Acknowledgement. The author would like to thank the Associate Editor and two referees for insightful comments and suggestions.

2 Main results

2.1 Posterior convergence rates

Our first result shows that, in the problem of density estimation, if a Pólya tree is used as a prior distribution on the density, the optimal minimax convergence rate with respect to the supremum norm is attained by the posterior distribution, provided the parameters of the tree are well chosen. Note that the choice (6) in the Theorem satisfies the summability condition $\sum_l a_l^{-1}$ in (4) above, which ensures that the prior distribution has a density. The notation Π is used for the distribution on densities induced by the considered Pólya tree.

We also note that the $\|\cdot\|_\infty$-norm in the result is, as defined above, the essential supremum with respect to Lebesgue measure on $[0,1]$: the proof of the result is based on a Haar-wavelet analysis of the posterior density f, which identifies f Lebesgue-almost everywhere. Let, for any reals a, b, denote $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$.

Theorem 1. *Let $X^{(n)} = (X_1, \dots, X_n)$ be i.i.d. from law P_0 with density f_0. Let f_0 belong to $\mathcal{C}^\alpha[0,1]$, for $\alpha \in (0,1]$ and suppose f_0 is bounded away from 0 on $[0,1]$. Let Π be the prior on densities generated by a Pólya tree random measure with respect to the canonical dyadic partition of $[0,1]$ with parameters $\mathcal{A} = \{\alpha_\varepsilon,\ \varepsilon \in \mathcal{E}\}$ chosen as $\alpha_\varepsilon = a_{|\varepsilon|} \vee 8$ for any $\varepsilon \in \mathcal{E}$, with*

$$a_l = l 2^{2l\alpha}, \qquad l \geq 0. \tag{6}$$

Then as $n \to \infty$, for any $M_n \to \infty$, it holds

$$E^n_{f_0} \Pi[f :\ \|f - f_0\|_\infty \leq M_n \varepsilon^*_{n,\alpha} \,|\, X^{(n)}] \to 1.$$

This result implies that for the considered prior, most of the mass of the posterior distribution concentrates in a $\|\cdot\|_\infty$ ball around f_0 of radius the minimax rate of convergence. It immediately implies rates for all L^q-norms, $1 \leq q < \infty$, that are minimax optimal up to a logarithmic factor. The choice of parameters (6) realises an adequate 'bias-variance' trade-off for which the optimal minimax rate $\varepsilon^*_{n,\alpha}$ is attained.

Theorem 1 assumes that $\log f_0$ is bounded. This is for simplicity of presentation and could be improved, though it would not add to the ideas we want to expose here: we preferred to keep a simple condition to make proofs more transparent.

We also have the following result.

Proposition 1. *Under the same assumptions as in Theorem 1, let* Π *a Pólya tree prior* Π *defined in the same way except that one now sets*

$$a_l = l2^{2l\delta}, \qquad l \geq 0, \tag{7}$$

for some $\delta \in (0, 1]$ *possibly different from the Hölder–regularity* α *of* f_0*. Set*

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

Then as $n \to \infty$*, for any* $M_n \to \infty$*, it holds*

$$E^n_{f_0}\Pi[f : \ \|f - f_0\|_\infty \leq M_n\varepsilon^*_{n,\alpha,\delta} \,|\, X^{(n)}] \to 1.$$

The proof of these results can be adapted to handle different choices of parameters; we do not elaborate on this in details here but only note that

1. the presence of the factor l in (6) corresponds to the fact that we looked for a sharp optimal minimax rate (up to a constant) in the supremum norm. Removing this factor in the choice of a_l leads to a rate $(\log^\eta n)\varepsilon^*_{n,\alpha}$ for some $\eta > 0$, with an extra logarithmic term, instead of $\varepsilon^*_{n,\alpha}$ as above (something similar happens in the Gaussian white noise model with series priors, see [21] and [5]). On the other hand, the presence of an 'l' or not does not affect results for most smooth functionals, as will be seen below.
2. results for truncated priors can be obtained similarly, for instance the choice

$$\alpha_\varepsilon^{-1} = \begin{cases} 1 & \text{if } |\varepsilon| = l \text{ and } l \leq l_n \\ 0 & \text{if } |\varepsilon| = l \text{ and } l > l_n, \end{cases} \tag{8}$$

where l_n is defined in (12) below with $\delta = \alpha$, and with the convention that $\text{Beta}(\infty, \infty)$ is the Dirac mass $\delta_{1/2}$ distribution, leads to the same posterior contraction rate as in Theorem 1. The proof is similar, though easier, as one truncates high frequencies. However, it is a n-dependent prior; in contrast the prior (6) is canonical, in the sense that it does not depend on n. We discuss this further in Section 2.4 below.

So far, only a few results on posterior convergence in the supremum norm have been obtained, see e.g. [21], [5] and [25]. In [5] we suggested a possible approach to obtain such results. One of the starting points for the present paper is a question of a referee of [5], who asked whether some results for non-n-dependent priors in density estimation could be obtained. The proof of Theorem 1 gives another illustration of the approach in [5] and answers the question positively.

2.2 Donsker-type theorem

Let us now consider the behaviour of the cumulative distribution function $F(x) = \int_0^x f(t)dt$ induced by the posterior distribution when a Pólya tree is used as prior. Given data $X_1, \ldots, X_n$, let F_n denote the empirical distribution function

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \leq t}.$$

For $\delta > 0$, define a sequence $\alpha_\varepsilon = a_{|\varepsilon|} \vee 8$ for $\varepsilon \in \mathcal{E}$, where

$$a_l = 2^{2l\delta}, \quad l \geq 0 \quad \text{or} \quad a_l = l2^{2l\delta}, \quad l \geq 0. \tag{9}$$

For a prior Π on densities induced by a Pólya tree distribution with parameters as in (9), let $\bar{f}_n$ denote the posterior mean $\int f d\Pi(f \,|\, X)$ and let $\bar{F}_n(t) = \int_0^t \bar{f}_n(u)du$ denote its distribution function. In the next result, $\mathcal{L}(G)$ denotes the law of a process G, and $\mathcal{L}(F \,|\, X)$ denotes the induced posterior distribution on F. Also, on a metric space S, such as the space of $C[0,1]$ continuous functions on $[0,1]$ equipped with the supremum norm, we denote by β_S the bounded-Lipschitz metric on S, which metrises weak convergence on S. The definition of β_S is recalled in (27) in the Appendix, where more details can be found.

Theorem 2 (Donsker's theorem for Pólya tree posteriors)**.** *Let $X = (X_1, \ldots, X_n)$ be i.i.d. from law P_0 with density f_0. Let f_0 belong to $\mathcal{C}^\alpha[0,1]$, for some $\alpha \in (0,1]$ and suppose f_0 is bounded away from 0 on $[0,1]$. Let Π be a Pólya tree $PT(\mathcal{A})$ with parameters $\mathcal{A} = \{\alpha_\varepsilon,\ \varepsilon \in \mathcal{E}\}$ such that $\alpha_\varepsilon = a_{|\varepsilon|}$ for all $\varepsilon \in \mathcal{E}$, with (a_l) is as in (9) for some $\delta > 0$.*

Let G_{P_0} be a P_0-Brownian bridge $G_{P_0}(t), t \in [0,1]$. For any parameters $\alpha \in (0,1], \delta > 0$, as $n \to \infty$,

$$\beta_{C[0,1]}\left(\mathcal{L}(\sqrt{n}(F - \bar{F}_n) \,|\, X), \mathcal{L}(G_{P_0})\right) \to^{P_{f_0}} 0.$$

Furthermore, for any $\alpha \in (0,1]$ and δ such that $\delta < 1/2 + \alpha$, as $n \to \infty$,

$$\beta_{L^\infty[0,1]}\left(\mathcal{L}(\sqrt{n}(F - F_n) \,|\, X), \mathcal{L}(G_{P_0})\right) \to^{P_{f_0}} 0.$$

In particular, the last display holds true if $\delta \leq 1/2$, regardless of the value of α.

This result parallels Lo's result [33] for the Dirichlet process, here in a regime where the Pólya tree as well as the true law P_0 have a density. A few results of this type have been obtained in the literature since then, mostly for priors whose realisations are discrete measures, like the Dirichlet process, see the introduction of [7] for some references.

A possible route for proving such a result is a direct analysis of the induced posterior on $F(\cdot)$. Here we use the approach proposed in [7] and obtain it as a fairly direct consequence of a more general result on the shape of the posterior distribution stated in the next section.

2.3 Limiting shape of the posterior distribution

We focus now on limiting shape results for (aspects of) the posterior density f. For this we follow the approach to nonparametric Bernstein–von Mises (BvM) theorems introduced in [7]. The idea is to formulate convergence in distribution of the posterior density to a Gaussian process in a large enough space $\mathcal{M}_0$ defined below that enables convergence at rate $\sqrt{n}$. Once convergence in distribution on $\mathcal{M}_0$ is obtained, it will typically be possible to deduce results via continuous mapping for continuous functionals $\psi : \mathcal{M}_0 \to \mathcal{Y}$ for some given space $\mathcal{Y}$.

First we need a sequence of 'weights' $w := \{w_l\}_{l \geq 0}$, such that $w_l/\sqrt{l} \uparrow \infty$. The space $\mathcal{M}_0 = \mathcal{M}_0(w)$ is defined as the multiscale sequence space

$$\mathcal{M}_0 = \Big\{x = \{x_{lk}\} : \lim_{l \to \infty} \max_k \frac{|x_{lk}|}{w_l} = 0\Big\}, \tag{10}$$

equipped with the norm $\|x\|_{lk} := \sup_l \max_k |x_{lk}|/w_l$. It is a separable Banach space. A (possibly generalised) function f is said to belong to $\mathcal{M}_0$ if the sequence of its wavelet coefficients $\langle f, \psi_{lk}\rangle$ over the Haar basis $\{\psi_{lk}\}$ belongs to $\mathcal{M}_0$.

Now we define the limiting process. For P a given probability distribution on $[0,1]$, let $\mathbb{G}_P$ be the Gaussian process indexed by the Hilbert space $L^2(P) \equiv \{f : [0,1] \to \mathbb{R} : \int_0^1 f^2 dP < \infty\}$ with covariance function

$$\mathbb{E}\left[\mathbb{G}_P(g)\mathbb{G}_P(h)\right] = \int_0^1 (g - Pg)(h - Ph)dP.$$

We call $\mathbb{G}_P$ the *P-white bridge process*. It can be checked, see [7], that $\mathbb{G}_P$, provided $w_l/\sqrt{l} \uparrow \infty$, is a tight Borel Gaussian variable in $\mathcal{M}_0$.

The first statement in Theorem 3 automatically recenters the posterior distribution around the posterior mean. Typically, one may wish to center instead around a 'canonical' centering, in that its definition does not depend on the posterior. This can be achieved by comparing $\bar{f}_n$, for instance, to a smoothed version of the empirical measure. Let P_n denote the empirical measure $n^{-1}\sum_{i=1}^n \delta_{X_i}$ associated to the observed data X and let, for L_n defined in (13) below,

$$\langle T_n, \psi_{lk}\rangle = \begin{cases} \langle P_n, \psi_{lk}\rangle & \text{if } \ l \le L_n \\ 0 & \text{if } \ l > L_n, \end{cases} \tag{11}$$

and T_n is a tight random variable in $\mathcal{M}_0$. For a given $\delta > 0$, let $j_n = j_n(\delta)$ and $l_n = l_n(\delta)$ be the *largest* integers such that

$$2^{j_n} \le n^{\frac{1}{2\delta+1}}, \qquad 2^{l_n} \le \left(\frac{n}{\log n}\right)^{\frac{1}{2\delta+1}}. \tag{12}$$

and set, in slight abuse of notation, either

$$L_n = j_n \ \ (\forall n \ge 1) \quad \text{or} \quad L_n = l_n \ \ (\forall n \ge 1). \tag{13}$$

As before, let $\bar{f}_n$ denote the posterior mean $\int f d\Pi(f \,|\, X)$.

In the next result, $\beta_{\mathcal{M}_0(w)}$ denotes the bounded-Lipschitz metric on $\mathcal{M}_0(w)$, see (27) in the Appendix for a definition.

Theorem 3. *Let $X = (X_1, \ldots, X_n)$ be i.i.d. from law P_0 with density f_0. Let f_0 belong to $\mathcal{C}^\alpha[0,1]$, for some $\alpha \in (0,1]$ and suppose f_0 is bounded away from 0 on $[0,1]$. Let Π be a Pólya tree $PT(\mathcal{A})$ with parameters $\mathcal{A} = \{a_l,\ l \ge 1\}$, where a_l is as in (9) for some $\delta > 0$.*

Let $\tau_{\bar{f}_n} : f \to \sqrt{n}(f - \bar{f}_n)$ and let $z = \{z_l\}_l$ be a weighting sequence such that $z_l/\sqrt{l} \uparrow \infty$. For any parameters $\alpha \in (0,1], \delta > 0$, as $n \to \infty$,

$$\beta_{\mathcal{M}_0(z)}(\Pi(\cdot \,|\, X) \circ \tau_{\bar{f}_n}^{-1}, \mathbb{G}_{P_0}) \to^{P_{f_0}} 0.$$

If $\delta \le \alpha$, if T_n is given by (11) and $\tau_{T_n} : f \to \sqrt{n}(f - T_n)$, then as $n \to \infty$,

$$\beta_{\mathcal{M}_0(z)}(\Pi(\cdot \,|\, X) \circ \tau_{T_n}^{-1}, \mathbb{G}_{P_0}) \to^{P_{f_0}} 0.$$

This result has several applications relevant for statistics. One application, that we only mention, is the construction of confident credible bands for

fixed regularities, see [7] Section 4.2. Another application is the derivation of Bernstein-von Mises theorems for semiparametric functionals via the continuous mapping theorem. A prototypical example is the map $f \to \int_0^{\cdot} f = F(\cdot)$, leading to a Donsker-type result for the distribution function F, as stated in the previous section. Indeed, it has been shown in [7] that the map $f \to F$ is continuous from $\mathcal{M}_0(w)$ to $\mathcal{C}[0,1]$. Theorem 2 then essentially follows from Theorem 3 combined with Theorem 4 in [7], see Section 4 for a detailed proof. Results for smooth linear functionals also fairly directly follow from Theorem 3. For details on this and several other examples of functionals, we refer to [6]-[7].

