Electronic Journal of Statistics
Vol. 16 (2022) 6267–6312
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2086

Optional Pólya trees: Posterior rates and uncertainty quantification[*]

Ismaël Castillo

Sorbonne Université, LPSM, UMR 8001, F-75005 Paris, France,
Institut Universitaire de France
e-mail: ismael.castillo@sorbonne-universite.fr
url: https://www.lpsm.paris/pageperso/castillo//

and

Thibault Randrianarisoa

Sorbonne Université, LPSM, UMR 8001, F-75005 Paris, France
e-mail: thibault.randrianarisoa@sorbonne-universite.fr
url: https://thibaultrandrianarisoa.netlify.app

Abstract: We consider statistical inference in the density estimation model using a tree–based Bayesian approach, with Optional Pólya trees as prior distribution. We derive near-optimal convergence rates for corresponding posterior distributions with respect to the supremum norm. For broad classes of Hölder–smooth densities, we show that the method automatically adapts to the unknown Hölder regularity parameter. We consider the question of uncertainty quantification by providing mathematical guarantees for credible sets from the obtained posterior distributions, leading to near–optimal uncertainty quantification for the density function, as well as related functionals such as the cumulative distribution function. The results are illustrated through a brief simulation study.

MSC2020 subject classifications: 62G20, 62G07, 62G05.
Keywords and phrases: Bayesian nonparametrics, Pólya trees, posterior convergence rates, supremum norm, uncertainty quantification, frequentist coverage of credible sets.

Received December 2021.

Contents

[*]This work was funded by the ANR, project ANR-17-CE40-0001 (BASICS).

1. Introduction

Tree–based methods are among the most broadly used algorithms in statistics and machine learning. This goes from single tree algorithms such as CART [4] or Bayesian CART [13, 15], to the use of random forests [3, 12], that is ensembles of trees. Due in particular to their ability to quantify uncertainty, there has been much interest in Bayesian tree–based methods. While for frequentist methods there is a by now well–established theory in quadratic loss for CART and related algorithms, advances on the mathematical understanding of Bayesian counterparts are very recent. In [38, 27], L^2–posterior contraction rates are obtained for both trees and forests in a regression setting. Still in regression, the work [11] addresses the case of the stronger supremum norm loss for Bayesian CART–type priors. The present paper can be seen as a continuation of [11], investigating the density estimation setting. In Bayesian density estimation, a classical tree–method is that of Pólya trees (henceforth PTs, see e.g. [18], Chapter 3). For well–chosen parameters, PTs' samples are random densities, and contraction rates for the corresponding posterior densities have been obtained in [7]. The idea behind Pólya tree is to grow a fixed, infinite, tree; this is typically not flexible enough to address refined statistical goals such as adaptation. Notably, Wong and Ma introduced in [39] a flexible alternative to standard PTs that they call Optional Pólya Trees (OPTs in the sequel), which have been successfully extended and applied to a number of settings in e.g. [31, 25, 30, 28, 14]. Yet, from the theoretical point of view, only posterior consistency was established in [39] and follow-up works. Not based on (flexible) trees, we also note the different construction of spike–and–slab Pólya trees introduced in [8].

There are two main goals in the present paper. The first is to continue the investigations of [11] for tree–methods in order to obtain inference in the practically very desirable supremum norm loss, but in the model of density estimation, and the second to elaborate a theory for rates and uncertainty quantification (henceforth, UQ) for Optional Pólya Trees. In fact, our methods enable to cover also more general priors, although for simplicity we will mostly stick to OPTs in this work. We now briefly review a number of related results. While the use of a general theory based on prior mass and testing [17, 18] made a relatively broad L^2–theory possible [27, 38], results for the supremum norm are typically more delicate, as uniform testing rates required in [17] appear to be slower [20]. Recent advances on this front include [6, 23, 34, 33, 40]. The first supremum norm posterior rates for tree methods, optimal up to a logarithmic factor, were obtained in [11] in regression models; we refer to [11] for more context and references on rates for tree–based methods.

The main results of the paper are as follows

1. we prove that Optional Pólya Trees (OPTs) achieve optimal supremum–norm posterior contraction rates (up to a logarithmic factor) in density estimation: this provides an optimal rate–theory for the consistency results of [39], who introduced the OPT prior, for the computationally efficient case of dyadic splits.
2. we show that tree–based inference with OPTs leads to (near–) optimal uncertainty quantification in terms of confidence bands, both for the density f and the distribution function $F = \int_0^{\cdot} f$, in an adaptive way.

Those constitute the first results, to the best of our knowledge, showing that tree–based methods in density estimation lead to near–optimal uncertainty quantification in terms of the supremum norm. Apart from making the consistency results of [39] precise, this work shows that the programme for inference with tree–priors outlined in [11], who considered regression settings only, carries over to density estimation; the techniques presented could also be used for other tree priors beyond OPTs.

The paper is organized as follows. Section 2 introduces a class of tree–based priors on density functions, of which OPTs are a special case. Section 3 states our main result on tree–based supremum norm contraction, while Section 4 focuses on Uncertainty Quantification, both for the density function and smooth functionals thereof. Section 5 illustrates our findings numerically through a simulation study. Section 6 briefly summarises and discusses the results and future research directions. Proofs are gathered in Section 7 and the Appendix.

2. Dyadic tree–based random densities and Optional Pólya trees (OPTs)

2.1. Bayesian framework

Adopting a Bayesian point of view, the density estimation model on $[0,1)$ consists in observing

$$\begin{aligned} X = (X_1,\dots,X_n)\,|\, f &\sim P_f^{\otimes n} \\ f &\sim \Pi, \end{aligned} \tag{1}$$

where P_f is the distribution on $[0,1)$ with density f with respect to Lebesgue measure: $dP_f = f d\mu$, and where Π is a prior distribution on densities f to be defined below. The posterior distribution is then the conditional distribution of f given X and is denoted $\Pi[\cdot\,|\,X]$.

Frequentist analysis of Bayesian posteriors. To analyse mathematically the behaviour of the posterior distribution $\Pi[\cdot\,|\,X]$, once the posterior is formed using the Bayesian model, we make the frequentist assumption that the data X has actually been generated from a 'true' parameter value f_0, that is, in the density estimation setting, $X \sim P_{f_0}^{\otimes n}$. In the sequel, we thus study the behaviour of $\Pi[\cdot\,|\,X]$ in probability under $P_{f_0} = P_{f_0}^{\otimes n}$. For more details and context, we refer the reader to the book [18].

Motivated by recent work [11] on Bayesian CART in regression settings (see e.g. the discussion in Section 5 of [11]), we introduce a family of tree-based prior distributions on density functions. For simplicity, we mostly consider the case of densities on the unit interval, but our results could be extended to higher dimensions up to using slightly more complex notation, which we refrain to do here − see, though, the discussion in Section 6 for more on this −.

Informal prior description. The prior on densities is defined in three steps, which will be more formally introduced below

Step 1 a random tree $\mathcal{T}$ is sampled from a prior $\Pi_{\mathbb{T}}$ on trees;

Step 2 given $\mathcal{T}$, a partition $I_{\mathcal{T}}$ of the unit interval is produced, built recursively in a tree fashion 'along' $\mathcal{T}$ with breakpoints placed at midpoints of the successive intervals;

Step 3 given $I_{\mathcal{T}}$, the output density f is a histogram with random heights whose distribution follows a Pólya tree–type law.

2.2. Priors $\Pi_{\mathbb{T}}$ on full binary trees

Definition 1. A *full binary tree* is a set of nodes $\mathcal{T} = \{(l,k),\ l\geq 0,\ 0\leq k\leq 2^l-1\}$ verifying the condition

$$(l,k)\in\mathcal{T} \implies \text{if } l>0,\ \Big(l-1,\lfloor k/2\rfloor\Big)\in\mathcal{T} \text{ and } \big(l,k+(-1)^k\big)\in\mathcal{T}.$$

One then says that $\Big(l-1,\lfloor k/2\rfloor\Big)$ is the *parent* node of its *children* (l,k) and $\big(l,k+(-1)^k\big)$, and a node with no children is called an external node or leaf;

$(0,0)$ belongs to every non-empty tree and is called the tree *root*. We denote by $\mathcal{T}_{\text{int}}$ the set of non-terminal – or 'internal' – nodes in $\mathcal{T}$ (i.e. those with children), and $\mathcal{T}_{\text{ext}} = \mathcal{T} \setminus \mathcal{T}_{\text{int}}$ the set of 'leaves' – also called 'external' nodes –.

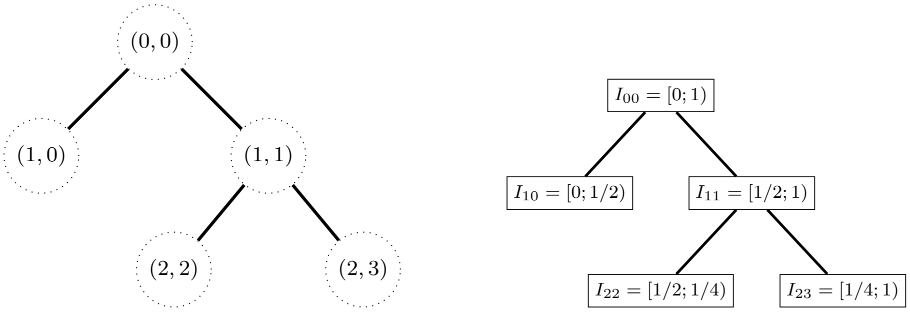
The parent-child relationship of the pairs in a tree gives rise to the tree representation depicted on Figure 1a. This justifies the following terminology as we define the *depth* of $\mathcal{T}$ as the integer

$$d(\mathcal{T}) \coloneqq \max_{(l,k)\in\mathcal{T}} l.$$

One further denotes by $\mathbb{T}$ the set of all binary trees and, putting a slight restriction on the maximum depth,

$$\mathbb{T}_n \coloneqq \{\mathcal{T} \in \mathbb{T} : \ d(\mathcal{T}) \le L_{\max}\}, \qquad \text{with } L_{\max} \coloneqq \left\lfloor \log_2\left(n/\log^2(n)\right) \right\rfloor. \tag{2}$$

The prior distributions considered below put mass 1 to the subset $\mathbb{T}_n$ of $\mathbb{T}$.



(a) Tree pairs. (b) Tree partitioning $I_{\mathcal{T}}$.

Fig 1: Tree $\mathcal{T} = \{(0,0), (1,0), (1,1), (2,2), (2,3)\}$.

Next we give two examples of priors $\Pi_{\mathbb{T}}$ on full binary trees. Both are actually considered in actual Bayesian CART implementations [13, 15].

Example 1 (GW(p) Markov process on tree)**.** *A random tree is recursively defined by the following process. First, let us attribute to each possible pair* (l,k) *a deterministic parameter* $p_{lk} \in [0,1]$*. Starting at the root node* $(0,0)$*, either the tree with only* $(0,0)$ *as node is returned with probability* $1 - p_{00}$*, or there is a split and the tree contains not only* $(0,0)$ *but at least also* $(1,0)$ *and* $(1,1)$*. The construction process then continues recursively until either there are no further nodes to split, or a maximum depth* $L_{\max}$ *is reached, after which (i.e. for* $l \ge L_{\max}$*) we do not further grow the tree. More precisely, the recursion is from up to down (*l *grows) and left to right (*k *grows), as follows: given the tree contains* (l,k)*, with probability* $1-p_{lk}$ *the node* (l,k) *is a leaf; and with probability* p_{lk}*, the tree further has a split at* (l,k)*, i.e. the node* (l,k) *has* $(l+1,2k)$ *and* $(l+1,2k+1)$ *as children in the tree.*

The process producing such a random tree $\mathcal{T}$ is Markov (along the complete dyadic tree) in the sense that the probability that a node (l,k) further splits only depends on the fact that the node is present or not and on the parameter p_{lk}, but not on the rest of the tree built so far (above and to the left of (l,k)). By analogy to Galton–Watson processes, with here nodes having either two or zero children with probabilities p_{lk} and $1-p_{lk}$ respectively, we call $\Pi_{\mathbb{T}}$ as above a $\mathrm{GW}(p)$ prior, with parameters $p=(p_{lk})=(p_\epsilon)$ (we define the link between ϵ and (l,k) below, in Section 2.3), $p_{L_{max}k}=0$.

Example 2 (Conditioning on the number of leaves)**.** *In this construction, one samples first a number K of leaves according to a prior on integers and given K one then samples uniformly from the set of all full binary trees with K leaves and depth at most $L_{\max}$.*

2.3. Partitioning $I_{\mathcal{T}}$

Let us first introduce notation on dyadic numbers and intervals. For any binary sequence $\epsilon \in \{0,1\}^l$, its length is $|\epsilon| = l > 0$. For any dyadic number $r = k/2^l$ in $[0,1)$ with $0 \le k < 2^l$, $l > 0$, one writes $\epsilon(k,l) = \epsilon_1(r)\cdots\epsilon_l(r) \in \{0,1\}^l$, such that $r = \sum_{k=1}^{l} \epsilon_k(r)2^{-k}$, its unique decomposition in base 2^{-1} with $|\epsilon| = l$. Accordingly, one introduces the dyadic intervals, for $\epsilon = \epsilon(k,l)$,

$$I_\epsilon := I_{lk} := \left[\frac{k}{2^l}, \frac{k+1}{2^l}\right),$$

and one sets $I_\varnothing = I_{0,0} = [0,1)$. In addition, for any ϵ and $0 < i \le |\epsilon|$, one writes $\epsilon^{[i]} = \epsilon_1 \dots \epsilon_i$. Also, we introduce $\mathcal{E}^* = \cup_{l=0}^{\infty} \{0;1\}^l$ where $\{0;1\}^0 = \{\varnothing\}$.

To each full binary tree encoded as above as the collection of its nodes (l,k), we associate a partition $I_{\mathcal{T}}$ of the unit interval given by, with $\mathcal{T}_{ext}$ the external nodes of $\mathcal{T}$ as in Definition 1,

$$[0,1) = \bigcup_{(l,k)\in\mathcal{T}_{ext}} I_{lk}.$$

Such a tree-based recursive partitioning of $[0,1)$ is illustrated on Figure 1b. The deeper the tree locally, the more refined the corresponding partition becomes. By definition of I_{lk}, note that the partition has split-points at dyadic numbers. The final partition $I_{\mathcal{T}}$ can also be seen as being obtained from recursively splitting $[0,1)$ in halves, continuing to split locally only if the tree continues further down at that location. For this reason we talk about *splitting at midpoints*. Note that, still using full binary trees $\mathcal{T}$, one could make splits at a different, possibly random, location. Although this makes the construction even more flexible, we shall not consider this here for simplicity (we note in passing that computationally the split–at–midpoint construction appears often to be among the easiest to simulate from, as it does not require to draw split locations; we refer to [11], Section 4, for more on ‘unbalanced’ splits).

2.4. Prior values given tree and partitioning

Once a tree $\mathcal{T}$ and partitioning $I_{\mathcal{T}}$ are given, we draw a random histogram over the partition given by $I_{\mathcal{T}}$ by sampling heights over each sub-interval in such a way that the overall histogram is a positive density f (i.e. $f > 0$ and $\int_0^1 f = 1$). To do so, we use a mass–splitting process along the tree $\mathcal{T}$, which actually coincides with that of Pólya trees – we refer to the Appendix A for more on those –. This choice is for simplicity but we could consider other choices too (in this vein, the Beta(a, a) law at the end of Definition 2 could be taken to depend on (l, k) or be a different distribution).

Definition 2 (Prior Π)**.** Let $\Pi_{\mathbb{T}}$ be a prior on full binary trees. Let (Y_ε) be a sequence of independent variables of distribution Beta$(a_{\varepsilon 0}, a_{\varepsilon 1})$, for some $a_{\varepsilon 0}, a_{\varepsilon 1} \in [0, 1]$, indexed by $\varepsilon \in \mathcal{E}^*$. The prior Π draws a random tree–based histogram f as follows

$$\mathcal{T} \sim \Pi_{\mathbb{T}} \tag{3}$$

$$f \,|\, \mathcal{T} \sim \sum_{\varepsilon \equiv (l,k) \in \mathcal{T}_{ext}} h_\varepsilon 1\!\!1_{I_{lk}}, \qquad \text{with} \;\; h_\varepsilon = 2^l \prod_{i=1}^{l} Y_{\varepsilon[i]}\,. \tag{4}$$

The distribution $f \,|\, \mathcal{T} = T$ for a given $T \in \mathbb{T}$ is called a T–Pólya tree with parameters (a_ε). In the sequel we set $a_\varepsilon = a$ for some fixed $a > 0$, in which case the distribution is denoted as T–PT(a).

It results from the definition that the overall prior Π is a mixture of T–Pólya trees. When the mixing distribution $\Pi_{\mathbb{T}}$ is a GW(p) prior, it turns out that Π coincides with Optional Pólya trees introduced in [39], in the case of splits at midpoints.

Proposition 1. *Let Π be the mixture distribution induced on densities f constructed as*

$$\mathcal{T} \sim \text{GW}(p)$$
$$f \,|\, \mathcal{T} \sim \mathcal{T}\text{–PT}(a).$$

Then Π coincides with the Optional Pólya tree of [39] corresponding to the recursive partitioning $\{I_\epsilon,\ \epsilon \in \mathcal{E}^*\}$ with splits at midpoints and parameters $M(I_\epsilon) = \lambda(I_\epsilon) = 1, K_1(I_\epsilon) = 2$, stopping probabilities $\rho(I_\epsilon) = 1 - p_\epsilon$ for any $\epsilon \in \mathcal{E}^*$ and parameters for mass allocation $\alpha_1^1 = \alpha_1^2 = a$.

The proof of Proposition 1 is presented in Appendix B. Our notation differs slightly from [39] (which does not make the tree connection) for two reasons: first, the tree–setting enables one to use the framework of [11] and second, although in what follows we stick to OPTs for simplicity, the same proofs work nearly unmodified for other tree–priors, such as the one in Example 2.

2.5. Posterior distribution

Let us recall that the prior Π in Definition 2 is the mixture

$$\begin{aligned} \mathcal{T} &\sim \Pi_{\mathbb{T}} \\ f \,|\, \mathcal{T} &\sim \Pi(\cdot \,|\, \mathcal{T}), \end{aligned} \tag{5}$$

where $\Pi(\cdot \,|\, \mathcal{T})$ is, given $\mathcal{T}$, a $\mathcal{T}$–Pólya tree. For a given dyadic interval I, let $N_X(I)$ denote the number of points X_i that fall in I. The next result is proved in Appendix C.

