WORKING PAPER · NO. 2019-127

Multiscale Analysis of Bayesian CART

Ismael Castillo and Veronika Rockova

OCTOBER 2019

Becker Friedman Institute | Big Data Initiative

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Multiscale Analysis of Bayesian CART

Ismaël Castillo[*] **and Veronika Ročková** [†]

Sorbonne Université
Laboratoire de Probabilités, Statistique et Modélisation, LPSM
4, Place Jussieu, 75252 Paris cedex 05, France
e-mail: ismael.castillo@upmc.fr

University of Chicago
Booth School of Business, 5807 S. Woodlawn Avenue
Chicago, IL, 60637, USA
e-mail: veronika.rockova@chicagobooth.edu

Abstract: This paper affords new insights about Bayesian CART in the context of *structured* wavelet shrinkage. We show that practically used Bayesian CART priors lead to adaptive rate-minimax posterior concentration in the supremum norm in Gaussian white noise, performing optimally up to a logarithmic factor. To further explore the benefits of structured shrinkage, we propose the *g-prior* for trees, which departs from the typical wavelet product priors by harnessing correlation induced by the tree topology. Building on supremum norm adaptation, an adaptive non-parametric Bernstein–von Mises theorem for Bayesian CART is derived using multiscale techniques. For the fundamental goal of uncertainty quantification, we construct *adaptive* confidence bands with uniform coverage for the regression function under self-similarity.

MSC 2010 subject classifications: Primary 62G20, 62G15.
Keywords and phrases: Bayesian CART, Posterior Concentration, Nonparametric Bernstein–von Mises theorem, Recursive Partitioning, Regression Trees.

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.

This work sheds light on Bayesian CART [20, 23] which is a popular learning tool based on ideas of recursive partitioning and which forms an integral constituent of BART [19]. 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 non-parametric regression [19], classification [48], variable selection [7, 47, 45], shape constrained inference [18],

[*] The author gratefully acknowledges support from the Institut Universitaire de France and from the ANR grant ANR-17-CE40-0001 (BASICS).

[†] The author gratefully acknowledges support from the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

causal inference [40, 39], to name a few. While theory for random forests, the frequentist counterpart, has been developed to fruition [65, 5, 57, 43, 64], theory for BART has not kept pace with its application. With the first theoretical results (Hellinger convergence rates) emerging very recently [56, 46, 55], many fundamental questions pertaining to, e.g., *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 Bayesian CART using multiscale techniques. As a jumping off point, we first show that Bayesian CART reaches a (nearly-)optimal posterior convergence rate under the *supremum-norm* loss, a natural loss for UQ of smooth regression functions. We are actually not aware of any supremum-norm convergence rate result for CART methods in the literature. Second, we provide an *adaptive* non-parametric Bernstein-von Mises theorem to finally construct an *adaptive* credible band for the unknown regression function with (nearly, up a to logarithmic term) optimal uniform coverage under self-similarity. With these new results, our paper makes an important 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 [35, 13, 42] including specific models [59, 70, 51, 50]. However, Bayesian *adaptation* for the supremum loss has been obtained, to the best of our knowledge, *only* through spike-and-slab priors (the work [69] uses Gaussian process priors, but adaptation is obtained via the Lespki's method). In particular, [42] show that spike-and-slab priors on wavelet coefficients yield the *exact* adaptive minimax rate in the white noise model and [68] considers the anisotropic case in a regression framework. For density estimation, [14, 15] derive optimal $\|\cdot\|_\infty$–rates for Pólya tree priors, while [49] considers adaptation for log-density spike and slab priors. In this work, we consider Bayesian CART priors which are widely used practice.

Bayesian CART is a method of function estimation based on ideas of recursive partitioning of the predictor space. The work [25] highlighted the link between dyadic CART and best ortho-basis selection using Haar wavelets in two dimensions; [29] furthered this connection by considering unbalanced Haar wavelets of [37]. CART methods have been also studied in the machine learning literature, see e.g. [6, 58, 66] and references therein. We note that, unlike plain wavelet shrinkage methods and standard spike-and-slab priors, general Bayesian CART priors have extra flexibility by allowing for 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. The mathematical development throughout this paper is done under the Gaussian white noise model, which is an idealized version of non-parametric regression with fixed equispaced observations [9]. This model is defined through the following stochastic differential equation

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

where $X(t)$ is an observation process, $W(t)$ is the standard Wiener process on $[0,1]$ and $f_0 \in L^2[0,1]$ is an unknown bounded function on $[0,1]$ to be estimated. 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

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

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

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

obtained with orthonormal dilation-translations of $\psi = \mathbb{I}_{(0,1/2]} - \mathbb{I}_{(1/2,1]}$. Later in the text, we also consider weakly balanced Haar wavelet relaxations, as well as smooth wavelet bases.