2.4 Discussion

First let us address two natural questions about the results.

Are Pólya trees not too 'rough' as a prior to obtain nontrivial posterior rates of contraction? A reason to ask is that one can show that any version of the posterior density Pólya tree has a jump at any point of the subdivisions of $[0,1]$ corresponding to the successive partitions, that is at all the dyadic rationals for regular dyadic partitions. But this of course does not prevent the object to have good approximation properties, similar to the fact that histograms can be used to approximate e.g. Lipschitz functions (note that Pólya tree densities are not histograms though), and our results show that this is indeed the case. Even more, it can be checked that at any *non-dyadic* point x_0 of $[0,1]$, the density induced by a Pólya tree with parameters is locally α–Hölder at point x_0. This is of course in line with the result of Theorem 1 that for such choice of a_l the posterior has optimal concentration around α–Hölder functions.

Is the choice $a_l = 2^{2l\alpha}$, or $l2^{2l\alpha}$, reasonable in practice? Indeed, one may think that such an increase in the Beta-parameters, that is exponential, could be 'hard to fit' in practice. The theoretical results show good behaviour of the posterior for this choice though, and we claim that this exponential behaviour is the 'correct one' if one whishes to model all frequencies. Indeed, the exponential growth corresponds to the exponentially fast decrease of the width of dyadic intervals I_ε. It is simply that wavelet coefficients, which here are modelled through products of Beta variables, naturally decrease exponentially fast for Hölder classes, see Eq. (5). Similarly, in regression, typical Gaussian processes used as prior distributions have variances decreasing as a power of 2^{-l}, which is exponentially fast, too. Of course, one may also consider priors that truncate high frequencies as in Eq. (8), in which case the contraction rate is essentially driven by the cut-off point, not so much by the individual variance parameters, similar to what has been noted e.g. in [39].

The results are also part of a more general programme linked to obtaining Bernstein-von Mises results, as well as posterior contraction in strong losses such as the supremum norm. For instance, the results are of interest for

1. Bernstein-von Mises theorems for 'smooth' functionals. In [8], we obtained limiting posterior shape results on a family of random histograms. One may note that the truncated version of the Pólya tree defined by (8) is also a random histogram, but with a quite different randomness in the weights. Similar to what is noted below Theorem 1, Theorems 2 and 3 above can be checked to hold for this prior as well. The histograms in [8] can be seen as histograms-projections of the Dirichlet process, giving Dirichlet weights.

Here the weights are not Dirichlet, but correspond to a tree-type product of Beta variables.

2. non-parametric BvM and posterior contraction in supremum norm. In [6], [7], [5], a multiscale approach was developed and a programme to obtain results of this type was proposed. Only a few examples of priors have been investigated within this framework so far, and investigating other classes of prior distributions is of great interest. One may note for instance that in the density estimation model, the priors considered in [7], [5] were all n-dependent (note that, in fact, there cannot be a non-n-dependent version of the random histograms as in [7], [5], as the corresponding underlying infinite dimensional prior would be a Dirichlet process, which has no density). The Pólya-tree class of priors in the present paper precisely provides an example of such canonical prior.

We plan to study further properties of Pólya trees in future work. Among others one can mention two natural questions. First, adaptation: here we have studied the case where the Hölder regularity parameter β of f_0 is given. Several constructions can be considered to build an adaptive prior, that automatically adapts to the unknown regularity β. Second, results for higher regularities, that is $\beta \geq 1$, would also be of interest.

Extensions to higher regularities. A first question is whether the analysis of the present paper could be carried out, for the same Pólya tree prior, to higher regularities $\alpha > 1$ using higher order wavelets. Indeed, one may think that choosing $a_l = l2^{2l\alpha}$ may continue to lead to optimal rates even when $\alpha > 1$. We believe this is not the case. One indication why this is presumably not the case is that under the *prior* distribution, it can be checked that locally around any point $x_0 \in [0,1]$, the density is not more than locally $\mathcal{C}^1$ even if $\alpha > 1$ (as noted above, when $\alpha \leq 1$ the prior density when $a_l = l2^{2l\alpha}$ is locally $\mathcal{C}^\alpha$ at any non-dyadic point). The intuition is that having all Y_ε independent at a same level $l = |\varepsilon|$ creates 'too much independence' between values at different points to produce a highly smooth density.

The second question is whether a different, cascade–like tree–induced scheme for sequentially defining random masses $P(I_\varepsilon)$ of intervals could produce a random density with a given arbitrary smoothness level α possibly larger than 1. Such a construction could for instance be inspired by the schemes defining wavelets bases $\{\psi_{lk}\}$ that enable to capture smoother regularities (e.g. Daubechies or boundary-corrected wavelets) compared to the Haar basis ψ^H_{lk}. The point is to understand whether this is could be done while still preserving a form of conjugacy: here conjugacy is obtained at a given level with a multinomial likelihood on the one hand (the data produce counts $N_X(I_\varepsilon)$ on each dyadic I_ε) and a finite-tree prior of Beta distributions for interval probabilities on the other hand. It is thus quite directly related to a definition of f via inner-products with indicators $\langle f, 1\!\!1_{I_\varepsilon}\rangle$, which naturally leads to the Haar basis. Would there exist a conjugate structure that would enable to define $\langle f, \psi_{lk}\rangle$ along a tree–like scheme ? constructions instead. This will be studied elsewhere.

3 Proofs

3.1 Preliminaries and notation

By the standard conjugacy property of Pólya trees, see [12], [31], if P follows a $PT(\mathcal{A})$ distribution, the posterior distribution $P \,|\, X_1, \ldots, X_n$ follows a Pólya tree distribution $PT(\mathcal{A}^*)$ with respect to the same partition and with updated parameters $\mathcal{A}^* = \{\alpha^*_\varepsilon,\ \varepsilon \in \mathcal{E}\}$, where

$$\alpha^*_\varepsilon = \alpha_\varepsilon + N_X(I_\varepsilon), \tag{14}$$

with $N_X(I_\varepsilon) = \sum_{i=1}^n I\{X_i \in I_\varepsilon\}$.

The following sets of notation will be used throughout the proofs.

1. *Tilded notation, posterior distribution.* We denote by $\tilde{P}$ a distribution sampled from the posterior distribution and by $\tilde{Y}$ the corresponding variables Y in (1). In particular, the variable $\tilde{Y}_{\varepsilon 0}$ is $\mathrm{Beta}(\alpha^*_{\varepsilon 0}, \alpha^*_{\varepsilon_1})$ distributed.
2. *Bar notation, posterior mean.* Let $\bar{f} = \int f d\Pi(f \,|\, X)$ denote the posterior mean density and $\overline{P}$ the corresponding probability measure. We use the notation $\bar{Y}$ for the variables defining $\overline{P}$ via (1).
3. *Paths along the tree.* A given $\varepsilon = \varepsilon_1 \cdots \varepsilon_l \in \mathcal{E}$ gives rise to a 'path' $\varepsilon_1 \to \varepsilon_1\varepsilon_2 \to \varepsilon_1\varepsilon_2 \cdots \varepsilon_l$. We denote

$$I^{[i]}_\varepsilon := I_{\varepsilon_1 \ldots \varepsilon_i},$$

for any i in $\{1, \ldots, l\}$. Similarly, denote, with E_X the expectation under the posterior distribution,

$$\tilde{Y}^{[i]}_\varepsilon = \tilde{Y}_{\varepsilon_1 \ldots \varepsilon_i}, \quad \bar{Y}^{[i]}_\varepsilon = E_X[\tilde{Y}^{[i]}_\varepsilon].$$

Conversely, any pair (l, k) with $l \geq 0$ and $k \in \{0, \ldots, 2^l - 1\}$ is associated with a unique $\varepsilon = \varepsilon(l, k)$, the expression of length l in base $1/2$ of $k2^{-l}$.

For a given distribution P with distribution function F and density f on $[0, 1]$, denote $P(B) = F(B) = \int_B f$, for any measurable subset B of $[0, 1]$. In particular under the 'true' distribution, we denote $P_0(B) = F_0(B) = \int_B f_0$. In the sequel C denotes a universal constant whose value only depends on other fixed quantities of the problem.

For a function f in L^2, and L_n an integer, denote by f^{L_n} the L^2-projection of f onto the linear span of all elements of the basis $\{\psi_{lk}\}$ up to level $l = L_n$. Also, denote $f^{L_n^c}$ the projection of f onto the orthocomplement $\mathrm{Vect}\{\psi_{lk},\ l > L_n\}$. In the proofs, we shall use the decomposition $f = f^{L_n} + f^{L_n^c}$, which holds in L^2 and L^∞ under prior and posterior: this follows from Lemma 7, which gives sufficient conditions in terms of the sequence (a_l) for both L^2 and L^∞ statements. Both conditions of the Lemma are satisfied for (a_l) of the form (6) or (9).

3.2 Proof of Theorem 1

Proof. Define L_n to be the integer such that

$$2^{L_n} = \lfloor c_0 \left(\frac{n}{\log n} \right)^{\frac{1}{1+2\alpha}} \rfloor, \tag{15}$$

for c_0 a small enough constant to be chosen below.

Step 0. Haar decomposition and an event $\mathcal{B}$. First, we define an event $\mathcal{B}$ on the data space. For any integer l, set $\Lambda_n(l)^2 := (l + L_n)n/2^l$. Recall the notation I_k^l from Section 1.1. Define $\mathcal{B}$ as the event on which, *simultaneously* for the countable family of indexes $l \geq 1, 0 \leq k < 2^l$, for M large enough to be chosen,

$$M^{-1}|N_X(I_k^l) - nF_0(I_k^l)| \leq \Lambda_n(l) \vee (l + L_n), \tag{16}$$

where as before $N_X(I)$ is the number of data points in I. By Lemma 4, we have

$$P_{f_0}^n(\mathcal{B}^c) = o(1). \tag{17}$$

Let us now decompose, using the notation above for the projection,

$$f - f_0 = (f^{L_n} - \bar{f}^{L_n}) + (\bar{f}^{L_n} - f_0^{L_n}) + f^{L_n^c} - f_0^{L_n^c}, \tag{18}$$

which holds in L^∞ (Lebesgue-almost surely) and in L^2.

For any given l, k, for $\varepsilon = \varepsilon(l,k)$ the expression in base $1/2$ of $k2^{-l}$, let us write $I_{\varepsilon(l,k)} = I_{\varepsilon 0} \cup I_{\varepsilon 1}$. For P a probability measure of density f and $\{\psi_{lk}\}$ the Haar basis, by definition $\langle f, \psi_{lk}\rangle_2 = 2^{l/2}(P(I_{\varepsilon 1}) - P(I_{\varepsilon 0}))$. For a function g, denote by g_{lk} its coefficients onto the Haar basis. If P follows a Pólya tree distribution with density f, we thus have the equality in law

$$f_{lk} := \langle f, \psi_{lk}\rangle_2 = 2^{l/2}P(I_\varepsilon)(1 - 2Y_{\varepsilon 0}). \tag{19}$$

To start with, let us note that

$$\begin{aligned}\|f_0^{L_n^c}\|_\infty &= \Big\| \sum_{l>L_n,k} f_{0,lk}\psi_{lk}\Big\|_\infty \\ &\leq \sum_{l>L_n} \left\{\max_k |f_{0,lk}|\right\} \left\| \sum_k |\psi_{lk}|\right\|_\infty \lesssim \sum_{l>L_n} 2^{-l\alpha} \lesssim \varepsilon_{n,\alpha}^*,\end{aligned}$$

using that f_0 is Hölder and the definition of L_n. We now focus successively on each of the remaining terms in the decomposition (18), before putting the bounds together and concluding.