Proposition 2 (Posterior given $\mathcal{T}$)**.** *Suppose the prior is given by* (5)*, where the prior given $\mathcal{T}$ is a $\mathcal{T}$–Pólya tree with parameters (a_ε). Then, in the density estimation model* (1)*, the posterior $\Pi[\cdot \,|\, X, \mathcal{T}]$ is a $\mathcal{T}$–Pólya tree with parameters (a_ε^X) given by, for any $\varepsilon \in \mathcal{E}^*$,*

$$a_\varepsilon^X = a_\varepsilon + N_X(I_\varepsilon).$$

Let us now move on to describe the posterior induced on trees. We denote

$$N_T(X) = \int \prod_{i=1}^n f(X_i) d\Pi(f \,|\, \mathcal{T} = T) \tag{6}$$

the marginal distribution of X given $\mathcal{T} = T$. It follows from Bayes' formula that $\Pi[\cdot \,|\, X]$ induces a posterior distribution on trees given as: for any $T \in \mathbb{T}$, and $N_T(X)$ as in (6),

$$\Pi[\mathcal{T} = T \,|\, X] = \frac{\Pi_{\mathbb{T}}[\mathcal{T} = T] N_T(X)}{\sum_{T \in \mathbb{T}} \Pi_{\mathbb{T}}[\mathcal{T} = T] N_T(X)}. \tag{7}$$

This is in general a fairly complicated distribution with no closed–form expression. In case the prior $\Pi_{\mathbb{T}}$ on trees is GW (p), it turns out that the posterior on trees is GW (p^X) for updated parameters p^X. Let, for $a > 0$,

$$\nu_\varepsilon^X = 2^{N_X(I_\varepsilon)} \frac{B(a + N_X(I_{\varepsilon 0}), a + N_X(I_{\varepsilon 1}))}{B(a, a)}. \tag{8}$$

Let us now consider parameters (p_ε^X) given by the equations

$$\frac{p_\varepsilon^X}{1 - p_\varepsilon^X}(1 - p_{\varepsilon 0}^X)(1 - p_{\varepsilon 1}^X) = \frac{p_\varepsilon}{1 - p_\varepsilon}(1 - p_{\varepsilon 0})(1 - p_{\varepsilon 1})\nu_\varepsilon^X, \tag{9}$$

Equations (9) together admit a unique solution (p_ε^X) obtained by a bottom–up recursion noting that for $|\varepsilon| = L_{\max}$, $p_\varepsilon^X = p_\varepsilon = 0$. This is verified along the proof of Proposition 3 below.

Proposition 3 (Special case of OPTs)**.** *In the setting of Proposition 2, suppose further that the distribution $\Pi_{\mathbb{T}}$ on trees is* GW(p) *with split probabilities* (p_ε)*. Then the posterior distribution can be described as*

$$\Pi[\mathcal{T} = \cdot \mid X] \sim \mathrm{GW}(p_\varepsilon^X)$$
$$\Pi[\cdot \mid X, \mathcal{T}] \sim \mathcal{T}\text{–PT}(a_\varepsilon^X)$$

with splits probabilities (p_ε^X) verifying the recursion (9) and a_ε^X as in Proposition 2. In other words the posterior follows an OPT distribution with corresponding hyperparameters as specified in Proposition 1.

The proof of this proposition is presented in Appendix C.

2.6. Notation and function spaces

Below we shall consider the Hölder class of functions with support in $[0, 1)$ and smoothness parameter $0 < \alpha \leq 1$, defined as

$$\mathcal{C}^\alpha[0,1) := \left\{ f : [0,1) \mapsto \mathbb{R}, \quad \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^\alpha} < +\infty \right\}$$

and we similarly define Hölder balls with parameters $\alpha > 0$ and $K \geq 0$ as

$$\Sigma(\alpha, K) := \left\{ f : [0,1) \mapsto \mathbb{R}, \quad \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^\alpha} \leq K \right\}.$$

Bounded Lipschitz metric. Let $(\mathcal{S}, d)$ be a metric space. The bounded Lipschitz metric $\beta_{\mathcal{S}}$ on probability measures of $\mathcal{S}$ is defined as, for any μ, ν probability measures of $\mathcal{S}$,

$$\beta_{\mathcal{S}}(\mu, \nu) = \sup_{F; \|F\|_{BL} \leq 1} \left| \int_{\mathcal{S}} F(x)(d\mu(x) - d\nu(x)) \right|, \tag{10}$$

where $F : \mathcal{S} \to \mathbb{R}$ and

$$\|F\|_{BL} = \sup_{x \in \mathcal{S}} |F(x)| + \sup_{x \neq y} \frac{|F(x) - F(y)|}{d(x,y)}. \tag{11}$$

This metric metrises the convergence in distribution, see e.g. [16], Theorem 11.3.3.

As shown in [7], it is also useful to introduce the Haar wavelet basis to carry out an analysis of Pólya tree-like posterior distributions. Indeed, one can relate the inclusion of a node (l, k) in a tree $\mathcal{T}$ to the fact that the coefficient corresponding to the Haar wavelet function ψ_{lk} in the decomposition of $f \sim \Pi[\cdot|\mathcal{T}]$ is non-zero almost surely. More precisely, the Haar basis of $L^2[0;1)$ is the family composed of the mother wavelet $\phi = \mathbb{1}_{[0;1)}$ and the functions

$$\psi_{lk}(\cdot) = 2^{l/2}\psi(2^l \cdot -k)$$

for $l \geq 0$ and $0 \leq k < 2^l$, where $\psi = \mathbb{1}_{[1/2;1)} - \mathbb{1}_{[0;1/2)}$. However, as we consider the problem of density estimation, maps f under scrutiny all verify $\langle f, \phi\rangle = \int_0^1 f(t)dt = 1$, so that we only focus on the wavelets ψ_{lk} and the corresponding coefficients $f_{lk} := \langle f, \psi_{lk}\rangle$ in the following. As for the true density, we define $f_{0,lk} := \langle f_0, \psi_{l,k}\rangle$.

3. Posterior contraction rates for OPTs

For any $\alpha > 0$, $\mu > 0$, $K \geq 0$, we define the regularity class of densities

$$\mathcal{F}(\alpha, K, \mu) := \left\{ f \geq \mu, \quad \int_0^1 f = 1, \quad f \in \Sigma(\alpha, K) \right\},$$

as well as the sequence

$$\varepsilon_n(\alpha) := \left(n^{-1}\log^2 n\right)^{\frac{\alpha}{1+2\alpha}}. \tag{12}$$

Up to a logarithmic factor, this corresponds to the minimax supremum norm rate of estimation over the class $\mathcal{F}(\alpha, K, \mu)$, which equals $(n/\log n)^{-\alpha/(1+2\alpha)}$ up to constants [24].

3.1. Supremum norm convergence for the whole posterior distribution

We now show that the posterior distribution $\Pi[\cdot \mid X]$ asymptotically concentrates most of its mass on a $\|\cdot\|_\infty$–ball of optimal radius.

Theorem 1. *Suppose that $f_0 \in \mathcal{F}(\alpha, K, \mu)$ for some $\mu > 0$, $0 < \alpha \leq 1$ and $K \geq 0$. Let Π be an OPT prior with split probabilities $p_{lk} = \Gamma^{-l}$, $l \geq 0$, $0 \leq k < 2^l$, $\Gamma > 0$, and parameter $a > 0$. Then, for Γ large enough, any sequence $M_n \to \infty$, as $n \to \infty$, and $\varepsilon_n = \varepsilon_n(\alpha)$ as in (12),*

$$E_{f_0}\Pi\Big[\, \|f - f_0\|_\infty > M_n\varepsilon_n \mid X\Big] \to 0.$$

Theorem 1 shows that an OPT posterior with split probabilities decreasing exponentially fast with nodes depth concentrates most of its mass in a supremum norm ball of (near–) minimax optimal radius, whenever the signal has regularity $\alpha \leq 1$. Some comments are in order. First, the regularity requirement $\alpha \leq 1$ is typical and expected for 'hard trees', which produce histogram-type estimators. An alternative would be to use 'soft trees', where individual learner are smooth [27, 11], see also the discussion in Section 6. Second, the slight loss of a logarithmic term in the convergence rate can be shown to be intrinsic to trees and is not due to a possible suboptimality of our rate upper–bounds: this has been formally shown in [11], Theorem 2, in a regression context; an analogous result could be shown in density estimation in a similar way.

A consequence of Theorem 1 is that a posterior draw is close with high probability to the true unknown density function of interest. This settles the *estimation* problem, but it does not yet say much about the quantification of uncertainty, i.e. the construction of *confidence sets*, a question addressed in Section 4.

3.2. Convergence rate for the median tree

While Theorem 1 entails convergence in probability of a draw from $\Pi[\cdot \mid X]$, one may ask what happens for aspects of such distribution, e.g. point estimators derived from it. A natural such estimator from the point of view of tree priors is the median tree estimator defined below, since there is a natural tree associated to it. Such an estimator will also turn helpful for uncertainty quantification as considered below.

The *median tree* is defined as the tree $\mathcal{T}^*$ whose interior nodes are

$$\mathcal{T}^*_{\text{int}} = \{(l,k): \ \Pi[(l,k) \in \mathcal{T}_{\text{int}}|X] > 1/2\}\,, \tag{13}$$

and which is actually a tree as defined previously (see [11], Lemma 13). One associates to it the *median tree density estimator*

$$\hat{f}_{\mathcal{T}^*} = 1 + \sum_{(l,k)\in\mathcal{T}^*_{\text{int}}} 2^{l/2} \frac{N_X\left(I_{(l+1)(2k+1)}\right) - N_X\left(I_{(l+1)(2k)}\right)}{n} \psi_{lk}. \tag{14}$$

Lemma 7 in the appendix shows that this estimator converges in probability to the actual density f_0 at the same almost-minimax rate ε_n in supnorm as in Theorem 1. In Section 5, examples of $\mathcal{T}^*$ and $\hat{f}_{\mathcal{T}^*}$ are presented in Figures 2 and 3.

4. Uncertainty quantification for OPTs

In nonparametrics the problem of uncertainty quantification is well–known to be more delicate than the one of estimation: first negative results to the ambitious goal of constructing confidence sets that both cover the unknown truth and have a diameter that adapts in an optimal way to the smoothness of the unknown function or density were due to [26] and [29]. The general picture that emerged in recent years following these early works is that the difficulty of the problem depends on the considered loss function and on certain testing rates of separation, see [21], Chapter 8. Notably, for the supremum norm, contrary to L^2–losses for which some ‘window’ of adaptation is possible, constructing adaptive confidence sets in full generality is impossible unless one restricts the set of possible functions by assuming e.g. self–similarity conditions. Such conditions can be shown to be essentially necessary; they are also fairly natural from the practical perspective given that self–similarity is itself quite wide–spread in natural phenomena.

Let us briefly describe the uncertainty quantification results we derive. A first confidence band based on the posterior median and using self–similarity is built in Section 4.2. Next, we prove in Section 4.3 that the quantile posterior credible set for the cumulative distribution function leads to optimal UQ; this is a consequence of a more general result, an (adaptive) nonparametric Bernstein–von Mises theorem, proved in Appendix E. Finally in Section 4.4 we construct a confidence band integrating further information from some functionals that is less conservative than the simple band constructed in Section 4.2 and achieves a target confidence level. Our results can be seen as counterparts in density estimation and for tree priors of the results in [36]. Another approach in density estimation would be to use spike–and–slab Pólya priors as recently considered by the second author in [8]. Nevertheless, the latter are expected to be less efficient to compute in high–dimensions (as they, e.g., require to explore all wavelet coefficients in the different dimensions), a setting that, while not investigated in the present paper, is particularly promising for OPTs, see also the discussion in Section 6.

4.1. A self-similarity condition

Here we take the same condition as in [36] (see also [21]). It is fairly simple to state, and can be only slightly improved (see [5]).

Definition 3 (Set $\mathcal{S}$ of self–similar functions)**.** Given an integer $j_0 > 0$ and $\alpha \in (0,1]$, we say that $f \in \Sigma(\alpha, K)$ is *self-similar* if, for some constant $\eta > 0$,

$$\|K_j(f) - f\|_\infty \geq \eta 2^{-j\alpha} \text{ for all } j \geq j_0,$$

where $K_j(f) = \sum_{l<j} \sum_k \langle f, \psi_{lk}\rangle \psi_{lk}$. The set of such f's is denoted $\mathcal{S} = \mathcal{S}(\alpha, K, \eta)$.

The condition assumes that at each resolution depth $j \geq j_0$, the overall 'energy' (measured in terms of supremum norm) of the wavelet coefficients at levels larger than j is lower bounded by a typical amount for α–Hölder functions. Indeed, for any $j \geq j_0$, the quantity $\|K_j(f) - f\|_\infty$ is itself also upper–bounded up to a constant by the same quantity (this follows from standard bounds on the supremum norm and the definition of the Hölder class).

4.2. Simple confidence band

A first construction consists in defining a band from a centering function and a radius. A first and simple possibility consists in defining those using the median tree (13): the resulting median tree estimator (14) can serve as center, while a radius can be defined as

$$\sigma_n = v_n \sqrt{\frac{\log n}{n}} 2^{d(\mathcal{T}^*)/2}, \tag{15}$$

where $d(\mathcal{T}^*)$ is the depth of the median tree $\mathcal{T}^*$, for some slowly diverging sequence (v_n) as specified below. This allows us to define the confidence band, for $\hat{f}_{\mathcal{T}^*}$ as in (14),

$$\mathcal{C}_n = \left\{ f : \left\| f - \hat{f}_{\mathcal{T}^*} \right\|_\infty \leq \sigma_n \right\}. \tag{16}$$

Under self–similarity as in Definition 3, the median tree can in particular be shown to have a depth of the order of the oracle cut–off $2^{L_n^*} \approx n^{1/(2\alpha+1)}$ (up to a logarithmic factor, see the Appendix for a precise statement in Lemma 5) which in turn implies desirable properties for the band $\mathcal{C}_n$ as is made explicit in the next theorem.

Theorem 2. *Let $0 < \alpha_1 < \alpha_2 \leq 1$, $K > 0$, $\mu > 0$ and $\eta > 0$. Let Π be the same prior as in Theorem 1, $\mathcal{C}_n$ as in (16) with $v_n / \log^{1/2} n \to \infty$, then uniformly on $f_0 \in \mathcal{S}(\alpha, K, \eta) \cap \mathcal{F}(\alpha, K, \mu)$, $\alpha \in [\alpha_1, \alpha_2]$,*

$$|\mathcal{C}_n|_\infty = O_{P_0}\left(v_n \left(\frac{\log n}{n} \right)^{\alpha/(2\alpha+1)} \right)$$

and

$$P_0 \left[f_0 \in \mathcal{C}_n \right] = 1 + o(1), \qquad \Pi[\mathcal{C}_n \mid X] = 1 + o_{P_0}(1).$$

For a slowly diverging sequence (v_n), the diameter of $\mathcal{C}_n$ is then within a logarithmic factor of the minimax rate of estimation on $\Sigma(\alpha, K)$ with high probability. It is attained adaptively (the definition of $\mathcal{C}_n$ does not depend on α) for any window $[\alpha_1; \alpha_2]$. The set $\mathcal{C}_n$ allows to quantify uncertainty on f_0 as it is an asymptotic confidence set, and it is also a credible set of credibility going to 1.

4.3. UQ for functionals: A Donsker–type theorem

OPTs with flat initialisation. Let us introduce a slight modification of the OPT prior where trees from the prior distribution are constrained to include all nodes of depth less than some number $l_0 = l_0(n)$, slowly diverging to ∞.

Definition 4. A prior on densities Π of the type (5) is said to have *flat initialisation up to level $l_0 = l_0(n)$* if the prior on trees $\Pi_{\mathbb{T}}$ verifies

$$\Pi_{\mathbb{T}} \left[\bigcap_{l \leq l_0(n), k} \{(l,k) \in \mathcal{T}\} \right] = 1.$$

The next result considers the behaviour of the induced posterior on $F(\cdot) = \int_0^{\cdot} f$, that is on the distribution function for an OPT prior on f. Let us also define, for $\hat{f}_{\mathcal{T}^*}$ the median tree estimator,

$$\hat{F}_n^{med}(t) = \int_0^t \hat{f}_{\mathcal{T}^*}(u) du. \tag{17}$$

Let us recall that for Q a probability measure on $[0,1]$ of distribution function H, a Q–Brownian bridge is a centered Gaussian process $Z(t)$ with covariance function $E[Z(s)Z(t)] = \min(H(s), H(t)) - H(s)H(t)$ and $0 \leq s, t \leq 1$.

Theorem 3 (Donsker's theorem for OPTs)**.** *Let $X = (X_1, \dots, X_n)$ be i.i.d. from law P_0 with density f_0. Let $f_0 \in \mathcal{F}(\alpha, K, \mu)$, for some $\alpha \in (0; 1]$, $K \geq 0$, $\mu > 0$. Let Π be an OPT prior with flat initialisation up to level $l_0(n)$ that verifies $\sqrt{\log n} \leq l_0(n) \leq \log n / \log\log n$, and other than that for $l > l_0(n)$ with same parameters as the prior in Theorem 1.*

Let G_{P_0} be a P_0-Brownian bridge $G_{P_0}(t), t \in [0, 1)$. For $\hat{F}_n^{med}$ as in (17), as $n \to \infty$,

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

Furthermore, for F_n the empirical distribution function, as $n \to \infty$,

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

This implies that the induced posterior distribution $\mathcal{L}(\sqrt{n}\|F - \hat{F}_n^{med}\|_\infty \mid X)$ converges weakly in probability to $\mathcal{L}(\|G_{P_0}\|_\infty)$. Furthermore, for $0 < \gamma < 1$, the credible set

$$\mathcal{F}_n = \{F : \ \|F - \hat{F}_n^{med}\|_\infty \leq \rho_n^X\},$$

with ρ_n^X chosen such that $\Pi[\mathcal{F}_n \mid X] = 1 - \gamma$, is an asymptotically optimal (efficient) confidence set of level $1 - \gamma$. We refer to [10] for more details on this; note that in the latter paper the results are for priors of fixed regularity only, whereas here the prior additionally enables adaptation to the smoothness of f. The behaviour of the credible set $\mathcal{F}_n$ is illustrated in Figure 5.

4.4. Multiscale confidence band

Here we follow the approach introduced in [9, 10] and first briefly recall the idea. One wishes to define a 'multiscale' space (i.e. defined from wavelet coefficients) with an associated metric that is weak enough so that convergence of the posterior distribution for f in that space converges at rate $1/\sqrt{n}$, instead of the slower nonparametric rate of order $n^{-\alpha/(2\alpha+1)}$. In such space one can then formulate a convergence of the posterior to a Gaussian limit, namely a nonparametric Bernstein–von Mises theorem. Below we only define the multiscale space as it is used in the definition of the credible band and postpone details on the precise statement of convergence to Appendix E.

