The Annals of Statistics
2021, Vol. 49, No. 6, 3482–3509
https://doi.org/10.1214/21-AOS2093

UNCERTAINTY QUANTIFICATION FOR BAYESIAN CART

BY ISMAËL CASTILLO[1] AND VERONIKA ROČKOVÁ[2]

[1]*Laboratoire de Probabilités, Statistique et Modélisation, Institut Universitaire de France, Sorbonne Université, ismael.castillo@upmc.fr*

[2]*Booth School of Business, University of Chicago, veronika.rockova@chicagobooth.edu*

This work affords new insights into Bayesian CART in the context of structured wavelet shrinkage. The main thrust is to develop a formal inferential framework for Bayesian tree-based regression. We reframe Bayesian CART as a *g-type* prior which departs from the typical wavelet product priors by harnessing correlation induced by the tree topology. The practically used Bayesian CART priors are shown to attain adaptive near rate-minimax posterior concentration in the *supremum norm* in regression models. For the fundamental goal of uncertainty quantification, we construct *adaptive* confidence bands for the regression function with uniform coverage under self-similarity. In addition, we show that tree-posteriors enable optimal inference in the form of efficient confidence sets for smooth functionals of the regression function.

1. Introduction. The widespread popularity of Bayesian tree-based regression has raised considerable interest in theoretical understanding of their empirical success. However, theoretical literature on methods such as Bayesian CART and BART is still in its infancy. In particular, statistical *inferential* theory for regression trees and forests (both frequentist and Bayesian) has been severely under-developed.

This work sheds light on Bayesian CART [20, 25] which is a popular learning tool based on ideas of recursive partitioning and which forms an integral constituent of BART [22]. Bayesian Additive Regression Trees (also known as BART) have emerged as one of today's most effective general approaches to predictive modeling under minimal assumptions. Their empirical success has been amply illustrated in the context of nonparametric regression [22], classification [45], variable selection [8, 41, 43], shape constrained inference [21], causal inference [37, 38], to name a few. The BART model deploys an additive aggregate of individual trees using Bayesian CART as its building block. While theory for random forests, the frequentist counterpart, has seen numerous recent developments [6, 44, 52, 57, 58], theory for Bayesian CART and BART has not kept pace with its application. With the first theoretical results (Hellinger convergence rates) emerging very recently [42, 50, 51], many fundamental questions pertaining to, see, for example, convergence in stronger losses such as the supremum norm, as well as *uncertainty quantification* (UQ), have remained to be addressed. This work takes a leap forward in this important direction by developing a formal frequentist statistical framework for uncertainty quantification with confidence bands for Bayesian CART.

We first show that Bayesian CART reaches a (near-)optimal posterior convergence rate under the *supremum-norm* loss, a natural loss for UQ of regression functions. Many methods that are adaptive for the L^2-loss actually fail to be adaptive in an L^∞-sense, as we illustrate below. We are actually not aware of any sharp supremum-norm convergence rate result for related machine learning methods in the literature, including CART, random forests and deep learning. Regarding inference, we provide a construction of an *adaptive* credible band for

Received November 2020; revised May 2021.
MSC2020 subject classifications. 62G20, 62G15.
Key words and phrases. Bayesian CART, posterior concentration, recursive partitioning, regression trees, nonparametric Bernstein–von Mises theorem.

the unknown regression function with (nearly, up a to logarithmic term) optimal uniform coverage under self-similarity. In addition, we provide efficient confidence sets and bands for a family of smooth functionals. Uncertainty quantification for related random forests or deep learning has been an open problem, with distributional results available only for point-wise prediction using bootstrap techniques [44]. Our results make a needed contribution to the literature on the widely sought-after UQ for (tree-based) machine learning methods.

Regarding supremum-norm (and its associated discrete ℓ_∞ version) posterior contraction rates, their derivation is typically more delicate compared to the more familiar testing distances (e.g., L^2 or Hellinger) for which general theory has been available since the seminal work [32]. Despite the lack of unifying theory, however, advances have been made in the last few years [14, 34, 39] including specific models [47, 48, 54, 63]. However, Bayesian *adaptation* for the supremum loss has been obtained, to the best of our knowledge, *only* through spike-and-slab priors (the work [62] uses Gaussian process priors, but adaptation is obtained via Lepski's method). In particular, [39] show that spike-and-slab priors on wavelet coefficients yield the *exact* adaptive minimax rate in the white noise model and [61] considers the anisotropic case in a regression framework. For density estimation, [15, 16] derive optimal $\|\cdot\|_\infty$-rates for Pólya tree priors, while [46] considers adaptation for log-density spike and slab priors. In this work, we consider Gaussian white noise and nonparametric regression with Bayesian CART which is widely used in practice.

Bayesian CART is a method of function estimation based on ideas of recursive partitioning of the predictor space. The work [26] highlighted the link between dyadic CART and best ortho-basis selection using Haar wavelets in two dimensions; [30] furthered this connection by considering unbalanced Haar wavelets of [36]. CART methods have been also studied in the machine learning literature; see, for example, [7, 53, 59] and references therein. Unlike plain wavelet shrinkage methods and standard spike-and-slab priors, general Bayesian CART priors have extra flexibility by allowing for (some) *basis selection*. First results in this direction are derived in Section 4. This aspect is particularly useful in higher-dimensional data, where CART methods have been regarded as an attractive alternative to other methods [27].

By taking the Bayesian point of view, we relate Bayesian CART to structured wavelet shrinkage using libraries of *weakly* balanced Haar bases. Each tree provides an underlying skeleton or a 'sparsity structure' which supervises the sparsity pattern (see, e.g., [2]). We show that Bayesian CART borrows strength between coefficients in the tree ancestry by giving rise to a variant of the *g-prior* [64]. Similarly as independent product priors, we show that these dependent priors *also* lead to adaptive supremum norm concentration rates (up to a logarithmic factor). To illustrate that local (internal) sparsity is a key driver of adaptivity, we show that dense trees are incapable of adaptation.

To convey the main ideas, the mathematical development will be performed through the lense of a Gaussian white noise model. Our techniques, however, also apply in nonparametric regression. Results in this setting are briefly presented in Section 3.5 with details postponed until the Supplementary Material (Section S-1.1). The white noise model is defined through the following stochastic differential equation, for an integer $n \geq 1$,

$$dX(t) = f_0(t)\,dt + \frac{1}{\sqrt{n}}\,dW(t), \quad t \in [0,1], \tag{1}$$

where $X(t)$ is an observation process, $W(t)$ is the standard Wiener process on $[0,1]$ and f_0 is unknown and belongs to $L^2[0,1]$, set of squared-integrable functions on $[0,1]$. The model (1) is observationally equivalent to a Gaussian sequence space model after projecting the observation process onto a wavelet basis $\{\psi_{lk} : l \geq 0, 0 \leq k \leq 2^l - 1\}$ of $L^2[0,1]$. This sequence model writes as

$$X_{lk} = \beta^0_{lk} + \frac{\varepsilon_{lk}}{\sqrt{n}}, \quad \varepsilon_{lk} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{2}$$

where the wavelet coefficients $\beta_{lk}^0 = \langle f_0, \psi_{lk}\rangle = \int_0^1 f_0(t)\psi_{lk}(t)\,dt$ of f_0 are indexed by a scale index $l \geq -1$ and a location index $k \in \{0, \ldots, (2^l - 1)_+\}$. A paradigmatic example is the standard Haar wavelet basis

$$\psi_{-10}(x) = \mathbb{I}_{[0,1]}(x) \quad \text{and} \quad \psi_{lk}(x) = 2^{l/2}\psi\big(2^l x - k\big) \quad (l \geq 0), \tag{3}$$

obtained with orthonormal dilation-translations of $\psi = \mathbb{I}_{(0,1/2]} - \mathbb{I}_{(1/2,1]}$, where $\mathbb{I}_A$ denotes the indicator of a set A. Later in the text, we also consider weakly balanced Haar wavelet relaxations (Section 4), as well as smooth wavelet bases (Section S-4.2).

One of the key motivations behind the Bayesian approach is the mere fact that the posterior is an actual distribution, whose limiting shape can be analyzed towards obtaining *uncertainty quantification* and inference. Our results in this direction can be grouped in two subsets. First, for uncertainty quantification for f_0 itself, we construct adaptive and honest confidence bands under self-similarity (with coverage converging to one). Exact asymptotic coverage is achieved through intersections with a multiscale credible band (along the lines of [49]). Confidence bands construction for regression surfaces is a fundamental task in nonparametric regression and can indicate whether there is empirical evidence to support conjectured features such as multi-modality or exceedance of a level. Results of this type are, to date, unavailable for classical CART, random forests and/or deep learning. Second, we consider inference for smooth functionals of f_0, including linear ones and the primitive functional $\int_0^{\cdot} f_0$, for which exact optimal confidence sets are derived from posterior quantiles. While these results for functionals are stated in the main paper (Theorem 4 below), their derivation is most naturally obtained through a general limiting shape result, stated and proved in the Supplementary Material (Theorem S-3). Such an adaptive Bernstein-von Mises theorem for Bayesian CART is obtained following the approach of [17, 18]; it is only the second result of this kind (providing *adaptation*) after the recent result of Ray [49].

The paper is structured as follows. Section 2 introduces regression tree-priors, as well as the notion of tree-shaped sparsity and the g-prior for trees. In Section 3, we state supremum-norm inference properties of Bayesian dyadic CART (estimation and confidence bands). Section 4 considers flexible partitionings allowing for basis choice. A brief discussion can be found in Section 5. The proof of our master Theorem 1 can be found in Section 6. The Supplementary Material [19] gathers the proofs of the remaining results. The sections and equations of this supplement are referred to with an additional symbol "S-" in the numbering.

Notation. Let $\mathcal{C}([0, 1])$ denote the set of continuous functions on $[0, 1]$ and let ϕ_σ denote the normal density with zero mean and variance σ^2. Let $\mathbb{N} = \{0, 1, 2, \ldots\}$ be the set of natural integers and $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. We denote by I_K the $K \times K$ identity matrix, Also, B^c denotes the complement of a set B. For an interval $I = (a, b] \subset (0, 1]$, let $|I| = b - a$ be its diameter and $a \vee b = \max(a, b)$. The notation $x \lesssim y$ means $x \leq Cy$ for C a large enough universal constant, and $:=$ (or $=:$) means "the left-hand side is defined as."

2. Trees and wavelets. In this section, we discuss multiscale prior assignments on functions $f \in L^2[0, 1]$ (i.e., priors on the sequence of wavelet coefficients $\beta_{lk} = \langle f, \psi_{lk}\rangle$) inspired by (and including) Bayesian CART. Such methods recursively subdivide the predictor space into cells where f can be estimated locally. The partitioning process can be captured with a tree object (a hierarchical collection of nodes) and a set of splitting rules attached to each node. Section 2.1 discusses priors on the tree object. The splitting rules are ultimately tied to a chosen basis, where the traditional Haar wavelet basis yields deterministic dyadic splits (as we explain in Section 2.1.2). Later in Section 4, we extend our framework to random unbalanced Haar bases which allow for more flexible splits. Beyond random partitioning, an

integral component of CART methods are histogram heights assigned to each partitioning cell. We flesh out connections between Bayesian histograms and wavelets in Section 2.2. Finally, we discuss Bayesian CART priors over histogram heights in Section 2.3.

2.1. *Priors on trees* $\Pi_{\mathbb{T}}(\cdot)$. First, we need to make precise our definition of a tree object which will form a skeleton of our prior on (β_{lk}) for each given basis $\{\psi_{lk}\}$. Throughout this paper, we will largely work with the Haar basis.

DEFINITION 1 (Tree terminology). We define a *binary tree* $\mathcal{T}$ as a collection of nodes (l,k), where $l \in \mathbb{N}$, $k \in \{0,\dots,2^l-1\}$, that satisfies

$$(l,k) \in \mathcal{T}, \quad l \geq 1 \quad \Rightarrow \quad \big(l-1, \lfloor k/2 \rfloor\big) \in \mathcal{T}.$$

In the last display, the node (l,k) is a *child* of its *parent* node $(l-1,\lfloor k/2\rfloor)$. A *full binary tree* consists of nodes with exactly 0 or 2 children. For a node (l,k), we refer to l as the *layer index* (or also *depth*) and k as the *position* in the lth layer (from left to right). The cardinality $|\mathcal{T}|$ of a tree $\mathcal{T}$ is its total number of nodes and the *depth* is defined as $d(\mathcal{T}) = \max_{(l,k)\in\mathcal{T}} l$.

A node $(l,k) \in \mathcal{T}$ belongs to the set $\mathcal{T}_{\text{ext}}$ of *external* nodes (also called *leaves*) of $\mathcal{T}$ if it has no children and to the set $\mathcal{T}_{\text{int}}$ of *internal* nodes, otherwise. By definition $|\mathcal{T}| = |\mathcal{T}_{\text{int}}| + |\mathcal{T}_{\text{ext}}|$, where, for full binary trees, we have $|\mathcal{T}| = 2|\mathcal{T}_{\text{int}}| + 1$. An example of a full binary tree is depicted in Figure 1(a). In the sequel, $\mathbb{T}$ denotes the set of full binary trees of depth no larger than $L = L_{\max} = \lfloor \log_2 n \rfloor$, a typical cut-off in wavelet analysis. Indeed, trees can be associated with certain wavelet decompositions, as will be seen in Section 2.2.2.

Before defining tree-structured priors over the entire functions f's, we first discuss various ways of assigning a prior distribution over $\mathbb{T}$, that is over trees themselves. We focus on the Bayesian CART prior [20], which became an integral component of many Bayesian tree regression methods including BART [22].

2.1.1. *Bayesian CART priors*. The Bayesian CART construction of [20] assigns a prior over $\mathbb{T}$ via the heterogeneous Galton-Watson (GW) process. The prior description utilizes the following top-down left-to-right exploration metaphor (see also [50]). Denote with Q a queue of nodes waiting to be explored. Each node (l,k) is assigned a random binary indicator $\gamma_{lk} \in \{0,1\}$ for whether or not it is split. Starting with $\mathcal{T} = \varnothing$, one initializes the exploration process by putting the root node $(0,0)$ tentatively in the queue, that is, $Q = \{(0,0)\}$. One then repeats the following three steps until $Q = \varnothing$:

(a) Pick a node $(l,k) \in Q$ with the highest priority (i.e., the smallest index $2^l + k$) and if $l < L_{\max}$, split it with probability

$$p_{lk} = \mathbb{P}(\gamma_{lk} = 1). \tag{4}$$

If $l = L_{\max}$, set $\gamma_{lk} = 0$.

(b) If $\gamma_{lk} = 0$, remove (l,k) from Q.