Step 1, term $\bar{f}^{L_n} - f_0^{L_n}$ in (18). Given indexes l, k, and $\varepsilon = \varepsilon(l,k)$,

$$\bar{f}_{lk} = 2^{l/2}\bar{P}(I_\varepsilon)(1 - 2\bar{Y}_{\varepsilon 0})$$

as well as, with $y_{\varepsilon 0} := F_0(I_{\varepsilon 0})/F_0(I_\varepsilon)$,

$$f_{0,lk} = 2^{l/2}P_0(I_\varepsilon)(1 - 2y_{\varepsilon 0}).$$

This leads to the expression

$$\bar{f}_{lk} - f_{0,lk} = f_{0,lk}\left[\frac{\bar{P}(I_\varepsilon)}{P_0(I_\varepsilon)} - 1\right] + 2^{\frac{l}{2}+1}\bar{P}(I_\varepsilon)(y_{\varepsilon 0} - \bar{Y}_{\varepsilon 0}).$$

Combining Lemmas 1 and 2 and the fact that $F_0(I_\varepsilon) \lesssim 2^{-l}$ leads to, on $\mathcal{B}$,

$$\begin{aligned}|\bar{f}_{lk} - f_{0,lk}| &\lesssim |f_{0,lk}| \left\{\sum_{i=1}^{l} \frac{a_i 2^i}{n} + \sqrt{\frac{L_n 2^l}{n}}\right\} + \left(\frac{2^l a_l}{n}|f_{0,lk}| + \sqrt{\frac{L_n}{n}}\right) \\ &\lesssim |f_{0,lk}| \left\{\frac{a_l 2^l}{n} + \sqrt{\frac{L_n 2^l}{n}}\right\} + \sqrt{\frac{L_n}{n}}.\end{aligned}$$

From this deduce that, on the event $\mathcal{B}$,

$$\begin{aligned}\|\bar{f}^{L_n} - f_0^{L_n}\|_\infty &\lesssim \sum_{l=0}^{L_n} 2^{l/2} \max_{0\le k<2^l-1} |\bar{f}_{lk} - f_{0,lk}| \\ &\lesssim \sum_{l=0}^{L_n} l 2^{l(\alpha+1)} n^{-1} + \sqrt{\frac{L_n 2^{L_n}}{n}} \\ &\lesssim 2^{-L_n\alpha} + \sqrt{\frac{L_n 2^{L_n}}{n}} \lesssim \varepsilon^*_{n,\alpha},\end{aligned}$$

where we have used $a_l \le l 2^{2l\alpha}$ and $|f_{0,lk}| \lesssim 2^{-l(1/2+\alpha)}$.

Step 2, term $f^{L_n} - \bar{f}^{L_n}$ in (18). We define an event $\mathcal{A}$ for which $\Pi[\mathcal{A}^c \,|\, X^{(n)}] = o(1)$. We aim at having each variable $\tilde{Y}_\varepsilon$ defining the posterior law not too far from its expectation. In terms of f, this means a control on $\int_{I_\varepsilon^{[i+1]}} f / \int_{I_\varepsilon^{[i]}} f$ for all admissible i, $\varepsilon = \varepsilon(l,k)$ with $l \le L_n$. Let $\mathcal{A}$ be the measurable set of densities f on which, simultaneously for all possible ε, i,

$$\begin{aligned}|\tilde{Y}_\varepsilon^{[i]} - \bar{Y}_\varepsilon^{[i]}| &= \Big| \int_{I_\varepsilon^{[i]}} f / \int_{I_\varepsilon^{[i-1]}} f - \int \left[\int_{I_\varepsilon^{[i]}} f / \int_{I_\varepsilon^{[i-1]}} f \right] d\Pi(f \,|\, X^{(n)}) \Big| \\ &\le M \sqrt{\frac{L_n}{n F_0(I_\varepsilon^{[i]})}} =: r_\varepsilon^{[i]}. \end{aligned} \tag{20}$$

Let us check that the complement of $\mathcal{A}$ has small posterior probability. By definition $\tilde{Y}_\varepsilon^{[i]}$ follows a Beta distribution of parameters $\varphi_i = a_i + N_X(I_{\varepsilon_1\dots\varepsilon_i})$ and $\psi_i = a_i + N_X(I_{\varepsilon_1\dots(1-\varepsilon_i)})$. So $\varphi_i \wedge \psi_i \ge a_i \ge 8$ and $\varphi_i + \psi_i = 2a_i + N_X(I_\varepsilon^{[i-1]})$. Also, one has $\bar{Y}_\varepsilon^{[i]} = \varphi_i/(\varphi_i + \psi_i)$. By Lemma 2, this ratio is bounded below by a constant times $F_0(I_\varepsilon^{[i]})/F_0(I_\varepsilon^{[i-1]})$, for n large enough. Indeed, the remainder term in Lemma 2 is a $o(1)$ for our choices of a_l, L_n and using the regularity of f_0. So $\varphi_i/(\varphi_i + \psi_i)$ is bounded away from 0 and 1.

Now one can apply Lemma 6, with $x = ML_n^{1/2}/2$ and M a constant to be chosen below. First one checks that

$$\begin{aligned}\varphi_i + \psi_i &\ge N_X(I_\varepsilon^{[i-1]}) \ge N_X(I_\varepsilon^{[i]}) \\ &\ge n F_0(I_\varepsilon^{[i]}) - \sqrt{2L_n n 2^{-i}} \\ &\ge n F_0(I_\varepsilon^{[i]})/2,\end{aligned}$$

as $nF_0(I_\varepsilon^{[i]})/2 \asymp n2^{-i}$ and $\sqrt{2L_n n 2^{-i}} = o(n2^{-i})$ uniformly in $i \le L_n$. Thus for n large enough, for $x = ML_n^{1/2}/2$, using Remark 1,

$$\mathbb{P}\left[|\tilde{Y}_\varepsilon^{[i]} - \bar{Y}_\varepsilon^{[i]}| > \frac{x}{\sqrt{n F_0(I_\varepsilon^{[i]})}} \right] \le D e^{-x^2/4}.$$

A union bound leads to, for $\mathcal{A}$ defined by (20) and d a small enough constant,

$$\Pi[\mathcal{A}^c \,|\, X^{(n)}] \lesssim \sum_{l \le L_n} 2^l e^{-dM^2 \log n},$$

which tends to 0 for M a large enough constant. Now

$$\begin{aligned} f_{lk} - \bar{f}_{lk} &= 2^{l/2}\bar{P}(I_\varepsilon)\left[\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)}(1-2\tilde{Y}_{\varepsilon 0}) - (1-2\bar{Y}_{\varepsilon 0})\right] \\ &= 22^{l/2}\bar{P}(I_\varepsilon)(\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}) + 2^{l/2}\bar{P}(I_\varepsilon)\left[\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right]\left[1 - 2\bar{Y}_{\varepsilon 0} + 2(\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0})\right] \\ &= 22^{l/2}\bar{P}(I_\varepsilon)(\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}) + \left[\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right]\left[\bar{f}_{lk} + 22^{l/2}\bar{P}(I_\varepsilon)(\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0})\right], \end{aligned}$$

where by definition

$$\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1 = \prod_{j=1}^{l} \frac{\tilde{Y}_\varepsilon^{[i]}}{\bar{Y}_\varepsilon^{[i]}} - 1.$$

By Lemma 2, the mean $\bar{Y}_{\varepsilon 0}$ is close to $F_0(I_{\varepsilon 0})/F_0(I_\varepsilon)$ when $l \le L_n$, and similarly for $\bar{Y}_{\varepsilon 1}$. In particular it is bounded away from 0 and 1. So on $\mathcal{B}$, one can replace $|\tilde{Y}_\varepsilon^{[i]} - \bar{Y}_\varepsilon^{[i]}|$ in (20) by $|\tilde{Y}_\varepsilon^{[i]}/\bar{Y}_\varepsilon^{[i]} - 1|$ up to multiplying the upper bound $r_\varepsilon^{[i]}$ in (20) by a universal constant. The conditions of Lemma 3 are satisfied, as $L_n 2^{L_n}/n$ is a $o(1)$. Deduce that on $\mathcal{A}$ and on the event $\mathcal{B}$,

$$\left|\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right| \lesssim \sum_{i=0}^{l-1} r_\varepsilon^{[i]} \lesssim \sqrt{\frac{L_n 2^l}{n}}.$$

On the other hand, we directly have with (20) that $|\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}| \lesssim \sqrt{L_n 2^l/n}$. Conclude that for any f in $\mathcal{A}$ and on the event $\mathcal{B}$,

$$\begin{aligned} |f_{lk} - \bar{f}_{lk}| &\lesssim |\bar{f}_{lk}|\sqrt{\frac{L_n 2^l}{n}} + 2^{l/2}\bar{P}(I_\varepsilon)\left[\sqrt{\frac{L_n 2^l}{n}} + \frac{L_n 2^l}{n}\right] \\ &\lesssim |\bar{f}_{lk}|\sqrt{\frac{L_n 2^l}{n}} + \sqrt{\frac{L_n}{n}}, \end{aligned}$$

using that, on $\mathcal{B}$, the mean $\bar{P}(I_\varepsilon)$ is within a constant of 2^{-l}. By the triangle inequality $|\bar{f}_{lk}| \le |\bar{f}_{lk} - f_{0,lk}| + |f_{0,lk}|$. Now it suffices to notice that the terms induced by $\bar{f}_{lk} - f_{0,lk}$ and $f_{0,lk}$ respectively have already been dealt with in Step 1 above. Deduce that, on $\mathcal{B}$, for f in $\mathcal{A}$,

$$\|f^{L_n} - \bar{f}^{L_n}\|_\infty \lesssim \varepsilon_{n,\alpha}^* + \sqrt{\frac{L_n 2^{L_n}}{n}} \lesssim \varepsilon_{n,\alpha}^*.$$

Step 3, term $f^{L_n^c}$ in (18). For any $R > 1$ to be chosen, denoting by E_X the expectation under the posterior distribution,

$$\begin{aligned} E_X\|f^{L_n^c}\|_\infty &\le \sum_{l > L_n} 2^{l/2} E_X\left[\max_k |f_{lk}|\right] \\ &\le \sum_{l > L_n} 2^{l/2}\left[\sum_{k=0}^{2^l-1} E_X|f_{lk}|^R\right]^{1/R}, \end{aligned}$$

where we have used Hölder's inequality and bounded the max by the sum. From (19), with $\varepsilon = \varepsilon(l,k)$, using independence of $\tilde{Y}$s given the data along a path in the tree,

$$E_X|f_{lk}|^R = 2^{lR/2} E_X \tilde{P}(I_\varepsilon)^R E_X|1 - 2\tilde{Y}_{\varepsilon 0}|^R. \tag{21}$$

First we deal with the last expectation in (21). We apply Lemma 5 with $a = a_l + N_X(I_{\varepsilon 0})$, $d = N_X(I_{\varepsilon 0}) - N_X(I_{\varepsilon 1})$ and

$$R = R_n := \log_2 n.$$

To do so, we check that the condition $2|d| \le a$ is satisfied on the event $\mathcal{B}$. Let $M_l = M((ln/2^l)^{1/2} \vee l)$ be the constant appearing in Lemma 4 when $l \ge L_n$. Since $\tilde{Y}_{\varepsilon 0} \sim \text{Beta}(a_l + N_X(I_{\varepsilon 0}), a_l + N_X(I_{\varepsilon 1}))$, we note that, on $\mathcal{B}$,

$$\begin{aligned} |N_X(I_{\varepsilon 0}) - N_X(I_{\varepsilon 1})| &\le n|F_0(I_{\varepsilon 0}) - F_0(I_{\varepsilon 1})| + M_l \\ &\lesssim n2^{-l(1+\alpha)} + M_l \lesssim M_l, \end{aligned}$$

where we have used that f_0 is α-Hölder, that $l > L_n$ and that

$$\begin{aligned} a_l + N_X(I_{\varepsilon 0}) &\ge a_l + nF_0(I_{\varepsilon 0}) - M_l \\ &\gtrsim a_l + n2^{-l} - M_l, \end{aligned}$$

so the condition is satisfied on $\mathcal{B}$ for $l > L_n$. Lemma 5 now implies that

$$\begin{aligned} E_X[|1 - 2\tilde{Y}_{\varepsilon 0}|^R] &\le (C|N_X(I_{\varepsilon 0}) - N_X(I_{\varepsilon 1})|/a_l)^R + (CR/a_l)^{R/2}, \\ &\le (CM_l/a_l)^R + (CR/a_l)^{R/2} \lesssim (CR/a_l)^{R/2} \end{aligned}$$

for n large enough so that $R = R_n \ge R_0 \vee M^2$. For the first expectation term in (21), the formula for the R-th moment of a Beta variable leads to

$$\begin{aligned} E_X[\tilde{P}(I_\varepsilon)^R] &= \prod_{i=0}^{l-1} \prod_{r=0}^{R-1} Q_{X,\varepsilon}(i,r), \\ Q_{X,\varepsilon}(i,r) &= \frac{a_i + N_X(I_\varepsilon^{[i+1]}) + r}{2a_i + N_X(I_\varepsilon^{[i]}) + r}. \end{aligned}$$

Let us distinguish the two regimes $i \le L_n$ and $L_n \le i \le l$. Let us write $N_X(I_\varepsilon^{[i]}) = nF_0(I_\varepsilon^{[i]}) + M_\varepsilon(i)$. When $i \le L_n$

$$\begin{aligned} Q_{X,\varepsilon}(i,r) &= \frac{F_0(I_\varepsilon^{i+1})}{F_0(I_\varepsilon^i)} \frac{1 + n^{-1}(a_i + M_\varepsilon(i+1) + r)/F_0(I_\varepsilon^{i+1})}{1 + n^{-1}(2a_i - M_\varepsilon(i) + r)/F_0(I_\varepsilon^i)} \\ &\le \frac{F_0(I_\varepsilon^{i+1})}{F_0(I_\varepsilon^i)} \Big[1 + \frac{a_i + M_\varepsilon(i+1) + r}{nF_0(I_\varepsilon^{i+1})}\Big] \Big[1 + \frac{M_\varepsilon(i)}{nF_0(I_\varepsilon^i)}\Big], \end{aligned}$$

where to obtain the last inequality we have bounded the denominator from below and used the inequality $1/(1-x) \le 1 + x$ for $x < 1$, as well as $nF_0(I_\varepsilon^i) \gtrsim n2^{-i}$ and, for any $i \le L_n$, by definition of L_n,

$$M_\varepsilon(i)\frac{2^i}{n} \le M\Big[\sqrt{\frac{i2^i}{n}} \vee \frac{i2^i}{n}\Big] \lesssim \sqrt{\frac{L_n 2^{L_n}}{n}} = o(1).$$