Our paper makes contributions on both methodological and theoretical fronts. On the methodological side, we propose tree-shaped sparsity priors which exert local and global sparsity for wavelet shrinkage. In order to borrow strength between coefficients in the tree ancestry, we then propose a variant of the *g-prior* [71] for structured wavelet shrinkage. Similarly as independent product priors, we show that these dependent priors *also* lead to adaptive ℓ_∞ concentration rates (up to a log factor). To illustrate that local (internal) sparsity is a key driver of adaptivity, we will show that dense trees are incapable of adaptation.

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. Regarding the limiting shape, the Bernstein-von Mises (BvM) phenomenon is known to be non-trivial to formulate in growing or infinite dimensions (see [62], Chapter 10, for BvM formulations and consequences in parametric settings and, e.g., Freedman's negative result [28] outlining some issues in non-parametric settings). A number of positive results have nevertheless appeared in [31, 44, 16, 17]. In particular, [16, 17] formalized a programme to derive non-parametric BvM in spaces with weak topologies that admit $1/\sqrt{n}$-consistent estimation, see Section 3 for precise definitions. Recently, Ray [54] proved that this approach could incorporate adaption to the unknown regularity. In particular, he showed that the spike-and-slab prior [42] achieves an adaptive BvM property when the coarsest scales that capture the gross signal are not shrunk. We show that the widely used Bayesian CART prior *also* achieves the BvM property. The first consequence implied by our BvM is the derivation of confidence sets for a variety of smooth functionals. Further, for uncertainty quantification of f_0 itself, we construct adaptive and honest adaptive credible bands under self-similarity. Confidence bands construction for regression surfaces is a fundamental task in non-parametric regression and can indicate whether there is empirical evidence to support conjectured features such as multi-modality or exceedance of a level.

Although for clarity of exposition we focus on (white noise) regression, the introduced ideas can be applied to other settings as well, including density estimation. In the latter setting, Pólya trees have been a popular class of non-parametric prior distributions (see e.g. [33], Chapter 3 and [67, 14]). While their construction is unrelated to Bayesian CART, there are interesting connections (see comments in our Discussion).

The paper is structured as follows. Section 2 introduces regression trees-priors, as well as the notion of tree-shaped sparsity and the g-prior for trees. In Section 3 we prove multiscale properties of Bayesian dyadic CART. Section 4 considers non-dyadic partitioning allowing basis choice. A brief discussion can be found in Section 5. The proof of our master Theorem 1 can be found in Section 6, while the proofs of further results and technical lemmata can be found in the appendix Section 7.

Notation. Let $L^2[0,1]$ denote the set of square integrable functions on $[0,1]$ and by $\mathcal{C}([0,1])$ the set of continuous functions on $[0,1]$. Let $\phi_\sigma(x)$ denote the normal density with zero mean and variance σ^2. Let $\mathbb{N}=\{0,1,2,\ldots\}$ 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 and $A\vee B=\max(A,B)$. For an interval $I=(a,b]\subset(0,1]$, let $|I|=b-a$ be its diameter. The notation $x\lesssim y$ means $x\leq Cy$ for C a large enough universal constant.

2. Tree-based Prior Distributions

The multiscale setup enables one to assign a prior on $f\in L^2[0,1]$ directly through a prior on the sequence of its coefficients (β_{lk}). CART-based 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. 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 2.2). Later in Section 4, we will 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. Fleshing out connections between Bayesian histograms and wavelets in Section 2.3 and 2.4, we discuss several Bayesian CART priors over histogram heights in Section 2.5.

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 $\{\psi_{lk}\}$, which for now in this section is taken to be 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*

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

In the last display, the node (l,k) is a child of its parent node $(l-1,\lfloor\frac{k}{2}\rfloor)$. 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 l^{th} 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 $\mathcal{T}_{ext}$ of *external* nodes (also called *leaves*) of $\mathcal{T}$ if it has no children and to the set $\mathcal{T}_{int}$ of *internal* nodes, otherwise. By definition $|\mathcal{T}| = |\mathcal{T}_{int}|+|\mathcal{T}_{ext}|$, where, for full binary trees, we have $|\mathcal{T}| = 2|\mathcal{T}_{int}|+1$. 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 $L_{max} = \lfloor \log_2 n \rfloor$. The choice of full binary trees is traditional and mostly for simplicity of presentation, and L_{max} is a typical cut-off in wavelet analysis, as indeed trees can be associated with certain wavelet decompositions.

2.1. Priors on Trees

There are various ways of assigning a prior distribution over $\mathbb{T}$. We discuss two conventional Bayesian CART choices [20, 23], which became an integral component of many Bayesian tree regression methods including BART [19].

2.1.1. The Galton-Watson Process Prior (à la [20])

One of the earliest Bayesian CART proposals is due to [20], who suggest assigning a prior over $\mathbb{T}$ via the heterogeneous Galton-Watson (GW) process. We now provide an algorithmic description of this process (see also [55] for further discussion).