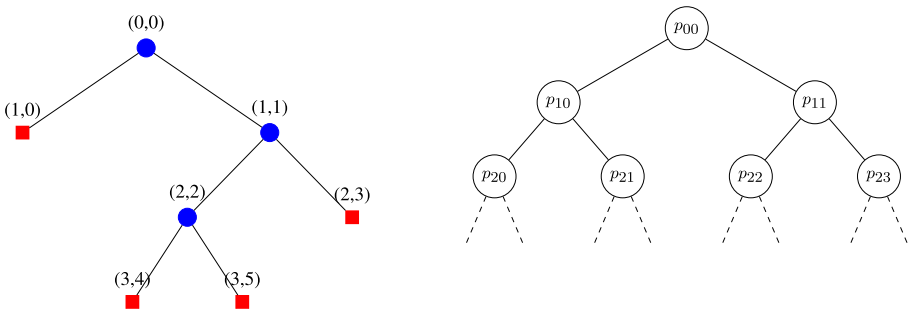
(c) If $\gamma_{lk} = 1$, then:

(i) add (l,k) to the tree, that is, $\mathcal{T}_{\text{int}} \leftarrow \mathcal{T}_{\text{int}} \cup \{(l,k)\}$,

(ii) remove (l,k) from Q and if $l < L_{\max}$ add its children to Q, that is,

$$Q \leftarrow Q \setminus \big\{(l,k)\big\} \cup \big\{(l+1,2k),(l+1,2k+1)\big\}.$$

The tree skeleton is probabilistically underpinned by the cut probabilities (p_{lk}) which are typically assumed to decay with the depth l as a way to penalise too complex trees. While [20] suggest $p_{lk} = \alpha/(1+l)^\gamma$ for some $\alpha \in (0,1)$ and $\gamma > 0$, [50] point out that this decay



(a) A full binary tree (b) Binary tree of prior cut probabilities

FIG. 1. (*Left*) *A full binary tree* $\mathcal{T} = \mathcal{T}_{\text{int}} \cup \mathcal{T}_{\text{ext}}$. *Red nodes are external nodes* $\mathcal{T}_{\text{ext}}$ *and blue nodes are internal nodes* $\mathcal{T}_{\text{int}}$. (*Right*) *A binary tree of cut probabilities* p_{lk} *in* (4).

may not be fast enough and suggest instead $p_{lk} = \Gamma^{-l}$ for some $2 < \Gamma < n$, which leads to a (near) optimal empirical L^2-convergence rate. We use a similar assumption in our analysis, and also assume that the split probability depends only on l, and simply denote $p_l = p_{lk}$.

Independently of [20, 25] proposed another variant of Bayesian CART, which first draws the number of leaves (i.e., external nodes) $K = |\mathcal{T}_{\text{ext}}|$ at random from a certain prior on integers, for example, a Poisson distribution (say, conditioned to be nonzero). Then, a tree $\mathcal{T}$ is sampled uniformly at random from all full binary trees with K leaves. Noting that there are $\mathbb{C}_{K-1}$ such trees, with $\mathbb{C}_K$ the Kth Catalan number (see Lemma S-3), this leads to $\Pi(\mathcal{T}) = (\lambda^K/[K!(e^\lambda - 1)]) \cdot \mathbb{C}_{K-1}^{-1}$. As we restrict to trees in $\mathbb{T}$, that is, with depth at most $L = L_{\max}$, we slightly update the previous prior choice by setting, for some $\lambda > 0$, with $K = |\mathcal{T}_{\text{ext}}|$,

$$\Pi_{\mathbb{T}}(\mathcal{T}) \propto \frac{\lambda^K}{(e^\lambda - 1)K!} \frac{1}{\mathbb{C}_{K-1}} \mathbb{I}_{\mathcal{T} \in \mathbb{T}}, \tag{5}$$

where $\propto$ means 'proportional to.' We call the resulting prior $\Pi_{\mathbb{T}}$ the 'conditionally uniform prior' with a parameter λ.

2.1.2. *Trees and random partitions.* Trees provide a structured framework for generating random partitions of the predictor space (here we choose $(0, 1]$ for simplicity of exposition). In CART methodology, each node $(l, k) \in \mathcal{T}$ is associated with a partitioning interval $I_{lk} \subseteq (0, 1]$. Starting from the trivial partition $I_{00} = (0, 1]$, the simplest way to obtain a partition is by successively dividing each I_{lk} into $I_{lk} = I_{l+12k} \cup I_{l+12k+1}$. One central example is *dyadic* intervals I_{lk} which correspond to the domain of the balanced Haar wavelets ψ_{lk} in (3), that is,

$$I_{00} = (0, 1], \qquad I_{lk} = \big(k2^{-l}, (k+1)2^{-l}\big] \quad \text{for } l \geq 0 \text{ and } 0 \leq k < 2^l. \tag{6}$$

For any fixed depth $l \in \mathbb{N}$, the intervals $\bigcup_{0 \leq k < 2^l} I_{lk}$ form a deterministic regular (equispaced) partition of $(0, 1]$. Trees, however, generate *more flexible* partitions $\bigcup_{(l,k) \in \mathcal{T}_{\text{ext}}} I_{lk}$ by keeping only those intervals I_{lk} attached to the leaves of the tree. Since $\mathcal{T}$ is treated as random with a prior $\Pi_{\mathbb{T}}$ (as defined in Section 2.1), the resulting partition will also be random.

EXAMPLE 1. Figure 1(a) shows a full binary tree $\mathcal{T} = \mathcal{T}_{\text{int}} \cup \mathcal{T}_{\text{ext}}$, where $\mathcal{T}_{\text{int}} = \{(0, 0), (1, 1), (2, 2)\}$ and $\mathcal{T}_{\text{ext}} = \{(1, 0), (2, 3), (3, 4), (3, 5)\}$, resulting in the partition of $(0, 1]$ given by

$$(I_{lk})_{(l,k) \in \mathcal{T}_{\text{ext}}} = \big\{(0, 1/2], (1/2, 5/8], (5/8, 3/4], (3/4, 1]\big\}. \tag{7}$$

The set of possible *split points* obtained with (6) is confined to dyadic rationals. One can interpret the resulting partition as the result of recursive splitting where, at each level l, intervals I_{lk} for each internal node $(l,k) \in \mathcal{T}_{\text{int}}$ are cut in half and intervals I_{lk} for each external node $(l,k) \in \mathcal{T}_{\text{ext}}$ are left alone. We will refer to such a recursive splitting process as *dyadic CART*. There are several ways to generalize this construction, for instance by considering arbitrary splitting rules that iteratively dissect the intervals at values other than the midpoint. We explore such extensions in Section 4.

2.2. *Tree-shaped priors on f.* This section outlines two strategies for assigning a tree-shaped prior distribution on f underpinned by a tree skeleton $\mathcal{T} \in \mathbb{T}$. Each tree $\mathcal{T} = \mathcal{T}_{\text{int}} \cup \mathcal{T}_{\text{ext}}$ can be associated with two sets of coefficients: (a) *internal* coefficients β_{lk} attached to wavelets ψ_{lk} for $(l,k) \in \mathcal{T}_{\text{int}}$ and (b) *external* coefficients $\widetilde{\beta}_{lk}$ attached to partitioning intervals I_{lk} for $(l,k) \in \mathcal{T}_{\text{ext}}$ (see Section 2.1.2). While wavelet priors (Section 2.2.1) assign the prior distribution internally on β_{lk}, Bayesian CART priors [20, 25] (Section 2.2.2) assign the prior externally on $\widetilde{\beta}_{lk}$. We discuss and relate these two strategies in more detail below.

2.2.1. *Tree-shaped wavelet priors.* Traditional (linear) Haar wavelet reconstructions for f deploy *all* wavelet coefficients β_{lk} with resolutions l smaller than some $d > 0$. This strategy amounts to fitting a *flat tree* with d layers (i.e., a tree that contains all nodes up to a level d, see Figure 2) or, equivalently, a regular dyadic regression histogram with 2^d bins. This construction can be made more flexible by selecting coefficients prescribed by trees that are not necessarily flat. Given a full binary tree $\mathcal{T} \in \mathbb{T}$, one can build the following wavelet reconstruction of f using *only* active wavelet coefficients that are *inside* a tree $\mathcal{T}$:

$$f_{\mathcal{T},\boldsymbol{\beta}}(x) = \beta_{-10}\psi_{-10}(x) + \sum_{(l,k)\in\mathcal{T}_{\text{int}}} \beta_{lk}\psi_{lk}(x) = \sum_{(l,k)\in\mathcal{T}'_{\text{int}}} \beta_{lk}\psi_{lk}(x), \tag{8}$$

where $\boldsymbol{\beta} = (\beta_{-10}, (\beta_{lk})_{0\le l\le L-1, 0\le k<2^l})'$ is a vector of wavelet coefficients and where $\mathcal{T}'_{\text{int}} = \mathcal{T}_{\text{int}} \cup \{(-1,0)\}$ is the 'rooted' tree with the index $(-1,0)$ added to $\mathcal{T}_{\text{int}}$. Note that $|\mathcal{T}'_{\text{int}}| = |\mathcal{T}_{\text{ext}}|$.

Define a *tree-shaped wavelet prior* on $f_{\mathcal{T},\boldsymbol{\beta}}$ as the prior induced by the hierarchical model

$$\begin{aligned} \mathcal{T} &\sim \Pi_{\mathbb{T}}, \\ (\beta_{lk})_{lk} \mid \mathcal{T} &\sim \bigotimes_{(l,k)\in\mathcal{T}'_{\text{int}}} \pi(\beta_{lk}) \otimes \bigotimes_{(l,k)\notin\mathcal{T}'_{\text{int}}} \delta_0(\beta_{lk}), \end{aligned} \tag{9}$$

where $\Pi_{\mathbb{T}}$ is a prior on trees as described in Section 2.1.1 and where the active wavelet coefficients β_{lk} for $(l,k) \in \mathcal{T}_{\text{int}}$ follow a distribution with a bounded and positive density $\pi(\beta_{lk})$ on $\mathbb{R}$. The prior (9) is seen as a distribution on $\mathbb{R}^{2^L}$, where all remaining coefficients, that is, β_{lk}'s for $l \ge L$, are set to 0.

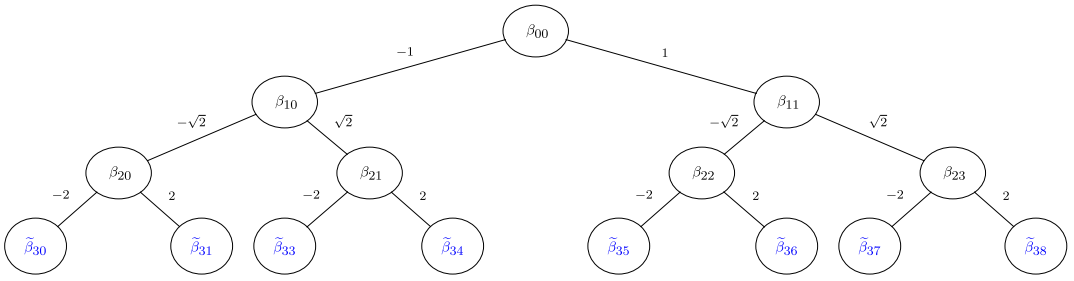


FIG. 2. *Flat tree with edges weighted by the amplitude of the Haar wavelets.*

The prior (9) contains the so-called *sieve priors* [18] (i.e., flat trees) as a special case, where the sieve is with respect to the approximating spaces $\text{Vect}\{\psi_{lk}, l < d\}$ for some $d \geq 0$. For nonparametric estimation of f_0, it is well known that sieve priors can achieve (nearly) adaptive rates in the L^2-sense (see, e.g., [33]). In turns out, however, that sieve priors (and therefore flat tree priors) are too rigid to enable adaptive results for stronger losses such as the supremum norm, as we demonstrate in Theorem 5 in Section 3.4 (Supplementary Material). This theorem illustrates that supremum norm adaptation using Bayesian (or other likelihood-based) methods is a delicate phenomenon that is not attainable by many typical priors.

By definition, the prior (9) weeds out all wavelet coefficients β_{lk} that are not supported by the tree skeleton (i.e., are not *internal* nodes in $\mathcal{T}$). This has two shrinkage implications: global and local. First, the global level of truncation (i.e., the depth of the tree) in (9) is not fixed but random. Second, unlike in sieve priors, only some low resolution coefficients are active depending on whether or not the tree splits the node (l, k). These two shrinkage aspects create hope that tree-shaped wavelet priors (9) attain adaptive supremum norm rates (up to log factors) and enable construction of adaptive confidence bands. We see later in Section 3 that this optimism is indeed warranted.

For adaptive wavelet shrinkage, [23] propose a Gaussian mixture spike-and-slab prior on the wavelet coefficients. The point mass spike-and-slab incarnation of this prior was studied by [39] and [49]. Independently for each wavelet coefficient β_{lk} at resolutions larger than some $l_0(n)$ (strictly increasing sequence), the prior in [49] can be written in the standard spike-and-slab form

$$\pi(\beta_{lk} \mid \gamma_{lk}) = \gamma_{lk}\pi(\beta_{lk}) + (1 - \gamma_{lk})\delta_0(\beta_{lk}), \tag{10}$$

where $\gamma_{lk} \in \{0, 1\}$ for whether or not the coefficient is active with $\mathbb{P}(\gamma_{lk} = 1 \mid \theta_l) = \theta_l$. Moreover, the prior on all coefficients at resolutions no larger than $l_0(n)$ is dense, that is, $\theta_l = 1$ for $l \leq l_0(n)$. The value θ_l can be viewed as the probability that a given wavelet coefficient β_{lk} at resolution l will contain 'signal.'

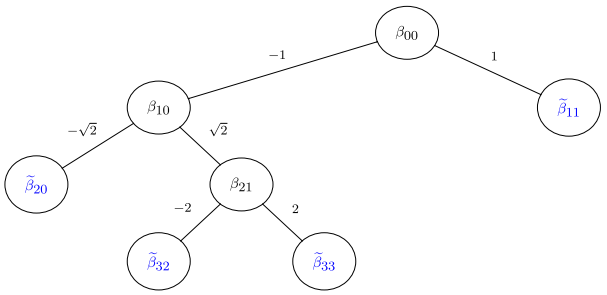
There are undeniable similarities between (9) and (10), in the sense that the binary inclusion indicator γ_{lk} in (10) can be regarded as the node splitting indicator γ_{lk} in (4). While the indicators γ_{lk} in (10) are *independent* under the spike-and-slab prior, they are hierarchically constrained under the CART prior, where the pattern of nonzeroes encodes the tree oligarchy. The seeming resemblance of the CART-type prior (9) to the spike-and-slab prior (10) makes one naturally wonder whether, unlike sieve-type priors, CART posteriors attain adaptive supremum-norm inference.

2.2.2. *Bayesian CART priors*. A perhaps more transparent approach to assigning a tree-shaped prior on f is through histograms (as opposed to wavelet reconstructions from Section 2.2.1). Each tree $\mathcal{T} \in \mathbb{T}$ generates a random partition via intervals I_{lk} (see Section 2.1.2) and gives rise to the following histogram representation:

$$\widetilde{f}_{\mathcal{T},\widetilde{\boldsymbol{\beta}}}(x) = \sum_{(l,k)\in\mathcal{T}_{\text{ext}}} \widetilde{\beta}_{lk}\mathbb{I}_{I_{lk}}(x), \tag{11}$$

where $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_{lk} : (l, k) \in \mathcal{T}_{\text{ext}})'$ is a vector of reals interpreted as step heights and where I_{lk}'s are obtained from the tree $\mathcal{T}$ as in Section 2.1.2 (and as illustrated in Example 1). We now define the *(Dyadic) Bayesian CART prior* on f using the following hierarchical model on the *external* coefficients rather than *internal* coefficients (compare with (9)):

$$\begin{aligned} \mathcal{T} &\sim \Pi_{\mathbb{T}}, \\ (\widetilde{\beta}_{lk})_{(l,k)\in\mathcal{T}_{\text{ext}}} \mid \mathcal{T} &\sim \bigotimes_{(l,k)\in\mathcal{T}_{\text{ext}}} \widetilde{\pi}(\widetilde{\beta}_{lk}), \end{aligned} \tag{12}$$



(a)

$$\begin{pmatrix} \widetilde{\beta}_{11} \\ \widetilde{\beta}_{20} \\ \widetilde{\beta}_{32} \\ \widetilde{\beta}_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & -\sqrt{2} & 0 \\ 1 & -1 & \sqrt{2} & -2 \\ 1 & -1 & \sqrt{2} & 2 \end{pmatrix} \begin{pmatrix} \beta_{-10} \\ \beta_{00} \\ \beta_{10} \\ \beta_{21} \end{pmatrix}$$

(b)

FIG. 3. (*a*) *Example of a full binary tree, edges weighted by the amplitude of the Haar wavelets.* (*b*) *Pinball matrix of the tree in* (*a*).

where $\Pi_{\mathbb{T}}$ is as in Section 2.1, and where the height $\widetilde{\beta}_{lk}$ at a specific $(l,k) \in \mathcal{T}_{\text{ext}}$ has a bounded and positive density $\widetilde{\pi}(\widetilde{\beta}_{lk})$ on $\mathbb{R}$. This model *coincides* with the widely used Bayesian CART priors using a midpoint dyadic splitting rule (as we explained in Section 2.1.2). In practice, the density $\widetilde{\pi}$ is often chosen as centered Gaussian with some variance $\sigma^2 > 0$ [20, 25].