Let us call the sequence $w = (w_l)_{l \geq 0}$ 'admissible' if $w_l/\sqrt{l} \to \infty$ as $l \to \infty$. For such a sequence, let us define

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x = (x_{lk})_{l,k}, \quad \lim_{l \to \infty} \max_{0 \leq k < 2^l} \frac{|x_{lk}|}{w_l} = 0 \right\}. \tag{18}$$

Equipped with the norm $\|x\|_{\mathcal{M}_0} = \sup_{l \geq 0} \max_{0 \leq k < 2^l} |x_{lk}|/w_l$, this is a separable Banach space [10]. In a slight abuse of notation, we write $f \in \mathcal{M}_0$ if the sequence of its Haar wavelet coefficients $(\langle f, \psi_{lk} \rangle)_{l,k}$ belongs to that space.

Let us consider a credible ball in the space $\mathcal{M}_0$: recalling the definition (14) of the median tree estimator $\hat{f}_{\mathcal{T}^*}$, let us choose $R_n = R_n(X)$ in such a way that

$$\Pi[\|f - \hat{f}_{\mathcal{T}^*}\|_{\mathcal{M}_0(w)} \leq R_n/\sqrt{n} \mid X] = 1 - \gamma \tag{19}$$

(or possibly $\geq 1-\gamma$ if the equation has no exact solution, in which case the limit in the confidence statement of the next proposition is replaced by a liminf and equality by $\geq$).

Let us define, for R_n as in (19), σ_n as in (15) and $f_{\mathcal{T}^*}$ the median tree estimator (14),

$$\mathcal{C}_n^{\mathcal{M}} = \left\{f:\ \|f-\hat{f}_{\mathcal{T}^*}\|_\infty \leq \sigma_n\right\} \bigcap \left\{f:\ \|f-\hat{f}_{\mathcal{T}^*}\|_{\mathcal{M}_0(w)} \leq R_n/\sqrt{n}\right\}. \quad (20)$$

The next result states that $\mathcal{C}_n^{\mathcal{M}}$ is under self–similarity asympotically a confidence band of prescribed level $1-\gamma$.

Proposition 4. *Let $0<\alpha_1<\alpha_2\leq 1$, $K>0$, $\mu>0$ and $\eta>0$. Let $\mathcal{C}_n^{\mathcal{M}}$ be defined by (20), for $v_n/\log^{1/2} n\to\infty$, and Π an OPT prior with flat initialisation up to level $l_0(n)$ that verifies $\sqrt{\log n}\leq l_0(n)\leq \log n/\log\log n$, and other than that for $l>l_0(n)$ with same parameters as the prior in Theorem 1. First, for the admissible sequence $w_l=l^{2+\delta}$ for some $\delta>0$, the set $\mathcal{C}_n^{\mathcal{M}}$ is a $(1-\gamma)$–credible band as, uniformly on $\alpha\in[\alpha_1,\alpha_2]$ and $f_0\in\mathcal{S}(\alpha,K,\eta)\cap\mathcal{F}(\alpha,K,\mu)$,*

$$\Pi[\mathcal{C}_n^{\mathcal{M}}\,|\,X] = 1-\gamma+o_{P_0}(1).$$

Further, under the same conditions,

$$\left|\mathcal{C}_n^{\mathcal{M}}\right|_\infty = O_{P_0}\left(v_n\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}\right),$$

$$P_0\left[f_0\in\mathcal{C}_n^{\mathcal{M}}\right] = 1-\gamma+o(1).$$

Proposition 4 quite directly follows from combining Theorem 2, which concerns $\mathcal{C}_n$ and the nonparametric BvM Theorem 4 proved in the Appendix, which concerns the second part of the intersection in (20). Compared to $\mathcal{C}_n$ the advantage of $\mathcal{C}_n^{\mathcal{M}}$ is that it uses more 'posterior information' by intersecting with the $\mathcal{M}_0(w)$ credible ball, resulting in a credible ball with both credibility and confidence close to a given user–specified confidence level $1-\gamma$. By contrast, $\mathcal{C}_n$ was more 'conservative' in this respect, having credibility and confidence both going to 1. The behaviour of the credible band $\mathcal{C}_n^{\mathcal{M}}$, in particular in comparison to $\mathcal{C}_n$ from (16), is illustrated in simulations in the following Section 5.

Remark 1 (The choice of $\hat{f}_{\mathcal{T}^*}$ as centering)**.** The median tree density estimator can also be written

$$\hat{f}_{\mathcal{T}^*} = 1+\sum_{l\geq 0,k} 1_{\Pi[(l,k)\in\mathcal{T}_{\text{int}}|X]\geq 1/2}\cdot\hat{f}_{lk}\cdot\psi_{lk}, \quad (21)$$

where $\hat{f}_{lk} = 2^{l/2}(N_X\left(I_{(l+1)(2k+1)}\right) - N_X\left(I_{(l+1)(2k)}\right)/n$. The median–tree estimator is (in part) Bayesian: it uses the median tree built from the posterior distribution $\Pi[\cdot\,|\,X]$, and then a simple estimate $\hat{f}_{lk}$ of the wavelet coefficient f_{lk}. Note that this last choice is for simplicity, and other choices could be considered:

for instance, one could take the mean of the posterior distribution induced on the wavelet coefficient f_{lk} given the node (l,k) belongs to the tree. Yet, $\hat{f}_{lk}$ above makes for more transparent and less technical arguments, so we preferred it over other possible choices (a similar in spirit median–tree–estimator is considered in the regression context in [11], see Eq. (30)).

One main intuition and interpretation behind the median–tree–estimator is that it performs a wavelet thresholding, with the thresholding quantile selected automatically by the posterior. Lemma 4 proves that this thresholding is performed at level $O\left(n^{-1/2}\log n\right)$ thanks to the exponential decay of the splitting probabilities in the tree. In other words, (21) above automatically performs a *variable selection* along a certain tree, the median tree. We find this enhances the interpretability of the estimator (over other estimators such as the posterior mean, which does not come with variable selection; other than this the posterior mean in itself could possibly used as well, see below), which can be viewed as a pair $(\mathcal{T}^*, \hat{f})$, the median tree $\mathcal{T}^*$ specifying which nodes (l,k) are kept in the wavelet decomposition.

Instead of $\mathcal{T}^*$, we could think of using a 'mean tree'. Seeing that a tree is defined as a set of pairs of non-negative integers, a possibility is to interpret it as a *weighted tree*, that is, every node (l,k) is assigned a weight $w_{lk} \in (0,1)$. For instance, set

$$w_{lk}(X) = \Pi\left[(l,k) \in \mathcal{T}_{\text{int}} | X\right],$$

the weights of the 'mean-tree' and define a 'mean-tree' estimator as

$$\tilde{f} = 1 + \sum_{l \geq 0, k} w_{lk}(X)\hat{f}_{lk}\psi_{lk}.$$

While $\hat{f}_{\mathcal{T}^*}$ operates a 'hard' thresholding of wavelet coefficients, the mean tree estimator $\tilde{f}$ is more akin to a 'soft' selection. Though this 'soft' selection procedure would be of particular interest, we focus on the median tree in view of its properties (see Appendix D), as well as for its 'variable selection'/'tree selection' interpretation.

Replacing the median–tree density estimator $\hat{f}_{\mathcal{T}^*}$ by the posterior mean $\int f d\Pi(f \mid X)$ should conceivably lead to the same properties as in Proposition 4 (minimax adaptive L^∞-diameter and exact confidence/credible level). Our proposed confidence set is defined as the intersection of two balls centered on $\hat{f}_{\mathcal{T}^*}$, in supremum and multiscale norms. For the multiscale ball, our results rely on obtaining a Bernstein-von Mises theorem (see Appendix E). This requires a centering on an efficient estimator, converging at rate $n^{-1/2}$ in multiscale distance. Remark 2 of [10] gives conditions to verify that the posterior mean is actually an efficient estimator as well, notably implying that

$$\left\| E\left(f|X\right) - \hat{f}_{\mathcal{T}^*} \right\|_{\mathcal{M}_0(w)} = o_{P_{f_0}}\left(n^{-1/2}\right).$$

One could also prove the latter directly. As for the supremum ball, our result essentially relies on the fact that $\hat{f}_{\mathcal{T}^*}$ is an adaptive minimax estimator (in probability, see Lemma 7) of f_0 for this norm (up to some logarithmic term). Any other adaptive estimator should work. However, the posterior mean proves to have a more involved explicit expression and to be more complex to analyze than $\hat{f}_{\mathcal{T}^*}$. Other estimators, potentially frequentist ones with thresholds based on empirical wavelet coefficients, should work as well. Once again, the advantage of the median–tree–estimator is its simplicity and interpretability, so that we focus on this choice here.

Remark 2 (Necessity for flat initialisation)**.** Our approach to uncertainty quantification mainly relies on the obtention of a shape result for the posterior, an adaptive Bernstein-von Mises theorem (Theorem 4). It states that a $\sqrt{n}$-rescaling of the posterior converges weakly to a Gaussian limit, in a multiscale space. Without the flat initialisation, selecting all nodes (or equivalent Haar wavelet coefficients) up to level $l_0(n)$, the prior proves to be too sparse and there can be 'holes' in the limit. This was observed first by K. Ray in [36] (Proposition 3.7) in a white noise model setting, where the author proves that without flat intitialisation there exists self-similar densities such that, with high probability, the posterior (on the $\sqrt{n}$ scale) allocates vanishing mass to any multiscale ball, if the prior is too sparse. As a consequence, it becomes impossible to perform uncertainty quantification in this case, even under the good frequentist self-similarity property which is a central assumption for the construction of adaptive confidence regions. As explained in [36], the rescaled posterior behaves like this because it actually selects wavelet coefficients by thresholding at level $n^{-1/2}\log n$, instead of level $n^{-1/2}$ (see Lemma 4). For large l's, it turns out that the weighting sequence $(w_l)_{l\geq 0}$ can actually regularize the extra log factor. As for small l's, forcing the inclusion of all first nodes in the prior and the fitting of the first wavelet coefficients corrects this shortcoming.

5. Simulation study

We consider the credible sets $\mathcal{C}_n$ and $\mathcal{C}_n^{\mathcal{M}}$ defined in (16) and (20) respectively and illustrate their coverage and diameter properties numerically through a simulated study.

We focus on a prior as in Proposition 4, with parameters $\Gamma = 1.1$, $a = 1$ and $l_0(n) = \sqrt{\log n}$. We take four fairly different densities f_0, illustrating different aspects of inference and UQ with Optional Pólya trees:

- The triangular density $x \mapsto (.5+2*x)1_{0\leq x<0.5}+(1.5-2*(x-.5))1_{0.5\leq x<1}$ that is Lipschitz regular.
- The density
$$t \mapsto \frac{e^{W_t}}{\int_0^1 e^{W_s} ds}$$
where $(W_t)_{t\in[0;1)}$ is a Brownian motion that is almost surely $(1/2-\delta)$–Hölder regular for any $0<\delta<1/2$.
- The density
$$t \mapsto C\left(e^{W_t}1_{0\leq t<0.5} + c1_{0.5\leq x<1}\right\}$$

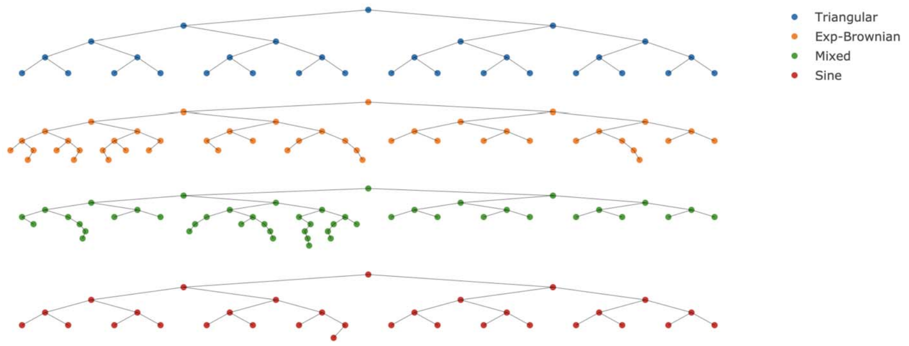


Fig 2: Interior nodes $\mathcal{T}^*_{\text{int}}$ of the median tree - $n = 10^5$.

for $(W_t)_{t\in[0;1)}$ a Brownian motion and C, c real numbers such that this actually defines a continuous density function. In this case, the regularity is different and of a higher order on the second half of the interval.

- The sine density $t \mapsto 1 + 0.5 * \sin(2\pi x) \in C^\infty([0;1))$.

We first illustrate the behaviour of the median tree $\mathcal{T}^*$ and the associated estimator $\hat{f}_{\mathcal{T}^*}$ defined in (14) in these different situations. In Figure 2, we observe how this tree adapts to the regularity of the underlying sampling density f_0 via the interior nodes it selects. First, in the case of the smoother sine and triangular densities, fewer nodes are included, while the tree grows deeper with the other two more irregular signals. Indeed, as mentioned before and explicited in Lemma 5, the median tree can be shown to have a depth close to the oracle cut-off L_n^*, satisfying $2^{L_n^*} \approx n^{1/(2\alpha+1)}$. However, although the sine density is even more regular than the triangular one, their respective median trees have a similar behaviour and grow at the same pace. Indeed, since we use a piecewise constant tree estimator which relates to the Haar wavelet basis, our method cannot leverage additional regularities, beyond $\mathcal{C}^1[0,1)$. Finally, when it comes to the mixed density, the median tree has a spatial-dependent behaviour. It includes much more nodes in regions that corresponds to the first half of the sampling space, where the target regularity is that of the exp-Brownian density. As for the other half of the sampling space, it doesn't get deeper than $l_0(n)$. It highlights a desirable feature of tree-based methods, that is their spatial adaptivity. While we consider adaptation to global regularity in our theoretical results, one could also consider local adaptation, as was recently considered in [37], where results on local adaptation for tree–based priors (among others) are obtained in a regression setting.

In Figure 3, for the four sampling densities, we illustrate the estimator $\hat{f}_{\mathcal{T}^*}$ (orange) and the bounds of the credible set $\mathcal{C}_n$ (red), where we took $v_n = (\log n)^{0.501}$ in (15). The estimator (14) struggles to approximate the 'spiky' portions of the most irregular signals. Still, in any case, the credible band covers the true density f_0 as expected.

Then, to illustrate the intersected set $\mathcal{C}_n^{\mathcal{M}}$, defined in (20) via a multiscale

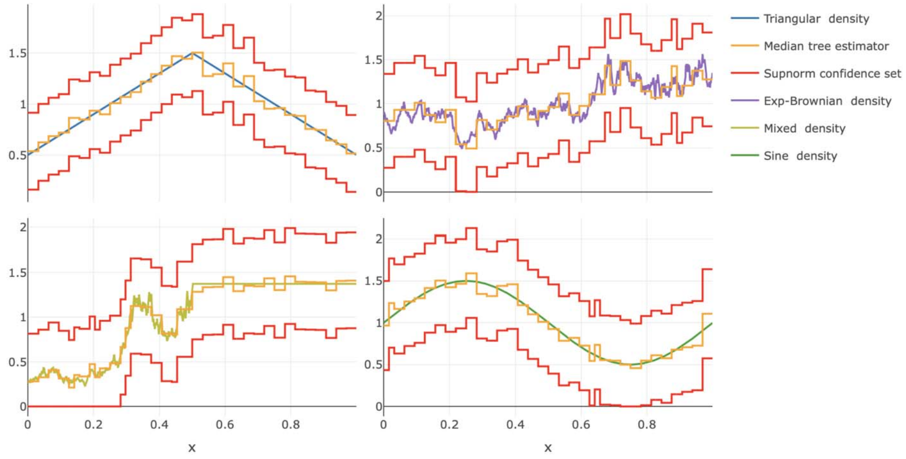


Fig 3: Median tree estimator $\hat{f}_{\mathcal{T}^*}$ and credible set $\mathcal{C}_n$ - $n = 10^4$

Table 1
Credibility of sets $C_n^{L^\infty}$ and $\mathcal{C}_n^{\mathcal{M}}$ for the triangular density f_0.

Chosen significance γ	0.99	0.95	0.9	0.85
	$n = 10^4$			
Credibility of $C_n^{L^\infty}$	0.99	0.95	0.9	0.85
Credibility of $\mathcal{C}_n^{\mathcal{M}}$	0.99	0.95	0.8981	0.85
Credibility of $C_n^{L^\infty} \cap \mathcal{C}_n^{\mathcal{M}}$	0.9801	0.9029	0.8108	0.725
Credibility of the intersection if independence	0.9801	0.9025	0.81	0.7225
	$n = 10^5$			
Credibility of $C_n^{L^\infty}$	0.99	0.95	0.9	0.85
Credibility of $\mathcal{C}_n^{\mathcal{M}}$	0.9894	0.9494	0.8994	0.8494
Credibility of $C_n^{L^\infty} \cap \mathcal{C}_n^{\mathcal{M}}$	0.9801	0.9028	0.8118	0.7254
Credibility of the intersection if independence	0.9795	0.9019	0.8095	0.722

condition, we sampled 10000 draws from the posterior and plotted, in Figure 4, 100 of those belonging to the confidence band (blue), for $\gamma = 0.05$. We use the admissible sequence $w_l = l^{2+\delta}$ for some $\delta = 0.01$ to compute the multiscale distance, complying with the requirements of Proposition 4. Most of those samples do not seem to lie close to the bounds of $\mathcal{C}_n$ which is consistent with the fact that $\mathcal{C}_n$, resp. $\mathcal{C}_n^{\mathcal{M}}$, has a posterior mass close to 1, respectively 0.95. Though our illustrations concern the intersection of $\mathcal{C}_n^{\mathcal{M}}$ with the support the posterior, via the representation of posterior draws, it appears that $\mathcal{C}_n^{\mathcal{M}}$ is actually smaller than $\mathcal{C}_n$.

As for the confidence sets $\mathcal{F}_n$ on the cumulative distribution function $F_0(\cdot) = \int_0^{\cdot} f_0(t)dt$, we illustrate an example in Figure 5 for a smaller sample size of $n = 10^3$ and $\gamma = 0.05$. The bounds of $\mathcal{F}_n$ follow tightly the true signal and the set covers it, in spite of the fewer number of observations available compared to previous plots. Indeed, following the discussion after Theorem 3, $\mathcal{F}_n$ has a radius decreasing at the parametric rate $\sqrt{n}^{-1}$.