Deduce, for C_3, C_4 large enough constants,

$$\begin{aligned} Q_{X,\varepsilon}(i,r) &\le \frac{F_0(I_\varepsilon^{i+1})}{F_0(I_\varepsilon^i)}\Big[1 + C_3\frac{2^i a_i + \sqrt{in}2^{i/2} + 2^i R_n}{n}\Big]\Big[1 + C_3\frac{\sqrt{in}2^{i/2}}{n}\Big], \\ &\le \frac{F_0(I_\varepsilon^{i+1})}{F_0(I_\varepsilon^i)}\Big[1 + C_4\frac{2^i a_i + \sqrt{in}2^{i/2} + 2^i R_n}{n}\Big]. \end{aligned}$$

This implies that, for some $C_5 > 0$,

$$\begin{aligned} \prod_{i=0}^{L_n} Q_{X,\varepsilon}(i,r) &\le F_0(I_\varepsilon^{L_n+1})\prod_{i=0}^{L_n}\Big[1 + C_4\frac{2^i a_i + \sqrt{in}2^{i/2} + 2^i R_n}{n}\Big] \\ &\lesssim 2^{-L_n}\exp\left\{C_4\sum_{i=0}^{L_n}\frac{2^i a_i + \sqrt{in}2^{i/2} + 2^i R_n}{n}\right\} \\ &\lesssim 2^{-L_n}\exp\left\{C_5\frac{2^{L_n} a_{L_n} + \sqrt{L_n n}2^{L_n/2} + 2^{L_n} R_n}{n}\right\} \lesssim 2^{-L_n}, \end{aligned}$$

where the last exponential term is bounded due to the definitions of L_n, α_{L_n} and R_n. Now in the regime $L_n \le i \le l$,

$$\begin{aligned} Q_{X,\varepsilon}(i,r) &= \frac{a_i\Big(1 + \frac{N_X(I_\varepsilon^{[i+1]})}{a_i} + \frac{r}{a_i}\Big)}{2a_i\Big(1 + \frac{N_X(I_\varepsilon^{[i]})}{2a_i} + \frac{r}{2a_i}\Big)} \\ &\le \frac{1}{2}\Big(1 + \frac{N_X(I_\varepsilon^{[i+1]})}{a_i} + \frac{r}{a_i}\Big). \end{aligned}$$

This implies that, on $\mathcal{B}$, for some $C_6 > 0$, with $M_\varepsilon(i) \le \sqrt{in}2^{-i/2} + i$,

$$\begin{aligned} \prod_{i=L_n+1}^{l} Q_{X,\varepsilon}(i,r) &\le 2^{-(l-L_n)}\exp\left\{\sum_{i=L_n+1}^{l}\frac{nF_0(I_\varepsilon^{i+1}) + M_\varepsilon(i) + R_n}{a_i}\right\} \\ &\le 2^{-(l-L_n)}\exp\left\{C_6\sum_{i=L_n+1}^{l}\frac{n2^{-i} + \sqrt{in}2^{-i/2} + i + R_n}{a_i}\right\} \\ &\le 2^{-(l-L_n)}\exp\left\{C_6\Big(\frac{n}{L_n 2^{L_n(1+2\alpha)}} + \frac{\sqrt{n}}{\sqrt{L_n}2^{L_n(\frac{1}{2}+2\alpha)}} + \frac{2R_n}{2^{2L_n\alpha}}\Big)\right\} \\ &\lesssim 2^{-(l-L_n)}, \end{aligned}$$

using again the definition of L_n. This leads to the bound

$$E_X[\tilde{P}(I_\varepsilon)^R] \lesssim \prod_{r=0}^{R-1}(C2^{-L_n}2^{L_n - l}) \lesssim (C2^{-l})^R.$$

Conclude that, with $R = R_n = \log_2 n$ and some $c > 0$,

$$\begin{aligned} E_X\|f^{L_n^c}\|_\infty &\le \sum_{l>L_n} 2^{l/2}\left[2^l 2^{lR/2} C^R 2^{-lR}(CR/a_l)^{R/2}\right]^{1/R} \\ &\lesssim \sum_{l>L_n} 2^{cl/R_n} R_n^{1/2} a_l^{-1/2} \lesssim \sum_{l>L_n} 2^{c\frac{l}{\log_2 n}}\sqrt{\frac{\log n}{l}}2^{-l\alpha} \\ &\lesssim \sum_{L_n < l < \log_2 n} C2^{-l\alpha} + \sum_{l>\log_2 n} 2^{(\frac{c}{\log_2 n} - \alpha)l} \end{aligned}$$

For n large we have $c(\log_2 n)^{-1} \leq \nu\alpha$, for any fixed $\nu > 0$. The first sum in the last display is less than a constant times $\varepsilon^*_{n,\alpha}$ and the second sum is less than $n^{-(1-\nu)\alpha}$. By choosing $\nu < (2\alpha)/(2\alpha+1)$, the second sum is thus of smaller order. Conclude that $E_X \|f^{L_n^c}\|_\infty \lesssim \varepsilon^*_{n,\alpha}$.

Now putting together the different bounds obtained, for any $M_n \to \infty$, setting $\mathcal{T}_n := \{f : \|f - f_0\|_\infty \leq M_n \varepsilon^*_{n,\alpha}\}$ and using Markov's inequality,

$$\begin{aligned} E^n_{f_0} \Pi[\mathcal{T}_n^c \,|\, X] &\leq E^n_{f_0} \Pi[\mathcal{T}_n^c \,|\, X] \mathbb{1}_{\mathcal{B}} + E^n_{f_0} \Pi[\mathcal{T}_n^c \,|\, X] \mathbb{1}_{\mathcal{B}^c} \\ &\leq E^n_{f_0} \Pi[\mathcal{T}_n^c \cap \mathcal{A} \,|\, X] \mathbb{1}_{\mathcal{B}} + E^n_{f_0} \Pi[\mathcal{A}^c \,|\, X] + o(1) \\ &\leq M_n^{-1} {\varepsilon^*_{n,\alpha}}^{-1} E^n_{f_0} E_X[\mathbb{1}_{f \in \mathcal{A}} \|f - f_0\|_\infty] \mathbb{1}_{\mathcal{B}} + o(1) \\ &\lesssim M_n^{-1} + o(1) = o(1). \end{aligned}$$

This concludes the proof of Theorem 1. □

Lemma 1. *Let $\varepsilon \in \mathcal{E}$ with $|\varepsilon| = l$, for some $l \leq L_n$ and L_n defined by (15). Suppose, for any $i \leq l \leq L_n$, that $a_i \leq i 2^{2\alpha i}$. Then, on the event $\mathcal{B}$ defined by (16), for c_0 in (15) small enough, for n large enough,*

$$\left| \frac{\bar{P}(I_\varepsilon)}{P_0(I_\varepsilon)} - 1 \right| \leq C \left[\sum_{i=1}^{l} \frac{a_i 2^i}{n} + \sqrt{\frac{L_n 2^l}{n}} \right].$$

Proof. Notice that $\bar{P}(I_\varepsilon)$ and $P_0(I_\varepsilon)$ can be written as the products $\prod_{i=1}^l w_i$ and $\prod_{i=1}^l y_i$ respectively, with

$$w_i = \frac{a_i + N_X(I_\varepsilon^{[i]})}{2a_i + N_X(I_\varepsilon^{[i-1]})}, \qquad y_i = \frac{F_0(I_\varepsilon^{[i]})}{F_0(I_\varepsilon^{[i-1]})}.$$

On the event $\mathcal{B}$, we have $N_X(I_\varepsilon^{[i]}) = nF_0(I_\varepsilon^{[i]}) + \delta_{i,\varepsilon}$ where $\delta_{i,\varepsilon}$ is controlled below. That is,

$$w_i = y_i \frac{1 + n^{-1}(a_i + \delta_{i,\varepsilon})/F_0(I_\varepsilon^{[i]})}{1 + n^{-1}(2a_i + \delta_{i-1,\varepsilon})/F_0(I_\varepsilon^{[i-1]})}.$$

By definition of $\mathcal{B}$, for $l \leq L_n$ we have $|\delta_{i,\varepsilon}| \leq C\sqrt{nL_n 2^{-i}}$. Since $L_n 2^{L_n} = o(n)$, this bound is always of smaller order than $n2^{-i} \lesssim nF_0(I_\varepsilon^{[i]})$, since f_0 is bounded away from 0. So the denominator of the last expression is bounded away from 0. Deduce, for any $1 \leq i \leq L_n$,

$$\left| \frac{w_i}{y_i} - 1 \right| \leq C \left[\frac{a_i 2^i}{n} + \sqrt{\frac{L_n 2^i}{n}} \right].$$

For large n and c_0 in (15) small enough, the last display is smaller than 1. Also,

$$\sum_{i=1}^{l} \left| \frac{w_i}{y_i} - 1 \right| \leq C \left[\sum_{i=1}^{l} \frac{a_i 2^i}{n} + \sqrt{\frac{L_n 2^l}{n}} \right],$$

which remains bounded. An application of Lemma 3 completes the proof. □

Lemma 2. *Let $\varepsilon \in \mathcal{E}$ with $|\varepsilon| = l$ and $I_\varepsilon = I_k^l$, for some admissible indexes l, k. Then, on the event $\mathcal{B}$ defined by (16), for any $l \le L_n$,*

$$\Big|\bar{Y}_{\varepsilon 0} - \frac{F_0(I_{\varepsilon 0})}{F_0(I_\varepsilon)}\Big| \le C\frac{2^{l/2}}{n}\Big(2^l a_{l+1}|f_{0,lk}| + \sqrt{nL_n}\Big).$$

Proof. Similar to the proof of Lemma 1, let $N_X(I_{\varepsilon 0}) = nF_0(I_{\varepsilon 0}) + \delta_{l+1,\varepsilon}$, so that

$$\begin{aligned}\Big|\frac{\bar{Y}_{\varepsilon 0}}{y_{\varepsilon 0}} - 1\Big| &= \Big|\frac{1 + (a_{l+1} + \delta_{l+1,\varepsilon})/nF_0(I_{\varepsilon 0})}{1 + (2a_{l+1} + \delta_{l,\varepsilon})/nF_0(I_\varepsilon)} - 1\Big| \\ &\le C\frac{a_{l+1}}{n}|F_0(I_{\varepsilon 0})^{-1} - 2F_0(I_\varepsilon)^{-1}| + \frac{C}{n}\Big(\frac{|\delta_{l,\varepsilon}|}{F_0(I_\varepsilon)} + \frac{|\delta_{l+1,\varepsilon}|}{F_0(I_{\varepsilon 0})}\Big) \\ &\le C\frac{2^{2l}a_{l+1}}{n}2^{-l/2}|f_{0,lk}| + C\frac{2^{l/2}(nL_n)^{1/2}}{n},\end{aligned}$$

on $\mathcal{B}$, where we have used the bound $|\delta_{l,\varepsilon}| + |\delta_{l+1,\varepsilon}| \le C(nL_n 2^{-l})^{1/2}$. □

Lemma 3. *Let $\{y_i\}_{1\le i\le L}, \{w_i\}_{1\le i\le L}$ be two sequences of positive real numbers such that there are constants c_1, c_2 with*

$$\max_{1\le i\le L}\Big|\frac{w_i}{y_i} - 1\Big| \le c_1 < 1, \qquad \sum_{i=1}^{L}\Big|\frac{w_i}{y_i} - 1\Big| \le c_2 < \infty.$$

Then there exists c_3 depending on c_1, c_2 only such that

$$\Big|\prod_{i=1}^{L}\frac{w_i}{y_i} - 1\Big| \le c_3\sum_{i=1}^{L}\Big|\frac{w_i}{y_i} - 1\Big|.$$

Proof. It suffices to bound $e^\zeta - 1$ from above and below, where $\zeta = \sum \log(w_i/y_i)$. For the upper bound, one uses $\log(1+u) \le |u|$ followed by $e^{|v|} - 1 \le e^{c_2}|v|$ for $|v| \le c_2$. For the lower bound, one uses $\log(1+u) \ge -(1-c_1)^{-1}|u|$ if $|u| \le c_1 < 1$ followed by $e^{-C|u|} - 1 \ge -C|u|$. □

3.3 Proof of Theorem 3

Proof. By Lemma 8, it is enough to check that finite-dimensional projections converge (28), and the tightness-type property (29) at rate $1/\sqrt{n}$.

Finite-dimensional projections. First, let us formulate the problem in terms of convergence for histograms.