The histogram prior (11) can be rephrased in terms of wavelets. Indeed, the histogram representation (11) can be rewritten in terms of the *internal* coefficients, that is, $\widetilde{f}_{\mathcal{T},\widetilde{\boldsymbol{\beta}}}(x) = f_{\mathcal{T},\boldsymbol{\beta}}(x)$ as in (8), with β_{lk}'s and $\widetilde{\beta}_{lk}$'s linked via

$$\widetilde{\beta}_{lk} = \beta_{-10} + \sum_{j=0}^{l-1} s_{\lfloor k/2^{l-j-1} \rfloor} 2^{j/2} \beta_{j \lfloor k/2^{l-j} \rfloor}, \tag{13}$$

where $s_k = (-1)^{k+1}$. The identity (13) follows the fact that for $x \in I_{lk}$ we obtain $\widetilde{\beta}_{lk} = \sum_{(l',k') \in P_{lk}} \beta_{l'k'} \psi_{l'k'}$ from (11), where $P_{lk} \equiv \{(j, \lfloor k/2^{l-j} \rfloor) : j = 0, \ldots, l-1\}$ are the ancestors of the bottom node (l,k). Note that $\psi_{j \lfloor k/2^{l-j} \rfloor} = 2^{j/2} s_{\lfloor k/2^{l-j-1} \rfloor}$ where $s = (-1)^{k+1}$ for whether x belongs to the left (positive sign) or right (negative sign) of the wavelet piece. There is a pinball game metaphor behind (13). A ball is dropped through a series of dyadically arranged pins of which the ball can bounce off to the right (when $s_k = +1$) or to the left (when $s_k = -1$). The ball ultimately lands in one of the histogram bins I_{lk} whose coefficient $\widetilde{\beta}_{lk}$ is obtained by aggregating β_{lk}'s of those pins (l,k) that the ball encountered on its way down. The pinball aggregation process can be understood from Figure 3. The duality between the equivalent representations (11) and (8) through (13) provides various avenues for constructing prior distributions, and enables an interesting interpretation of Bayesian CART [20, 25] as a correlated wavelet prior, as we now see.

2.3. *The g-prior for trees.* We now discuss various ways of assigning a prior distribution on the bottom node histogram heights $\widetilde{\beta}_{lk}$ and, equivalently, the internal Haar wavelet coefficients β_{lk}. This section also describes an interesting connection between the widely used Bayesian CART prior [20, 25] and a g-prior [64] on wavelet coefficients. For a given tree $\mathcal{T}$, let $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}'_{\text{int}})'$ denote the vector of *ordered internal* node coefficients β_{lk} including the extra root node $(-1,0)$ (and with ascending ordering according to $2^l + k$). Similarly, $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}_{\text{ext}})'$ is the vector of *ordered external* node coefficients $\widetilde{\beta}_{lk}$. The duality between $\boldsymbol{\beta}_{\mathcal{T}}$ and $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}}$ is apparent from the pinball equation (13) written in matrix form,

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} = A_{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{T}}, \tag{14}$$

where $A_{\mathcal{T}}$ is a square $|\mathcal{T}_{\text{ext}}| \times |\mathcal{T}'_{\text{int}}|$ matrix (noting $|\mathcal{T}_{\text{ext}}| = |\mathcal{T}'_{\text{int}}|$), further referred to as the *pinball matrix*. Each row of $A_{\mathcal{T}}$ encodes the ancestors of the external node, where the nonzero entries correspond to the internal nodes in the family pedigree. The entries are rescaled, where younger ancestors are assigned more weight. For example, the tree $\mathcal{T}$ in Figure 3(a) induces a pinball matrix $A_{\mathcal{T}}$ in Figure 3(b). The pinball matrix $A_{\mathcal{T}}$ can be easily expressed in terms of a diagonal matrix and an orthogonal matrix as

$$A_{\mathcal{T}} A'_{\mathcal{T}} = \boldsymbol{D}_{\mathcal{T}} \quad \text{where } \boldsymbol{D}_{\mathcal{T}} = \text{diag}\{\widetilde{d}_{lk,lk}\}_{(l,k)\in\mathcal{T}_{\text{ext}}}, \widetilde{d}_{lk,lk} = 2^l. \tag{15}$$

This results from the fact that the collection $(2^{l/2}\mathbb{I}_{lk}, (l,k) \in \mathcal{T}_{\text{ext}})$ is an orthonormal system spanning the same space as $(\psi_{jk}, (j,k) \in \mathcal{T}'_{\text{int}})$, so $\boldsymbol{D}_{\mathcal{T}}^{-1/2} A_{\mathcal{T}}$ is an orthonormal change-of-basis matrix. We now exhibit precise connections between the theoretical wavelet prior (9) which draws $\beta_{lk} \sim \pi$ and the practical Bayesian CART histogram prior which draws $\widetilde{\beta}_{lk} \sim \widetilde{\pi}$.

Recall that the wavelet prior (9) assumes independent wavelet coefficients, for example, through the standard Gaussian prior $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, I_{|\mathcal{T}_{\text{ext}}|})$. Starting *from within* the tree, this translates into the following independent product prior on the bottom coefficients $\widetilde{\beta}_{lk}$ through (14):

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} \sim \mathcal{N}(0, \boldsymbol{D}_{\mathcal{T}}) \quad \text{where } \boldsymbol{D}_{\mathcal{T}} \text{ was defined in (15)}, \tag{16}$$

that is, $\text{var}\,\widetilde{\beta}_{lk} = 2^l$ where the variances increase with the resolution l.

The Bayesian CART prior [20, 25], on the other hand, starts *from outside* the tree by assigning $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} \sim \mathcal{N}(0, g_n I_{|\mathcal{T}_{\text{ext}}|})$ for some $g_n > 0$, ultimately setting the bottom node variances equal. This translates into the following 'g-prior' on the *internal* wavelet coefficients through the duality (14).

DEFINITION 2. Let $\mathcal{T} \in \mathbb{T}$ with a pinball matrix $A_{\mathcal{T}}$ and denote with $\boldsymbol{\beta}_{\mathcal{T}}$ the internal wavelet coefficients. We define the *g-prior* for trees as

$$\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}\big(0, g_n (A'_{\mathcal{T}} A_{\mathcal{T}})^{-1}\big) \quad \text{for some } g_n > 0. \tag{17}$$

Note that, except for very special cases (e.g., flat trees) $A'_{\mathcal{T}} A_{\mathcal{T}}$ is in general not diagonal, unlike $A_{\mathcal{T}} A'_{\mathcal{T}}$. This means that the correlation structure induced by the Bayesian CART prior on internal wavelet coefficients is nontrivial, although $A'_{\mathcal{T}} A_{\mathcal{T}}$ admits some partial sparsity. We characterize basic properties of the pinball matrix in Section S-2.1 in the Supplementary Material. For example, Proposition S-3 shows that matrices $A'_{\mathcal{T}} A_{\mathcal{T}}$ and $A_{\mathcal{T}} A'_{\mathcal{T}}$ have the same eigenspectrum consisting of values 2^l where l corresponds to the depth of the bottom nodes. This means that the g-prior variances (diagonal elements of $g_n (A'_{\mathcal{T}} A_{\mathcal{T}})^{-1}$) are lower-bounded by the minimal eigenvalue of $g_n (A'_{\mathcal{T}} A_{\mathcal{T}})^{-1}$ which equals $g_n 2^{-l}$ (where l is the depth of the deepest external node) which is lower-bounded by g_n/n. Since the traditional wavelet prior assumes variance 1, the choice $g_n = n$ matches the lower bound 1 by under-smoothing all possible variance combinations. While other choices could be potentially used (see [28, 29, 40] in the context of linear regression), we will consider $g_n = n$ in our results below.

We regard (17) as the 'g-prior for trees' due to its apparent similarity to g-priors for linear regression coefficients [64]. The g-prior has been shown to have many favorable properties in terms of invariance or predictive matching [4, 5]. Here, we explore the benefits of the g-type correlation structure in the context of structured wavelet shrinkage where each 'model' is defined by a tree topology. The correlation structure (17) makes this prior very different from any other prior studied in the context of wavelet shrinkage.

3. Inference with (dyadic) Bayesian CART. In this section, we investigate the inference properties of tree-based posteriors, showing that (a) they attain the minimax rate of posterior concentration in the supremum-norm sense (up to a log factor), and (b) enable uncertainty quantification: for f in the form of adaptive confidence bands, and for smooth functionals thereof, in terms of Bernstein-von Mises type results. For clarity of exposition, we focus now on the one-dimensional case, but the results readily extend to the multi-dimensional setting with $\mathbb{R}^d$, $d \geq 1$ fixed, as predictor space; see Section S-1.4 for more details.

3.1. *Posterior supremum-norm convergence.* Let us recall the standard inequality (see, e.g., (60) below), for f_0 a continuous function and f a Haar histogram (8), with coefficients β_{lk}^0 and β_{lk},

$$\|f - f_0\|_\infty \leq |\beta_{-10} - \beta_{-10}^0| + \sum_{l \geq -1} 2^{l/2} \max_{0 \leq k < 2^l} |\beta_{lk} - \beta_{lk}^0| =: \ell_\infty(f, f_0). \tag{18}$$

As ℓ_∞ dominates $\|\cdot\|_\infty$, it is enough to derive results for the ℓ_∞-loss.

Given a tree $\mathcal{T} \in \mathbb{T}$, and recalling that trees in $\mathbb{T}$ have depth at most $L := L_{\max} = \lfloor \log_2 n \rfloor$, we consider a generalized tree-shaped prior Π on the *internal wavelet* coefficients, recalling the notation $\mathcal{T}'_{\text{int}}$ from Section 2.2,

$$\begin{aligned} &\mathcal{T} \sim \Pi_{\mathbb{T}}, \\ &(\beta_{lk})_{l \leq L, k < 2^l} \,|\, \mathcal{T} \sim \pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k) \notin \mathcal{T}'_{\text{int}}} \delta_0(\beta_{lk}), \end{aligned} \tag{19}$$

where $\pi(\boldsymbol{\beta}_{\mathcal{T}})$ is a law to be chosen on $\mathbb{R}^{|\mathcal{T}'_{\text{int}}|}$, not necessarily of a product form. This is a generalization of (9), which allows for *correlated* wavelet coefficients (e.g., the g-prior). Let $\boldsymbol{X}_{\mathcal{T}}$ denote the vector of ordered responses X_{lk} in (2) for $(l,k) \in \mathcal{T}'_{\text{int}}$. From the white noise model, we have

$$\boldsymbol{X}_{\mathcal{T}} = \boldsymbol{\beta}_{\mathcal{T}} + \frac{1}{\sqrt{n}} \boldsymbol{\varepsilon}_{\mathcal{T}} \quad \text{with } \boldsymbol{\varepsilon}_{\mathcal{T}} \sim \mathcal{N}(0, I_{|\mathcal{T}_{\text{ext}}|}) \text{ (given } \mathcal{T}).$$

By Bayes' formula, the posterior distribution $\Pi[\cdot \,|\, X]$ of the variables $(\beta_{lk})_{l \leq L, k}$ has density

$$\sum_{\mathcal{T} \in \mathbb{T}} \Pi[\mathcal{T} \,|\, X] \cdot \pi(\boldsymbol{\beta}_{\mathcal{T}} \,|\, X) \cdot \prod_{(l,k) \notin \mathcal{T}'_{\text{int}}} \mathbb{I}_0(\beta_{lk}), \tag{20}$$

where, denoting as shorthand $N_X(\mathcal{T}) = \int \mathrm{e}^{-\frac{n}{2}\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 + n \boldsymbol{X}'_{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{T}}} \pi(\boldsymbol{\beta}_{\mathcal{T}}) d\boldsymbol{\beta}_{\mathcal{T}}$,

$$\pi(\boldsymbol{\beta}_{\mathcal{T}} \,|\, X) = \frac{\mathrm{e}^{-\frac{n}{2}\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 + n \boldsymbol{X}'_{\mathcal{T}} \boldsymbol{\beta}_{\mathcal{T}}} \pi(\boldsymbol{\beta}_{\mathcal{T}})}{N_X(\mathcal{T})}, \tag{21}$$

$$\Pi[\mathcal{T} \,|\, X] = \frac{W_X(\mathcal{T})}{\sum_{\mathcal{T} \in \mathbb{T}} W_X(\mathcal{T})} \quad \text{with } W_X(\mathcal{T}) = \Pi_{\mathbb{T}}(\mathcal{T}) N_X(\mathcal{T}). \tag{22}$$

Let us note that the sum in the last display is finite, as we restrict to trees of depth at most $L = L_{\max}$. Note that the classes of priors $\Pi_{\mathbb{T}}$ from Section 2 are nonconjugate, the posterior on trees is given by the somewhat intricate expression (22) and does not belong to one of the classes of $\Pi_{\mathbb{T}}$ priors. While the posterior expression (21) allows for general priors $\pi(\boldsymbol{\beta}_{\mathcal{T}})$, we will focus on conditionally conjugate Gaussian priors for simplicity. This assumption is not essential and can be relaxed. For instance, in case $\pi(\boldsymbol{\beta}_{\mathcal{T}})$ is of a product form, one could use a product of, for example, Laplace distributions, using similar ideas as in [17], Theorem 5.

Our first result exemplifies the potential of tree-shaped priors by showing that Dyadic Bayesian CART achieves the minimax rate of posterior concentration over Hölder balls in the

sup-norm sense, that is, $\varepsilon_n = (n/\log n)^{-\alpha/(2\alpha+1)}$, up to a logarithmic term. Define a Hölder-type ball of functions on $[0,1]$ as

$$\mathcal{H}(\alpha, M) := \Big\{ f \in \mathcal{C}[0,1] : \max_{l \geq 0, 0 \leq k < 2^l} 2^{l(\frac{1}{2}+\alpha)} |\langle f, \psi_{lk}\rangle| \vee |\langle f, \psi_{-10}\rangle| \leq M \Big\}. \tag{23}$$

For balanced Haar wavelets ψ_{lk} as in (3), $\mathcal{H}(\alpha, M)$ contains the a standard α-Hölder (resp. Lipschitz when $\alpha = 1$) ball of functions for any $\alpha \in (0,1]$, defined as

$$\mathcal{H}_M^\alpha := \Big\{ f : \|f\|_\infty \leq M, \frac{|f(x) - f(y)|}{|x-y|^\alpha} \leq M \ \forall x, y \in [0,1] \Big\}. \tag{24}$$

Our master rate-theorem, whose proof can be found in Section 6, is stated below. It will be extended in various directions in the sequel.

THEOREM 1. *Let $\Pi_{\mathbb{T}}$ be the Galton-Watson process prior from Section 2.1 with $p_{lk} = \Gamma^{-l}$ and $\Gamma > 2e^3$. Consider the tree-shaped wavelet prior (19) with $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$, where $\Sigma_{\mathcal{T}}$ is either $I_{|\mathcal{T}'_{\text{int}}|}$ or $g_n(A'_{\mathcal{T}}A_{\mathcal{T}})^{-1}$ with $g_n = n$. Define*

$$\varepsilon_n = \left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \quad \text{for } \alpha > 0. \tag{25}$$

Then for any $\alpha \in (0,1]$, $M > 0$, any sequence $M_n \to \infty$ we have for $n \to \infty$

$$\sup_{f_0 \in \mathcal{H}(\alpha, M)} E_{f_0} \Pi\big[f_{\mathcal{T},\boldsymbol{\beta}} : \ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}}, f_0) > M_n \varepsilon_n \,|\, X \big] \to 0. \tag{26}$$

By (18), the statement (26) also holds for the supremum loss $\|\cdot\|_\infty$.