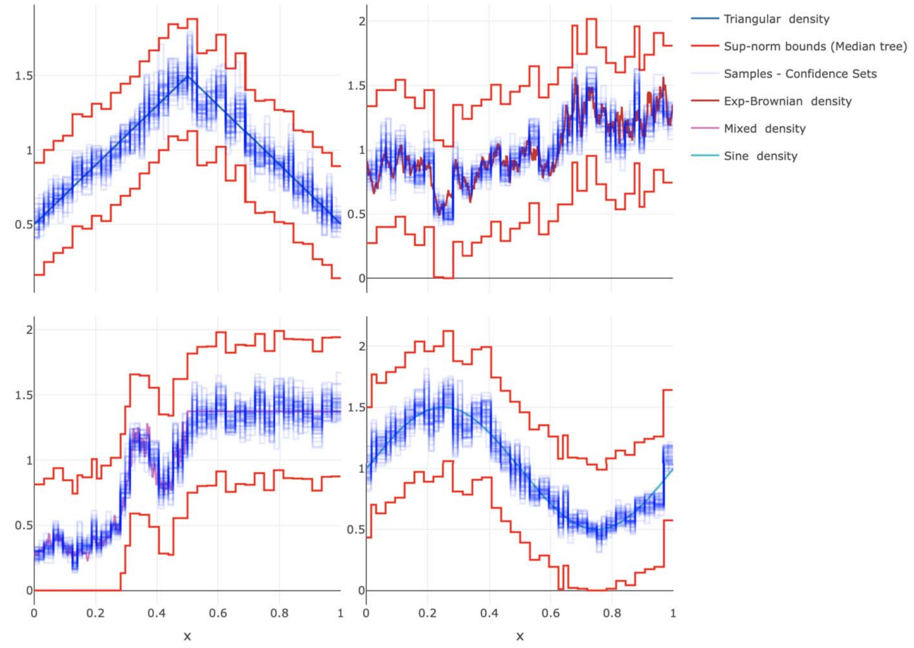


Fig 4: Posterior sample in the confidence band $\mathcal{C}_n^{\mathcal{M}}$ - $\gamma = 0.05$ and $n = 10^4$.

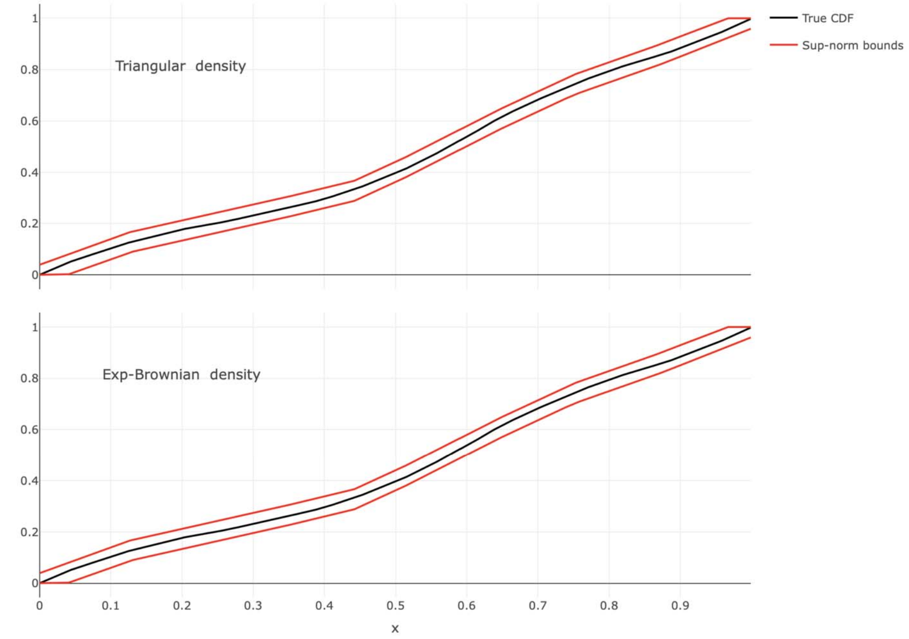


Fig 5: Posterior samples in the confidence set $\mathcal{F}_n$ - $n = 10^3$.

We end this section with an illustration of a phenomenon that was noticed and established in [36] for a spike-and-slab prior in a regression setting. Namely, since we constructed an adaptive $(1-\gamma)$-confidence bands whose diameter in supnorm shrinks at an almost optimal rate, one may wonder how much it differs from the $(1-\gamma)$-credible band in the supremum norm $C_n^{L^\infty} := \left\{f : \left\|f - \hat{f}_{\mathcal{T}^*}\right\|_\infty \leq Q_n(\gamma)\right\}$, where $Q_n(\gamma)$ is chosen such that $\Pi\left[C_n^{L^\infty}|X\right] \geq 1-\gamma$. In a white noise regression setting, [36] proved that these two sets are asymptotically independent (see Theorem 5.3 therein), in the sense that $\Pi\left[C_n^{L^\infty} \cap \mathcal{C}_n^{\mathcal{M}}|X\right] \overset{P_{f_0}}{\to} (1-\gamma)^2$. As above, we sampled 10^4 draws from the posterior to estimate de posterior credibility of the different sets, which we present in Table 1. The results seem to indicate that the independence phenomenon of the credible sets as described above still hold in the present density estimation setting, as the margin of difference observed is of the order of the Monte-Carlo error. Intuitively speaking, this independence under the posterior if true (at least asymptotically) would mean that the two credible sets reflect *different* aspects of the posterior distribution. Although this result from [36] is seemingly verified in density estimation with an OPT prior, we did not investigate this question from a theoretical point of view in the present paper; we expect the proof to be significantly more involved than in the (conjugate) Gaussian white noise setting and we leave this point for future work.

6. Discussion

In the present work we establish an inference theory for Optional Pólya trees introduced in [39] by deriving posterior contraction rates as well as confidence bands for the problem of uncertainty quantification. By contrast, only posterior consistency had been previously obtained until now for such priors. Although we focus on this class of prior distributions, we point out that our proofs and results also apply to different tree priors, such as ones conditioning on the number of leaves as in Example 2. The results and proofs highlight how beneficial a multiscale approach to study tree-based methods, as introduced in [11], can be.

As for related priors in density estimation, non-adaptive contraction rates were obtained in [7] for Pólya trees for carefully chosen regularity-dependent parameters of the Beta random variables. The addition of a hyperprior on the tree structure in OPTs allows for adaptation, so that the Beta parameters can be set as an arbitrary constant (a similar comment can be done about Spike-and-slab Pólya trees [8]). The Beta variables in the Pólya-like mass allocation mechanism could be replaced by another distribution, but we sticked to them for simplicity of analysis and presentation.

In Section 5, we mentioned some further results on OPTs to be investigated. First, tree-based methods have a natural ability to adapt to the local regularity. While this has been proved in [37] in a regression setting, this should also be the case with OPTs in density estimation. Another expected advantage of trees is that in high-dimensional settings, they induce a 'tree–structured' sparsity, which could help in addressing the curse of dimensionality. As original OPTs [31] have been introduced in arbitrary dimensions, it is natural to further our theoretical analysis in this direction, Also, the interesting alleged posterior independence of sets $C_n^{L^\infty}$ and $\mathcal{C}_n^{\mathcal{M}}$ still needs to be proven and would confirm that the two constructions rely on somewhat different aspects of the posterior distribution in density estimation too.

Finally, since we use a 'hard'-tree construction (i.e. a histogram), it is quite expected that similar limitations arise as with typical histograms estimators, for which adaptation to smoothness is often limited to $\alpha \in (0,1]$. But indeed, it is natural to wonder if one could possibly adapt beyond regularity one. In order to achieve faster rates for smoother densities, one possibility explored in [27] consists in replacing 'hard' (histogram) trees with 'smooth' trees. Another promising possibility is to look at forests priors. Indeed, the aggregation of many trees tends to result in estimators that are more 'regular' and thereby more suitable to the estimation of smoother objects: for frequentist estimators in regression, this was noted in [1, 32] for regularities $\alpha \leq 2$. We also recently considered this question in the paper [35] for the Hellinger loss, where it is shown that a certain aggregation of Pólya trees can adapt to any Hölder–regularity $\alpha > 0$ under this loss. Note that in the present paper we consider the stronger supremum norm loss, which is typically more challenging (e.g. the general Ghosal–Ghosh–van der Vaart concentration theorems [17], as used in [35], do not directly apply).

7. Proof of the main results

Below, the depth $L_n = L_n(\alpha)$ defined as

$$2^{L_n} = c_0(n/\log n)^{\frac{1}{1+2\alpha}}, \tag{22}$$

for some $c_0 > 0$, will be helpful in our theoretical analysis. Also, C stands for a generic constant whose precise value we do not track and can change from line to line.

7.1. Proof of Theorem 1

Let's write $T_n = \{\mathcal{T} \,|\, d(\mathcal{T}) \leq L_n,\ S(f_0,\tau) \subset \mathcal{T}\}$, $S(f_0,\tau)$ as in Lemma 2, and $\mathcal{E}_n = \{f :\ \exists \mathcal{T} \in T_n,\ f \text{ piecewise constant on } I_{\mathcal{T}}\}$. Moreover, we write, for L_n as in (22) and any tree $\mathcal{T} \in \mathbb{T}_n$, the following othogonal projections of f_0: $f_0^{\mathcal{T}}$ onto the span $\{\psi_{lk} \,|\, (l,k) \in \mathcal{T}\}$, $f_0^{L_n^c}$ onto $\{\psi_{lk} \,|\, l > L_n\}$, and $f_0^{\mathcal{T}^c, L_n}$ onto the orthocomplement of the union of the two last spans. For $f_0 \in \mathcal{C}^\alpha[0,1)$, $0 < \alpha \leq 1$, we have in particular that

$$||f_0^{L_n^c}||_\infty \leq \sum_{l > L_n} 2^{l/2} \max_{0 \leq k < 2^l} |f_{0,lk}| \lesssim \sum_{l > L_n} 2^{-l\alpha} \lesssim \left(n^{-1}\log n\right)^{\frac{\alpha}{1+2\alpha}}, \tag{23}$$

(see for instance [7]). Then, for any density f_0, we have the upper bound, for $\mathcal{B}_M$ as in Lemma 8,

$$\begin{aligned}&\Pi\left[\|f-f_0\|_\infty > M_n\epsilon_n \,|\, X\right]\\&\le \Pi[\mathcal{E}_n^c \,|\, X]\mathbb{1}_{\mathcal{B}_M} + \Pi\left[\|f-f_0\|_\infty > M_n\epsilon_n,\ f\in\mathcal{E}_n \,|\, X\right]\mathbb{1}_{\mathcal{B}_M} + \mathbb{1}_{\mathcal{B}_M^c}.\end{aligned}$$

On one hand, Lemma 8 guarantees that $\mathbb{P}_0\left(\mathcal{B}_M^c\right) = o(1)$ for M large enough and Lemmas 1 and 2 ensures that

$$E_{f_0}\left\{\Pi_f[\mathcal{E}_n^c \,|\, X]\mathbb{1}_{\mathcal{B}_M}\right\} = o(1).$$

On the other hand, we also have the inequality $\|f-f_0\|_\infty \le \left\|f - f_0^{\mathcal{T}}\right\|_\infty + \left\|f_0^{\mathcal{T}^c,L_n}\right\|_\infty + \left\|f_0^{L_n^c}\right\|_\infty$. This allows us to control the last term in the above upper bound by mean of the Markov inequality:

$$\begin{aligned}\Pi\Big[f\in\mathcal{E}_n,\quad \|f-f_0\|_\infty > M_n\epsilon_n \,|\, X\Big]\mathbb{1}_{\mathcal{B}_M} &\le (M_n\epsilon_n)^{-1}\int_{\mathcal{E}_n}\|f-f_0\|_\infty\, d\Pi[f,\mathcal{T}\,|\,X]\mathbb{1}_{\mathcal{B}_M}\\ \le (M_n\epsilon_n)^{-1}\Big[\int_{\mathcal{E}_n}&\left\|f-f_0^{\mathcal{T}}\right\|_\infty d\Pi[f,\mathcal{T}\,|\,X]\mathbb{1}_{\mathcal{B}_M}\\ &+ \int_{\mathcal{E}_n}\left\|f_0^{\mathcal{T}^c,L_n}\right\|_\infty d\Pi[\mathcal{T}\,|\,X]\mathbb{1}_{\mathcal{B}_M} + \left\|f_0^{L_n^c}\right\|_\infty\Big],\end{aligned} \tag{24}$$

and (23) ensures that the last term above is $o(1)$. Similarly, for the second term, using the definition of $\mathcal{E}_n$ and denoting L^* the largest integer such that $2^{-L^*(\alpha+1/2)} \ge n^{-1/2}\log n$,

$$\begin{aligned}\|f_0^{\mathcal{T}^c,L_n}\|_\infty &\le \sum_{l\le L_n} 2^{l/2}\max_{k:(l,k)\notin\mathcal{T}}|f_{0,lk}| \lesssim \sum_{l\le L_n} 2^{l/2}\left(\max_{0\le k<2^l}|f_{0,lk}| \wedge \log n \,/\, \sqrt{n}\right)\\ &\lesssim \sum_{l\le L^*} 2^{l/2}\frac{\log n}{\sqrt{n}} + \sum_{L^*<l\le L_n} 2^{l/2}2^{-l(1/2+\alpha)} \lesssim 2^{L^*/2}\frac{\log n}{\sqrt{n}} + 2^{-L^*\alpha} \lesssim 2^{-L^*\alpha}.\end{aligned}$$

This allows us to conclude that the second term in the bound (24) is also of the order $o(1)$. It remains to bound the first term in the bound that is also of order $o(1)$ according to Lemma 3. This concludes our proof.

It remains to prove the different lemmas we used to upper bound the different terms above.

Lemma 1. *Suppose $f_0 \in \mathcal{F}(\alpha, K, \mu)$, for some $\mu > 0$, $0 < \alpha \le 1$, $K > 0$, and assume f follows a prior as in Theorem 1. Then, for any $M > 0$ as in Lemma 8 and Γ large enough, on events $\mathcal{B}_M$, we have, as $n \to \infty$,*

$$\Pi[d(\mathcal{T}) > L_n \,|\, X] \to 0,$$

where L_n is as in (22).

Proof. Let $\mathcal{T}$ be a tree of depth $L_n < d\left(\mathcal{T}\right) = l \le L_{\max}$. Then, for

$$\tilde{k} = \min_{(2k,l)\in\mathcal{T}} k, \qquad \epsilon = \epsilon\left(\tilde{k}, l-1\right),$$

let $\mathcal{T}^-$ be the corresponding tree whose nodes $(l, 2\tilde{k})$ and $(l, 2\tilde{k}+1)$ have been removed, i.e. $\mathcal{T} = \mathcal{T}^- \cup \big\{(l, 2\tilde{k}), (l, 2\tilde{k}+1)\big\}$. From (8) and (9), we have

$$\begin{aligned}\Pi[\mathcal{T} \mid X] &= \Pi[\mathcal{T}^- \mid X] \frac{p_\epsilon^X}{1 - p_\epsilon^X} \left(1 - p_{\epsilon 0}^X\right)\left(1 - p_{\epsilon 1}^X\right) \\ &= \Pi[\mathcal{T}^- \mid X] p_\epsilon \frac{(1 - p_{\epsilon 0})(1 - p_{\epsilon 1})}{1 - p_\epsilon} \nu_\varepsilon^X \\ &\leq \left(1 - \Gamma^{-L_n}\right)^{-1} \Pi[\mathcal{T}^- \mid X] \frac{2^{N_X(I_\varepsilon)}}{\Gamma^{l+1}} \underbrace{\frac{B(a + N_X(I_{\varepsilon 0}), a + N_X(I_{\varepsilon 1}))}{B(a,a)}}_{=:Q}.\end{aligned} \tag{25}$$

Then, from Lemma 10, we have for $\tilde{n}_0 = N_X(I_{\epsilon 0})$, $\tilde{n}_1 = N_X(I_{\epsilon 1})$ and $\tilde{n} = N_X(I_\epsilon)$, that

$$Q \lesssim \underbrace{\frac{(2a + \tilde{n}_1 - 1/2)^{\tilde{n}_1} (2a + \tilde{n}_2 - 1/2)^{\tilde{n}_2}}{(2a + \tilde{n} - 1/2)^{\tilde{n}}}}_{=:Q_1} \underbrace{\frac{(2a + \tilde{n}_1 - 1/2)^{a-1/2} (2a + \tilde{n}_2 - 1/2)^{a-1/2}}{(2a + \tilde{n} - 1/2)^{2a-1/2}}}_{=:Q_2}.$$

Under our assumptions on f_0, on the event $\mathcal{B}_M$ and for n large enough,

$$n_X(I_{l,k}) \geq \frac{\mu}{2} n 2^{-l} \to \infty$$

for any $l \leq L_{\max}$. Under the same conditions,

$$|\tilde{n}_1 - \tilde{n}_2| \leq n \left|P_0(I_{\varepsilon 0}) - P_0(I_{\varepsilon 1})\right| + 2MM_{n,l} \leq nK2^{-l(1+\alpha)} + 2MM_{n,l}.$$

The last inequality stems from the fact that f_0 is α-Hölder regular. Therefore, on $\mathcal{B}_M$, for n large enough, if we note $v_{\tilde{n}_1,\tilde{n}_2} = \tilde{n}_1 - \tilde{n}_2$, since $\tilde{n} = \tilde{n}_1 + \tilde{n}_2$ and $\log(1+x) \leq x$ for $x > -1$,

$$\begin{aligned}&Q_1 \\ &= \exp\left(\tilde{n}_1 \log\left(\frac{1}{2} + \frac{\tilde{n}_1 - \tilde{n}_2 + 2a - 1/2}{2(2a - 1/2 + \tilde{n})}\right) + \tilde{n}_2 \log\left(\frac{1}{2} - \frac{\tilde{n}_1 - \tilde{n}_2 - 2a + 1/2}{2(2a - 1/2 + \tilde{n})}\right)\right) \\ &= \frac{1}{2^{\tilde{n}}} \exp\left(\tilde{n}_1 \log\left(1 + \frac{v_{\tilde{n}_1,\tilde{n}_2} + 2a - 1/2}{2a - 1/2 + \tilde{n}}\right) + \tilde{n}_2 \log\left(1 - \frac{v_{\tilde{n}_1,\tilde{n}_2} - 2a + 1/2}{2a - 1/2 + \tilde{n}}\right)\right) \\ &\leq \frac{1}{2^{\tilde{n}}} \exp\left(\frac{v_{\tilde{n}_1,\tilde{n}_2}^2}{2a - 1/2 + \tilde{n}} + \frac{\tilde{n}(2a - 1/2)}{2a - 1/2 + \tilde{n}}\right) \\ &\leq \frac{C}{2^{\tilde{n}}} \exp\left(\frac{8K^2 n^2 2^{-2l(1+\alpha)}}{\mu n 2^{-l}} + \frac{16M^2 M_{n,l}^2}{\mu n 2^{-l}}\right) \\ &\leq \frac{C}{2^{\tilde{n}}} \exp\left(\left(8K^2 \mu^{-1} c_0^{-1-2\alpha} + 32M^2 (\mu \log 2)^{-1}\right) \log n\right).\end{aligned}$$

The last inequality stems from $l > L_n$ and the definition of L_n. The last factor is even easier to control as, on $\mathcal{B}_M$,

$$Q_2 \lesssim \left[n2^{-l}\right]^{-1/2} \lesssim n^{-\frac{\alpha}{1+2\alpha}} \log(n)^{-\frac{1/2}{1+2\alpha}}.$$