The finite-dimensional subspace V_J is $\mathrm{Vect}\{\varphi, \psi_{lk}, 0 \le k < 2^l, l \le J\}$. Note that, if $K = J + 1$, it coincides with the space of all histograms on the dyadic regular grid of $[0,1]$ of meshwidth 2^{-K}. So, if $J_i = ((i-1)2^{-K}, i2^{-K})$, one also has $V_J = \mathrm{Vect}\{2^K \mathbb{1}_{J_i}, 1 \le i \le 2^K\}$ and $\pi_{V_J} f$ has the explicit expression

$$\pi_{V_J} f = 2^K \sum_{i=1}^{2^K}\Big(\int_{J_i} f\Big)\mathbb{1}_{J_i} = 2^K \sum_{i=1}^{2^K} F(J_i)\mathbb{1}_{J_i}.$$

The distribution of $\pi_{V_J} f$ can be equivalently specified by the joint distribution of $(F(J_1), \dots, F(J_{2^K}))$.

Below we show that if f is a draw from the posterior distribution $\Pi[\cdot \,|\, X]$,

$$\Big(\sqrt{n}(\int_{J_i} f - \mathbb{P}_n(J_i))\Big)_{1\le i\le 2^K} \to (\mathbb{G}_{P_0} 1\!\!1_{J_i}))_{1\le i\le 2^K}\,, \tag{22}$$

where the convergence is in distribution in $\mathbb{R}^{2^K}$, in probability under P_0, and $\mathbb{P}_n(I) = N_X(I)/n$ is the mass the empirical measure associated to the data $X_1, \ldots, X_n$ puts on an interval I.

To prove (22), we exhibit a parametric model where the same distributions as in (22) arise, and where the convergence holds under P_0. Set $\Theta \equiv \mathcal{S}_{2^K} = \{(\theta_1, \ldots, \theta_{2^K}) \in (0,1)^{2^K},\ \sum_i \theta_i = 1\}$ the interior of the unit simplex in $\mathbb{R}^{2^K}$. Consider the parametric model

$$\mathcal{P} \equiv \mathcal{P}_K = \Big\{P = P_g,\quad g = 2^K \sum_{i=1}^{2^K} g_i 1\!\!1_{J_i},\quad (g_1, \ldots, g_{2^K}) \equiv \theta \in \Theta\Big\}.$$

It consists of positive densities that are regular histograms with 2^K bins. As usual the unit simplex Θ can be identified to the subset of $\mathbb{R}^{2^K-1}$ consisting of $(\theta_1, \ldots, \theta_{2^K-1})$ such that $0 < \theta_i < 1$ for all $1 \le i \le 2^K - 1$ and

$$\sum_{i=1}^{2^K-1} \theta_i < 1.$$

Define a prior distribution Π_K on Θ viewed as a subset of $\mathbb{R}^{2^K-1}$ by 'cutting' the Pólya tree distribution at level K. That is, define the joint law of $(g_i)_{1\le i\le 2^K}$ as the joint law of $(F(J_i))_{1\le i\le 2^K}$, where F is sampled from a Pólya tree with the prescribed parameters.

The algebraic expression, given data $X_1, \ldots, X_n$, of the induced posterior distribution on $(F(J_1), \ldots, F(J_{2^K}))$ in the original model with the original Pólya tree prior, and the posterior distribution of $(g_1, \ldots, g_{2^K})$ in model $\mathcal{P}_K$ with prior Π_K, are the same: this follows by the conjugacy properties of the beta-distributions with respect to the likelihood, which is of multinomial type. The posterior distribution has, under both models, a tree-type structure: the posterior of $F(I_\varepsilon)$ has same law as a product of $\tilde{Y}_\varepsilon^{[i]}$'s, $i \le K$, which are Beta variables with updated parameters $\alpha_\varepsilon^* = \alpha_\varepsilon + N_X(I_\varepsilon)$. In particular, note that the joint posterior distribution of $(F(J_1), \ldots, F(J_{2^K}))$ only depends on the data through the counts $N_X(J_i)$, for $1 \le i \le 2^K$.

Now, the counts $(N_X(J_i), 1 \le i \le 2^K)$, have the same distribution under $X \sim P_{f_0}$, the original true model, and under $X \sim P_{f_0^{[K]}}$, where

$$f_0^{[K]} = 2^K \sum_{i=1}^{2^K} F_0(J_i) 1\!\!1_{J_i}$$

is the L^2-projection of f onto $\mathcal{P}_K$. This is because the counts are multinomially distributed with parameters $F_0(J_i) = F_0^{[K]}(J_i)$, both under P_{f_0} and $P_{f_0^{[K]}}$.

Deduce that the induced posterior distribution on $\mathcal{P}_K$ has same law under P_{f_0} as the posterior in model $\mathcal{P}_K$ with prior Π_K and under $P_{f_0^{[K]}}$. To the latter distribution one can apply the parametric Bernstein-von Mises theorem, as

stated e.g. in [42] Chapter 10, and we now check the corresponding assumptions. The model $\mathcal{P}_K$ is smoothly parameterised, and in particular differentiable in quadratic mean. The testing condition (10.3) from [42] is easily verified using Hellinger-type tests: denoting by P_θ an arbitrary element of $\mathcal{P}_K$, one checks that for any $\theta, \theta' \in \Theta$, the squared-Hellinger distance between P_θ and $P_{\theta'}$ verifies $(\theta - \theta')^2 \lesssim h^2(P_\theta, P_{\theta'}) \lesssim (\theta - \theta')^2$. In this context, the existence of appropriate Hellinger tests follows from the works by Le Cam [32] and Birgé [3], see e.g. [4] Corollary 1. Finally, the prior Π_K has a positive density in a neighborhood of θ_0, the element of the simplex corresponding to $f_0^{[K]}$, since all parameters α_ε are strictly positive. Conclude that the posterior in model $\mathcal{P}_K$ converges in the total variation distance

$$\|\Pi_K[\cdot \,|\, X] \circ \tau^{-1} - N(0, \Gamma_K)\|_{TV} \to^{P_{f_0^{[K]}}} 0,$$

where Γ_K is the inverse Fisher information matrix in the model $\mathcal{P}_K$ at θ_0, the element of the simplex in $\mathbb{R}^{2^K}$ with coordinates $\theta_{0,j} = F_0(J_i)$ (recall that one may identify the simplex Θ with a subset of $\mathbb{R}^{2^K-1}$, dropping out the last coordinate, as usual; also, $(\Gamma_K)_{i,j} = \theta_{0,i} \mathbb{1}_{i=j} - \theta_{0,i}\theta_{0,j}$) and

$$\tau : \theta \to \sqrt{n}(\theta - \hat{\theta}_K),$$

with $\hat{\theta}_K := (N_X(J_1)/n, \ldots, N_X(J_{2^K-1})/n) = (\mathbb{P}_n(J_1), \ldots, \mathbb{P}_n(J_{2^K-1}))$.

Deduce that, in the model $\mathcal{P}_K$, for any real numbers $b_1, \ldots, b_{2^K}$, the posterior distribution of

$$\sqrt{n} \sum_{i=1}^{2^K} b_i(\theta_i - \hat{\theta}_i) = \sqrt{n} \sum_{i=1}^{2^K-1} (b_i - b_{2^K})(\theta_i - \hat{\theta}_i)$$

converges to $(b_\cdot - b_{2^K})^T \Gamma_K (b_\cdot - b_{2^K})$. This coincides with

$$E_{P_0}(\mathbb{G}_{P_0} \sum_{i=1}^{2^K} b_i \mathbb{1}_{J_i})^2 = \mathrm{Var}_{P_0}(\sum_{i=1}^{2^K} b_i \mathbb{1}_{J_i}).$$

Indeed, rewriting the expression using $1 = \sum_{i=1}^{2^K} \mathbb{1}_{J_i}$,

$$\begin{aligned}\mathrm{Var}_{P_0}(\sum_{i=1}^{2^K} b_i \mathbb{1}_{J_i}) &= \mathrm{Var}_{P_0}(\sum_{i=1}^{2^K-1} b_i \mathbb{1}_{J_i} + b_{2^K}(1 - \sum_{i=1}^{2^K-1} \mathbb{1}_{J_i})) \\ &= \sum_{1 \le i,j \le 2^K-1} (b_i - b_{2^K})(b_j - b_{2^K}) \mathrm{Cov}_{P_0}(\mathbb{1}_{J_i}, \mathbb{1}_{J_j}),\end{aligned}$$

where

$$\begin{aligned}\mathrm{Cov}_{P_0}(\mathbb{1}_{J_i}, \mathbb{1}_{J_j}) &= \int (\mathbb{1}_{J_i} - P_0(J_i))(\mathbb{1}_{J_i} - P_0(J_i)) dP_0 \\ &= \mathbb{1}_{i=j}\theta_{0,i} - \theta_{0,i}\theta_{0,j} = (\Gamma_K)_{i,j},\end{aligned}$$

recalling that here θ_0 is the vector of $\theta_{0,i} = F_0(J_i) = P_0 \mathbb{1}_{J_i}$, which leads to

$$\mathrm{Var}_{P_0}(\sum_{i=1}^{2^K} b_i \mathbb{1}_{J_i}) = (b_\cdot - b_{2^K})^T \Gamma_K (b_\cdot - b_{2^K}).$$

By Cramér-Wold, this shows that the left hand-side of (22) converges in distribution to a centered normal limit, with the same covariance structure as that of the right-hand-side of (22), in P_0-probability. This establishes (28), with centering T_n given by (11). Instead of centering at the empirical counts, one can also center at the posterior mean, as can be checked by a simple computation (this also follows from the bound on $\bar{f}_{lk} - \hat{f}_{lk}$ obtained below and applied for finite $l \leq J$).

Tightness. One now needs to check (29). We will exploit several intermediate results obtained along the proof of Theorem 1. Those are obtained under a specific choice of a_l that depends on the regularity of f_0. Nevertheless, it is easy to check that most statements in that proof remain valid when a_l is one of the two sequences in (9), provided the cut-off level L_n is redefined, for each choice as in (9), as in (12)-(13), namely

$$2^{L_n} := J_n := \lfloor n^{\frac{1}{1+2\delta}} \rfloor \quad \text{or} \quad 2^{L_n} := J_n := \lfloor \Big(\frac{n}{\log n} \Big)^{\frac{1}{1+2\delta}} \rfloor$$

respectively. The events $\mathcal{B}$ on the data-space and $\mathcal{A}$ from the proof of Theorem 1 are redefined accordingly, using this definition of L_n.

In particular, we will repeatedly use the bound, established in the proof of Theorem 1, on the set $\mathcal{A}$ and on $\mathcal{B}$, for any $l \leq L_n$,

$$\left| \frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1 \right| \lesssim \sqrt{\frac{L_n 2^l}{n}} \lesssim \sqrt{\frac{L_n 2^{L_n}}{n}} = o(1). \tag{23}$$

Note that establishing this bound did not require any specific smoothness conditions on f_0.

We apply Lemma 8 below. First we note that one can work with the posterior conditioned to the set $\mathcal{A}$, that is $\Pi[\cdot \,|\, X, \mathcal{A}]$. This is allowed thanks to Remark 2 below Lemma 8. Along the proof, one can go back to the original posterior by using $\Pi[\cdot \,|\, X, \mathcal{A}] = \Pi[\cdot \cap \mathcal{A} \,|\, X] \Pi[\mathcal{A} \,|\, X]^{-1}$. As $\Pi[\mathcal{A} \,|\, X] = 1 + o_P(1)$, this does not affect the following argument. To simplify the notation, in the sequel we omit the conditioning on $\mathcal{A}$ when writing posterior quantities.

The quantity under expectation on the left-hand side of (29) in Lemma 8 is $\|f - \bar{f}^{L_n}\|_{\mathcal{M}_0(z)}$, that we split into $\|f^{L_n} - \bar{f}^{L_n}\|_{\mathcal{M}_0(z)}$ and $\|f^{L_n^c}\|_{\mathcal{M}_0(z)}$. We have, for say $M > 1$, and E_X denoting expectation under the posterior,

$$\begin{aligned}
&E_X \left[\sqrt{n} \max_{l \leq L_n} z_l^{-1} \max_k |f_{lk} - \bar{f}_{lk}| \right] \\
&\quad \leq M + \int_M^\infty \Pi \left[\sqrt{n} \max_{l \leq L_n} z_l^{-1} \max_k |f_{lk} - \bar{f}_{lk}| > u \,|\, X \right] du \\
&\quad \leq M + \sum_{l \leq L_n, k} \int_M^\infty \Pi \left[|f_{lk} - \bar{f}_{lk}| > u z_l / \sqrt{n} \,|\, X \right] du.
\end{aligned}$$

The difference $f_{lk} - \bar{f}_{lk}$ can be bounded in terms of $\tilde{P}(I_\varepsilon)$ and Ys: using the identities linking the function f to the variables Ys obtained in the proof of

Theorem 1, one obtains

$$
\begin{aligned}
|f_{lk} - \bar{f}_{lk}| &\leq \left|1 + \left[\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right]\right| 2^{l/2+1}\bar{P}(I_\varepsilon)|\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}| + \left|\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right||\bar{f}_{lk}| \\
&\lesssim 2^{-l/2}|\bar{Y}_{\varepsilon 0} - \tilde{Y}_{\varepsilon 0}| + \left|\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right||\bar{f}_{lk}| \\
&=: \qquad (a) \qquad + \qquad (b),
\end{aligned}
$$

where to bound the first term we have used (23) and the fact that $\bar{P}(I_\varepsilon) \lesssim 2^{-l}$ holds on $\mathcal{B}$ thanks to Lemma 1 (which, again, holds for the adapted choices of (a_l) and L_n as above, and α in the statement replaced by δ). We now bound (a) and (b) successively.