EXTENSION 1. While Theorem 1 is formulated for Bayesian CART obtained with Haar wavelets, the concept of tree-shaped sparsity extends to general wavelets that give rise to smoother objects than just step functions. With $\{\psi_{lk}\}$ an S-regular wavelet basis on $[0,1]$, for example, the boundary-corrected wavelet basis of [24] (see [35], Chapter 4, with adaptation of the range of indices l), and with $f_0 \in \mathcal{H}(\alpha, M)$ defined in (23) for some $M > 0$ and *arbitrary* $0 < \alpha \leq S$, one indeed obtains the statement (26) by choosing $\Gamma \geq \Gamma_0(S) > 0$ or $c \geq c_0 > 0$ large enough, see Section S-4.2.

Theorem 1 encompasses both original Bayesian CART proposals for priors on bottom coefficients $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} \sim \mathcal{N}(0, I_{|\mathcal{T}_{\text{ext}}|})$ (the case $\Sigma_{\mathcal{T}} = g_n(A_{\mathcal{T}}A'_{\mathcal{T}})^{-1}$ discussed in Section 2.3) as well as the mathematically slightly simpler wavelet priors $\Sigma_{\mathcal{T}} = I_{|\mathcal{T}_{\text{ext}}|}$ (discussed in Section 2.2.1). We did not fully optimize the constants in the statement; for instance, one can check that $\Gamma > 2$ for the g-prior works. The rate ε_n in (25) coincides with the minimax rate for the supremum norm in the white noise model up to a logarithmic factor $(\log n)^{\frac{\alpha}{2\alpha+1}}$. We next show that this logarithmic factor is in fact real, that is, *not* an artifact of the upper-bound proof. We state the results for smooth-wavelet priors, which enable to cover arbitrarily large regularities, but a similar result could also be formulated for the Haar basis.

THEOREM 2. *Let $\Pi_{\mathbb{T}}$ be one of the Bayesian CART priors from Theorem 1. Consider the tree-shaped wavelet prior (19) with $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$, where $\Sigma_{\mathcal{T}}$ is $I_{|\mathcal{T}_{\text{ext}}|}$ and $\{\psi_{lk}\}$ an S-regular wavelet basis, $S \geq 1$. Let ε_n be the rate defined in (25) for a given $0 < \alpha \leq S$. Let the parameters of $\Pi_{\mathbb{T}}$ verify either $\Gamma \geq \Gamma_0(S)$ a large enough constant, or $c \geq c_0 > 0$ large enough. For any $M > 0$, there exists $m > 0$ such that, as $n \to \infty$,*

$$\inf_{f_0 \in \mathcal{H}(\alpha, M)} E_{f_0} \Pi\big[\ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}}, f_0) \leq m \varepsilon_n \,|\, X \big] \to 0. \tag{27}$$

In other words, there exists a sequence of elements of $\mathcal{H}(\alpha, M)$ along which the posterior convergence rate is *slower* than $m\varepsilon_n$ in terms of the ℓ_∞-metric. In particular, the upper-bound rate of Theorem 1 *cannot* hold uniformly over $\mathcal{H}(\alpha, M)$ with a rate faster than ε_n, which shows that the obtained rate is sharp (note the reversed inequality in (27) with respect to (26); we refer to [13] for more details on the notion of posterior rate lower bound). The proof of Theorem 2 can be found in Section S-4.3.

EXTENSION 2. Theorem 1 holds for a variety of other tree priors. This includes the conditionally uniform prior mentioned in Section 2.1.1 with $\lambda = 1/n^c$ in (5), or an exponential-type prior $\Pi_{\mathbb{T}}(\mathcal{T}) \propto \mathrm{e}^{-c|\mathcal{T}_{\mathrm{ext}}|\log n}\mathbb{I}_{\mathcal{T}\in\mathbb{T}}$ for some $c > 0$. One can also assume a general Gaussian prior on active wavelet coefficients with an unstructured covariance matrix $\Sigma_{\mathcal{T}}$ which satisfies $\lambda_{\min}(\Sigma_{\mathcal{T}}) \gtrsim 1/\sqrt{\log n}$ and $\lambda_{\max}(\Sigma_{\mathcal{T}}) \lesssim n^a$ for some $a > 0$. Detailed proofs can be found in the Supplementary Material (Section S-4.1).

Only very few priors (actually *only* point mass spike-and-slab based priors, as discussed in the Introduction) were shown to attain adaptive posterior sup-norm concentration rates. Theorem 1 now certifies Dyadic Bayesian CART as one of them. The logarithmic penalty in the rate (25) reflects that Bayesian CART priors occupy the middle ground between flat trees (with only a depth cutoff) and spike-and-slab priors (with general sparsity patterns). As mentioned earlier, flat trees are incapable of supremum-norm adaptation, as we formally prove in Section 3.4. The fact that the more flexible Bayesian CART priors still achieves supremum-norm adaptation in a near-optimal way is a rather notable feature. From a more general perspective, we note that while general tools are available to derive adaptive L^2- or Hellinger-rate results in broad settings (e.g., model selection techniques, or the theory of posterior rates in [32]), deriving adaptive L^∞-results is often obtained in a case-by-case basis; two possible techniques are wavelet thresholding (when empirical estimates of wavelet coefficients are available) and Lepski's method (which requires some 'ordered' set of estimators, typically in terms of variance; for tree-estimators for instance it would not readily be applicable). The fact that tree methods enable for supremum-norm adaptation in nonparametric settings is one of the main take-away messages of this work.

3.2. *Adaptive honest confidence bands for f_0.* We now turn to the ultimate landing point of this paper, uncertainty quantification for f_0 and its functionals. The existence of adaptive confidence sets in general is an interesting and delicate question (see Chapter 8 of [35]). In the present context of regression function estimation under the supremum norm loss, it is in fact impossible to build adaptive confidence bands without further restricting the parameter space. We do so by imposing some classical self-similarity conditions (see [35, 49] for more details).

DEFINITION 3 (Self-similarity). Given an integer $j_0 > 0$, we say that $f \in \mathcal{H}(\alpha, M)$ is *self-similar* if, for some constant $\varepsilon > 0$,

$$\|K_j(f) - f\|_\infty \geq \varepsilon 2^{-j\alpha} \quad \text{for all } j \geq j_0, \tag{28}$$

where $K_j(f) = \sum_{l\leq j-1}\sum_k \langle \psi_{lk}, f\rangle \psi_{lk}$ is the wavelet projection at level j. The class of all such self-similar functions will be denoted by $\mathcal{H}_{\mathrm{SS}}(\alpha, M, \varepsilon)$.

Section 8.3.3 in [35] describes self-similar functions as typical representatives of the Hölder class. As shown in Proposition 8.3.21 of [35], self-dissimilar functions are nowhere dense in the sense that they cannot approximate any open set in $\mathcal{H}(\alpha, M)$. In addition,

Bayesian nonparametric priors for Hölder functions charge self-similar functions with probability 1. Finally, self-similarity does not affect the difficulty of the statistical estimation problem, where the (ℓ_∞) minimax rate is not changed after adding this assumption. A variant of the self-similarity condition was shown to be *necessary* for adaptive inference, in that such condition cannot essentially be weakened for uniform coverage with an optimal rate to hold [11].

Following [49], we construct adaptive honest credible sets by first defining a pivot centering estimator, and then determining a data-driven radius.

DEFINITION 4 (The median tree). Given a posterior $\Pi_{\mathbb{T}}[\cdot \mid X]$ over trees, we define the *median tree* $\mathcal{T}_X^* = \mathcal{T}^*(\Pi_{\mathbb{T}}[\cdot \mid X])$ as the set of nodes

$$\mathcal{T}_X^* = \big\{(l,k), l \le L_{\max}, \Pi\big[(l,k) \in \mathcal{T}_{\text{int}} \mid X\big] \ge 1/2\big\}. \tag{29}$$

Similarly, as in the median probability model [3, 4], a node belongs to $\mathcal{T}_X^*$ if its (marginal) posterior probability to be selected by a tree estimator exceeds $1/2$. Interestingly, as the terminology suggests, $\mathcal{T}_X^*$ *is an actual tree*, that is, the nodes follow hereditary constraints (see Lemma S-10 in the Supplementary Material). We define the resulting median tree estimator as

$$\widehat{f}_T(x) = \sum_{(l,k)\in\mathcal{T}_X^*} X_{lk}\psi_{lk}(x). \tag{30}$$

Moreover, we define a *radius*, for some $v_n \to \infty$ to be chosen, as

$$\sigma_n = \sigma_n(X) = \sup_{x\in[0,1]} \sum_{l=0}^{L_{\max}} v_n \sqrt{\frac{\log n}{n}} \sum_{k=0}^{2^l-1} \mathbb{I}_{(l,k)\in\mathcal{T}_X^*} |\psi_{lk}(x)|. \tag{31}$$

A credible band with a radius $\sigma_n(X)$ as in (31) and a center $\widehat{f}_T$ as in (30) is

$$\mathcal{C}_n = \big\{f : \|f - \widehat{f}_T\|_\infty \le \sigma_n(X)\big\}. \tag{32}$$

Theorem 3, proved in Section S-4.4, shows that valid frequentist uncertainty quantification with Bayesian CART is attainable (up to log factors). Indeed, the confidence band (32) has a near-optimal diameter and a uniform frequentist coverage under self-similarity.

THEOREM 3. *Let $0 < \alpha_1 \le \alpha_2 \le 1$, $M \ge 1$ and $\varepsilon > 0$. Let Π be any prior as in the statement of Theorem 1. Let σ_n be as in (31) with v_n such that $(\log n)^{1/2} = o(v_n)$ and let $\widehat{f}_T$ denote the median tree estimator (30). Then for $\mathcal{C}_n$ defined in (32), uniformly over $\alpha \in [\alpha_1, \alpha_2]$, as $n \to \infty$,*

$$\inf_{f_0\in\mathcal{H}_{\text{SS}}(\alpha,M,\varepsilon)} P_{f_0}(f_0 \in \mathcal{C}_n) \to 1.$$

For every $\alpha \in [\alpha_1, \alpha_2]$ and uniformly over $f_0 \in \mathcal{H}_{\text{SS}}(\alpha, M, \varepsilon)$, the diameter $|\mathcal{C}_n|_\infty = \sup_{f,g\in\mathcal{C}_n} \|f - g\|_\infty$ and the credibility of the band verify, as $n \to \infty$,

$$|\mathcal{C}_n|_\infty = O_{P_{f_0}}\big((n/\log n)^{-\alpha/(2\alpha+1)} v_n\big), \tag{33}$$

$$\Pi[\mathcal{C}_n \mid X] = 1 + o_{P_{f_0}}(1). \tag{34}$$

Similarly as for Theorem 1, the results of Theorem 3 carry over to wavelet priors over a smooth wavelet basis, leading to the construction of confidence sets with arbitrary regularities $0 < \alpha_1 \le \alpha_2 < \infty$. The undersmoothing factor v_n is commonplace in the context of confidence bands, with the condition $v_n \gg (\log n)^{1/2}$ reflecting the slight logarithmic price

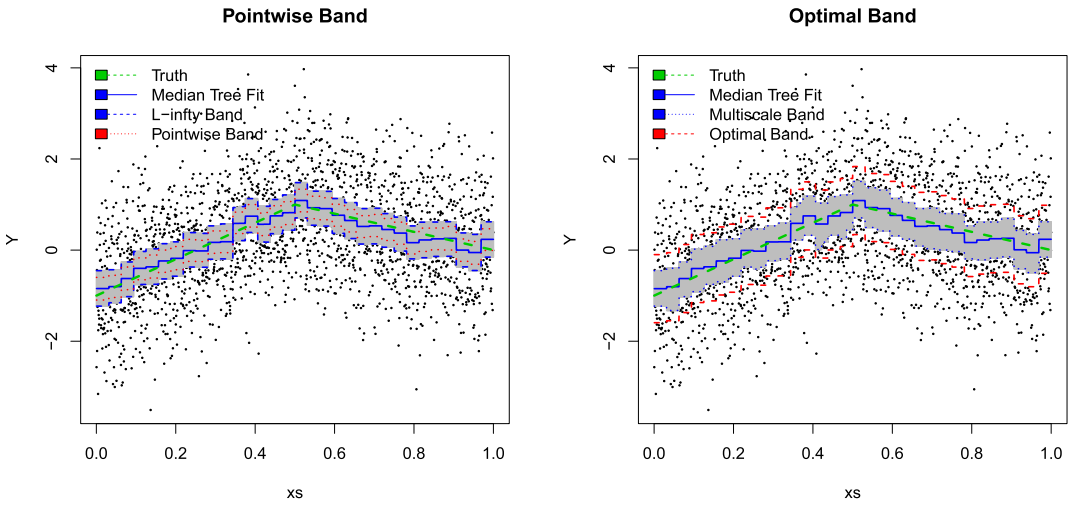


FIG. 4. (*Left*) *Pointwise* 0.95% *credible intervals together with a* 95% L^∞*-credible band* (*gray area*). (*Right*) *Not-intersected multiscale* 0.95% *credible band* (*S*-14) (*gray area*) *using* $w_l = l^{1/2+0.01}$ (*see Supplementary Material*, *Section S*-1.3.5) *together with the 'optimal' set* (32) *obtained with* $v_n = 1$. *The true function is* $f_0(x) = (4x-1)\mathbb{I}(x \le 1/2) + (-2x+2)\mathbb{I}(x > 1/2)$.

to pay for trees noted earlier in terms of L^∞-estimation accuracy. In the previous statement both confidence and credibility of $\mathcal{C}_n$ tend to 1. It is possible to achieve exact coverage by intersecting $\mathcal{C}_n$ further with another ball. A natural way to do so (from the 'estimating many functionals' perspective, see [18]) is to intersect with a multiscale ball (we refer to Sections S-1.3 and S-1.2 in the Supplementary Material for details and demonstrations). For stability reasons, this intersection-band seems also preferable in practice and we present in Figure 4 on the right an illustration of coverage of such a band in nonparametric regression. Apart from the intersection band, another natural choice is an L^∞-credible band. Namely, given a centering estimator $\hat{f}$ (such as the median-tree estimator), one can consider an L^∞-ball around $\hat{f}$ that captures 0.95% of the posterior mass (see Figure 4 on the left). We are not aware of any frequentist validation results for such bands in the adaptive L^∞-setting. Results for such type of credible sets have been obtained in the L^2-setting, for instance, in [55]. To guarantee coverage, the authors need to incorporate a 'blow-up' factor (diverging to infinity) to the radius of the set (see [49] for more discussion). Finally, another possibility would be to 'paste together' marginal pointwise credible intervals (see Figure 4 on the left). It is not clear how much 'blow-up' would be needed to guarantee frequentist coverage under self-similarity and, again, we are not aware of any theoretical results for such sets.

3.3. *Inference for functionals of* f_0: *Bernstein–von Mises theorems*. By slightly modifying the Bayesian CART prior on the coarsest scales, it is possible to obtain asymptotic normality results, in the form of Bernstein-von Mises theorems, that imply that posterior quantile-credible sets are optimal-size confidence sets. In the next result, β_S denotes the bounded-Lipschitz metric on the metric space S (see also the Supplementary Material Section S-1.3).

THEOREM 4. *Assume the Bayesian CART priors* $\Pi_{\mathbb{T}}$ *from Theorem* 1 *constrained to trees that fit* $j_0(n)$ *layers, that is,* $\gamma_{lk} = 1$ *for* $l \le j_0(n)$, *for* $j_0(n) \asymp \sqrt{\log n}$.