Finally, this leads us to

$$\Pi[\mathcal{T} \mid X] = o\left(\Pi[\mathcal{T}^- \mid X] \frac{n^{\left(8K^2\mu^{-1}c_0^{-1-2\alpha}+32M^2(\mu\log 2)^{-1}-\frac{\alpha}{1+2\alpha}\right)}}{\Gamma^l}\right)$$

uniformly on $\mathcal{T}$ such that $L_n < d(\mathcal{T}) = l \leq L_{\max}$. The application $\mathcal{T} \longrightarrow \mathcal{T}^-$ defined above is surjective and is such that each tree $\mathcal{T}^-$ is the image of at most 2^{l-1} trees $\mathcal{T}$. Then, the event of interest verifies for $\Gamma > 2$ and $\bar{C} = 8K^2\mu^{-1}c_0^{-1-2\alpha} + 32M^2(\mu\log 2)^{-1} - \frac{\alpha}{1+2\alpha}$,

$$\begin{aligned}\Pi[d(\mathcal{T}) > L_n \mid X] &= \sum_{l=L_n+1}^{L_{\max}} \Pi[d(\mathcal{T}) = l \mid X] = \sum_{l=L_n+1}^{L_{\max}} \sum_{\mathcal{T}:d(\mathcal{T})=l} \Pi[\mathcal{T} \mid X] \\ &= o\left(\sum_{l=L_n+1}^{L_{\max}} \sum_{\mathcal{T}:d(\mathcal{T})=l} \Pi[\mathcal{T}^- \mid X] \frac{n^{\bar{C}}}{\Gamma^l}\right) \qquad (26) \\ &= o\left(\sum_{l=L_n+1}^{L_{\max}} \sum_{\mathcal{T}^-} \Pi[\mathcal{T}^- \mid X] \frac{2^l n^{\bar{C}}}{\Gamma^l}\right) = o\left(\frac{2^{L_n} n^{\bar{C}}}{\Gamma^{L_n}}\right)\end{aligned}$$

which is $o(1)$ whenever $\{\log\Gamma/\log 2 - 1\}/(1+2\alpha) \geq \bar{C}$, that is, if $\Gamma \geq 2^{1+\bar{C}(1+2\alpha)}$. □

Lemma 2. *Under the same assumptions on f_0 as in Lemma 1, for Π as in Theorem 1 and on the events $\mathcal{B}_M$ from Lemma 8, for $\tau > 0$ large enough and L_n as in (22), the set*

$$S(f_0, \tau) := \left\{(l,k) : |f_{0,lk}| \geq \tau \frac{\log n}{\sqrt{n}}\right\}$$

satisfies, as $n \to \infty$,

$$\Pi[\{\mathcal{T} : S(f_0,\tau) \not\subset \mathcal{T}_{int}\} \mid X] \to 0.$$

Proof. First, since $f_0 \in \Sigma(\alpha, K)$ for some $\alpha, K > 0$, there exists $C > 0$ such that, for any $l \geq 0, 0 \leq k < 2^l$, $|f_{0,lk}| \leq C2^{-l(\alpha+1/2)}$. Thus, for τ large enough, $(l,k) \in S(f_0,\tau)$ implies $l \leq L_n$.

Now, let's take (l_S, k_S) a node in $S(f_0,\tau)$. Then, let's define

$$\mathbb{T}_{n,(l_S,k_S)} := \{\mathcal{T} \in \mathbb{T}_n \mid (l_S,k_S) \notin \mathcal{T}_{\text{int}}\},$$

the set of trees in the support of our prior distribution on tree structures that do not have (l_S,k_S) as an internal node, and $\epsilon = \epsilon(k_S,l_S)$. To any tree $\mathcal{T} \in \mathbb{T}_{n,(l_S,k_S)}$, it is possible to associate the full binary tree $\mathcal{T}^+$ which is the smallest extension of $\mathcal{T}$ with (l_S,k_S) as an interior node,

$$\mathcal{T}^+ = \underset{\mathcal{T}' \in \mathbb{T}_n:\ \mathcal{T} \subset \mathcal{T}',\ (l_S,k_S) \in \mathcal{T}'_{\text{int}}}{\arg\min} |\mathcal{T}'|.$$

This new tree is realized with the completion of the route from the root to the node (l_S, k_S), starting from the leaf node (l_0, k_0) of this route which is included in $\mathcal{T}$. Then, as in (25) and using Lemma 10, we now have for some constant $C > 1$,

$$\begin{aligned}
&\frac{\Pi[\mathcal{T} \,|\, X]}{\Pi[\mathcal{T}^+ \,|\, X]} \\
&\quad \le C^{{l_S}^2} \prod_{l=l_0}^{l_S} \left(2^{n_X\left(I_{\epsilon^{[l]}}\right)} \mathrm{B}\Big(a + n_X\left(I_{\epsilon^{[l]}0}\right), a + n_X\left(I_{\epsilon^{[l]}1}\right)\Big) \right)^{-1} \\
&\quad \le C^{{l_S}^2} \underbrace{\prod_{l=l_0}^{l_S} \frac{(2a + n_X\left(I_{\epsilon^{[l]}}\right) - 1/2)^{2a-1/2}}{(a + n_X\left(I_{\epsilon^{[l]}0}\right) - 1/2)^{a-1/2} (a + n_X\left(I_{\epsilon^{[l]}1}\right) - 1/2)^{a-1/2}}}_{=:\, Q_1} \\
&\quad \underbrace{\prod_{l=l_0}^{l_S} \frac{(2a + n_X\left(I_{\epsilon^{[l]}}\right) - 1/2)^{n_X\left(I_{\epsilon^{[l]}}\right)}}{2^{n_X\left(I_{\epsilon^{[l]}}\right)} (a + n_X\left(I_{\epsilon^{[l]}0}\right) - 1/2)^{n_X\left(I_{\epsilon^{[l]}0}\right)} (a + n_X\left(I_{\epsilon^{[l]}1}\right) - 1/2)^{n_X\left(I_{\epsilon^{[l]}1}\right)}}}_{=:\, Q_2}.
\end{aligned} \tag{27}$$

where we recall that $\epsilon^{[l]}$ denotes the l first elements of the sequence ϵ. On the event $\mathcal{B}_M$, for all $l \le L_n + 1$ and possible k, we have, using that $f_0 \ge \mu > 0$, $N_X\left(I_{l,k}\right) \gtrsim n2^{-l} \gtrsim n2^{-L_n} \to \infty$ as $n \to \infty$. Since it is also upper bounded (as f_0 is a Hölder density), we have $N_X\left(I_{l,k}\right) \lesssim n2^{-l}$. Therefore, since these bounds are uniform on $l \le L_n + 1$,

$$Q_1 \le \prod_{l=l_0}^{l_S} C\left(n2^{-l}\right)^{1/2} \le C^{l_S} \sqrt{n}^{l_S}.$$

Also, in Q_2, the factor at index l is equal to, writing $\tilde{n}_0 = N_X\left(I_{\epsilon^{[l]}0}\right), \tilde{n}_1 = N_X\left(I_{\epsilon^{[l]}1}\right), \tilde{n} = N_X\left(I_{\epsilon^{[l]}}\right)$,

$$\exp\left[\tilde{n}_0 \log\left(\frac{2a - 1/2 + \tilde{n}}{2a - 1 + 2\tilde{n}_0}\right) + \tilde{n}_1 \log\left(\frac{2a - 1/2 + \tilde{n}}{2a - 1 + 2\tilde{n}_1}\right)\right].$$

If we write $KL(a;b)$ the Kullback-Leibler divergence between Bernoulli distributions of parameters $0 \le a, b \le 1$, then, for n large enough, on $\mathcal{B}_M$, this is bounded by

$$\exp\left[-C\tilde{n}KL\Big(\frac{a - 1/2 + \tilde{n}_0}{2a - 1 + \tilde{n}}; 1/2\Big)\right] \exp\left[\tilde{n} \log\left(1 + \frac{1}{4a - 2 + 2\tilde{n}}\right)\right].$$

The second factor can be bounded by a constant, uniformly on $l \le L_n + 1$. The first factor can be bounded by 1 for $l < l_S$, while for $l = l_S$, we can use the bound $KL(a;b) \ge \|\mathrm{Be}(a) - \mathrm{Be}(b)\|_1^2/2$ to write

$$\exp\left[-C\tilde{n}KL\Big(\frac{a - 1/2 + \tilde{n}_0}{2a - 1 + \tilde{n}}; 1/2\Big)\right] \le \exp\left[-C\tilde{n}^{-1}(\tilde{n}_0 - \tilde{n}_1)^2\right].$$

By definition $|f_{0,l_S k_S}| = 2^{l_S/2}\left|P_0(I_{(l_S+1)(2k+1)}) - P_0(I_{(l_S+1)(2k)})\right|$, so that on $\mathcal{B}_M$, $|\tilde{n}_0 - \tilde{n}_1| \geq n\,|f_{0,l_S k_S}|\,2^{-l_S/2} - 2MM_{n,l_S+1}$, hence the upper bound for τ large enough:

$$\exp\left[-C(\tau\log n - 2M\sqrt{l_S+1+L_n})^2\right] \leq \exp\left[-C\tau^2\log^2 n\right],$$

where we used the definition of S, M_{n,l_S+1}, L_n and $l_S \leq L_n$.

Finally, for τ large enough and using that $l_S \leq L_n \leq \log n$, we can conclude that there exists constants $C_1, C_2 > 0$ such that

$$\frac{\Pi[\mathcal{T}\,|\,X]}{\Pi[\mathcal{T}^+\,|\,X]} \leq C_1^{l_S^2} n^{-(C_2\tau^2-1/2)\log n} \leq n^{-(C_2\tau^2-1/2-\log C_1)\log n}. \tag{28}$$

Since any tree verifying $(l_S,k_S) \in \mathcal{T}$ is the image of at most l_S+1 trees by the map

$$\begin{aligned} \mathbb{T}_{n,(l_S,k_S)} &\to \{\mathcal{T}' \in \mathbb{T}_n :\ (l_S,k_S) \in \mathcal{T}'_{\text{int}}\} \\ \mathcal{T} &\mapsto \mathcal{T}^+ \end{aligned} ,$$

as it is the length of the path from the root to the node (l_S,k_S) in a tree $\mathcal{T} \in \mathbb{T}_n$,

$$\begin{aligned} \Pi[(l_S,k_S) \notin \mathcal{T}\,|\,X] &= \sum_{\mathcal{T}:(l_S,k_S)\notin\mathcal{T}} \frac{\Pi[\mathcal{T}\,|\,X]}{\Pi[\mathcal{T}^+\,|\,X]}\Pi[\mathcal{T}^+\,|\,X] \\ &\leq n^{-(C_2\tau^2-1/2-\log C_1)\log n}(l_S+1)\sum_{\mathcal{T}:(l_S,k_S)\in\mathcal{T}}\Pi[\mathcal{T}\,|\,X] \\ &\leq n^{-(C_2\tau^2-1/2-\log C_1)\log n}\log n, \end{aligned}$$

which allows us in conjunction with the definition of L_n to conclude that

$$\begin{aligned} \Pi[\{\mathcal{T} : S(f_0,\tau) \not\subset \mathcal{T}\}\,|\,X] &\leq \sum_{(l,k)\in S(f_0,\tau)} \Pi[(l,k) \notin \mathcal{T}\,|\,X] \\ &\leq 2^{L_n+1} n^{-(C_2\tau^2-1/2-\log C_1)\log n}\log n \\ &\to 0 \end{aligned}$$

as $n \to \infty$ for τ large enough. □

Lemma 3. *Let $T_n = \{\mathcal{T} \in \mathbb{T}_n :\ d(\mathcal{T}) \leq L_n,\ S(f_0,\tau) \subset \mathcal{T}\}$ for L_n as in (22), $c_0 > 0$ small enough, and $\tau > 0$ as in Lemma 2. Then, under the conditions of Lemma 1 and on the event $\mathcal{B}_M$ for $M > 0$ large enough, there exists a constant $C > 0$ such that for n sufficiently large, uniformly on $\mathcal{T} \in T_n$,*

$$\int \max_{(l,k)\in\mathcal{T}_{int}} |f_{lk} - f_{0,lk}|\, d\Pi[f\,|\,\mathcal{T},X] \leq C\sqrt{\frac{\log n}{n}}.$$

Proof. Given a tree $\mathcal{T}$, let us define the map $\bar{f}_{\mathcal{T}}$ such that, for each terminal node (l,k) in $\mathcal{T}_{\text{ext}}$ and $x \in I_{lk}$,

$$\bar{f}_{\mathcal{T}}(x) = 2^l \prod_{i=1}^{l} \bar{Y}_{\epsilon[i]}, \quad \epsilon = \epsilon\,(k,l)\,,$$

where

$$\bar{Y}_\epsilon = E\,[Y_\epsilon \,|\, X, \mathcal{T}] = \frac{a + N_X(I_{\epsilon 0})}{2a + N_X(I_\epsilon)}.$$

This defines the mean posterior density given the tree structure $\mathcal{T}$. Similarly, for each $(l,k) \in \mathcal{T}$, with $\epsilon = \epsilon\,(k,l)$, the mean probability measure of I_ϵ is

$$\bar{P}(I_\epsilon) = \prod_{i=1}^{|\epsilon|} \bar{Y}_{\epsilon[i]} =: \bar{p}_\epsilon.$$

Then, expressing the coefficients of the decomposition in the Haar wavelet basis of this mean posterior density, we obtain that for each $(l,k) \in \mathcal{T}_{\text{int}}$, $\epsilon = \epsilon\,(k,l)$,

$$\bar{f}_{\mathcal{T},lk} := \langle \bar{f}_{\mathcal{T}}, \psi_{lk} \rangle = 2^{l/2}\left(\bar{p}_\epsilon - 2\bar{p}_{\epsilon 0}\right) = 2^{l/2}\bar{p}_\epsilon\left(1 - 2\bar{Y}_{\epsilon 0}\right),$$

while $\bar{f}_{\mathcal{T},lk} = 0$ for $(l,k) \notin \mathcal{T}_{\text{int}}$. When it comes to the true sampling density f_0, we obtain the similar expression, denoting $p_{0,\epsilon} := P_0(I_\epsilon)$ and $y_{\epsilon 0} := \frac{P_0(I_{\epsilon 0})}{P_0(I_\epsilon)}$,

$$f_{0,lk} = 2^{l/2} p_{0,\epsilon}(1 - 2y_{\epsilon 0}),$$

and, for densities f sampled from the posterior distribution given $\mathcal{T}$, with $p_\epsilon := \prod_{i=1}^{|\epsilon|} Y_{\epsilon[i]}$,

$$f_{lk} = 2^{l/2}\tilde{p}_\epsilon\left(1 - 2Y_{\epsilon 0}\right)\mathbb{1}_{(l,k)\in\mathcal{T}_{\text{int}}}.$$

From now on, for simplicity of notations, $\epsilon = \epsilon(k,l)$ as the context will make it clear what the pair (l,k) is. For any $\mathcal{T} \in \mathbb{T}_n$, one can bound $|f_{lk} - f_{0,lk}| \le \left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| + \left|\bar{f}_{\mathcal{T},lk} - f_{0,lk}\right|$. Using the above expressions, the second term is rewritten as

$$\left|\bar{f}_{\mathcal{T},lk} - f_{0,lk}\right| = \left| f_{0,lk}\left[\frac{\bar{p}_\epsilon}{p_{0,\epsilon}} - 1\right] + 2^{l/2+1}(y_{\epsilon 0} - \bar{Y}_{\epsilon 0})\right|.$$

Then, as we are on the event $\mathcal{B}_M$, we bound the two terms above by means of Lemmas 1 and 2 from [7] (which are valid for some c_0 small enough) and the bound $p_{0,\epsilon} \lesssim 2^{-|\epsilon|}$ (as f_0 is upper bounded), which give uniformly on $\mathcal{T} \in \mathbb{T}_n$ and $(l,k) \in \mathcal{T}_{\text{int}}$,

$$\begin{aligned}\left|\bar{f}_{\mathcal{T},lk} - f_{0,lk}\right| &\lesssim |f_{0,lk}|\left[a\frac{2^l}{n} + \sqrt{\frac{L_n 2^l}{n}}\right] + \left[|f_{0,lk}|\frac{a2^l}{n} + \sqrt{\frac{L_n}{n}}\right] \\ &\lesssim |f_{0,lk}|\left[a\frac{2^l}{n} + \sqrt{\frac{L_n 2^l}{n}}\right] + \sqrt{\frac{\log n}{n}} \qquad \text{as } L_n \lesssim \log n.\end{aligned} \tag{29}$$

Since f_0 is α-Hölder, $|f_{0,lk}| \lesssim 2^{-l(1/2+\alpha)}$, and the last quantity in the above inequality is smaller (up to a constant) than $\sqrt{n^{-1}\log n}$ as $l \leq L_n$. It then remains to bound the term

$$\int \max_{(l,k)\in\mathcal{T}_{\text{int}}} \left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| d\Pi[f \mid \mathcal{T}, X].$$

To do so, let's first define the event

$$\mathcal{A} = \bigcap_{\epsilon:|\epsilon|<L_n} \left\{ |\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}| \leq M' \sqrt{\frac{L_n}{nP_0(I_{\epsilon 0})}} \right\}$$

for $M' > 0$. By Lemma 9, it follows that, for d a small constant,

$$\Pi\left[\mathcal{A}^c \mid \mathcal{T}, X\right] \lesssim \sum_{l \leq L_n} 2^l \exp(-CM'^2 \log n) \lesssim 2^{L_n} \exp(-CM'^2 \log n), \tag{30}$$

which is smaller than $(n/\log n)^{1/(1+2\alpha)}\, n^{-CM'^2}$. Then,

$$\left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| = \left| 2^{l/2+1}\bar{p}_\epsilon \left(\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}\right) + \left[\frac{p_\epsilon}{\bar{p}_\epsilon} - 1\right] \left(\bar{f}_{\mathcal{T},lk} + 2^{l/2+1}\bar{p}_\epsilon(\bar{Y}_{\epsilon 0} - Y_{\epsilon 0})\right) \right|.$$

Applying Lemmas 2 and 3 from [7] (valid once again for some c_0 small enough), on the events $\mathcal{B}_M$ and $\mathcal{A}$, uniformly on ϵ such that $|\epsilon| = l$ for some $l \leq L_n$,

$$\left|\frac{p_\epsilon}{\bar{p}_\epsilon} - 1\right| \lesssim \sum_{i=1}^{l} \sqrt{\frac{L_n}{nP_0(I_{\epsilon[i]})}} \lesssim \sqrt{\frac{L_n 2^l}{n}}.$$