On the event $\mathcal{B}$, the variable $\tilde{Y}_{\varepsilon 0}$ is Beta-distributed, with parameters that are within constants of $nF_0(I_{\varepsilon 0})$ and $nF_0(I_{\varepsilon 1})$ respectively. Both are thus bounded above and below by multiples of $n2^{-|\varepsilon|} = n2^{-l}$. It now follows from Lemma 6 that, for $u \geq M > 1$,

$$
\begin{aligned}
&\Pi\left[(a) \geq \frac{z_l}{2\sqrt{n}}u \mid X\right] \\
&\leq \Pi\left[|\tilde{Y}_{\varepsilon 0} - \bar{Y}_{\varepsilon 0}| \geq \frac{2^{\frac{l}{2}}z_l}{4\sqrt{n}}u + \frac{2^{\frac{l}{2}}z_l}{4\sqrt{n}} \mid X\right] \\
&\leq \Pi\left[|\tilde{Y}_{\varepsilon 0} - \bar{Y}_{\varepsilon 0}| \geq \frac{2^{\frac{l}{2}}z_l}{4\sqrt{n}}u + \frac{2}{n2^{-l}} \mid X\right] \\
&\lesssim e^{-Cz_l^2u^2},
\end{aligned}
$$

where the second inequality holds because $z_l \gtrsim 8/(n2^{-l})^{1/2}$ uniformly in $l \leq L_n$, for n large enough.

We now deal with the term (b) and use the following intermediate bound obtained in the proof of Theorem 1, on $\mathcal{B}$,

$$
|\bar{f}_{lk}| \leq |f_{0,lk}|\left\{\frac{2^l a_{l+1}}{n} + \sqrt{\frac{L_n 2^l}{n}}\right\} + \sqrt{\frac{L_n}{n}}.
$$

Note that the last bound is always smaller than $C2^{-l/2}$ for $l \leq L_n$. So, when we evaluate the posterior probability

$$
\Pi\left[\left|\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right||\bar{f}_{lk}| > \frac{z_l}{2\sqrt{n}}u \mid X\right],
$$

one can assume that $u \leq \sqrt{n}c_0 2^{-l/2}z_l^{-1}$, for c_0 an arbitrarily small fixed constant, otherwise the posterior probability in the last display is 0 (recall that (23) implies that the term in factor of $|\bar{f}_{lk}|$ in the last display goes to 0). Denote

$$
\mathcal{A}_i(u) = \left\{f : \ |\tilde{Y}_\varepsilon^{[i]}(f) - \bar{Y}_\varepsilon^{[i]}| \leq uz_l\sqrt{\frac{2^i}{n}}\right\}.
$$

Denoting Pr as a shorthand for $\Pi[\cdot \,|\, X]$,

$$\begin{aligned}
\Pr\Big(\bigcap_{i\le l}\mathcal{A}_i(u)\Big) &\le \Pr\Big(f\in\bigcap_{i\le l}\mathcal{A}_i(u),\quad \sum_{i\le l}|\tilde{Y}_\varepsilon^{[i]}(f)-\bar{Y}_\varepsilon^{[i]}|\le C\frac{2^{\frac{l}{2}}z_l}{2\sqrt{n}}u\Big)\\
&\le \Pr\Big(f\in\bigcap_{i\le l}\mathcal{A}_i(u),\quad \sum_{i\le l}\Big|\frac{\tilde{Y}_\varepsilon^{[i]}(f)}{\bar{Y}_\varepsilon^{[i]}}-1\Big|\le C'\frac{2^{\frac{l}{2}}z_l}{2\sqrt{n}}u\Big)\\
&\le \Pr\Big(\Big|\prod_{i=1}^{l}\frac{\tilde{Y}_\varepsilon^{[i]}(f)}{\bar{Y}_\varepsilon^{[i]}}-1\Big|\le C'\frac{2^{\frac{l}{2}}z_l}{2\sqrt{n}}u\Big),
\end{aligned}$$

where we have used that $\bar{Y}_\varepsilon^{[i]}$ is bounded away from 0 and 1, as follows from Lemma 2, and for the last inequality we have used Lemma 3 together with the fact that $uz_l(2^i/n)^{1/2}\le c_0$ can be made as small as desired for c_0 small enough. This implies, using again that $|\bar{f}_{lk}|\lesssim 2^{-l/2}$, that

$$\begin{aligned}
&\Pi\left[(b)>\frac{z_l}{2\sqrt{n}}u\,|\,X\right]\\
&\quad\le \Pi\left[\left|\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)}-1\right|>\frac{2^{\frac{l}{2}}z_l}{2\sqrt{n}}u\,|\,X\right]\\
&\quad\le \sum_{i=1}^{l}\Pi[\mathcal{A}_i(u)^c\,|\,X]\le le^{-cz_l^2u^2},
\end{aligned}$$

as for the term (a) above.

Combining the obtained bounds on (a) and (b) leads to, for some $c>0$,

$$\begin{aligned}
&E_X\left[\sqrt{n}\max_{l\le L_n}z_l^{-1}\max_k|f_{lk}-\bar{f}_{lk}|\right]\\
&\quad\lesssim M+\sum_{l\le L_n,k}l\int_M^\infty e^{-cz_l^2u^2}du\\
&\quad\lesssim M+\sum_{l\le L_n}l2^le^{-cz_l^2M^2}.
\end{aligned}$$

The last quantity is bounded as soon as M is chosen large enough (note that the previous bound holds in probability, since we work on the event $\mathcal{B}$ and have restricted to the set $\mathcal{A}$ with $\Pi[\mathcal{A}\,|\,X]=1+o_P(1)$).

Now we focus on bounding the part $l>L_n$. Proceeding as in the proof of Theorem 1, one can obtain bounds for $E_X[\max_k|f_{lk}|]$. First, using Lemma 5,

$$E_X[|1-2\tilde{Y}_{\varepsilon 0}|^R\lesssim\left(\frac{CR}{a_l}\right)^{\frac{R}{2}}+\left(C\frac{n2^{-l(1+\alpha)}+M_l}{a_l}\right)^R.$$

From this via (21) and as in the proof of Theorem 1 it follows

$$E_X[\max_k|f_{lk}|]\lesssim 2^{-l/2}\left[\left(\frac{l}{a_l}\right)^{1/2}+\frac{n2^{-l(1+\alpha)}+M_l}{a_l}\right]2^{cl/\log_2 n}. \tag{24}$$

We distinguish the two cases $\delta\le\alpha$ and $\delta>\alpha$. In the undersmoothing case $\delta\le\alpha$, the first term in the last bracket dominates, as $n2^{-l(1+\alpha)}\lesssim M_l$ and

$M_l \le (la_l)^{1/2}$, since both $\delta \le \alpha$ and $l > L_n$. From this we directly deduce that, on the event $\mathcal{B}$, when $\delta \le \alpha$, and for any $\{z_l\}$ with $z_l \ge \sqrt{l}$,

$$\begin{aligned} E_X &\left[\sqrt{n} \max_{l>L_n} z_l^{-1} \max_k |f_{lk}| \right] \\ &\lesssim \sqrt{n} \sum_{l>L_n} z_l^{-1} \sqrt{l} 2^{-l/2} a_l^{-1/2} 2^{cl/\log_2 n} \\ &\lesssim \sqrt{n} \sum_{l>L_n} 2^{-l/2} a_l^{-1/2}, \end{aligned}$$

which is bounded by 1 for both choices of (a_l) in (9) given our choice of L_n. Also, as $E_X f_{lk} = \bar{f}_{lk}$ and

$$\sqrt{n} \max_{l>L_n} z_l^{-1} \max_k |\bar{f}_{lk}| \le E_X \left[\sqrt{n} \max_{l>L_n} z_l^{-1} \max_k |f_{lk}| \right],$$

the bound in the last but one display also holds with a different constant when f_{lk} is replaced by $f_{lk} - \bar{f}_{lk}$.

In the oversmoothing case $\delta > \alpha$, the first term in the bracket in (24) dominates if $n2^{-l(1+\alpha)} + M_l = o(a_l^{1/2})$. This is the case for $l \ge \lambda_n$, for $\lambda_n := C \log_2 n$ with C large enough. So for $l \ge \lambda_n$, one can use the same argument as in the case $\delta \le \alpha$. For $L_n \le l \le \lambda_n$, one should work with $f_{lk} - \bar{f}_{lk}$ as a whole instead of separating both terms. It follows from (19) that

$$\begin{aligned} f_{lk} - \bar{f}_{lk} &= -2^{l/2+1} \tilde{P}(I_\varepsilon)(\tilde{Y}_{\varepsilon 0} - \bar{Y}_{\varepsilon 0}) \; + \; 2^{l/2} (\tilde{P}(I_\varepsilon) - \bar{P}(I_\varepsilon))(1 - 2\bar{Y}_{\varepsilon 0}) \\ &= -2^{l/2+1} \tilde{P}(I_\varepsilon)(\tilde{Y}_{\varepsilon 0} - \bar{Y}_{\varepsilon 0}) \; + \; \left[\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1 \right] \bar{f}_{lk} \\ &= \qquad\qquad (i) \qquad\qquad + \qquad\qquad (ii). \end{aligned}$$

As $\bar{Y}_{\varepsilon 0} = E_X[\tilde{Y}_{\varepsilon 0}]$, the term (i) nearly coincides with f_{lk}, except that the bias has been substracted from $\tilde{Y}_{\varepsilon 0}$, so one can use (21) combined with the estimate of $E|Y - EY|^R$ obtained in Lemma 5. This leads to the same estimate as in the undersmoothing case, that is

$$E_X \left[\sqrt{n} \max_{l>L_n} z_l^{-1} \max_k |(i)| \right] \lesssim \sqrt{n} \sum_{l>L_n} 2^{-l/2} a_l^{-1/2} \lesssim 1.$$

For the term (ii), to control the term $\tilde{P}(I_\varepsilon)/\bar{P}(I_\varepsilon) - 1$, one proceeds as in the proof of Theorem 1, extending the event $\mathcal{A}$ to an event $\mathcal{A}'$ on which, for T large enough to be chosen below,

$$|\tilde{Y}_\varepsilon^{[i]} - \bar{Y}_\varepsilon^{[i]}| \le T \frac{L_n^{1/2}}{\sqrt{nF_0(I_\varepsilon^{[i]}) + a_i}} =: \rho_\varepsilon^{[i]},$$

for any $i \le |\varepsilon|$ and any ε such that $|\varepsilon| \le \lambda_n$. As in the proof of Theorem 1, one checks that, on the event $\mathcal{B}$,

$$\Pi[(\mathcal{A}')^c \,|\, X] \lesssim \sum_{l \le \lambda_n} 2^l e^{-CT^2 \log n} = o(e^{-cT^2 \log n})$$

for T large enough, as well as the fact that on $\mathcal{A}'$ and on the event $\mathcal{B}$,

$$\left|\frac{\tilde{P}(I_\varepsilon)}{\bar{P}(I_\varepsilon)} - 1\right| \lesssim \sum_{i=0}^{l-1} \rho_\varepsilon^{[i]} \lesssim \sum_{i=0}^{l-1} \frac{L_n^{1/2}}{\sqrt{n2^{-i} + a_i}} \lesssim \sqrt{\frac{L_n 2^{L_n}}{n}},$$

since we are in the regime $l > L_n$ (one also uses the fact that $\bar{Y}_\varepsilon^{[i]}$ is bounded away from 0 and 1 for both $i \leq L_n$ and $i > L_n$). Using the expression (ii) one deduces

$$\begin{aligned} E_X\left[\sqrt{n} \max_{L_n<l<\lambda_n} z_l^{-1} \max_k |(ii)| \mathbb{1}_{\mathcal{A}'}\right] &\lesssim \sqrt{n}\sqrt{\frac{L_n 2^{L_n}}{n}} \max_{L_n<l<\lambda_n} z_l^{-1} \max_k |\bar{f}_{lk}| \\ &\leq \sqrt{L_n 2^{L_n}} E_X\left[\max_{L_n<l<\lambda_n} z_l^{-1} \max_k |f_{lk}|\right] \end{aligned}$$

Next one bounds the maximum in l by the sum and uses the general bound (24). This shows that the display is a $o(1)$. Finally, using the rough bound $(ii) \leq C2^l \leq C2^{\lambda_n}$ and bounding probabilities and $\bar{Y}_\varepsilon$ by 1, one gets on $\mathcal{B}$

$$E_X\left[\sqrt{n} \max_{L_n<l<\lambda_n} z_l^{-1} \max_k |(ii)| \mathbb{1}_{\mathcal{A}'^c}\right] \lesssim \sqrt{n} 2^{\lambda_n} \Pi[(\mathcal{A}')^c \,|\, X] = o(1)$$

as $\Pi[(\mathcal{A}')^c \,|\, X]$ can be made an arbitrary large power of n^{-1} by choosing T large enough.

Gathering the bounds obtained for $l \leq L_n$ and $l > L_n$ leads to (29) with centering at the posterior mean. This proves the first assertion of Theorem 3.

To derive the second assertion of Theorem 3, first note that $\langle T_n, \psi_{lk}\rangle = \langle P_n, \psi_{lk}\rangle$ equals $\hat{f}_{lk} := 2^{l/2}(\hat{F}(I_{\varepsilon 1}) - \hat{F}(I_{\varepsilon 0})) := 2^{l/2}(N_X(I_{\varepsilon 1}) - N_X(I_{\varepsilon 0}))$, for $\varepsilon = \varepsilon(l,k)$ and $l \leq L_n$. It suffices to show that $\|\bar{f}^{L_n} - \hat{f}^{L_n}\|_{\mathcal{M}_0(z)}$ is a $o_P(n^{-1/2})$.