1. BvM for smooth functionals $\psi_b(f) := \langle f, b\rangle$. *Let* $b \in L^\infty[0,1]$ *with coefficients* ($b_{lk} = \langle b, \psi_{lk}\rangle$). *Assume* $\sum_k |b_{lk}| \le c_l$ *for all* $l \ge 1$ *with* $\sum_l l^2 c_l < \infty$. *Then, in* P_{f_0}*-probability,*

$$\beta_{\mathbb{R}}\big(\mathcal{L}\big(\sqrt{n}\big(\psi_b(f) - \hat{\psi}_b\big) \mid X\big), \mathcal{L}\big(\mathcal{N}\big(0, \|b\|_2^2\big)\big)\big) \to 0.$$

2. Functional BvM for the primitive $F(\cdot) = \int_0^{\cdot} f$. *Let* $(G(t) : t \in [0,1])$ *be a Brownian motion. Then, in* P_{f_0}*-probability,*

$$\beta_{C([0,1])}\left(\mathcal{L}\left(\sqrt{n}\left(F(\cdot) - \int_0^{\cdot} dX^{(n)} \,\middle|\, X\right)\right), \mathcal{L}(G)\right) \to 0$$

As a consequence of this result, quantile credible sets for the considered functionals are optimal confidence sets. For $\alpha \in (0,1)$, let $q^{\psi_b}_{\alpha/2}(X)$ and $q^{\psi_b}_{1-\alpha/2}(X)$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the induced posterior distribution on the functional $\psi_b = \int_0^1 f(u)b(u)\,du$ and set $I_b(X) := [q^{\psi_b}_{\alpha/2}(X), q^{\psi_b}_{1-\alpha/2}(X)]$. Theorem 4 (part 1) then implies (see [18] for a proof) that

$$P_{f_0}\big[\psi_b(f_0) \in I_b(X)\big] \to 1-\alpha.$$

Similarly, let $R_n(X)$ be the data-dependent radius chosen from the induced posterior distribution on $F(\cdot) = \int_0^{\cdot} f$ as follows, for $\hat{F}(\cdot) = \int_0^{\cdot} dX^{(n)}$,

$$\Pi\big[\|F - \hat{F}\|_\infty \le R_n(X) \mid X\big] = 1-\alpha. \tag{35}$$

Consider the band $\mathcal{C}^F(X) := \{F : \|F - \hat{F}\|_\infty \le R_n(X)\}$. Then Theorem 4 (part 2) implies (see [18], Corollary 2 for a related statement and proof), for $F_0(\cdot) = \int_0^{\cdot} f_0$,

$$P_{f_0}\big[F_0 \in \mathcal{C}^F(X)\big] \to 1-\alpha.$$

In other words, the band (35) has exact asymptotic coverage. It can also be checked that it is optimal efficient in semiparametric terms (that is, its width is optimal asymptotically). We derive Theorem 4 as a consequence of an adaptive nonparametric BvM (Theorem S-3 in the Supplementary Material; see Section S-4.5 for a proof, where other possible choices for $j_0(n)$ are discussed), only obtained so far for *adaptive* priors in the work of Ray [49], which considered (conjugate) spike and slab priors. Derivation of the band (35) in practice is easily obtained once posterior samples are available. Theorem 4 is illustrated, in the regression framework studied in Section S-1.1, on a numerical example with a piece-wise linear regression function (details on the implementation are in Section S-1.2) in Figure 5. The left panel presents a histogram of posterior samples (together with 2.5% and 97.5% quantiles) of the rescaled primitive functional $\widetilde{F}(x) = nF(x) = \sum_{t_i \le x} f(t_i)$ for $x = 0.8$ with true value is marked with a red solid line. The right panel portrays the confidence band (35) which uniformly captures the true functional (dotted line).

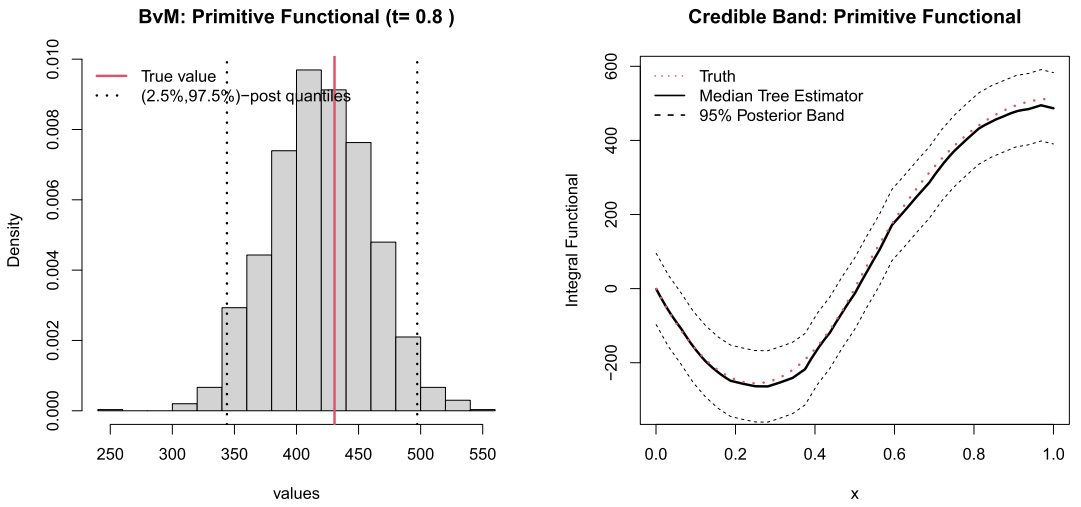


FIG. 5. *(Left) 0.95% credible interval for the (rescaled) primitive functional* $\widetilde{F}(x)$ *with* $x = 0.8$*; (Right) the confidence band* (35) *obtained for* $f_0(x) = (4x-1)\mathbb{I}(x \le 1/2) + (-2x+2)\mathbb{I}(x > 1/2)$.

3.4. *Lower bound*: *Flat trees are* (*grossly*) *suboptimal for the* $\|\cdot\|_\infty$*-loss*. Recall that the spike-and-slab prior achieves the *actual* ℓ_∞-minimax rate *without* any additional factor. Interestingly, the very same prior misses the ℓ_2-minimax rate by a log factor [39]. This illustrates that ℓ_2 and ℓ_∞ adaptations require different desiderata when constructing priors. Product priors that correspond to separable rules *do not* yield adaptation with exact rates in the ℓ_2 sense [12]. Mixture priors that are adaptive in ℓ_2, on the other hand, may not yield ℓ_∞ adaptation. We now provide one example of this phenomenon in the context of flat (complete binary) trees.

The *flat* tree of depth $d = d(\mathcal{T})$ is the binary tree which contains all possible nodes until level d, that is, $\gamma_{lk} = \mathbb{I}_{l<d}$. An example of a flat tree with $d = 3$ layers is in Figure 2. The simplest possible prior on tree topologies (confined to symmetric trees) is just the Dirac mass at a given flat tree of fixed depth $d = D$; an adaptive version thereof puts a prior D and samples from the set of all flat trees. Such priors coincide with so-called *sieve* priors, where the sieve spans the expansion basis (e.g., Haar) up to level D. Flat dyadic trees only keep Haar wavelet coefficients at resolutions smaller than some $d > 0$ (i.e., $\gamma_{lk} = 0$ for $l \geq d$). The implied prior on $(\beta_{lk})_{lk}$ can be written as, with $\pi(\beta_{lk}) \propto \sigma_l^{-1}\phi(\beta_{lk}/\sigma_l)$,

$$(\beta_{lk}) \mid d \sim \bigotimes_{l<d,k} \pi(\beta_{lk}) \otimes \bigotimes_{l\geq d,k} \delta_0(\beta_{lk}), \tag{36}$$

where $\phi(\cdot)$ is some bounded density that is strictly positive on $\mathbb{R}$ and σ_l are fixed positive scalars. The sequence (σ_l) is customarily chosen so as it decays with the resolution index l, for example, $\sigma_l = 2^{-l(\beta+1/2)}$ for some $0 < \beta \leq \alpha$. This "undersmoothing" prior requires the knowledge of (a lower bound on) α and yields a *nonadaptive* nonparametric BvM behavior [18].

A tempting strategy to manufacture adaptation is to treat the threshold d as random through a prior $\pi(d)$ on integers (and take constant σ_l), which corresponds to the hierarchical prior on regular regression histograms [51, 56]. It is not hard to check that the flat-tree prior (36) with random d has a marginal mixture distribution similar to the one of the spike-and-slab prior on each coordinate (l, k). Despite marginally similar, the probabilistic structure of these two priors is very different. Zeroing out signals internally, the spike-and-slab prior (10) *is* ℓ_∞*-adaptive* [39]. The flat tree prior (36), on the other hand, fits a few dense layers *without* internal sparsity and *is* ℓ_2*-adaptive* (up to a log term) [56]. However, as shown in the following theorem, flat trees fall short of ℓ_∞-adaptation.

THEOREM 5. *Assume the flat tree prior* (36) *with random* d, *where* $\pi(d)$ *is nonincreasing and where the active wavelet coefficients* β_{lk} *are Gaussian i.i.d.* $\mathcal{N}(0, 1)$. *Moreover, assume* $\{\psi_{lk}\}$ *is an* S*-regular wavelet basis for some* $S \geq 1$. *For any* $0 < \alpha \leq S$ *and* $M > 0$, *there exists* $f_0 \in \mathcal{H}(\alpha, M)$ *such that*

$$E_{f_0}\Pi\big[\ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}}, f_0) < \zeta_n \mid X\big] \to 0,$$

where the lower-bound rate ζ_n *is given by* $\zeta_n = (\frac{\log n}{n})^{\frac{\alpha}{2\alpha+2}}$.

Theorem 5, proved in Section S-4.6, can be applied to standard priors $\pi(d)$ with exponential decrease, proportional to e^{-d} or $\mathrm{e}^{-d\log d}$, or to a uniform prior over $\{1, \ldots, L_{\max}\}$. In [1], a negative result is also derived for sieve-type priors, but only for the posterior mean and for Sobolev classes instead of the, here arguably more natural, Hölder classes for supremum losses (which leads to different rates for estimating the functional-at-a-point). Here, we show that when the target is the ℓ_∞-loss for Hölder classes the sieve-prior is severely suboptimal.

3.5. *Nonparametric regression*: *Overview of results*. Our results obtained under the white noise model can be transported to the more practical nonparametric regression model. While these two models are asymptotically equivalent [10] (under uniform smoothness assumptions satisfied, for example, by α-Hölderian functions with $\alpha > 1/2$), it is not automatic that the knowledge of a (wavelet shrinkage/nonlinear) minimax procedure in one model implies the optimality in the other. It turns out, however, that our results *can* be carried over to fixed-design regression without necessarily assuming $\alpha > 1/2$. We assume outcomes $Y = (Y_1, \ldots, Y_n)'$ arising from

$$Y_i = f_0(t_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), i = 1, \ldots, n = 2^{L_{\max}+1} \tag{37}$$

where f_0 is an unknown regression function and $\{t_i \in [0,1] : 1 \le i \le n\}$ are fixed design points. For simplicity, we consider a regular grid, that is, $t_i = i/n$ for $1 \le i \le n$ and assume n is a power of 2. In Section S-1.1, we show that most results for Bayesian CART obtained earlier in white noise carry over to the model (37) with a few minor changes. One minor modification concerns the loss function. We mainly consider the 'canonical' supremum-norm loss for the fixed design setting, that is, the 'max-norm' defined for given functions f, g by

$$\|f - g\|_{\infty,n} = \max_{1 \le i \le n} |f(t_i) - g(t_i)|,$$

but it is also possible to consider the whole supremum-norm loss $\|\cdot\|_\infty$. We postpone statements and proofs to the Supplementary Material, Sections S-1.1 and S-6.1. In a numerical study (Section S-1.2), we illustrate that the implementation of Bayesian CART [20, 25] and the construction of our confidence bands is rather straightforward. For example, Figure 4 shows how inference can be carried out with Bayesian CART posteriors in nonparametric regression with a piece-wise linear regression function using the intersecting band construction (detailed in Section S-1.3.5). Contrary to point-wise credible intervals (on the left) that are easy to produce but do not cover, our multiscale confidence band (on the right) uniformly captures the true regression function. More details on this example are presented in Section S-1.2.

4. Nondyadic Bayesian CART. A limitation of midpoint splits in dyadic trees is that they treat the basis as fixed, allowing the jumps to occur *only* at pre-specified dyadic locations even when not justified by data. General CART regression methodology [9, 31] avoids this restriction by treating the basis as *unknown*, where the partitioning cells shrink and stretch with data. In this section, we leave behind 'static' dyadic trees to focus on the analysis of Bayesian (nondyadic) CART [20, 25] and its connection to Unbalanced Haar (UH) wavelet basis selection.

4.1. *Unbalanced Haar wavelets*. UH wavelet basis functions [36] are *not* necessarily translates/dilates of any mother wavelet function and, as such, allow for different support lengths and design-adapted split locations. Here, we particularize the constructive definition of UH wavelets given by [30]. Assume that possible values for splits are chosen from a set of $n = 2^{L_{\max}}$ breakpoints $\mathcal{X} = \{x_i : x_i = i/n, 1 \le i \le n\}$. Using the scale/location index enumeration, pairs (l,k) in the tree are now equipped with (a) a *breakpoint* $b_{lk} \in \mathcal{X}$ and (b) *left and right brackets* $(l_{lk}, r_{lk}) \in \mathcal{X} \cup \{0,1\}$. Unlike balanced Haar wavelets (3), where $b_{lk} = (2k+1)/2^{l+1}$, the breakpoints b_{lk} are *not required* to be regularly dyadically constrained and are chosen from $\mathcal{X}$ in a hierarchical fashion as follows. One starts by setting $l_{00} = 0$, $r_{00} = 1$. Then:

(a) The first breakpoint b_{00} is selected from $\mathcal{X} \cap (0,1)$.

(b) For each $1 \le l \le L_{\max}$ and $0 \le k < 2^l$, set

$$
\begin{aligned}
l_{lk} &= l_{(l-1)\lfloor k/2 \rfloor}, \qquad r_{lk} = b_{(l-1)\lfloor k/2 \rfloor} \quad \text{if } k \text{ is even,} \\
l_{lk} &= b_{(l-1)\lfloor k/2 \rfloor}, \qquad r_{lk} = r_{(l-1)\lfloor k/2 \rfloor} \quad \text{if } k \text{ is odd.}
\end{aligned} \tag{38}
$$

If $\mathcal{X} \cap (l_{lk}, r_{lk}] \neq \varnothing$, choose b_{lk} from $\mathcal{X} \cap (l_{lk}, r_{lk}]$.

Let A denote the set of *admissible* nodes (l, k), in that (l, k) is such that $\mathcal{X} \cap (l_{lk}, r_{lk}] \neq \varnothing$, obtained through an instance of the sampling process described above and let

$$B = (b_{lk})_{(l,k) \in A}$$

be the corresponding set of breakpoints. Each collection of split locations B gives rise to nested intervals

$$L_{lk} = (l_{lk}, b_{lk}] \quad \text{and} \quad R_{lk} = (b_{lk}, r_{lk}].$$

Starting with the mother wavelet $\psi^B_{-10} = \psi_{-10} = \mathbb{I}_{(0,1)}$, one then recursively constructs wavelet functions ψ^B_{lk} with a support $I^B_{lk} = L_{lk} \cup R_{lk}$ as

$$\psi^B_{lk}(x) = \frac{1}{\sqrt{|L_{lk}|^{-1} + |R_{lk}|^{-1}}} \left(\frac{\mathbb{I}_{L_{lk}}(x)}{|L_{lk}|} - \frac{\mathbb{I}_{R_{lk}}(x)}{|R_{lk}|} \right). \tag{39}$$

By construction, the system $\Psi^B_A = \{\psi^B_{-10}, \psi^B_{lk} : (l, k) \in A\}$ is orthonormal in $L^2[0, 1]$.With UH wavelets, the decay of wavelet coefficients $\beta_{lk} = \langle f, \psi^B_{lk} \rangle$ for a α-Hölder function f verifies $|\beta^B_{lk}| \lesssim \max\{|L_{lk}|, |R_{lk}|\}^{\alpha+1/2}$, see Lemma S-6. [30] points out that the computational complexity of the discrete UH transform could be unnecessarily large and imposes the balancing requirement $\max\{|L_{lk}|, |R_{lk}|\} \le E(|L_{lk}| + |R_{lk}|)\ \forall (l, k) \in A$, for some $1/2 \le E < 1$. Similarly, in order to control the combinatorial complexity of the basis system, we require that the UH wavelets are *weakly balanced* in the following sense.