Therefore, we directly have that on the events $\mathcal{B}_M$ and $\mathcal{A}$,

$$\begin{aligned} \left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| &\lesssim \left|\bar{f}_{\mathcal{T},lk}\right| \sqrt{\frac{L_n 2^l}{n}} + 2^{l/2}\bar{p}_\epsilon \left[\sqrt{\frac{L_n}{nP_0(I_{\epsilon 0})}} + \frac{L_n}{n}\sqrt{\frac{2^l}{P_0(I_{\epsilon 0})}} \right] \\ &\lesssim \left|\bar{f}_{\mathcal{T},lk}\right| \sqrt{\frac{L_n 2^l}{n}} + \sqrt{\frac{L_n}{n}}, \end{aligned} \tag{31}$$

where we used that on $\mathcal{B}_M$, $\bar{p}_\epsilon \lesssim 2^{-|\epsilon|}$ for n large enough as f_0 is upper bounded, and $P_0(I_{\epsilon 0}) \gtrsim 2^{-|\epsilon|}$. Finally, with $\left|\bar{f}_{\mathcal{T},lk}\right| \leq \left|\bar{f}_{\mathcal{T},lk} - f_{0,lk}\right| + |f_{0,lk}|$ and using the same computation as for (29), we have $\left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| \lesssim \sqrt{\frac{\log n}{n}}$. This gives

$$\begin{aligned} &\int \max_{(l,k)\in\mathcal{T}_{\text{int}}} |f_{lk} - f_{0,lk}|\, d\Pi\left[f \mid \mathcal{T}, X\right] \lesssim \sqrt{\frac{\log n}{n}} + \int_{\mathcal{A}^c} \max_{(l,k)\in\mathcal{T}_{\text{int}}} \left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| d\Pi[f \mid \mathcal{T}, X] \\ &\lesssim \sqrt{\frac{\log n}{n}} + 2^{L_n/2}\Pi[\mathcal{A}^c \mid \mathcal{T}, X] \lesssim \sqrt{\frac{\log n}{n}} + \left(\frac{n}{\log n}\right)^{\frac{\alpha/2}{2\alpha+1}} \left(\frac{n}{\log n}\right)^{\frac{1}{1+2\alpha}} n^{-dM'^2} \\ &\lesssim \sqrt{\frac{\log n}{n}} \qquad \text{for } M' \text{ large enough,} \end{aligned} \tag{32}$$

where the second inequality comes from the fact that, for a density f, $|\langle f, \psi_{lk}\rangle| \leq 2^{l/2}$. This concludes the proof as this bound holds uniformly on $\mathcal{T} \in T_n$. □

7.2. Proofs for confidence bands

Proof of Proposition 2. On the event $\mathcal{E}$ from Lemma 4, the bound on the median tree depth implies that for any $h, g \in C_n$,

$$\begin{aligned}\|h-g\|_\infty &\le \|h - f_{\mathcal{T}^*}\|_\infty + \|g - f_{\mathcal{T}^*}\|_\infty\\ &\le 2\sigma_n\\ &\le 2A^{1/2} v_n \sqrt{\frac{\log n}{n}} 2^{L_n/2} \lesssim v_n \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}.\end{aligned}$$

Also, Lemma 7 ensures that

$$\|\hat{f}_{\mathcal{T}^*} - f_0\|_\infty = O_{P_0}\left(\left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right).$$

Then, according to the proof of Proposition 3 in [22], for any $f_0 \in \mathcal{S}(\alpha, K, \eta)$ and l_1 large enough

$$\sup_{(l,k):\ l\ge l_1} |\langle f_0, \psi_{lk}\rangle| \ge C 2^{-l_1(\alpha+1/2)}.$$

For $\Delta_n > 0$ and $\zeta > 0$ such that

$$\zeta\left(\frac{n}{\log^2 n}\right)^{1/(2\alpha+1)} \le 2^{\Delta_n} \le 2\zeta\left(\frac{n}{\log^2 n}\right)^{1/(2\alpha+1)},$$

this implies that

$$\sup_{(l,k):\ l\ge \Delta_n} |\langle f_0, \psi_{lk}\rangle| \ge C\zeta^{-\alpha-1/2}\frac{\log n}{\sqrt{n}}.$$

Therefore, if ζ is small enough, there exists $l \ge \Delta_n$ and $0 \le k < 2^l$ such that $|\langle f_0, \psi_{lk}\rangle| > A\log n/\sqrt{n}$, and then $(l,k) \in \mathcal{T}^*$ on $\mathcal{E}$ according to Lemma 4. As a consequence,

$$\sigma_n \ge v_n \sqrt{\frac{\log n}{n}} 2^{\Delta_n/2} \ge C' \frac{v_n}{\log^{1/2} n}\left(\frac{\log^2 n}{n}\right)^{\alpha/(2\alpha+1)}, \tag{33}$$

and since $\log^{1/2} n = o(v_n)$, $\|f_0 - f_{\mathcal{T}^*}\|_\infty \le \sigma_n/2$ for n large enough. This allows us to conclude that

$$P_0\left[f_0 \in \mathcal{C}_n\right] = P_0\left[\{f_0 \in \mathcal{C}_n\} \cap \mathcal{E}\right] + o(1) = 1 + o(1).$$

It remains to determine the credibility level of the set $\mathcal{C}_n$. From Theorem 1 and Lemma 7, the posterior contracts towards f_0 and the $\hat{f}_{\mathcal{T}^*}$ converges to f_0 on an asymptotically certain event $\mathcal{E}$, both at a faster rate than σ_n (see (33)). Therefore, an application of the triangular inequality gives

$$\Pi\left[\mathcal{C}_n \,|\, X\right] \ge \Pi\left[\|f - f_0\|_\infty \le \sigma_n/2 \,|\, X\right] \mathbb{1}_{\mathcal{E}} + \Pi\left[\mathcal{C}_n \,|\, X\right] \mathbb{1}_{\mathcal{E}^c} = 1 + o_{P_0}(1). \qquad \square$$

Proof of Proposition 4. The credibility statement follows from the fact that $\mathcal{C}_n$ (respectively the multiscale ball) has credibility 1 (respectively $1-\gamma$) asymptotically. The diameter statement follows from the inclusion $\mathcal{C}_n^{\mathcal{M}} \subset \mathcal{C}_n$. For coverage, one combines Theorem 2 which gives that $\mathcal{C}_n$ has asymptotic coverage 1, with Theorem 5 in [10] which from the nonparametric BvM (Theorem 4) enables to deduce frequentist coverage of $\|\cdot\|_{\mathcal{M}_0(w)}$–balls (hence the multiscale ball in the intersection defining $\mathcal{C}_n^{\mathcal{M}}$ has asymptotic coverage $1-\gamma$). □

Appendix A: The classical Pólya tree and *T*–Pólya trees

Let us partition the sample space $I_\varnothing = [0,1)$ as $I_{1,0} \cup I_{1,1}$, these two subsets being the level-1 elementary regions. These can in turn be partitioned as $I_{1,0} = I_{2,0} \cup I_{2,1}$ and $I_{1,1} = I_{2,2} \cup I_{2,3}$, involving level-2 elementary regions. Continuing this partitioning scheme gives the general level-k elementary region, $k \geq 1$, whose set will be written as $\mathcal{A}^k$. More precisely, we partition $I_{l,k} = I_{l+1,2k} \cup I_{l+1,2k+1}$, $l \geq 0, 0 \leq k \leq 2^l - 1$. From this recursive partitioning scheme, one defines a random recursive partition of $I_\varnothing$ and an associated random density.

The Pólya Tree prior corresponding to the partitioning $\cup_{l=1}^{\infty} \mathcal{A}^l$ is the distribution on probability measure on $[0;1)$, whose samples are defined by the conditional probabilities

$$\epsilon \in \mathcal{E}^*, \;\; P\left(I_{\epsilon 0} | I_\epsilon\right) = V_{\epsilon 0} \sim \mathrm{Beta}(\nu_{\epsilon 0}, \nu_{\epsilon 0}). \tag{34}$$

For an appropriate choice of Beta parameters $\nu_\epsilon, \;\epsilon \in \mathcal{E}^*$, samples from this prior actually extends almost surely to an absolutely continuous measure, so that it can be seen as a prior on densities. The Beta random variables $V_{\epsilon 0}$ then corresponds to the share of the mass on I_ϵ that is allocated to $I_{\epsilon 0}$. This mass allocation scheme is illustrated on Figure 6: the random mass of each interval I_ϵ is the product of Beta variables on the edges of the path from the root to the corresponding node. As a consequence, the random mass on I_ϵ, $\epsilon \in \mathcal{E}^*$, is equal to $\prod_{i=1}^{|\epsilon|} V_{\epsilon[i]}$.

A simpler related prior on densities, the truncated Pólya Tree prior, stops the splitting of the mass at some level $L < \infty$ and has sampled densities which are constant on each set I_ϵ in $\mathcal{A}^L$, with value $\mu\left(I_\epsilon\right)^{-1} \prod_{i=1}^{|\epsilon|} V_{\epsilon[i]}$. If one introduces the tree T as

$$T = \left\{(k,l), \;\; l \leq L, 0 \leq k < 2^l\right\},$$

that is the complete binary tree of depth $d(T) = L$, it corresponds to a T-Pólya tree distribution with $\Pi_{\mathbb{T}} = \delta_T$.

Appendix B: Tree posteriors: the Galton–Watson/Pólya tree case

As shown in Subsection 2.3, the Markov process on trees $GW(p)$ can be seen as a distribution on partitions. We first show that it corresponds to the distribution introduced in [39].

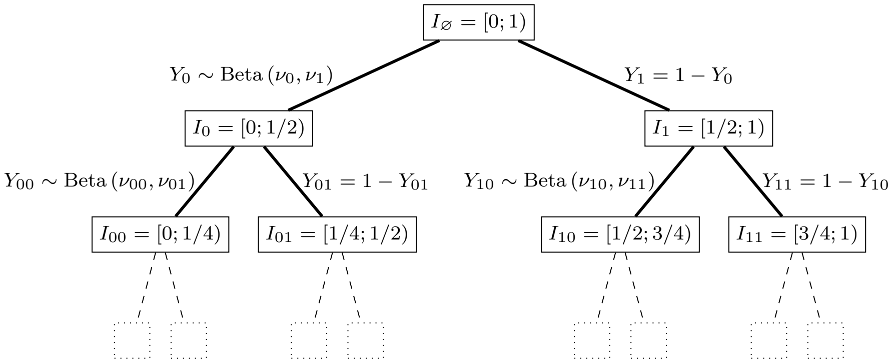


Fig 6: Pólya Tree process on the dyadic recursive partitioning, with splits at midpoints.

In the Optional Pólya Tree (OPT) construction, different recursive partitioning mechanism are allowed: each level-k elementary region $A \in \mathcal{A}^k$ can be split in $M(A)$ different ways, the j-th being written as

$$A = \cup_{i=1}^{K_j(A)} A_k^j, \tag{35}$$

where the A_k^j are level-$(k+1)$ elementary regions (see Appendix A). Then, a random partition of the sample space $[0;1)$ is produced recursively. For $0 \leq \rho([0,1)) \leq 1$, the partition is the sample space itself with probability $\rho([0,1))$. Otherwise, one of the $M([0,1))$ partitions are drawn according to probability vector $\lambda([0;1)) = \big(\lambda_1, \dots, \lambda_{M([0,1))}\big)$. The partitioning then continues: each elementary region A stays intact with probability $\rho(A)$, otherwise it is split (a decision encoded by the variable $S(A) \sim \mathrm{B}(\rho(A))$) and its partition is chosen according to probability vector $\lambda(A)$. Following the discussion in Subsection 2.3, the $GW(p)$ is a particular case, where $M(A) = 1$, $\lambda(A) = 1$ and $K_1(A) = 2$, as the intervals are only ever split at their midpoints,

$$I_{l,k} = I_{l+1,2k} \cup I_{l+1,2k+1}. \tag{36}$$

The level-k elementary regions are the I_ϵ with $|\epsilon| = k$. Also, it corresponds to the choice of

$$\rho(I_{l,k}) = 1 - p_{lk},\ l < L_{\max}, \qquad \rho(I_{L_{\max},k}) = 0.$$

Given a partition $\mathcal{I}$, in OPT, a probability measure Q is defined by the conditional probabilities, for A an elementary region split as in (35),

$$\Big(Q(A_1^j|A), \dots, Q(A_{K_j(A)}^j|A)\Big) = Q(A)\theta(A), \theta(A) \sim \mathrm{Dir}\left(\alpha_1^j(A), \dots, \alpha_{K_j(A)}^j(A)\right),$$

with Dirichlet random variables θ mutually independent and independent from the variables $S(A')$ for $A \not\subset A'$, and $Q([0,1)) = 1$. For $M(A) = 1$ and $K_1(A) = 2$, it is similar to the mass allocation mechanism in (34) when $\alpha_1^1 = \alpha_2^1 = a$. However, whenever the recursive partitioning stops and gives a finite partition, these equations do not completely characterize a measure on Borelians of $[0,1)$, so that the measure Q is defined on Borelians B as

$$Q(B) = \sum_{A \in \mathcal{I}} Q(A) \frac{\mu\left(A \cap B\right)}{\mu(A)}.$$

This corresponds to the absolutely continuous measure with density constant on the elements of $\mathcal{I}$. Therefore, the distribution from Proposition 1 is actually a special case of OPT.

Appendix C: The OPT posterior on trees

In the following, we prove Propositions 2 and 3. We first obtain a general formula for the posterior on trees, which implies an explicit formulation of $\Pi[\cdot \mid X, \mathcal{T}]$, and then focus on the OPT prior. The posterior distribution on trees is given for $T \in \mathbb{T}_n$ by Bayes' formula as

$$\Pi\left[T|X\right] = \frac{\int \Pi\left[X, T|f\right] d\Pi\left[f\right]}{\int \Pi\left[X|f\right] d\Pi\left[f\right]}.$$

Since $\Pi\left[X, T|f\right] = \mathbb{1}_{\mathcal{T}=T} \prod_{i=1}^n f\left(X_i\right)$, the numerator is equal to

$$\sum_{T' \in \mathbb{T}_n} \Pi\left[\mathcal{T} = T'\right] \mathbb{1}_{\mathcal{T}=T} \int \prod_{i=1}^n f\left(X_i\right) d\Pi\left[f|T'\right] = \Pi\left[\mathcal{T} = T\right] \int \prod_{i=1}^n f\left(X_i\right) d\Pi\left[f|T\right].$$

Writing $N_T(X) \coloneqq \int \prod_{i=1}^n f\left(X_i\right) d\Pi\left[f|T\right]$ the marginal likelihood, the denominator can be expressed as

$$\sum_{T' \in \mathbb{T}_n} \Pi\left[\mathcal{T} = T'\right] \int \Pi\left[X, \mathcal{T} = T'|f\right] d\Pi\left[f\right] = \sum_{T' \in \mathbb{T}_n} \Pi\left[\mathcal{T} = T'\right] N_{T'}(X).$$

Let's compute $N_T(X)$. By definition, for any $i = 1, \dots, n$,

$$f\left(X_i\right) = \prod_{(l,k) \in T_{\text{ext}}} \left(\prod_{j=1}^l 2Y_{\epsilon(k,l)^{[j]}} \right)^{\mathbb{1}_{X_i \in I_{lk}}},$$

and

$$\begin{aligned} \prod_{i=1}^n f\left(X_i\right) &= \prod_{(l,k) \in T_{\text{ext}}} \left(\prod_{j=1}^l 2Y_{\epsilon(k,l)^{[j]}} \right)^{N_X(I_{lk})} \\ &= \prod_{(l,k) \in T \setminus \{(0,0)\}} \left(2Y_{\epsilon(k,l)}\right)^{N_X(I_{lk})} \end{aligned}$$

$$= \prod_{(l,k)\in T_{\mathrm{int}}} \left(2Y_{\epsilon(k,l)0}\right)^{N_X\left(I_{\epsilon(k,l)0}\right)} \left(2(1-Y_{\epsilon(k,l)0})\right)^{N_X\left(I_{\epsilon(k,l)1}\right)}.$$

On the one hand, we obtain that

$$\begin{aligned}
\Pi[f \mid X, \mathcal{T}] &= N_T(X)^{-1}\Pi[f, X \mid \mathcal{T}] = N_T(X)^{-1}\Pi[X \mid f, \mathcal{T}]\Pi[f \mid \mathcal{T}] \\
&= C(X,T)\prod_{i=1}^{n} f\left(X_i\right) \prod_{(l,k)\in T_{\mathrm{ext}}} \prod_{j=1}^{l} Y^a_{\epsilon(k,l)^{[j]}} \left(1-Y_{\epsilon(k,l)^{[j]}}\right)^a \\
&= C(X,T) \prod_{(l,k)\in T_{\mathrm{int}}} Y_{\epsilon(k,l)0}^{a+N_X\left(I_{\epsilon(k,l)0}\right)} \left((1-Y_{\epsilon(k,l)0})\right)^{a+N_X\left(I_{\epsilon(k,l)1}\right)},
\end{aligned}$$

for $C(X,T)$ a constant depending on X and T only, which proves the claim of Proposition 2. On the other hand, for any variable $Y \sim \mathrm{Beta}(a,a)$, one obtains

$$E\left[Y^N(1-Y)^M\right] = \int_0^1 y^N(1-y)^M \frac{y^a(1-y)^a}{B(a.a)}dy = \frac{B(a+N, a+M)}{B(a,a)}.$$

Therefore,

$$N_T(X) = \prod_{(l,k)\in T_{\mathrm{int}}} 2^{N_X\left(I_{\epsilon(k,l)}\right)} \frac{B\left(a+N_X\left(I_{\epsilon(k,l)0}\right), a+N_X\left(I_{\epsilon(k,l)1}\right)\right)}{B(a,a)}.$$

Let's now focus on the special case of the GW(p) tree prior, as in Proposition 3. For any possible pair (l,k), take $T \in \mathbb{T}_n$ such that $(l,k) \in T_{\mathrm{ext}}$ and let

$$T^+ = T \cup \{(l+1, 2k), (l+1, 2k+1)\}.$$

Then,

$$\Pi[T^+] = \Pi[T]\frac{p_{lk}}{1-p_{lk}}(1-p_{l+1,2k})(1-p_{l+1,2k+1}), \tag{37}$$

and

$$\begin{aligned}
\frac{\Pi[T^+|X]}{\Pi[T|X]} &= \frac{\Pi\left[\mathcal{T}=T^+\right]L_{T^+}(X)}{\Pi\left[\mathcal{T}=T\right]L_T(X)} \\
&= \frac{p_{lk}}{1-p_{lk}}(1-p_{l+1,2k})(1-p_{l+1,2k+1}) \\
&\quad 2^{N_X\left(I_{\epsilon(k,l)}\right)} \frac{B\left(a+N_X\left(I_{\epsilon(k,l)0}\right), a+N_X\left(I_{\epsilon(k,l)1}\right)\right)}{B(a,a)}.
\end{aligned} \tag{38}$$

This last quantity is independent of T and T^+ and depends only on (l,k). Therefore, if we can find $p^X_{lk}, p^X_{l+1,2k}, p^X_{l+1,2k+1}$ such that the last quantity in (38) is equal to

$$\frac{p^X_{lk}}{1-p^X_{lk}}(1-p^X_{l+1,2k})(1-p^X_{l+1,2k+1}),$$

for any appropriate (l,k), we obtain a formula similar to (37) and the posterior on trees is a $GW(p^X)$ process. This defines a set of equations that has a solution, as for any $0 \leq k < 2^{L_{\max}}$, we necessarily have $p_{L_{\max}k} = 0$ and the equations can be solved to obtain p^X, starting from $l = L_{\max}$ and solving the successive equations in a "bottom–up" way up to the level $l = 0$.