Given indexes l, k, and $\varepsilon = \varepsilon(l,k)$, we have $\bar{f}_{lk} = 2^{l/2}\bar{P}(I_\varepsilon)(1 - 2\bar{Y}_{\varepsilon 0})$. On the other hand,

$$\hat{f}_{lk} = 2^{l/2}(\hat{F}(I_{\varepsilon 1}) - \hat{F}(I_{\varepsilon 0})) = 2^{l/2}\hat{F}(I_\varepsilon)(1 - 2\frac{\hat{F}(I_{\varepsilon 0})}{\hat{F}(I_\varepsilon)}).$$

One controls the difference $\bar{f}_{lk} - \hat{f}_{lk}$ in a similar way as for $\bar{f}_{lk} - f_{0,lk}$ in the proof of Theorem 1. Similar considerations as in Lemmas 1 and 2, but this time with $\bar{F}(I_\varepsilon)$ playing the role of $P_0(I_\varepsilon)$, lead to, on the event $\mathcal{B}$ as before,

$$\begin{aligned} |\bar{f}_{lk} - \hat{f}_{lk}| &\lesssim \frac{2^l}{n} a_l |\bar{f}_{lk}| + 2^{\frac{l}{2}} \hat{F}(I_k^l) 2^{\frac{3l}{2}} n a_{l+1} \sqrt{\frac{L_n}{n}} \\ &\lesssim \frac{2^l}{n} a_l |\bar{f}_{lk}| + (2^{-l} + \sqrt{\frac{L_n}{n}}) \frac{2^{2l}}{n} a_{l+1} \sqrt{\frac{L_n}{n}} \\ &\lesssim \frac{2^l}{n} a_l |f_{0,lk}| + \frac{2^l}{n} a_{l+1} \sqrt{\frac{L_n}{n}}, \end{aligned}$$

where we have used that $|\bar{f}_{lk}| \lesssim (2^l a_l/n)|f_{0,lk}| + \sqrt{L_n/n}$ on $\mathcal{B}$, as in the proof of Theorem 1.

Consider the case $a_l = 2^{2l\delta}$, the case $a_l = l2^{2l\delta}$ being dealt with similarly. The last term in the above display verifies,

$$\max_{l \leq L_n}\left[z_l^{-1} \frac{2^l}{n} a_{l+1} \sqrt{\frac{L_n}{n}}\right] \leq z_{L_n}^{-1} \frac{2^{L_n}}{n} a_{L_n+1} \sqrt{\frac{L_n}{n}} = o(n^{-1/2}).$$

As $|f_{0,lk}| \lesssim 2^{-l(1/2+\alpha)}$, deduce, when $\delta \leq \alpha$, on the event $\mathcal{B}$,

$$\|\bar{f}^{L_n} - \hat{f}^{L_n}\|_{\mathcal{M}_0(z)} \lesssim \max_{l \leq L_n} \left[z_l^{-1} \frac{2^l}{n} 2^{2l\delta} 2^{-l(\frac{1}{2}+\delta)} \right] \lesssim \frac{2^{L_n(\frac{1}{2}+\delta)}}{\sqrt{L_n} n} = o(n^{-1/2}). \quad \square$$

4 Appendix: remaining proofs

4.1 Proof of Proposition 1

One follows the steps of the proof of Theorem 1. The cut-off is taken equal to the cut-off l_n in (12). For low frequencies $l \leq l_n$, one uses the same arguments as in the proof of Theorem 1 with the new cut-off, similar to what is done in the proof of Theorem 3. For high-frequencies, one separates $f_0^{l_n^c}$ and $f^{l_n^c}$. For the latter, one uses Lemma 5 and this time both terms on the right-hand side of the inequality in Lemma 5 matter, depending on how large l is. The proof is largely similar to that of Theorem 1, so details are left to the reader.

4.2 Proof of Theorem 2

Proof. One applies Theorem 4 in [7], in the space $\mathcal{M}_0(z)$, where we take the sequence (z_l) to be $z_l = 2^{l/2}/l^2$, and the centering $T_n = \hat{f}^{L_n}$, with $\hat{f}$ defined in the proof of Theorem 3 above. To do so, let us check that the conditions of Theorem 4 in [7] are satisfied. By definition $\sum_l z_l 2^{-l/2}$ is finite. Also, it follows from (the proof of) Theorem 3 that the posterior recentered at $\bar{f}_n^{L_n}$ satisfies the Bernstein-von Mises theorem in $\mathcal{M}_0(z)$. One now checks that, for the above choice of z_l, one has $\|\bar{f}^{L_n} - \hat{f}^{L_n}\|_{\mathcal{M}_0(z)} = o_P(1)$. This is done in a similar way as in the proof of Theorem 3. The difference is in the estimate involving f_0: the last estimate in that proof then becomes

$$\max_{l \leq L_n} \left[z_l^{-1} \frac{2^l}{n} 2^{2l\delta} 2^{-l(\frac{1}{2}+\alpha)} \right] \lesssim \frac{2^{L_n(\frac{1}{2}+\delta)}}{n} L_n^2 2^{L_n(\delta-\alpha-\frac{1}{2})}.$$

As $\delta < \alpha + 1/2$, this bound is a $o_P(n^{-1/2})$, which leads to the first statement of Theorem 2. The second statement follows from the fact that, by a direct computation one can show that $\sqrt{n}\|\hat{F}_n^{L_n} - F_n\|_\infty$ is a $o_P(1)$ whenever $\delta < \alpha + 1/2$, as in the proof leading to Remark 9 in [20]. $\square$

4.3 An event of small probability

Let $\mathcal{B}_l$ be the collection of events defined by, for $l \geq 0$ and a sequence $L_n \to \infty$,

$$\mathcal{B}_l = \left\{ \max_{0 \leq k < 2^l} |N_X(I_k^l) - nF_0(I_k^l)| \leq M(\sqrt{l + L_n}\sqrt{\frac{n}{2^l}} \vee (l + L_n)) \right\}$$

Lemma 4. *Let $X_1, \ldots, X_n$ be i.i.d. of density f_0 on $[0,1]$, with f_0 bounded away from 0 and infinity. Then for M large enough and any $L_n \to \infty$, as $n \to \infty$,*

$$P_{f_0}^n \left[\bigcup_{l \geq 0} \mathcal{B}_l^c \right] = o(1).$$

Proof. For given fixed indexes $k, l \geq 1$, Bernstein's inequality applied to the variables $1\!\!1_{\{X_i \in I_k^l\}}$ gives the bound, for $y > 0$,

$$P_{f_0}^n \left[|N_X(I_k^l) - nF_0(I_k^l)| > ny \right] \leq 2 \exp\left(- \frac{n^2 y^2/2}{nF_0(I_k^l)(1 - F_0(I_k^l)) + ny/3} \right)$$
$$\leq 2 \exp\left(- \frac{Cny^2}{2^{-l} + y} \right).$$

Set $ny = ny_n$ to be the term appearing on the right hand side of the $\leq$ sign in the definition of $\mathcal{B}_l$. There are two regimes for the index l, depending on which one of the quantities in the maximum on the right hand side of the definition of $\mathcal{B}_l$ dominates. In each of the regimes, the last display is bounded by $C' e^{-CM(l+L_n)}$, for any k. Conclude using that $\sum_l 2^l e^{-CM(l+L_n)} = o(1)$ for large M. □

4.4 Lemmas on Beta variables

Lemma 5. *There exist universal constants a_0, R_0, c_1 such that, for any $a \geq a_0$, any integer $R \geq R_0$ and d a real number such that $|d| \leq a/2$, if Y follows a* $\mathrm{Beta}(a, a+d)$ *distribution,*

$$E|1 - 2Y|^R \leq 2^{R-1} \left[|EY - \frac{1}{2}|^R + E|Y - EY|^R \right]$$
$$\leq (c_1 d/a)^R + (c_1 R/a)^{R/2}$$

Proof. The first inequality follows by convexity of $u \to u^R$ on $\mathbb{R}^+$. For the first term on the right hand side one uses that by definition of Y, $|2EY - 1| = d/(2a+d) \leq d/(3a)$. For the second term, one writes

$$E|Y - EY|^R = R \int_0^\infty \mathbb{P}[|Y - EY|^R \geq u] u^{p-1} du$$
$$\leq R \left(\frac{2}{2a+d} \right)^R + R \int_{\frac{2}{2a+d}}^\infty \mathbb{P}[|Y - EY|^R \geq u] u^{p-1} du.$$

The last integral is bounded by, using Lemma 6 below and $1/3 \leq a/(2a+d) \leq 1/2$, and denoting $s = 2a + d$,

$$\int_0^\infty \mathbb{P}\left[|Y - EY|^R \geq \frac{w}{\sqrt{s}} + \frac{2}{s} \right] \left(\frac{w}{\sqrt{s}} + \frac{2}{s} \right)^{R-1} \frac{dw}{\sqrt{s}}$$
$$\leq \frac{1}{\sqrt{s}} \int_0^\infty \left(\frac{w}{\sqrt{s}} + \frac{2}{s} \right)^{R-1} D e^{-\frac{w^2}{4}} dw$$
$$\leq C 2^R \left[s^{-\frac{R}{2}} \int_0^w w^{R-1} e^{-\frac{w^2}{4}} dw + s^{\frac{1}{2} - R} \right].$$

By the standard formula on absolute moments of normal variables,

$$\int_0^w w^{R-1} e^{-\frac{w^2}{4}} dw \lesssim \Gamma\left(\frac{R-1}{2} \right) C^{\frac{R}{2}} \lesssim (CR)^{\frac{R}{2}}.$$

Combining the previous bounds leads to the result. □

Lemma 6. *Let φ, ψ belong to $(0, \infty)$. Let Z follow a* Beta(φ, ψ) *distribution. Suppose, for some reals c_0, c_1,*

$$0 < c_0 \leq \varphi/(\varphi + \psi) \leq c_1 < 1 \tag{25}$$

$$\varphi \wedge \psi > 8. \tag{26}$$

Then there exists $D > 0$ depending on c_0, c_1 only such that for any $x > 0$,

$$\mathbb{P}\left[|Z - E[Z]| > \frac{x}{\sqrt{\varphi + \psi}} + \frac{2}{\varphi + \psi}\right] \leq De^{-x^2/4}.$$

Remark 1. The bound in Lemma 6 can be read as a sub-Gaussian bound on Beta variables with 'balanced' (φ and ψ are roughly of the same order via (25)) and 'large enough' (via (26)) parameters. Under (26) and if $x \geq 1$, which is the case for the applications considered here, the term $2/(\varphi + \psi)$ can always be absorbed in the first term, up to a change in the constants.

Proof. The density of Z can be written $e^g / \int_0^1 e^g$, where for $u \in (0, 1)$,

$$g(u) = (\varphi - 1) \log u + (\psi - 1) \log(1 - u).$$

The mean of Z is $E[Z] = \varphi/(\varphi + \psi)$ and its mode m is the mode of g on $(0, 1)$, noting that g has a unique maximum on $(0, 1)$ if (26) is assumed. Solving $g'(m) = 0$ and simple algebra reveal that

$$\begin{aligned} |E[Z] - m| &= \left| \frac{\varphi - \psi}{(\varphi + \psi - 2)(\varphi + \psi)} \right| \\ &\leq \frac{2}{\varphi + \psi} \frac{|\varphi - \psi|}{\varphi + \psi} \leq \frac{2}{\varphi + \psi}. \end{aligned}$$

That is, to prove the inequality, it is enough to bound

$$\mathbb{P}\left[|Z - m| > B\right] = \frac{\int_{|u-m|>B} e^{g(u)} du}{\int_0^1 e^{g(u)} du},$$

where $B \equiv B(x) = x/\sqrt{\varphi + \psi}$. To do so, we bound numerator and denominator in the last expression by deriving two bounds on g.

The first bound is $g(u) - g(m) < -(\varphi + \psi)(u - m)^2/4$, for any u in $(0, 1)$, which follows from Taylor's formula together with the fact that

$$-g''(u) = \frac{\varphi - 1}{u^2} + \frac{\psi - 1}{(1 - u)^2} > \frac{(\varphi + \psi)}{2}(u^{-2} \wedge (1 - u)^2) > \frac{\varphi + \psi}{2}.$$

The second bound controls g close to m. First suppose $m \leq 1/2$ and let us bound g on $J := (m, m + 1/\sqrt{\varphi + \psi})$. We claim that J is contained in $(c_0/(1 + 8^{-1}), 3/4)$. The right boundary follows from combining $m \leq 1/2$ and (26). The left boundary is obtained from

$$m = \frac{\varphi + 1}{\varphi + \psi + 2} \geq \frac{\varphi}{\varphi + \psi} \frac{1}{1 + 8^{-1}} \geq \frac{c_0}{1 + 8^{-1}},$$

where the first bound uses (26) and the second bound uses (25). Now for any u in J, we may write $g(u) = g(m) + g''(\zeta)(u-m)^2/2$, for some $\zeta \in J$. But

$$|g''(\zeta)| \leq (\varphi + \psi)(\zeta^{-2} + (1-\zeta)^{-2}).$$

Using the previous bounds on the endpoints of J, one deduces that

$$|g(u) - g(m)| \leq c, \qquad \forall\, u \in J,$$

where c depends on c_0 only. In the case that $m > 1/2$, we instead bound g on $J' := (m - 1/\sqrt{\varphi+\psi}, m)$. Using the symmetric bound

$$m \leq 1 - \frac{\psi}{\varphi+\psi}\frac{1}{1+8^{-1}} \leq 1 - \frac{1-c_1}{1+8^{-1}},$$

one has $J' \subset (1/4, 1-(1-c_1)/(1+8^{-1}))$ from which we deduce as before that $|g(u) - g(m)| \leq c'$ for any $u \in J'$, where c' depends on c_1 only.