DEFINITION 5. A system $\Psi^B_A = \{\psi^B_{-10}, \psi^B_{lk} : (l, k) \in A\}$ of UH wavelets is *weakly balanced* with balancing constants $E, D \in \mathbb{N}^*$ if, for any $(l, k) \in A$,

$$\max\big(|L_{lk}|, |R_{lk}|\big) = \frac{M_{lk}}{2^{l+D}} \quad \text{for some } M_{lk} \in \{1, \ldots, E + l\}. \tag{40}$$

Note that in the actual BART implementation, the splits are chosen from sample quantiles to ensure balancedness (similar to our condition (40)). Quantile splits (Example 2 below) are a natural way to generate many weakly balanced systems, providing a much increased flexibility compared to dyadic splits, which correspond to uniform quantiles. Other examples together with a graphical depiction of the unbalanced Haar wavelets for certain nondyadic choices of split points b_{lk} are in the Supplementary Material (Figure S-3 in Section S-3).

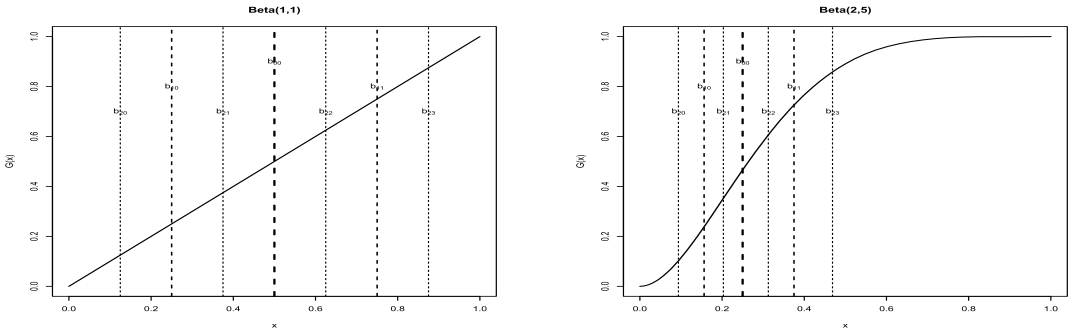
EXAMPLE 2 (Quantile splits). Denote with G a c.d.f with a density g on $[0, 1]$ that satisfies $\|g\|_\infty \le 2^{D-1}/(2E)$ for $E, D > 0$ chosen below and $\|1/g\|_\infty \le C_q$ for some $C_q > 0$. Let us define a dyadic projection of G as

$$G_l^{-1}(x) := 2^{-l} \lfloor 2^l G^{-1}(x) \rfloor,$$

and next define the breakpoints, for $l \le L_{\max}$ and $0 \le k < 2^l$, as

$$b_{lk} = G^{-1}_{L_{\max}+D}\big[(2k + 1)/2^{l+1}\big]. \tag{41}$$

The system Ψ^B_A obtained from steps (a) and (b) with splits (41) is weakly balanced for $E = 2 + 3C_q 2^{D-1}$. This is verified in Lemma S-9 in the Supplementary Material (Section S-3.4). Moreover, Figure 6 in illustrates the implementation of the quantile system, where splits are placed more densely in areas where $G(x)$ changes more rapidly.



(a) $g(x) \sim \mathcal{B}(1,1)$ (b) $g(x) \sim \mathcal{B}(2,5)$

FIG. 6. *Example of quantile splits for a uniform density $g(x)$ and a nonuniform beta density $g(x)$ using $L_{\max} = 6$.*

The *nondyadic Bayesian CART* prior is then defined as follows:

- *Step 1 (Basis Generation)*. Sample $B = (b_{lk})_{0 \le k < 2^l - 1, l \le L}$ from $\Pi_{\mathbb{B}}$ by following the steps (a)–(b) around (38) subject to satisfying the *balancing condition* (40).
- *Step 2 (Tree Generation)*. Independently of B, sample a binary tree $\mathcal{T}$ from one of the priors $\Pi_{\mathbb{T}}$ described in Section 2.1.
- *Step 3 (Step Heights Generation)*. Given $\mathcal{T}$, we obtain the coefficients (β_{lk}^B) from the tree-shaped prior (19). Using the UH wavelets, the prior on the internal coefficients β_{lk}^B can be translated into a model on the histogram heights $\widetilde{\beta}_{lk}^B$ through (8).

An example of such a prior is obtained by first randomly drawing quantiles (e.g., by drawing a density at random verifying conditions as in Example 2) to generate the breakpoints for Step 1 and then following the construction from Section 2 for Steps 2–3. The following theorem is proved in Section S-5.

THEOREM 6. *Let $\Pi_{\mathbb{B}}$ be any prior on breakpoint collections that satisfy weak balancedness according to Definition 5. Let $\Pi_{\mathbb{T}}$ be the Galton-Watson process prior from Section 2.1 with $p_{lk} = \Gamma^{-l^4}$. Consider the tree-shaped wavelet prior (19) with $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0, I_{|\mathcal{T}_{\text{ext}}|})$. Let $f_0 \in \mathcal{H}_M^\alpha$ as in (24) for some $M > 0$ and $0 < \alpha \le 1$ and define*

$$\varepsilon_n = (\log n)^{1+\frac{3}{2}} \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}. \tag{42}$$

Then, there exist $\Gamma_0, c_0 > 0$ depending only on the constants E, D in the weak balancedness condition such that, for any $\Gamma \ge \Gamma_0$ and $c \ge c_0$, for any $M_n \to \infty$, we have, for $n \to \infty$

$$E_{f_0} \Pi\big[\ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}}, f_0) \ge \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > M_n \varepsilon_n \mid X\big] \to 0. \tag{43}$$

In the context of piecewise constant priors, Theorem 6 allows further flexibility in the choice of the prior as compared to Theorem 1 in that the location of the breakpoints, on the top of their structure given by the tree prior, can vary in their location according to its own specific prior. Whether one can further weaken the balancing condition to still get optimal multiscale results is an interesting open question that goes beyond the scope of this paper. In addition, the log-factor in (42) could be further optimized, similarly as in Theorem 1.

5. Discussion. In this paper, we explored connections between Bayesian tree-based regression methods and structured wavelet shrinkage. We demonstrated that Bayesian tree-based methods attain (almost) optimal convergence rates in the supremum norm and obtain

limiting results for functionals, that follow from a nonparametric and adaptive Bernstein–von Mises theorem. The developed framework also allows us to construct adaptive credible bands around f_0 under self-similarity. To allow for nondyadically organized splits, we introduced weakly balanced Haar wavelets (an elaboration on unbalanced Haar wavelets of [36]) and showed that Bayesian CART performs basis selection from this library and attains a near-minimax rate of posterior concentration under the sup-norm loss.

Although for clarity of exposition we focused on the white noise model, our results can be extended to the more practical regression model for fixed regular designs (Section S-1.1 in the Supplementary Material) or possibly more general designs under some conditions. We note that the techniques of proof are nonconjugate in their key tree aspect, which opens the door to applications in many other statistical settings. A version of Bayesian CART for density estimation following the ideas of the present work is currently investigated by T. Randrianarisoa as part of his Ph.D. thesis. More precisely, using the present techniques, it is possible to develop multiscale rate results for Pólya trees with 'optional stopping' along a tree, in the spirit of [60]. Our confidence set construction can be also shown to have local adaptation properties. The ability of Bayesian CART to spatially adapt in this way will be investigated in a followup work. Further natural extensions include high-dimensional versions of the model, extending the multi-dimensional version briefly presented here, as well as forest priors. These will be considered elsewhere.

6. Proof of Theorem 1. The proof proceeds in three steps. In Section 6.1, we first show that the posterior concentrates on not too deep trees. In Section 6.2, we then show that the posterior probability of missing signal vanishes and, finally, in Section 6.3 we show that the posterior distribution concentrates around signals. To better convey main ideas, we present the proof for the independent prior $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$ with $\Sigma_{\mathcal{T}} = I_K$ for $K = |\mathcal{T}_{\mathrm{ext}}|$ and the Galton-Watson (GW) tree prior from Section 2.1.1 with a split probability p_l. The proof for the g-prior $\Sigma_{\mathcal{T}} = g_n(A_{\mathcal{T}}'A_{\mathcal{T}})^{-1}$ is more technically involved and is presented in Section S-4.1 in the Supplementary Material.

We will be working conditionally on the event

$$\mathcal{A} = \Big\{ \max_{-1 \le l \le L, 0 \le k < 2^l} \varepsilon_{lk}^2 \le 2\log\big(2^{L+1}\big) \Big\}, \tag{44}$$

where $L = L_{\max} = \lfloor \log_2 n \rfloor$. Since $\varepsilon_{lk} \sim \mathcal{N}(0,1)$, this event has a large probability in the sense that $P(\mathcal{A}^c) \lesssim (\log n)^{-1}$, which follows from $P[\max_{1 \le i \le N} |Z_i| > \sqrt{2\log N}] \le c_0/\sqrt{\log N}$ for some $c_0 > 0$ when $Z_i \sim \mathcal{N}(0,1)$ for $1 \le i \le N$.

6.1. *Posterior probability of deep trees.* The first step is to show that, on the event $\mathcal{A}$, the posterior concentrates on reasonably small trees, that is, trees whose depth $d(\mathcal{T})$ is no larger than an 'optimal' depth which depends on the unknown smoothness α. Let us define such a depth $\mathcal{L}_c = \mathcal{L}_c(\alpha, M)$ as

$$\mathcal{L}_c = \left\lceil \log_2 \left((8M)^{\frac{1}{\alpha+1/2}} \left(\frac{n}{\log n} \right)^{\frac{1}{2\alpha+1}} \right) \right\rceil. \tag{45}$$

LEMMA 1. *Under the assumptions of Theorem 1, on the event $\mathcal{A}$,*

$$\Pi\big[d(\mathcal{T}) > \mathcal{L}_c \mid X \big] \to 0 \quad (n \to \infty). \tag{46}$$

PROOF. Consider one tree $\mathcal{T} \in \mathbb{T}$ such that $d(\mathcal{T}) \ge 1$ and denote with $\mathcal{T}^-$ a pruned subtree obtained from $\mathcal{T}$ by turning its deepest rightmost internal node, say (l_1, k_1), into a terminal node. Then $\mathcal{T}^- = \mathcal{T}_{\mathrm{int}}^- \cup \mathcal{T}_{\mathrm{ext}}^-$, where

$$\mathcal{T}_{\mathrm{int}}^- = \mathcal{T}_{\mathrm{int}} \setminus \{(l_1,k_1)\}, \qquad \mathcal{T}_{\mathrm{ext}}^- = \mathcal{T}_{\mathrm{ext}} \setminus \{(l_1+1, 2k_1), (l_1+1, 2k_1+1)\} \cup \{(l_1,k_1)\}.$$

Note that $\mathcal{T}^-$ is a full binary tree and that the mapping $\mathcal{T} \to \mathcal{T}^-$ is not necessarily injective. Indeed, there are up to $2^{d(\mathcal{T}^-)}$ trees $\mathcal{T}$ that give rise to the same pruned tree $\mathcal{T}^-$. Let $\mathbb{T}_d = \{\mathcal{T} \in \mathbb{T} : d(\mathcal{T}) = d\}$ denote the set of all full binary trees of depth *exactly* $d \geq 1$. Then, using the notation (22),

$$\Pi[\mathbb{T}_d \,|\, X] = \frac{\sum_{\mathcal{T}\in\mathbb{T}_d} W_X(\mathcal{T})}{\sum_{\mathcal{T}\in\mathbb{T}} W_X(\mathcal{T})} = \frac{\sum_{\mathcal{T}\in\mathbb{T}_d} \frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} W_X(\mathcal{T}^-)}{\sum_{\mathcal{T}\in\mathbb{T}} W_X(\mathcal{T})}$$

$$\text{where } \frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \frac{\int \prod_{(l,k)\in\mathcal{T}'_{\text{int}}} \mathrm{e}^{nX_{lk}\beta_{lk} - n\beta_{lk}^2/2}\, d\pi(\boldsymbol{\beta}_{\mathcal{T}})}{\int \prod_{(l,k)\in\mathcal{T}'^-_{\text{int}}} \mathrm{e}^{nX_{lk}\beta_{lk} - n\beta_{lk}^2/2}\, d\pi(\boldsymbol{\beta}_{\mathcal{T}^-})}. \tag{47}$$

Let $\boldsymbol{X}_{\mathcal{T}} = (X_{lk} : (l,k) \in \mathcal{T}'_{\text{int}})'$ and $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}'_{\text{int}})'$ be the top-down left-to-right ordered sequences (recall that we order nodes according to the index $2^l + k$). Assuming $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$, and denoting $K = |\mathcal{T}_{\text{ext}}| = |\mathcal{T}_{\text{int}}| + 1$,

$$\begin{aligned}\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} &= \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \sqrt{\frac{|\Sigma_{\mathcal{T}^-}|}{2\pi\,|\Sigma_{\mathcal{T}}|}} \frac{\int \mathrm{e}^{n\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}} - \boldsymbol{\beta}'_{\mathcal{T}}[nI_K + \Sigma_{\mathcal{T}}^{-1}]\boldsymbol{\beta}_{\mathcal{T}}/2}\, d\boldsymbol{\beta}_{\mathcal{T}}}{\int \mathrm{e}^{n\boldsymbol{X}'_{\mathcal{T}^-}\boldsymbol{\beta}_{\mathcal{T}^-} - \boldsymbol{\beta}'_{\mathcal{T}^-}[nI_{K-1} + \Sigma_{\mathcal{T}^-}^{-1}]\boldsymbol{\beta}_{\mathcal{T}^-}/2}\, d\boldsymbol{\beta}_{\mathcal{T}^-}} \\ &= \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \sqrt{\frac{|\Sigma_{\mathcal{T}^-}|}{|\Sigma_{\mathcal{T}}|}} \sqrt{\frac{|nI_{K-1} + \Sigma_{\mathcal{T}^-}^{-1}|}{|nI_K + \Sigma_{\mathcal{T}}^{-1}|}} \frac{\mathrm{e}^{n^2 \boldsymbol{X}'_{\mathcal{T}}(nI_K + \Sigma_{\mathcal{T}}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}}/2}}{\mathrm{e}^{n^2 \boldsymbol{X}'_{\mathcal{T}^-}(nI_{K-1} + \Sigma_{\mathcal{T}^-}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}^-}/2}}.\end{aligned} \tag{48}$$

Since $X_{l_1k_1}$ corresponds to the node (l,k) with the highest index $2^l + k$, one can write $\boldsymbol{X}_{\mathcal{T}} = (\boldsymbol{X}_{\mathcal{T}^-}, X_{l_1k_1})'$.