Appendix D: Median tree properties

Lemma 4. *Under the same prior and assumptions as in Theorem 1, there exists an event $\mathcal{E}$, such that $P_0\left[\mathcal{E}\right] = 1 + o(1)$, on which the following is true: for some constants $A > 0, B > 0$,*

- $2^{d(\mathcal{T}^*)} \leq A2^{L_n} \asymp (n/\log n)^{1/(2\alpha+1)}$*, L_n as in (22),*
- *For any (l,k) such that $|f_{0,lk}| \geq Bn^{-1/2}\log n$, $(l,k) \in \mathcal{T}^*_{int}$.*

Proof. On the event $\mathcal{B}_M$ from Lemma 8, Lemma 2 shows that the set $\mathbb{T}^{(2)}$ of trees satisfying the second condition in the lemma, for B large enough, is such that $\Pi\left[\mathbb{T}^{(2)}|\, X\right] \to 1$. Therefore the event

$$\tilde{\mathcal{E}} = \left\{\Pi\left[\mathbb{T}^{(2)}|\, X\right] \geq 3/4\right\} \supset \mathcal{B}_M$$

is asymptotically certain.

For any node (l,k) such that $|f_{0,lk}| \geq Bn^{-1/2}\log n$, since it belongs to the interior nodes of any tree in $\mathbb{T}^{(2)}$ by definition,

$$\Pi\left[(l,k) \in \mathcal{T}_{\text{int}}|\, X\right] = \sum_{\mathcal{T}\in\mathbb{T}_n:\ (l,k)\in\mathcal{T}_{\text{int}}} \Pi\left[\mathcal{T}|\, X\right] \geq \Pi\left[\mathbb{T}^{(2)}|\, X\right].$$

Then, on $\tilde{\mathcal{E}}$, $(l,k) \in \mathcal{T}^*$ by definition and $\mathcal{T}^*$ satisfies the second condition of the lemma.

Let's now turn to the set $\mathbb{T}^{(1)}$ of trees satisfying the first condition in the lemma. Using the same arguments as for (26), there exists $C > 0$ such that for any l such that $2^l \gtrsim 2^{L_n}$ and $\Gamma > 0$ large enough,

$$\Pi\left[d\big(\mathcal{T}\big) > l \mid X\right] \leq n^C\left(2/\Gamma\right)^l,$$

which holds on the event $\mathcal{B}_M$. Then, since

$$\Pi\left[(l,k) \in \mathcal{T}_{\text{int}}|\, X\right] \leq \Pi\left[d\left(\mathcal{T}\right) > l|\, X\right],$$

Markov's inequality implies

$$\begin{aligned}
&P_0\left[\left\{\mathcal{T}^* \notin \mathbb{T}^{(1)}\right\} \cap \mathcal{B}_M\right] \\
&= P_0\left[\left\{\exists(l,k) :\ 2^l > A2^{L_n},\ (l,k) \in \mathcal{T}^*\right\} \cap \mathcal{B}_M\right] \\
&\leq \sum_{l:\ 2^l > A2^{L_n}}^{L_{\max}} \sum_{0\leq k<2^l-1} P_0\left[\left\{\Pi[(l-1, \lfloor k/2\rfloor) \in \mathcal{T}_{\text{int}} \mid X] > 1/2\right\} \cap \mathcal{B}_M\right]
\end{aligned}$$

$$\leq \sum_{l:\ 2^l > A2^{L_n}}^{L_{\max}} 2 \sum_{0 \leq k < 2^l - 1} E_0 \left[\Pi[(l-1, \lfloor k/2 \rfloor) \in \mathcal{T}_{\text{int}} \mid X] \mathbb{1}_{\mathcal{B}_M} \right]$$
$$= o(1) \text{ for } \Gamma \text{ large enough.}$$

One concludes by noting that $\mathcal{B}_M$ is asymptotically certain according to Lemma 8, and $\mathcal{E} = \left\{ \mathcal{T}^* \in \mathbb{T}^{(1)} \right\} \cap \mathcal{B}_M$ satisfies the conditions of the lemma. □

Lemma 5. *Let $0 < \alpha \leq 1$, $K > 0$, $\mu > 0$ and $\eta > 0$. Let Π be the same prior as in Theorem 2, then for $f_0 \in \mathcal{S}(\alpha, K, \eta) \cap \mathcal{F}(\alpha, K, \mu)$,*

$$\left(n/\log^2 n\right)^{1/(2\alpha+1)} \lesssim 2^{d(\mathcal{T}^*)} \lesssim \left(n/\log n\right)^{1/(2\alpha+1)},$$

on an event of probability converging to 1.

Proof. Using the same argument as above (33), we obtain the lower bound. Lemma 4 gives the upper bound. □

Lemma 6. *Let f_0 and ℓ_0 be as in Theorem 4, Π as in Proposition 4 and $\hat{f}_{\mathcal{T}^*}$ as defined in (14). The median tree estimator then satisfies*

$$\max_{l > \ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*, lk} - f_{0,lk}| = O_{P_0}\left(\frac{\log n}{\sqrt{n}} \right).$$

Proof. Let $Q = \max_{l > \ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*, lk} - f_{0,lk}|$. On the event $\mathcal{E}$ from Proposition 4, one has for B as in the proposition,

$$Q \leq \left(B \frac{\log n}{n^{1/2}} \right) \vee \max_{(l,k) \in \mathcal{T}^*_{\text{int}}, l > \ell_0(n)} |\hat{f}_{\mathcal{T}^*, lk} - f_{0,lk}|.$$

Indeed, for $(l,k) \notin \mathcal{T}^*_{\text{int}}$, we necessarily have $\hat{f}_{\mathcal{T}^*, lk} = 0$ and $|f_{0,lk}| < Bn^{-1/2} \log n$ on $\mathcal{E}$. From (42), it also follows that for A as in the proposition and L_n defined in (22)

$$\max_{(l,k) \in \mathcal{T}^*_{\text{int}},\ l > \ell_0(n)} |\hat{f}_{\mathcal{T}^*, lk} - f_{0,lk}| \leq \max_{(l,k),\ 2^{\ell_0(n)} < 2^l < A2^{L_n}} |P_n \psi_{lk} - P_0 \psi_{lk}| =: Q_n.$$

We have that

$$|P_n \psi_{lk} - P_0 \psi_{lk}|$$
$$\leq 2^{l/2} n^{-1} \left(|N(I_{l+1,2k}) - nP_0(I_{l+1,2k})| + |N(I_{l+1,2k+1}) - nP_0(I_{l+1,2k+1})| \right).$$

Therefore, on the event $\mathcal{B}_M$ from Lemma 8, for some constant C depending on M, A, c_0 and α only, and any l as in the above supremum,

$$|P_n \psi_{lk} - P_0 \psi_{lk}| \leq C \sqrt{\frac{\log n}{n}}. \tag{39}$$

It follows that $Q \lesssim n^{-1/2} \log n$ on the event $\mathcal{E} \cap \mathcal{B}_M$ that is such that $P_0\left(\mathcal{E} \cap \mathcal{B}_M\right) = 1 + o(1)$. □

Lemma 7. *Let $\mathcal{T}^*$ as in (13) and $\hat{f}_{\mathcal{T}^*}$ as in (14). Then, for $f_0 \in \mathcal{F}(\alpha, K, \mu)$,*

$$\|\hat{f}_{\mathcal{T}^*} - f_0\|_\infty = O_{P_0}\left(\left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right).$$

Proof. Let $\mathcal{E}$ as in Lemma 4 and $\mathcal{B}_M$ as in Lemma 8. On $\mathcal{E} \cap \mathcal{B}_M$, for M large enough,

$$\begin{aligned}&\|f_0 - f_{\mathcal{T}^*}\|_\infty \\ &\quad \le \sum_{l:\ 2^l < A2^{L_n}} 2^{l/2} \max\left[\max_{0\le k<2^l,\ (l,k)\in\mathcal{T}^*_{\text{int}}} |\langle f_0 - f_{\mathcal{T}^*}, \psi_{lk}\rangle|, \max_{0\le k<2^l,\ (l,k)\notin\mathcal{T}^*_{\text{int}}} |\langle f_0, \psi_{lk}\rangle|\right] \\ &\qquad + \sum_{l:\ 2^l \ge A2^{L_n}} 2^{l/2} \max_{0\le k<2^l} |\langle f_0, \psi_{lk}\rangle|,\end{aligned}$$

using the usual inequality for densities h, g, $\|h-g\|_\infty \le \sum_{l\ge 0} 2^{l/2} \max_{0\le k<2^l} |\langle h-g, \psi_{lk}\rangle|$. Since $f_0 \in \Sigma(\alpha, K)$, the second term is smaller than $2^{-\alpha L_n} = O\left((n/\log n)^{-\alpha/(2\alpha+1)}\right)$ (up to a constant depending only on α, K and the constant A from Lemma 4). Then, the first term can itself be upper bounded by the sum of

$$\begin{aligned}&\sum_{l:\ 2^l < A2^{L_n}} 2^{l/2} \max_{0\le k<2^l,\ (l,k)\in\mathcal{T}^*_{\text{int}}} |\langle f_0 - f_{\mathcal{T}^*}, \psi_{lk}\rangle| \\ &\qquad \lesssim 2^{L_n/2}\sqrt{\frac{\log n}{n}} = o\left(\left(\frac{\log^2 n}{n}\right)^{\alpha/(2\alpha+1)}\right),\end{aligned}$$

where we used that the argument of 39 can be extended to $l \le \ell_0(n)$ on $\mathcal{E} \cap \mathcal{B}_M$, and the term

$$\sum_{l:\ 2^l < A2^{L_n}} 2^{l/2} \max_{0\le k<2^l,\ (l,k)\notin\mathcal{T}^*_{\text{int}}} |\langle f_0, \psi_{lk}\rangle|.$$

It remains to upper bound this last quantity. Let's introduce

$$L^* = \max\left\{l : \max_{0\le k<2^l} |\langle f_0, \psi_{lk}\rangle| \ge Bn^{-1/2}\log n\right\}$$

which is such that $2^{L^*} \asymp \left(\frac{n^{1/2}}{\log n}\right)^{1/(\alpha+1/2)}$ since $\max_{0\le k<2^l} |\langle f_0, \psi_{lk}\rangle| \lesssim 2^{-l(1/2+\alpha)}$. Then, on the event $\mathcal{E}$, the term in the above display is bounded by

$$\begin{aligned}\sum_{l:\ 2^l < A2^{L_n}} 2^{l/2}\left(B\frac{\log n}{\sqrt{n}}\right) \wedge \max_{0\le k<2^l} |\langle f_0, \psi_{lk}\rangle| &\le \sum_{l:\ l\le L^*} 2^{l/2}\left(B\frac{\log n}{\sqrt{n}}\right) \\ &\quad + \sum_{l:\ 2^{L^*} < 2^l < A2^{L_n}} 2^{l/2} \max_{0\le k<2^l} |\langle f_0, \psi_{lk}\rangle| \\ &\lesssim \sqrt{2^{L^*}\frac{\log^2 n}{n}} + 2^{-\alpha L^*} \lesssim \left(\frac{\log^2 n}{n}\right)^{\alpha/(2\alpha+1)}.\end{aligned}$$

Combining the previous bounds leads to, on $\mathcal{E} \cap \mathcal{B}_M$,

$$\|f_0 - f_{\mathcal{T}^*}\|_\infty \leq C\left(\log^2 n/n\right)^{\alpha/(2\alpha+1)}. \qquad \square$$

Appendix E: Nonparametric BvM theorem

E.1. Space $\mathcal{M}_0$ and limiting Gaussian process $\mathcal{N}$

Recall the definition of the space $\mathcal{M}_0$ from (18), using an 'admissible' sequence $w = (w_l)_{l\geq 0}$ such that $w_l/\sqrt{l} \to \infty$ as $l \to \infty$,

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x = (x_{lk})_{l,k} \ ; \ \lim_{l\to\infty} \max_{0\leq k<2^l} \frac{|x_{lk}|}{w_l} = 0 \right\}.$$

Equipped with the norm $\|x\|_{\mathcal{M}_0} = \sup_{l\geq 0} \max_{0\leq k<2^l} |x_{lk}|/w_l$, this is a separable Banach space. In a slight abuse of notation, we write $f \in \mathcal{M}_0$ if the sequence of its Haar wavelet coefficients belongs to that space $(\langle f, \psi_{lk}\rangle)_{l,k} \in \mathcal{M}_0$ and for a process $(Z(f),\ f \in L^2)$, we write $Z \in \mathcal{M}_0$ if the sequence $(Z(\psi_{lk}))_{l,k}$ belongs to $\mathcal{M}_0(w)$ almost surely.

White bridge process. For P a probability distribution on $[0,1]$, following [10] one defines the P-white bridge process, denoted by $\mathbb{G}_P$, as the centered Gaussian process indexed by the Hilbert space $L^2(P) = \{f : [0,1] \to \mathbb{R}; \int_0^1 f^2 dP < \infty\}$ with covariance

$$E[\mathbb{G}_P(f)\mathbb{G}_P(g)] = \int_0^1 (f - \int_0^1 f dP)(g - \int_0^1 g dP) dP. \tag{40}$$

We denote by $\mathcal{N}$ the law induced by $\mathbb{G}_{P_0}$ (with $P_0 = P_{f_0}$) on $\mathcal{M}_0(w)$. The sequence $(\mathbb{G}_P(\psi_{lk}))_{l,k}$ indeed defines a tight Borel Gaussian variable in $\mathcal{M}_0(w)$, by Remark 1 of [10].

Admissible sequences (w_l). The main purpose of the sequence (w_l) is to ensure that $(\mathbb{G}_P(\psi_{lk}))_{l,k}$ belongs to $\mathcal{M}_0$. We refer to [10], Section 2.1 and Remark 1, for more background on the choice of (w_l) in the present multiscale setting, and to [9], Section 1.2, for a similar discussion in an Hilbert space setting where the targeted loss is the L^2–norm.

To establish a nonparametric Bernstein–von Mises (BvM) result, following [10] one first finds a space $\mathcal{M}_0$ large enough to have convergence at rate $\sqrt{n}$ of the posterior density to a Gaussian process. One can then derive results for some other spaces $\mathcal{F}$ using continuous mapping for continuous functionals $\psi : \mathcal{M}_0 \to \mathcal{F}$.

Recentering the distribution. To establish the BvM result, one also has to find a suitable way to center the posterior distribution. A possible centering is the median tree estimator $\hat{f}_{\mathcal{T}^*}$ as in (14). Other centerings are possible, typically appropriately 'smoothed' versions of the empirical measure P_n associated to the sample $X_1, \ldots, X_n$

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}. \tag{41}$$

Let us now also note that another way to write the median tree estimator (14) is

$$f_{\mathcal{T}^*} = 1 + \sum_{(l,k)\in\mathcal{T}^*_{\text{int}}} (P_n\psi_{lk}) \cdot \psi_{lk}, \tag{42}$$

where $P_n\psi_{lk} = n^{-1}\sum_{i=1}^n \psi_{lk}(X_i)$ are the empirical wavelet coefficients, and only terms corresponding to interior nodes (l,k) in the median tree $\mathcal{T}^*$ are active in the sum from the last display. From this we see that the median tree estimator (14) can also be interpreted as a smoothed (or 'truncated') version of the empirical measure P_n in (41), with truncation occuring along the median tree $\mathcal{T}^*$. Note also that if the prior Π has flat initialisation up to level $l_0(n)$, then all nodes (l,k) with $l \le l_0(n)$ are present in the above sum over $(l,k) \in \mathcal{T}^*_{\text{int}}$.

E.2. Nonparametric BvM: Statement

For the following result, we work with OPTs with flat initialisation as defined in Section 4.3. This is discussed below the next statement.

We have the following Bernstein-von Mises phenomenon for f_0 in Hölder balls. For C_n a function to be specified, we denote by τ_{C_n} the map $\tau_{C_n} : f \to \sqrt{n}(f - C_n)$.

Theorem 4. *Let $\mathcal{N}$ denote the distribution induced on $\mathcal{M}_0(w)$ by the P_0–white bridge $\mathbb{G}_{P_0}$ as defined in (40) and let $C_n = \hat{f}_{\mathcal{T}^*}$ the median tree estimator as in (14). Let Π be an OPT prior with flat initialisation with $l_0(n)$ that verifies $\sqrt{\log n} \le l_0(n) \le \log n/\log\log n$, and other than that for $l > l_0(n)$ with same parameters as the prior in Theorem 1. Then for every $\alpha \in (0,1]$, for $\mu > 0$, $K \ge 0$ and $\eta > 0$,*

$$\sup_{f_0\in\mathcal{F}(\alpha,K,\mu)} E_{f_0}\left[\beta_{\mathcal{M}_0(w)}(\Pi(\cdot|X)\circ\tau_{C_n}^{-1}, \mathcal{N})\right] \to 0,$$

as $n \to \infty$, for the admissible sequence $w_l = l^{2+\delta}$ for some $\delta > 0$.

Remark 3. Recalling that the typical nonparametric cut–off sequence $\mathcal{L}$ verifies $2^{\mathcal{L}} \asymp n^{1/(1+2\alpha)}$, assuming $\ell_0(n) = o(\log n)$ amounts to say that $\ell_0(n)$ does not 'interfere' with the nonparametric cut-off $\mathcal{L}$. Similar choices are made in [36], Corollary 3.6. Other choices of sequence $\ell_0(n)$ would also be possible, up to adjusting the sequence (w_l) – one can check that it suffices to have an increasing sequence (w_l) such that $w_{l_0(n)}/\log n \to \infty$ (see, e.g. Theorem S–3 in the Supplement of [11]) –; we do not consider these refinements here.