Combining the previous bounds, one obtains, in the case $m \leq 1/2$,

$$\begin{aligned}\mathbb{P}\left[|Z-m| > B\right] &\leq \frac{\int_{|u-m|>B} e^{-(\varphi+\psi)(u-m)^2/4} du}{\int_J e^{-c} du} \\ &\leq e^c \int_{|v|>x} e^{-v^2/4} dv \leq D e^{-x^2/4},\end{aligned}$$

for D large enough, using the standard bound $\int_x^\infty e^{-u^2} du \leq A e^{-x^2}$, for any $x > 0$ and a universal constant A. The case $m > 1/2$ follows similarly. □

4.5 Membership to L^2 and L^∞

Lemma 7. *Let Π be the distribution on densities generated by a Pólya tree with parameters $\alpha_\varepsilon = a_l$, for any $|\varepsilon| = l$ and $l \geq 0$, and some sequence $(a_l)_{l\geq 0}$. Under condition (4), that is*

$$\sum_{l=0}^{\infty} a_l^{-1} < \infty,$$

a density f drawn from Π belongs to $L^2[0,1]$, Π-almost surely. In other words $\Pi[f : \int_0^1 f^2 < \infty] = 1$. Under the stronger condition that for some $\delta > 1/2$,

$$\sum_{l=0}^{\infty} 2^{l^\delta} a_l^{-1/2} < \infty,$$

a density f drawn from Π belongs to $L^\infty[0,1]$, Π-almost surely. Moreover, in this case, Π-almost surely, f is also (Lebesgue)-almost everywhere the sum of its Haar wavelet series. Also, all these statements hold under the posterior distribution $\Pi[\cdot \,|\, X]$ as well.

Proof. For convergence in L^2, it is enough to check that the sequence (f_{lk}), with $f_{lk} = \langle f, \psi_{lk}\rangle$ the Haar wavelet coefficients of f, is square integrable Π-a.s., which is implied if

$$E\Big[\sum_{l,k} f_{lk}^2\Big] < \infty,$$

where E denotes the expectation under the prior distribution. From the expression (19), for $I_\varepsilon = I_{\varepsilon(l,k)}$ one gets $E(f_{lk}^2) = 2^l E[P(I_\varepsilon)^2] E(1-2Y_{\varepsilon 0})^2$, and

$$E[P(I_\varepsilon)^2] = \prod_{i=0}^{l-1} \frac{1}{2}\frac{a_i+1}{2a_i+1} = 2^{-2l} \prod_{i=0}^{l-1} \frac{1+a_i^{-1}}{1+(2a_i)^{-1}} \leq 2^{-2l} e^{\sum_{i=0}^{l-1} a_i^{-1}},$$

as well as $E(1-2Y_{\varepsilon 0})^2 = 4\mathrm{Var}(Y_{\varepsilon 0}) = 1/(2a_l+1)$. Deduce that the expectation at stake is finite as soon as (4) holds.

For the supremum norm, one first checks that the series

$$\sum_{l\geq 1} \sum_{k=1}^{2^l} f_{lk}\psi_{lk}$$

is normally converging Π-almost surely. For this it is enough to verify, as $\|\sum_k |\psi_{lk}|\|_\infty \leq 2^{l/2}$, that, denoting by E the expectation under the prior distribution,

$$E\Big[\sum_l 2^{l/2} \max_k |f_{lk}|\Big] < \infty.$$

Using Hölder's inequality, and next bounding the maximum by the sum, this expectation is bounded, for some $R>1$, by

$$\sum_l 2^{l/2} \left[\sum_{k=0}^{2^l-1} E|f_{lk}|^R\right]^{\frac{1}{R}} \leq \sum_l 2^l \left[\sum_{k=0}^{2^l-1} EP(I_{\varepsilon(l,k)})^R E|1-2Y_{\varepsilon(l,k)0}|^R\right]^{\frac{1}{R}}.$$

The first expectation in the last display is bounded proceeding as for the L^2 norm above, but using the formula for the R-th moment of a Beta variable instead of the second. This yields

$$\begin{aligned} EP(I_{\varepsilon(l,k)})^R &= \prod_{i=0}^{l-1}\prod_{r=0}^{R-1} \frac{\alpha_i+r}{2\alpha_i+r} = 2^{-lR} \prod_{r=0}^{R-1}\prod_{i=0}^{l-1} \frac{1+r/\alpha_i}{1+r/(2\alpha_i)} \\ &\leq 2^{-lR} \prod_{r=0}^{R-1} e^{\sum_{i=0}^{l-1} r/\alpha_i} \leq 2^{-lR} e^{CR^2}. \end{aligned}$$

For the second expectation to evaluate, we use Lemma 5 with $a=a_l$ and $d=0$ to obtain $E|1-2Y_{\varepsilon(l,k)0}|^R \leq (c_1 R/a_l)^{R/2}$. Combining the previous bounds one obtains that the considered expectation is bounded by

$$\sum_l 2^l (2^l 2^{-lR} e^{cR^2} (c_1 R/a_l)^{R/2})^{1/R} \lesssim \sum_l 2^{l/R} \frac{e^{cR}}{a_l^{1/2}} \sqrt{R}.$$

Taking $R = C\sqrt{l}$, the last sum converges by assumption, which shows normal convergence Π-almost surely. Deduce that the Haar-wavelet series of f Π-a.s. converges in L^∞, to an element say $g \in L^\infty$. As the wavelet series converges in L^2 to f by the first part of the proof, deduce that $\int (f-g)^2 = 0$, so that $f=g$ a.e.. That is, f belongs to L^∞ and coincides with the sum of its wavelet series almost everywhere, Π-almost surely. Finally, the statement about the posterior distribution follows from the fact that, under (4), the Pólya tree posterior is absolutely continuous with respect to the prior, see e.g. [9]. □

4.6 Weak convergence and BvM phenomenon in $\mathcal{M}_0(w)$

Convergence in distribution of random variables $X_n \to^d X$ in a metric space (S, d) can be metrised by metrising weak convergence of the induced laws $\mathcal{L}(X_n)$ to $\mathcal{L}(X)$ on S. Here we work with the bounded-Lipschitz metric β_S: Let μ, ν be probability measures on (S, d) and define

$$\beta_S(\mu, \nu) \equiv \sup_{F:\|F\|_{BL}\leq 1} \left| \int_S F(x)(d\mu(x) - d\nu(x)) \right|, \tag{27}$$

$$\|F\|_{BL} = \sup_{x\in S} |F(x)| + \sup_{x\neq y, x,y\in S} \frac{|F(x) - F(y)|}{d(x, y)}.$$

Lemma 8 (Proposition 6 in [7]). *Let $\pi_{V_J}, J \in \mathbb{N}$, be the projection operator onto the finite-dimensional space spanned by the ψ_{lk}'s with scales up to $l \leq J$. Let $f \sim \Pi(\cdot|X)$, $T_n = T_n(X)$, let $\tilde{\Pi}_n$ denote the laws of $\sqrt{n}(f - T_n)$ conditionally on X. Assume that the finite-dimensional distributions converge, that is,*

$$\beta_{V_J}\left(\tilde{\Pi}_n \circ \pi_{V_J}^{-1}, \mathbb{G}_{P_0} \circ \pi_{V_J}^{-1}\right) \to^{P_0} 0, \quad \text{as } n \to \infty, \tag{28}$$

for all $J \in \mathbb{N}$, and that for some sequence $z = (z_l) \uparrow \infty$, $z_l/\sqrt{l} \geq 1$,

$$E\left[\sup_l z_l^{-1} \max_k |\langle f - T_n, \psi_{lk}\rangle| \middle| X\right] = O_{P_0}\left(\frac{1}{\sqrt{n}}\right). \tag{29}$$

Then, for any w such that $w_l/z_l \uparrow \infty$ we have, as $n \to \infty$,

$$\beta_{\mathcal{M}_0(w)}(\tilde{\Pi}_n, \mathbb{G}_{P_0}) \to^{P_0} 0.$$

Remark 2. The result still holds true if $f \sim \Pi(\cdot|X)$ is replaced by $f \sim \bar{\Pi}(\cdot \,|\, X)$ for random measures $\bar{\Pi}(\cdot \,|\, X)$ s.t.

$$\beta_{\mathcal{M}_0}(\bar{\Pi}(\cdot \,|\, X), \Pi(\cdot \,|\, X)) \to^{P_0} 0$$

as $n \to \infty$. Likewise, the posterior can be replaced by the conditional posterior $\Pi(\cdot \,|\, X, D_n)$ for any sequence of sets D_n such that $\Pi(D_n \,|\, X) \to^{P_0} 1$.

References

[1] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561, 1999.

[2] J. O. Berger and A. Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *J. Amer. Statist. Assoc.*, 96(453):174–184, 2001.

[3] L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.*, 3(2):259–282, 1984.

[4] L. Birgé. Robust tests for model selection. *IMS Collect.*, 9:47–64, 2013.

[5] I. Castillo. On Bayesian supremum norm contraction rates. *Ann. Statist.*, 42(5):2058–2091, 2014.

[6] I. Castillo and R. Nickl. Nonparametric Bernstein–von Mises Theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028, 2013.

[7] I. Castillo and R. Nickl. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, 42(5):1941–1969, 2014.

[8] I. Castillo and J. Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.*, 43(6):2353–2383, 2015.

[9] L. Drăghici and R. V. Ramamoorthi. A note on the absolute continuity and singularity of Polya tree priors and posteriors. *Scand. J. Statist.*, 27(2):299–303, 2000.

[10] J. Fabius. Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.*, 35:846–856, 1964.

[11] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[12] T. S. Ferguson. Prior distributions on spaces of probability measures. *Ann. Statist.*, 2:615–629, 1974.

[13] D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

[14] D. A. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.*, 36:454–456, 1965.

[15] S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In *Bayesian nonparametrics*, Camb. Ser. Stat. Probab. Math., pages 35–79. Cambridge Univ. Press, Cambridge, 2010.

[16] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference*, 77(2):181–193, 1999.

[17] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[18] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223, 2007.

[19] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. 2015. Forthcoming.

[20] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37(4):1605–1646, 2009.

[21] E. Giné and R. Nickl. Rates on contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 39(6):2883–2911, 2011.

[22] T. Hanson and W. O. Johnson. Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.*, 97(460):1020–1033, 2002.

[23] R. Z. Has′minskiĭ. A lower bound for risks of nonparametric density estimates in the uniform metric. *Teor. Veroyatnost. i Primenen.*, 23(4):824–828, 1978.

[24] N. Hjort and S. Walker. Quantile pyramids for Bayesian nonparametrics. *Ann. Statist.*, 37(1):105–131, 2009.

[25] M. Hoffmann, J. Rousseau, and J. Schmidt-Hieber. On adaptive posterior concentration rates. *Ann. Statist.*, 43(5):2259–2295, 2015.

[26] I. A. Ibragimov and R. Z. Has′minskiĭ. An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 98:61–85, 161–162, 166, 1980. Studies in mathematical statistics, IV.

[27] A. Jara and T. E. Hanson. A class of mixtures of dependent tail-free processes. *Biometrika*, 98(3):553–566, 2011.

[28] J.-P. Kahane and J. Peyrière. Sur certaines martingales de Benoit Mandelbrot. *Advances in Math.*, 22(2):131–145, 1976.

[29] C. H. Kraft. A class of distribution function processes which have derivatives. *J. Appl. Probability*, 1:385–388, 1964.

[30] M. Lavine. Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, 20(3):1222–1235, 1992.

[31] M. Lavine. More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, 22(3):1161–1176, 1994.

[32] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[33] A. Lo. Weak convergence for Dirichlet processes. *Sankhyā*, 45(1):105–111, 1983.

[34] R. D. Mauldin, W. D. Sudderth, and S. C. Williams. Pólya trees and random distributions. *Ann. Statist.*, 20(3):1203–1221, 1992.

[35] M. Métivier. Sur la construction de mesures aléatoires presque sûrement absolument continues par rapport à une mesure donnée. *Z. Wahrscheinlichkeitshtheorie und Verw. Gebiete*, 20:332–344, 1971.

[36] L. E. Nieto-Barajas and P. Müller. Rubbery Polya tree. *Scand. J. Stat.*, 39(1):166–184, 2012.

[37] G. Pólya. Sur quelques points de la théorie des probabilités. *Ann. Inst. H. Poincaré*, 1(2):117–161, 1930.

[38] S. Resnick, G. Samorodnitsky, A. Gilbert, and W. Willinger. Wavelet analysis of conservative cascades. *Bernoulli*, 9(1):97–135, 2003.

[39] V. Rivoirard and J. Rousseau. Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.*, 7(2):311–333, 2012.

[40] L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26, 1965.

[41] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

[42] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Univ.Press, Cambridge, 1998.

[43] S. Walker. New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043, 2004.

[44] W. H. Wong and L. Ma. Optional Pólya tree and Bayesian inference. *Ann. Statist.*, 38(3):1433–1459, 2010.