We focus on the GW prior from Section 2.1.1 and on the independent prior $\Sigma_{\mathcal{T}} = I_K$ and present proofs for the remaining priors in Section S-4.1. Using the expression (48) and since (l_1, k_1) is the deepest rightmost internal node in $\mathcal{T}$, and $\mathcal{T}$ is of depth $d = d(\mathcal{T}) = l_1 + 1$, using the definition of the GW prior,

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \prod_{(l,k)\in\mathcal{T}'_{\text{int}}\setminus\mathcal{T}'^-_{\text{int}}} \frac{\mathrm{e}^{\frac{n^2}{2(n+1)}X_{lk}^2}}{\sqrt{n+1}} = \frac{p_{d-1}(1-p_d)^2}{1-p_{d-1}} \frac{\mathrm{e}^{\frac{n^2}{2(n+1)}X_{l_1k_1}^2}}{\sqrt{n+1}}.$$

Suppose $\mathcal{T}$ has depth $d(\mathcal{T}) > \mathcal{L}_c$. Then $l_1 \geq \mathcal{L}_c$ and from the Hölder continuity (23), one gets $8|\beta_{l_1k_1}| \leq \sqrt{\log n/n}$, where $\mathcal{L}_c$ is as in (45). Then, conditionally on the event (44),

$$|X_{l_1k_1}| \leq \frac{1}{\sqrt{n}}\left[\frac{1}{8}\sqrt{\log n} + \sqrt{2\log n + \log 4}\right] \tag{49}$$

and thereby $2X_{l_1k_1}^2 \leq 5\log n/n$. Recall that, under the GW-prior, the split probability is $p_d = \Gamma^{-d}$. As $\Gamma > 2$, one has $p_d < 1/2$ and so, for any $d > \mathcal{L}_c$,

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} \leq 2p_{d-1} \exp\left(\frac{5n\log n}{4(n+1)} - \frac{1}{2}\log(1+n)\right) < 2n^{3/4} p_{d-1}.$$

Going back to the ratio (47), we now bound, with $a(n,d) =: 2n^{3/4}p_{d-1}$,

$$\frac{\Pi[\mathbb{T}_d \,|\, X]}{a(n,d)} \leq \frac{\sum_{\mathcal{T}\in\mathbb{T}_d} W_X(\mathcal{T}^-)}{\sum_{\mathcal{T}\in\mathbb{T}} W_X(\mathcal{T})} \leq \frac{\sum_{\mathcal{T}\in\mathbb{T}_d^-} 2^{d(\mathcal{T}^-)} W_X(\mathcal{T})}{\sum_{\mathcal{T}\in\mathbb{T}} W_X(\mathcal{T})} \leq 2^d,$$

where $\mathbb{T}_d^-$ is the image of $\mathbb{T}_d$ under the map $\mathcal{T} \to \mathcal{T}^-$, and using that at most $2^{d(\mathcal{T}^-)}$ trees are mapped to the same $\mathcal{T}^-$. Using this bound one deduces that, on the event $\mathcal{A}$, with $L =$

$L_{\max} = \log_2 n$,

$$\Pi\big[d(\mathcal{T}) > \mathcal{L}_c \mid X\big] = \sum_{d=\mathcal{L}_c+1}^{L} \Pi[\mathbb{T}_d \mid X] \le 4n^{3/4} \sum_{d=\mathcal{L}_c+1}^{L} 2^{d-1} p_{d-1}$$
$$< 4n^{3/4} L \exp\big[-\mathcal{L}_c \log(\Gamma/2)\big].$$

As $\mathcal{L}_c \asymp (\log n)/(1+2\alpha)$, the right-hand side goes to zero as soon as, for example, $\log(\Gamma/2) > 7(1+2\alpha)/8$ that is, for $\alpha \le 1$, $\Gamma > 2e^3$. □

6.2. *Posterior probability of missing signal.* The next step is showing that the posterior probability of missing a node with large enough signal vanishes.

LEMMA 2. *Let us denote, for $A > 0$ to be chosen suitably large,*

$$S(f_0; A) = \left\{(l,k) : |\beta_{lk}^0| \ge A \frac{\log n}{\sqrt{n}}\right\}. \tag{50}$$

Under the assumptions of Theorem 1, on the event $\mathcal{A}$ from (44),

$$\Pi\big[\{\mathcal{T} : S(f_0; A) \not\subseteq \mathcal{T}\} \mid X\big] \to 0 (n \to \infty). \tag{51}$$

PROOF. As before, we present the proof with the GW prior from Section 2.1.1 and for the independent prior with $\Sigma_{\mathcal{T}} = I_K$, referring to Section S-4.1 for the g-prior. Let us first consider a given node $(l_S, k_S) \in S(f_0; A)$, for A to be specified below, and note that the Hölder condition on f_0 implies $l_S \le \mathcal{L}_c$ (for n large enough). Let $\mathbb{T}_{\setminus(l_S,k_S)} = \{\mathcal{T} \in \mathbb{T} : (l_S, k_S) \notin \mathcal{T}_{\text{int}}\}$ denote the set of trees that miss the signal node in the sense that they *do not have a cut* at (l_S, k_S). For any such tree $\mathcal{T} \in \mathbb{T}_{\setminus(l_S,k_S)}$ we then denote by $\mathcal{T}^+$ the smallest full binary tree (in terms of the number of nodes) that contains $\mathcal{T}$ and that splits on (l_S, k_S). Such a tree can be constructed from $\mathcal{T} \in \mathbb{T}_{\setminus(l_S,k_S)}$ as follows. Denote by $(l_0, k_0) \in \mathcal{T}_{\text{ext}} \cap [(0,0) \leftrightarrow (l_S, k_S)]$ the external node of $\mathcal{T}$ which is *closest* to (l_S, k_S) on the route from the root to (l_S, k_S) in a flat tree (denoted by $[(0,0) \leftrightarrow (l_S, k_S)]$). Next, denote by $\mathcal{T}^+$ the extended tree obtained from $\mathcal{T}$ by sequentially splitting all $(l,k) \in [(l_0,k_0) \leftrightarrow (l_S,k_S)]$. Similarly as for $\mathcal{T} \to \mathcal{T}^-$ above, the map $\mathcal{T} \to \mathcal{T}^+$ is not injective and we denote by $\mathbb{T}_{(l_S,k_S)}$ the set of all extended trees $\mathcal{T}^+$ obtained from some $\mathcal{T} \in \mathbb{T}_{\setminus(l_S,k_S)}$. Now, the posterior probability $\Pi[\mathbb{T}_{\setminus(l_S,k_S)} \mid X]$ of missing the signal node (l_S, k_S) equals

$$\frac{\sum_{\mathcal{T} \in \mathbb{T}_{\setminus(l_S,k_S)}} W_X(\mathcal{T})}{\sum_{\mathcal{T} \in \mathbb{T}} W_X(\mathcal{T})} \le \frac{\sum_{\mathcal{T} \in \mathbb{T}_{\setminus(l_S,k_S)}} \frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} W_X(\mathcal{T}^+)}{\sum_{\mathcal{T} \in \mathbb{T}_{(l_S,k_S)}} W_X(\mathcal{T})}. \tag{52}$$

Let us denote by $\mathcal{T}^{(j)}$ for $j = -1, \dots, s$ the sequence of nested trees obtained by extending one branch of $\mathcal{T}$ towards (l_S, k_S) by splitting the nodes $[(l_0,k_0) \leftrightarrow (l_S,k_S)]$, where $\mathcal{T}^+ = \mathcal{T}^{(s)}$ and $\mathcal{T} = \mathcal{T}^{(-1)}$. Then

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} \prod_{j=0}^{s} \frac{N_X(\mathcal{T}^{(j-1)})}{N_X(\mathcal{T}^{(j)})}. \tag{53}$$

Under the GW process prior with $p_l = \Gamma^{-l}$ for some $\Gamma > 2$, the ratio of prior tree probabilities in the last expression satisfies

$$\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} = \frac{1-p_{l_0}}{p_{l_0}} \times \left(\prod_{l=l_0+1}^{l_S} \frac{1}{p_l(1-p_l)}\right) \times \frac{1}{(1-p_{l_S+1})^2}. \tag{54}$$

The first term is due to the fact that $\mathcal{T}^+$ splits the node (l_0, k_0) while $\mathcal{T}$ does not. The second term in the denominator is the extra prior probability of $\mathcal{T}^+$ over $\mathcal{T}$ that is due to the branch reaching out to (l_S, k_S). Along this branch (note that this is the smallest possible branch), one splits *only* one daughter node for each layer l (thereby the term p_l) and not the other (thereby the term $1 - p_l$). The third term above is due to the fact that the two daughters of (l_S, k_S) are not split. The quantity (54) is bounded by $2^{l_S - l_0 + 2}\Gamma^{(l_0+l_S)(l_S-l_0+1)/2} < 4\Gamma^{2l_S^2}$.

Assuming $\Sigma_{\mathcal{T}} = I_K$, we can write for any $\mathcal{T}$ in $\mathbb{T}_{\setminus(l_S,k_S)}$

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} \prod_{(l,k)\in\mathcal{T}^+\setminus\mathcal{T}} \frac{\sqrt{n+1}}{\mathrm{e}^{\frac{n^2}{2(n+1)}X_{lk}^2}}. \tag{55}$$

Using the definition of the model and the inequality $2ab \geq -a^2/2 - 2b^2$ for $a, b \in \mathbb{R}$, we obtain $X_{l_S k_S}^2 \geq (\beta_{l_S k_S}^0)^2/2 - \varepsilon_{l_S k_S}^2/n$. On the event $\mathcal{A}$, one gets

$$\exp\left\{-\frac{n^2}{2(n+1)}X_{l_S k_S}^2\right\} \leq \exp\left\{-\frac{n^2(\beta_{l_S k_S}^0)^2}{4(n+1)} + \frac{n(\log 2)(\log_2 n + 1)}{n+1}\right\}.$$

The term in (55) can be thus bounded, for any $\mathcal{T} \in \mathbb{T}_{\setminus(l_S,k_S)}$, by

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} \leq C\Gamma^{2l_S^2} \exp\left\{\frac{3(l_S - l_0 + 1)(\log_2 n + 1)}{2} - \frac{nA^2\log^2 n}{4(n+1)}\right\} =: b(n, l_S).$$

We now continue to bound the ratio (52). For each given $\mathcal{T}^+$, there are *at most* l_S trees $\widetilde{\mathcal{T}} \in \mathbb{T}_{\setminus(l_S,k_S)}$ which have the same extended tree $\widetilde{\mathcal{T}}^+ = \mathcal{T}^+$. This is because $\mathcal{T}^+$ is obtained by extending one given branch by adding no more than l_S nodes. Using this fact, (52), and the definition of $b(n, l_S)$ on the last display,

$$\frac{\Pi[\mathbb{T}_{\setminus(l_S,k_S)} \mid X]}{b(n, l_S)} \leq \frac{\sum_{\mathcal{T}\in\mathbb{T}_{\setminus(l_S,k_S)}} W_X(\mathcal{T}^+)}{\sum_{\mathcal{T}\in\mathbb{T}_{(l_S,k_S)}} W_X(\mathcal{T})} \leq l_S \frac{\sum_{\mathcal{T}\in\mathbb{T}_{(l_S,k_S)}} W_X(\mathcal{T})}{\sum_{\mathcal{T}\in\mathbb{T}_{(l_S,k_S)}} W_X(\mathcal{T})}.$$

By choosing $A = A(\Gamma) > 0$ large enough, this leads to

$$\Pi[\mathbb{T}_{\setminus(l_S,k_S)} \mid X] \lesssim \mathrm{e}^{(3/2+3\log\Gamma)(\log_2 n+1)^2 - \frac{A^2}{8}\log^2 n} \lesssim \mathrm{e}^{-\frac{A^2}{16}\log^2 n}.$$

Then the result follows as, on the event $\mathcal{A}$,

$$\sum_{(l_S,k_S)\in S(f_0,A)} \Pi[\mathbb{T}_{\setminus(l_S,k_S)} \mid X] \lesssim 2^{\mathcal{L}_c+1}\mathrm{e}^{-\frac{A^2}{16}\log^2 n} \lesssim \mathrm{e}^{-\frac{A^2}{32}\log^2 n} \to 0.$$

□

6.3. *Posterior concentration around signals.* Let us now show that the posterior does not distort large signals too much.

LEMMA 3. *Let us denote, for $\mathcal{L}_c$ as in (45) and $S(f_0; A)$ as in (50),*

$$\mathsf{T} = \{\mathcal{T} : d(\mathcal{T}) \leq \mathcal{L}_c, S(f_0; A) \subset \mathcal{T}\}. \tag{56}$$

Then, on the event $\mathcal{A}$, for some $C' > 0$, uniformly over $\mathcal{T} \in \mathsf{T}$,

$$\int \max_{(l,k)\in\mathcal{T}'_{\mathrm{int}}} |\beta_{lk} - \beta_{lk}^0| \, d\Pi[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}] < C'\sqrt{\frac{\log n}{n}}, \tag{57}$$

with $\boldsymbol{X}_{\mathcal{T}} = (X_{lk} : (l,k) \in \mathcal{T}'_{\mathrm{int}})'$ the ordered vector of active responses.

PROOF. For a given tree $\mathcal{T}$ with $K = |\mathcal{T}_{\text{ext}}|$ leaves, we denote by $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}'_{\text{int}})'$ the vector of wavelet (internal node) coefficients, with $\boldsymbol{X}_{\mathcal{T}}$ the corresponding responses and with $\boldsymbol{\varepsilon}_{\mathcal{T}}$ the white noise disturbances. It follows from (21) that, given $\boldsymbol{X}_{\mathcal{T}}$ (so for fixed ε_{lk}) and $\mathcal{T}$, the vector $\boldsymbol{\beta}_{\mathcal{T}}$ has a Gaussian distribution $\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{T}}, \widetilde{\Sigma}_{\mathcal{T}})$, where $\widetilde{\Sigma}_{\mathcal{T}} = (nI_K + \Sigma_{\mathcal{T}}^{-1})^{-1}$ and $\boldsymbol{\mu}_{\mathcal{T}} = n\widetilde{\Sigma}_{\mathcal{T}}(\boldsymbol{\beta}^0_{\mathcal{T}} + \frac{1}{\sqrt{n}}\boldsymbol{\varepsilon}_{\mathcal{T}})$. Next, using Lemma S-4, we have

$$\mathbb{E}\big[\|\boldsymbol{\beta}_{\mathcal{T}} - \boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty \mid \boldsymbol{X}_{\mathcal{T}}\big] \le \|\boldsymbol{\mu}_{\mathcal{T}} - \boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty + \sqrt{2\bar{\sigma}^2 \log K} + 2\sqrt{2\pi\bar{\sigma}^2}, \tag{58}$$

where $\bar{\sigma}^2 = \max \operatorname{diag}(\widetilde{\Sigma}_{\mathcal{T}})$. Focusing on the first term, we can write

$$\|\boldsymbol{\mu}_{\mathcal{T}} - \boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty \le \sqrt{n}\|\widetilde{\Sigma}_{\mathcal{T}}\boldsymbol{\varepsilon}_{\mathcal{T}}\|_\infty + \|(n\widetilde{\Sigma}_{\mathcal{T}} - I_K)\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty. \tag{59}$$

Using the fact $(I + B)^{-1} = I - (I + B^{-1})^{-1}$, we obtain $n\widetilde{\Sigma}_{\mathcal{T}} - I_K = -(I_K + n\Sigma_{\mathcal{T}})^{-1}$. From now on, we focus on the simpler case $\Sigma_{\mathcal{T}} = I_K$ and refer to Section S-4.1.3 (Supplementary Material) for the proof for the g-prior. With $\Sigma_{\mathcal{T}} = I_K$ we can write $\|(n\widetilde{\Sigma}_{\mathcal{T}} - I_K)\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty = \frac{\|\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty}{1+n} < C/n$. Using the fact that $\|\boldsymbol{\varepsilon}_{\mathcal{T}}\|_\infty \lesssim \sqrt{\log n}$ on the event $\mathcal{A}$, we obtain $\sqrt{n}\|\widetilde{\Sigma}_{\mathcal{T}}\boldsymbol{\varepsilon}_{\mathcal{T}}\|_\infty \lesssim \sqrt{\frac{\log n}{n}}$. The sum of the remaining two terms in (58) can be bounded by a multiple of $\sqrt{\log n/n}$ by noting that $\bar{\sigma}^2 = 1/(n+1)$. The statement (57) then follows from (58). □

6.4. *Supremum-norm convergence rate.* Let us write $f_0 = f_0^{\mathcal{L}_c} + f_0^{\setminus \mathcal{L}_c}$, where $f_0^{\mathcal{L}_c}$ is the L^2-projection of f_0 onto the first $\mathcal{L}_c$ layers of wavelet coefficients. Under the Hölder condition the equality holds pointwise and $\|f_0^{\setminus \mathcal{L}_c}\|_\infty \le \sum_{l > \mathcal{L}_c} 2^{l/2} 2^{-l(1/2+\alpha)} \lesssim (\log n/n)^{\alpha/(2\alpha+1)}$.