Theorem 4 states that the posterior limiting distribution is Gaussian after rescaling; note that, similar to the first such result recently obtained in [36], one slightly modifies the OPT prior to fit the first levels by assuming a flat initialisation. This is in fact necessary for the result to hold, as otherwise the posterior would not be tight at rate $1/\sqrt{n}$ in the space $\mathcal{M}_0(w)$, as was noted in the white noise model in [36], Proposition 3.7. Let us also briefly comment on the recentering C_n: as follows from the proof of Theorem 4, one can replace

$C_n = \hat{f}_{\mathcal{T}^*}$ by another estimator that fits all first wavelet coefficients up to $\ell_0(n)$ and such that $\|C_n - f_0\|_{\mathcal{M}_0(\bar{w})} = O_{P_0}(1/\sqrt{n})$, for $\bar{w}$ as in that proof, see also Remark 4 for more on this.

E.3. Nonparametric BvM: Implications

Using the methods of [10], this result leads to several applications. A first direct implication (this follows from Theorem 5 in [10]) is the derivation of a confidence set in $\mathcal{M}_0(w)$. Setting

$$\mathcal{D}_n = \left\{ f = (f_{lk}) : \; \|f - C_n\|_{\mathcal{M}_0(w)} \leq \frac{R_n}{\sqrt{n}} \right\}, \tag{43}$$

where R_n is chosen in such a way that $\Pi[\mathcal{D}_n \,|\, X] = 1 - \gamma$, for some $\gamma > 0$ (or taking the generalised quantile for the posterior radius if the equation has no solution) leads to a set $\mathcal{D}_n$ with the following properties: it is a credible set by definition which is also asymptotically a confidence set in $\mathcal{M}_0(w)$ and the rescaled radius R_n is bounded in probability. Other applications are BvM theorems for functionals, as given a continuous map $\psi : \mathcal{M}_0(w) \to \mathcal{E}$ for some metric space $\mathcal{E}$, convergence results in $\mathcal{M}_0(w)$ can be translated into convergence in $\mathcal{E}$ via the continuous mapping theorem, see [10]. This is also at the basis of the proof of the Donsker Theorem 3.

Appendix F: Proof of limiting shape results

In this section we prove the nonparametric BvM Theorem 4 and, as a fairly direct consequence given the results of [10], the Bayesian Donsker Theorem 3.

Proof of Theorem 4. The proof is similar to the corresponding proofs for Pólya trees or spike–and–slab Pólya trees, so we highlight only the few differences. The proof consists in two steps. First, proving convergence of finite–dimensional distributions and second, showing tightness of the rescaled posterior in a slightly smaller space.

Regarding convergence of finite–dimensional distributions, it suffices to note that for a fixed depth $L > 0$, the prior on wavelet coefficients of levels $l \leq L$ (for large enough n so that $\ell_0(n) > L$) coincides with the prior induced by a standard Pólya tree, for which the convergence of finite–dimensional distributions is shown in [7].

Regarding tightness, let $\bar{w} = (\bar{w}_l)$ be the sequence $\bar{w}_l = w_l / l^{\delta/2} = l^{2+\delta/2}$. This sequence is increasing in l and verifies $\bar{w}_l \gtrsim \sqrt{l}$, $\bar{w}_l = o(w_l)$ as $l \to \infty$, and $\bar{w}_{\ell_0(n)} \geq \log n$, using the assumption on $\ell_0(n)$. Now by the same argument as in the proof of Theorem 3 in [8], to establish the nonparametric BvM it suffices to prove that the distribution $\mathcal{L}(\sqrt{n}(f - C_n) \,|\, X)$ is tight in $\mathcal{M}_0(\bar{w})$, which is true if both laws $\mathcal{L}(\sqrt{n}(f - f_0) \,|\, X)$ and $\mathcal{L}(\sqrt{n}(f_0 - C_n))$ are tight.

Focusing first on the tightness of $\mathcal{L}(\sqrt{n}(f-f_0)\,|\,X)$, we wish to show that for any $\eta \in (0,1)$, one can find $M = M(\eta)$ large enough such that

$$E_{f_0}\Pi[\|f-f_0\|_{\mathcal{M}_0(\bar{w})} > M/\sqrt{n}\,|\,X] \leq \eta. \tag{44}$$

We split, for $g = f - f_0$,

$$\|g\|_{\mathcal{M}_0(\bar{w})} \leq \max_{l\leq \ell_0(n),k} |g_{lk}|/\bar{w}_l + \max_{l>\ell_0(n),k} |g_{lk}|/\bar{w}_l =: (I) + (II).$$

For the term (I), as noted above, since the prior has a flat initialisation up to level $\ell_0(n)$, the induced prior and posterior on the first layers $l \leq \ell_0(n)$ of wavelet coefficients coincide with the prior/posterior of a standard Pólya tree, for which the corresponding tightness is proved in [7] (proof of Theorem 3). For the term (II), it follows from the proof of Theorem 1 (noting that the proof goes through with a prior with flat initialisation) that for T_n as in that proof and given $l > \ell_0(n)$, for any $\mathcal{T} \in T_n$ and on the event $\mathcal{B}_M$,

$$\int \max_{k:\,(l,k)\in\mathcal{T}_{int}} |f_{lk} - f_{0,lk}| d\Pi(f\,|\,\mathcal{T}, X) \leq C\sqrt{\frac{\log n}{n}}$$

and

$$\max_{k:\,(l,k)\notin\mathcal{T}_{int}} |f_{0,lk}| \leq C\frac{\log n}{\sqrt{n}}.$$

Since $\bar{w}_{\ell_0(n)} \geq \log n$ as verified above, one deduces that for any $\mathcal{T} \in T_n$ and on $\mathcal{B}_M$ the term (II) above is $O(1/\sqrt{n})$. Putting pieces together what precedes implies, with $\mathcal{E} = \{f_\mathcal{T},\ \mathcal{T} \in T_n\}$ as in the proof of Theorem 1,

$$\int_\mathcal{E} \|f-f_0\|_{\mathcal{M}_0(\bar{w})} d\Pi(f\,|\,X) = O_{P_0}(1/\sqrt{n}),$$

which in turn implies (44) using $\Pi[\mathcal{E}^c\,|\,X] = o_{P_0}(1)$.

It remains to prove tightness of $\mathcal{L}(\sqrt{n}(f_0 - C_n))$ in $\mathcal{M}_0(\bar{w})$. Again, one splits along indices: for $l \leq \ell_0(n)$, the posterior median tree estimator has same wavelet coefficients as the empirical measure P_n, and the estimate

$$E_{P_0} \max_{l\leq\ell_0(n)} \max_k |\langle P_0 - P_n, \psi_{lk}\rangle|/\bar{w}_l \leq C/\sqrt{n}$$

follows from the proof of Theorem 1 in [10] (see equation (36) there and lines below). For $l > \ell_0(n)$, one invokes the properties of the median tree estimator, namely

$$\max_{l>\ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}| = O_{P_0}\left(\frac{\log n}{\sqrt{n}}\right), \tag{45}$$

as in Lemma 6, noting that the argument in that proof is unchanged for a prior with flat initialisation. This gives, using again $\bar{w}_{\ell_0(n)} \geq \log n$, that

$$\max_{l\leq\ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}| = O_{P_0}(1/\sqrt{n}),$$

which gives the desired tightness property and concludes the proof. □

Remark 4. It follows from the proof of Theorem 4 that there is quite some flexibility in the choice of the centering C_n. For instance, the projection $P_n(L_n)$ of the empirical measure P_n onto the first L_n levels of wavelet coefficients, with L_n the oracle supremum–norm cut–off $(n/\log n)^{1/(2\alpha+1)}$ can be used. This is because for $l \leq \ell_0(n)$ the projection $P_n(L_n)$ has by definition same wavelet coefficients as the empirical measure P_n, while for $l > \ell_0(n)$ equation (45) holds for $\langle P_n(L_n), \psi_{lk}\rangle$ instead of $f_{\mathcal{T}^*,lk}$ (with the even better bound $O_{P_0}(\sqrt{\log n/n})$), as in the proof of Theorem 1 in [10].

Proof of Theorem 3. The results follows by applying Theorem 4 in [10]: since the posterior distribution on f satisfies the nonparametric BvM theorem 4, it suffices to check that the sequence (w_l) satisfies the condition $\sum_l w_l 2^{-l/2} < \infty$, which clearly holds, and to note that the centering $C_n = f_{\mathcal{T}^*}$ belongs to L^2. This shows that the Bayesian Donsker holds with centering $\hat{F}_n^{med} = \int_0^{\cdot} f_{\mathcal{T}^*}$. By using remark 4, the same result also holds with $\hat{F}_n^{med}$ replaced by the primitive, say $\mathbb{Z}_n(\cdot)$, of $P_n(L_n)$. But as noted in the proof of Corollary 1 in [10] (see also Remark 9 in [19]), we have $\|\mathbb{Z}_n - F_n\|_\infty = o_{P_0}(1/\sqrt{n})$, which implies the result with centering at F_n. □

Appendix G: Miscellaneous

We quickly remind that

$$\bar{Y}_\epsilon = E[Y_\epsilon \,|\, X^{(n)}] = \frac{a + N_X(I_{\epsilon 0})}{2a + N_X(I_\epsilon)}$$

and we define L_n as in (22).

Lemma 8. *Let $\alpha > 0$, $K > 0$ and P_0 be a distribution with a bounded density $f_0 \in \Sigma(\alpha, K)$ w.r.t. Lebesgue density. Then, for any*

$$M > \frac{1}{3}\left(\sqrt{\log 2}\sqrt{18\,\|f_0\|_\infty + \log 2} + \log 2\right),$$

the event

$$\mathcal{B}_M := \Big\{\forall l \geq 0, \quad \forall 0 \leq k \leq 2^l - 1,$$
$$M^{-1}|N_X(I_{l,k}) - nP_0(I_{l,k})| \leq \sqrt{\frac{n(l+L_n)}{2^l}} \vee (l + L_n) =: M_{n,l}\Big\}$$

is asymptotically certain under the law P_0 of the observations, i.e.

$$P_0\left(\mathcal{B}_M^c\right) = o(1).$$

Proof. According to Bernstein's inequality, for any $l \geq 0,\ 0 \leq k \leq 2^l - 1$,

$$P_0\left(|N_X(I_{l,k}) - nP_0(I_{l,k})| > MM_{n,l}\right)$$

$$\leq 2\exp\left(-\frac{M^2 M_{n,l}^2/2}{nP_0(I_{l,k})(1-P_0(I_{l,k}))+MM_{n,l}/3}\right).$$

By assumption, $P_0(I_{l,k})(1-P_0(I_{l,k})) \leq \|f_0\|_\infty 2^{-l}$. Then, whenever $M_{n,l} = l+L_n$ (which is equivalent to $l+L_n \geq n2^{-l}$) or $M_{n,l} = \sqrt{\frac{n(l+L_n)}{2^l}}$, we can further upper bound the above quantity as

$$P_0\left(|N_X(I_{l,k}) - nP_0(I_{l,k})| > MM_{n,l}\right) \leq 2\exp\left(-\frac{M^2}{2\|f_0\|_\infty + 2M/3}(l+L_n)\right).$$

Therefore,

$$P_0\left(\mathcal{B}_M^c\right) \leq 2\sum_{l\geq 0} 2^l \exp\left(-\frac{M^2}{2\|f_0\|_\infty + 2M/3}(l+L_n)\right) = O(2^{-L_n})$$
$$= O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+1}}\right),$$

the latter equality being true whenever

$$\frac{M^2}{2\|f_0\|_\infty + 2M/3} > \log 2,$$

i.e. $M > \frac{1}{3}\left(\sqrt{\log 2}\sqrt{18\|f_0\|_\infty + \log 2} + \log 2\right)$. □

Lemma 9. *Suppose $f_0 \in \Sigma(K,\alpha)$, with $0 < \alpha \leq 1$. For $M' > 0$, on the event $\mathcal{B}_M$ from Lemma 8, the set*

$$\mathcal{A} = \bigcap_{\epsilon:|\epsilon|<L_n} \left\{|\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}| \leq M'\sqrt{\frac{L_n}{nP_0(I_{\epsilon 0})}}\right\}$$

is such that

$$\Pi[\mathcal{A}^c \mid X] \lesssim \sum_{l\leq L_n} 2^l e^{-M'^2 \log n/4}$$

Proof. This proof comes from Lemmas 4 and 5 of [8]. For completeness, we give here some details of the proof. We have already explained that

$$Y_{\epsilon 0} \sim \text{Beta}(a + N_X(I_{\epsilon 0}), a + N_X(I_{\epsilon 0})).$$

We also noticed that on the event $\mathcal{B}_M$, $N_X(I_\epsilon) \to \infty$ uniformly for all $|\epsilon| \leq L_n$ for $n \to \infty$. Therefore, for n sufficiently large, $a + N_X(I_{\epsilon 0}) \wedge a + N_X(I_{\epsilon 0}) \geq 8$ for $|\epsilon| < L_n$. Also, under our assumptions, Lemma [2] from [7] allows us to say that, for n large enough, there exist μ, ν such that

$$0 < \mu \leq \frac{a + N_X(I_{\epsilon 0})}{2a + N_X(I_{\epsilon 0}) + N_X(I_{\epsilon 1})} \leq \nu < 1$$

uniformly on all $|\epsilon| < L_n$. In addition, if $i = |\epsilon|$, we have that

$$2a + N_X(I_{\epsilon 0}) + N_X(I_{\epsilon 1}) \geq N_X(I_{\epsilon 0}) \geq nP_0(I_{\epsilon 0}) - M\sqrt{2nL_n 2^{-i}}.$$

Under our assumptions on f_0 and L_n, the last bound is itself lower bounded by $nP_0(I_{\epsilon 0})/2$ for n large enough. As a consequence, an application of Lemma 6 from [7] gives, for $x = M' L_n^{1/2}/2$,

$$\Pi\left[|\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}| > \frac{x}{\sqrt{nP_0(I_{\epsilon 0})}} \,\middle|\, X \right] \leq De^{-x^2/4}$$

for some constant D. Finally, a union bound helps us to conclude that

$$\Pi[\mathcal{A} \,|\, X] \lesssim \sum_{l \leq L_n} 2^l e^{-M'^2 \log n/4}. \qquad \square$$

Lemma 10 (Theorem 1.5 of [2])**.** *For any $x > 0$,*

$$a\left(\frac{x+1/2}{e}\right)^{x+1/2} \leq \Gamma(x+1) \leq b\left(\frac{x+1/2}{e}\right)^{x+1/2},$$

where Γ is usual Gamma function, and $a = \sqrt{2e}$ and $b = \sqrt{2\pi}$ are the best possible constants.

Acknowledgments

The authors would like to thank Li Ma for insightful comments. This work has been supported by the grant ANR-17-CE40-0001 (BASICS).

References

[1] Arlot, S. and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv e-prints* arXiv:1407.3939.

[2] Batir, N. (2008). Inequalities for the gamma function. *Arch. Math. (Basel)* **91** 554–563. MR2465874

[3] Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test* **25** 197–227. MR3493512

[4] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks. MR0726392

[5] Bull, A. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics* **6** 1490–1516. MR2988456

[6] Castillo, I. (2014). On Bayesian supremum norm contraction rates. *The Annals of Statistics* **42** 2058–2091. MR3262477

[7] Castillo, I. (2017). Pólya tree posterior distributions on densities. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** 2074–2102. MR3729648

[8] Castillo, I. and Mismer, R. (2021). Spike and slab Pólya tree posterior densities: Adaptive inference. *Ann. Inst. Henri Poincaré Probab. Stat.* **57** 1521–1548. MR4291462

[9] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises Theorems in Gaussian white noise. *Ann. Statist.* **41** 1999-2028. MR3127856

[10] Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473

[11] Castillo, I. and Ročková, V. (2021). Uncertainty quantification for Bayesian CART. *Ann. Statist.* **49** 3482–3509. MR4352538

[12] Chipman, H., George, E. I. and McCulloch, R. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics* **4** 266–298. MR2758172

[13] Chipman, H., George, E. I. and McCulloch, R. E. (2000). Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing* **10** 17–24.

[14] Christensen, J. and Ma, L. (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 127–153. MR4060979

[15] Denison, D., Mallick, B. and Smith, A. (1998). A Bayesian CART Algorithm. *Biometrika* **85** 363–377. MR1649118

[16] Dudley, R. M. (2002). *Real analysis and probability. Cambridge Studies in Advanced Mathematics* **74**. Cambridge University Press, Cambridge. Revised reprint of the 1989 original. MR1932358 (2003h:60001)

[17] Ghosal, S., Ghosh, J. and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–5311. MR1790007

[18] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge University Press, Cambridge. MR3587782

[19] Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.* **37** 1605–1646. MR2546757

[20] Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395

[21] Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models. Cambridge Series in Statistical and Probabilistic Mathematics, [40]*. Cambridge University Press, New York. MR3588285

[22] Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872

[23] Hoffmann, M., Rousseau, J. and Schmidt-Hieber, J. (2015). On Adaptive Posterior Concentration Rates. *The Annals of Statistics* **43** 2259–2295. MR3396985

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

IV. MR591862 (82a:62059)

[25] Jiang, H., Mu, J. C., Yang, K., Du, C., Lu, L. and Wong, W. H. (2016). Computational aspects of optional Pólya tree. *J. Comput. Graph. Statist.* **25** 301–320. MR3474049

[26] Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135

[27] Linero, A. and Yang, Y. (2018). Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity. *Journal of the Royal Statistical Association* **80** 1087–1110. MR3874311

[28] Liu, L., Li, D. and Wong, W. H. (2017). Convergence rates of a partition based Bayesian multivariate density estimation method. In *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) **30**. Curran Associates, Inc.

[29] Low, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412

[30] Ma, L. (2017). Adaptive shrinkage in Pólya tree type models. *Bayesian Anal.* **12** 779–805. MR3655876

[31] Ma, L. and Wong, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Amer. Statist. Assoc.* **106** 1553–1565. MR2896856

[32] Mourtada, J., Gaïffas, S. and Scornet, E. (2020). Minimax optimal rates for Mondrian trees and forests. *Ann. Statist.* **48** 2253–2276. MR4134794

[33] Naulet, Z. (2018). Adaptive Bayesian density estimation in sup-norm. arXiv preprint 1805.05816. MR4388939

[34] Nickl, R. and Ray, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.* **48** 1383–1408. MR4124327

[35] Randrianarisoa, T. (2022). Smoothing and adaptation of shifted Pólya tree ensembles. *Bernoulli* **28** 2492–2517. MR4474551

[36] Ray, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics* **45** 2511–2536. MR3737900

[37] Rockova, V. and Rousseau, J. (2021). Ideal Bayesian Spatial Adaptation. *arXiv e-prints* arXiv:2105.12793.

[38] Ročková, V. and van der Pas, S. (2020). Posterior concentration for Bayesian regression trees and forests. *Ann. Statist.* **48** 2108–2131. MR4134788

[39] Wong, W. H. and Ma, L. (2010). Optional Pólya tree and Bayesian inference. *Ann. Statist.* **38** 1433–1459. MR2662348

[40] Yoo, W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics* **44** 1069–1102. MR3485954