The following inequality bounds the supremum norm by the ℓ_∞-norm:

$$\begin{aligned}\|f - f_0\|_\infty &\le \sum_{l \ge -1} \max_{0 \le k < 2^l} |\beta_{lk} - \beta^0_{lk}| \cdot \Big\| \sum_{0 \le k < 2^{-l}} |\psi_{lk}| \Big\|_\infty \\ &\le |\langle f - f_0, \varphi\rangle| + \sum_{l \ge 0} 2^{l/2} \max_{0 \le k < 2^l} |\beta_{lk} - \beta^0_{lk}| = \ell_\infty(f, f_0).\end{aligned} \tag{60}$$

We use the notation $S(f_0; A)$, T as in (50) and (56) and

$$\mathcal{E} = \{f_{\mathcal{T},\boldsymbol{\beta}} : \mathcal{T} \in \mathsf{T}\}. \tag{61}$$

Using the definition of the event $\mathcal{A}$ from (44), one can write

$$\begin{aligned}&E_{f_0}\Pi\big[f_{\mathcal{T},\boldsymbol{\beta}} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \mid X\big] \\ &\quad \le P_{f_0}[\mathcal{A}^c] + E_{f_0}\Pi[\mathcal{E}^c \mid X] + E_{f_0}\big\{\Pi\big[f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{E} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \mid X\big]\mathbb{I}_{\mathcal{A}}\big\}.\end{aligned} \tag{62}$$

By Markov's inequality and the previous bound (60),

$$\begin{aligned}&\Pi\big[f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{E} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \mid X\big]\mathbb{I}_{\mathcal{A}} \\ &\quad \le \varepsilon_n^{-1} \int_{\mathcal{E}} \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty \, d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \mid X]\mathbb{I}_{\mathcal{A}} \\ &\quad \le \varepsilon_n^{-1} \sum_{l \le \mathcal{L}_c} 2^{l/2} \Big\{ \int_{\mathcal{E}} \max_{0 \le k < 2^l} |\beta_{lk} - \beta^0_{lk}| \, d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \mid X]\mathbb{I}_{\mathcal{A}} \Big\} + \varepsilon_n^{-1} \|f_0^{\setminus \mathcal{L}_c}\|_\infty.\end{aligned}$$

With T as in (56), the integral in the last display can be written, for $l \le \mathcal{L}_c$,

$$\begin{aligned}&\int_{\mathcal{E}} \max_{0 \le k < 2^l} |\beta_{lk} - \beta^0_{lk}| \, d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \mid X] \\ &\quad = \sum_{\mathcal{T} \in \mathsf{T}} \pi[\mathcal{T} \mid X] \int \max_{0 \le k < 2^l} |\beta_{lk} - \beta^0_{lk}| \, d\Pi[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\mathcal{T}\in\mathsf{T}} \pi[\mathcal{T}\mid X] \int \max\Big(\max_{0\le k<2^l,(l,k)\notin\mathcal{T}'_{\mathrm{int}}} |\beta^0_{lk}|, \max_{0\le k<2^l,(l,k)\in\mathcal{T}'_{\mathrm{int}}} |\beta_{lk}-\beta^0_{lk}| \Big)\, d\Pi[\boldsymbol{\beta}_{\mathcal{T}}\mid \boldsymbol{X}_{\mathcal{T}}] \\
&\le \min\Big(\max_{0\le k<2^l} |\beta^0_{lk}|, A\frac{\log n}{\sqrt{n}} \Big) \\
&\quad + \sum_{\mathcal{T}\in\mathsf{T}} \pi[\mathcal{T}\mid X] \int \max_{0\le k<2^l,(l,k)\in\mathcal{T}'_{\mathrm{int}}} |\beta_{lk}-\beta^0_{lk}|\, d\Pi[\boldsymbol{\beta}_{\mathcal{T}}\mid \boldsymbol{X}_{\mathcal{T}}],
\end{aligned}
$$

where we have used that on the set $\mathcal{E}$, selected trees cannot miss any true signal larger than $A\log n/\sqrt{n}$. This means that any node (l,k) that is *not* in a selected tree must satisfy $|\beta^0_{lk}| \le A\log n/\sqrt{n}$.

Let $L^* = L^*(\alpha)$ be the integer closest to the solution of the equation in L given by $M2^{-L(\alpha+1/2)} = A\log n/\sqrt{n}$. Then, using that $f_0 \in \mathcal{H}(\alpha, M)$,

$$
\begin{aligned}
\sum_{l\le\mathcal{L}_c} 2^{\frac{l}{2}} \min\Big(\max_{0\le k<2^l} |\beta^0_{lk}|, A\frac{\log n}{\sqrt{n}} \Big) &\le \sum_{l\le L^*} 2^{\frac{l}{2}} A\frac{\log n}{\sqrt{n}} + \sum_{L^*<l\le\mathcal{L}_c} 2^{\frac{l}{2}} M 2^{-l(\frac{1}{2}+\alpha)} \\
&\le C 2^{L^*/2} A\frac{\log n}{\sqrt{n}} + C 2^{-L^*\alpha} \\
&\le \widetilde{C} 2^{-L^*\alpha} \le c\big(n^{-1}\log^2 n\big)^{\frac{\alpha}{2\alpha+1}}.
\end{aligned} \tag{63}
$$

Using $P_{f_0}[\mathcal{A}^c] + E_{f_0}\Pi[\mathcal{E}^c \mid X] = o(1)$ and Lemma 3, one obtains

$$
\begin{aligned}
&E_{f_0}\Pi\big[f_{\mathcal{T},\boldsymbol{\beta}} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \mid X\big] \\
&\le o(1) + \varepsilon_n^{-1} \sum_{l\le\mathcal{L}_c} 2^{l/2} \Big[\min\Big(\max_{0\le k<2^l} |\beta^0_{lk}|, A\frac{\log n}{\sqrt{n}} \Big) + C'\sqrt{\frac{\log n}{n}} \Big] + \varepsilon_n^{-1} \big\| f_0^{\setminus\mathcal{L}_c} \big\|_\infty \\
&\le o(1) + \varepsilon_n^{-1} \Big[c\Big(\frac{\log^2 n}{n}\Big)^{\frac{\alpha}{2\alpha+1}} + 2C'\sqrt{\frac{2\mathcal{L}_c\log n}{n}} \Big] + \varepsilon_n^{-1} \big\| f_0^{\setminus\mathcal{L}_c} \big\|_\infty \\
&\le o(1) + \varepsilon_n^{-1} \big[c(\log n)^{\alpha/(2\alpha+1)} + 2C' \big] \Big(\frac{\log n}{n}\Big)^{\frac{\alpha}{2\alpha+1}} + \varepsilon_n^{-1} \big\| f_0^{\setminus\mathcal{L}_c} \big\|_\infty
\end{aligned}
$$

for some $C' > 0$. Choosing $\varepsilon_n = M_n((\log^2 n)/n)^{\frac{\alpha}{2\alpha+1}}$, the right-hand side goes to zero for any arbitrarily slowly increasing sequence $M_n \to \infty$.

Funding. The first author gratefully acknowledges support from the Institut Universitaire de France and from the ANR Grant ANR-17-CE40-0001 (BASICS).

The second author gratefully acknowledges support from the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business and the National Science Foundation (Grant DMS-1944740).

SUPPLEMENTARY MATERIAL

Supplement to "Uncertainty quantification for Bayesian CART" (DOI: 10.1214/21-AOS2093SUPP; .pdf). The supplement [19] contains additional material, including results for nonparametric regression, a simulation study, an adaptive nonparametric Bernstein–von Mises theorem, and details on tensor–multivariate versions of the considered prior distributions. It also contains all remaining proofs.

REFERENCES

[1] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697 https://doi.org/10.1002/sjos.12002

[2] BARANIUK, R. G., CEVHER, V., DUARTE, M. F. and HEGDE, C. (2010). Model-based compressive sensing. *IEEE Trans. Inf. Theory* **56** 1982–2001. MR2654489 https://doi.org/10.1109/TIT.2010.2040894

[3] BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

[4] BARBIERI, M. M., BERGER, J. O., GEORGE, E. I. and ROČKOVÁ, V. (2018). The median probability model and correlated variables. Preprint. Available at arXiv:1807.08336.

[5] BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. MR3015035 https://doi.org/10.1214/12-AOS1013

[6] BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. MR2930634

[7] BLANCHARD, G., SCHÄFER, C. and ROZENHOLC, Y. (2004). Oracle bounds and exact algorithm for dyadic classification trees. In *Learning Theory. Lecture Notes in Computer Science* **3120** 378–392. Springer, Berlin. MR2177922 https://doi.org/10.1007/978-3-540-27819-1_26

[8] BLEICH, J., KAPELNER, A., GEORGE, E. I. and JENSEN, S. T. (2014). Variable selection for BART: An application to gene regulation. *Ann. Appl. Stat.* **8** 1750–1781. MR3271352 https://doi.org/10.1214/14-AOAS755

[9] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees. Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[10] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958 https://doi.org/10.1214/aos/1032181159

[11] BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456 https://doi.org/10.1214/12-EJS720

[12] CAI, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *J. Multivariate Anal.* **99** 421–436. MR2396972 https://doi.org/10.1016/j.jmva.2006.11.010

[13] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287 https://doi.org/10.1214/08-EJS273

[14] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477 https://doi.org/10.1214/14-AOS1253

[15] CASTILLO, I. (2017). Pólya tree posterior distributions on densities. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** 2074–2102. MR3729648 https://doi.org/10.1214/16-AIHP784

[16] 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 https://doi.org/10.1214/20-aihp1132

[17] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856 https://doi.org/10.1214/13-AOS1133

[18] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246

[19] CASTILLO, I. and ROČKOVÁ, V. (2021). Supplement to "Uncertainty quantification for Bayesian CART." https://doi.org/10.1214/21-AOS2093SUPP

[20] CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (1997). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–960.

[21] CHIPMAN, H., GEORGE, E. I., MCCULLOCH, R. E. and SHIVELY, T. (2016). High-dimensional nonparametric monotone function estimation using BART. Preprint. Available at arXiv:1612.01619.

[22] CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

[23] CHIPMAN, H. A., KOLACZYK, E. D. and MCCULLOCH, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92** 1413–1421.

[24] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. MR1256527 https://doi.org/10.1006/acha.1993.1005

[25] DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. MR1649118 https://doi.org/10.1093/biomet/85.2.363

[26] DONOHO, D. L. (1997). CART and best-ortho-basis: A connection. *Ann. Statist.* **25** 1870–1911. MR1474073 https://doi.org/10.1214/aos/1069362377

[27] ENGEL, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **49** 242–254. MR1276437 https://doi.org/10.1006/jmva.1994.1024

[28] FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410 https://doi.org/10.1016/S0304-4076(00)00076-2

[29] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177 https://doi.org/10.1214/aos/1176325766

[30] FRYZLEWICZ, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Statist. Assoc.* **102** 1318–1327. MR2412552 https://doi.org/10.1198/016214507000000860

[31] GEY, S. and NÉDÉLEC, E. (2005). Model selection for CART regression trees. *IEEE Trans. Inf. Theory* **51** 658–670. MR2236074 https://doi.org/10.1109/TIT.2004.840903

[32] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[33] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[34] 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 https://doi.org/10.1214/11-AOS924

[35] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[36] GIRARDI, M. and SWELDENS, W. (1997). A new class of unbalanced Haar wavelets that form an unconditional basis for L_p on general measure spaces. *J. Fourier Anal. Appl.* **3** 457–474. MR1468375 https://doi.org/10.1007/BF02649107

[37] HAHN, P. R., MURRAY, J. S. and CARVALHO, C. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at arXiv:1706.09523.

[38] HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

[39] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341

[40] KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934. MR1354008

[41] LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. MR3832214 https://doi.org/10.1080/01621459.2016.1264957

[42] LINERO, A. R. and YANG, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1087–1110. MR3874311 https://doi.org/10.1111/rssb.12293

[43] LIU, Y., ROČKOVÁ, V. and WANG, Y. (2018). ABC variable selection with Bayesian forests. Preprint. Available at arXiv:1806.02304.

[44] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26, 41 pp. MR3491120

[45] MURRAY, J. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. Preprint. Available at arXiv:1706.09523.

[46] NAULET, Z. (2018). Adaptive Bayesian density estimation in sup-norm. Preprint. Available at arXiv:1805.05816.

[47] NICKL, R. and RAY, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.* **48** 1383–1408. MR4124327 https://doi.org/10.1214/19-AOS1851

[48] NICKL, R. and SÖHL, J. (2019). Bernstein–von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat.* **13** 3513–3571. MR4013745 https://doi.org/10.1214/19-ejs1609

[49] RAY, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** 2511–2536. MR3737900 https://doi.org/10.1214/16-AOS1533

[50] ROČKOVÁ, V. and SAHA, E. (2019). On theory for BART. In *Proceedings of Machine Learning Research*: 22*nd International Conference on Artificial Intelligence and Statistics* **89** 2839–2848.

[51] ROČKOVÁ, V. and VAN DER PAS, S. (2020). Posterior concentration for Bayesian regression trees and forests. *Ann. Statist.* **48** 2108–2131. MR4134788 https://doi.org/10.1214/19-AOS1879

[52] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. MR3357876 https://doi.org/10.1214/15-AOS1321

[53] SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inf. Theory* **52** 1335–1353. MR2241192 https://doi.org/10.1109/TIT.2006.871056

[54] SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.* **9** 475–520. MR3217004 https://doi.org/10.1214/14-BA863

[55] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861 https://doi.org/10.1214/14-AOS1270

[56] VAN DER PAS, S. and ROČKOVÁ, V. (2017). Bayesian dyadic trees and histograms for regression. In *Proceedings of the* 31*st International Conference on Neural Information Processing Systems* 2086–2096.

[57] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

[58] WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. Preprint. Available at arXiv:1503.06388.

[59] WILLETT, R. M. and NOWAK, R. D. (2007). Multiscale Poisson intensity and density estimation. *IEEE Trans. Inf. Theory* **53** 3171–3187. MR2417680 https://doi.org/10.1109/TIT.2007.903139

[60] WONG, W. H. and MA, L. (2010). Optional Pólya tree and Bayesian inference. *Ann. Statist.* **38** 1433–1459. MR2662348 https://doi.org/10.1214/09-AOS755

[61] YOO, W., RIVOIRARD, R. and ROUSSEAU, J. (2017). Adaptive supremum norm posterior contraction: Wavelet spike-and-slab and anisotropic Besov spaces. Preprint. Available at arXiv:1708.01909.

[62] YOO, W. and VAN DER VAART, A. (2017). The Bayes Lepski's method and credible bands through volume of tubular neighborhoods. Preprint. Available at arXiv:1711.06926.

[63] YOO, W. W. and GHOSAL, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.* **44** 1069–1102. MR3485954 https://doi.org/10.1214/15-AOS1398

[64] ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques. Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437