The prior description utilizes the following top-down left-to-right exploration metaphor. 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 $\mathcal{T} = \emptyset$, one initializes the exploration process by putting the root node $(0,0)$ tentatively in the queue, i.e. $Q = \{(0,0)\}$. One then repeats the following three steps until $Q = \emptyset$:

(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, i.e. $\mathcal{T} \leftarrow \mathcal{T} \cup \{(l,k)\}$,

   (ii) remove (l,k) from Q and if $l < L_{max}$ add its children to Q, i.e.

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

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 $\gamma > 0$, [55] point out that this decay may not be fast enough and suggest instead $p_{lk} = \Gamma^{-l}$ for some $2 < \Gamma < n$, which leads to a (near) optimal ℓ_2 convergence rate. We will use a similar assumption in our analysis. Each dyadic tree $\mathcal{T} \in \mathbb{T}$ can be uniquely identified by a collection of binary indicators $\Gamma(\mathcal{T}) = \{\gamma_{00}, \gamma_{10}, \dots, \gamma_{d(\mathcal{T}),2^{d(\mathcal{T})}-1}\}$ for whether or not each node (l,k) was split. We relate this representation to spike-and-slab wavelet shrinkage in Section 2.6.

2.1.2. The Bayesian CART Prior (à la [23])

Independently of [20], [23] proposed another variant of Bayesian CART, which first draws the number of leaves (i.e. external nodes) $K = |\mathcal{T}_{ext}|$ at random from a certain prior on integers, e.g. a Poisson distribution (say, conditioned to be non-zero). 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 K–th Catalan number (see Lemma 7), this leads to $\Pi(\mathcal{T}) = (\lambda^K/[K!(e^\lambda - 1)]) \cdot \mathbb{C}_{K-1}^{-1}$. As we restrict to trees in $\mathbb{T}$, i.e. with depth at most $L = L_{max}$, we slightly update the previous prior choice by setting, for some $\lambda > 0$, with $K = |\mathcal{T}_{ext}|$,

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

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

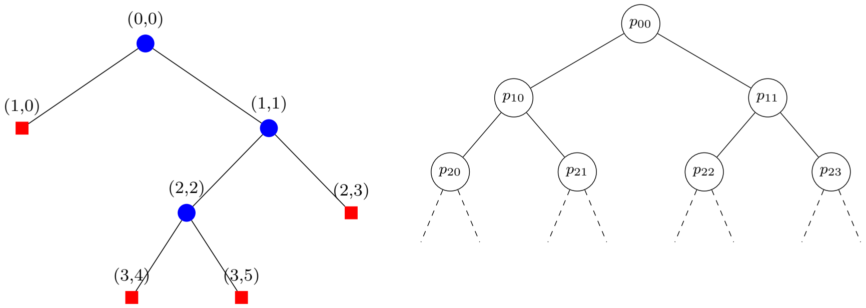
Another possibility, which can be just as easily implemented using Metropolis-Hasting strategies, is sampling the trees directly from a prior that penalizes larger trees

$$\Pi_{\mathbb{T}}(\mathcal{T}) \;\propto\; \mathrm{e}^{-c|\mathcal{T}_{ext}|\log n} \mathbb{I}_{\mathcal{T} \in \mathbb{T}}, \quad \text{for some } c > 0, \tag{6}$$

which we will refer to as the "exponential prior". The normalization is quite different as in the previous case, where smaller trees get higher probability.

2.1.3. Flat Trees

The *flat* tree of depth $d = d(\mathcal{T})$ is the binary tree which contains all possible nodes until level d, i.e. $\gamma_{lk} = \mathbb{I}_{l<d}$. An example of a flat tree with $d = 3$ layers is in



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

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

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.

2.2. 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 $I_{lk} = I_{l+1\,2k} \cup I_{l+1\,2k+1}$. One central example is *dyadic* intervals I_{lk} which correspond to the domain of the balanced Haar wavelets ψ_{lk} in (3), i.e.

$$I_{00} = (0, 1], \quad I_{lk} = (k2^{-l}, (k + 1)2^{-l}] \quad \text{for } l \geq 0 \text{ and } 0 \leq k < 2^l. \tag{7}$$

For any fixed depth $l \in \mathbb{N}$, the intervals $\cup_{0 \leq k < 2^l} I_{lk}$ form a deterministic regular (equispaced) partition of $(0, 1]$. Trees, however, generate *more flexible* partitions $\cup_{(l,k) \in \mathcal{T}_{ext}} I_{lk}$ 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 $\mathcal{T} = \mathcal{T}_{int} \cup \mathcal{T}_{ext}$, where $\mathcal{T}_{int} = \{(0, 0), (1, 1), (2, 2)\}$ and $\mathcal{T}_{ext} = \{(1, 0), (2, 3), (3, 4), (3, 5)\}$, resulting in the partition of $(0, 1]$ given by*

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

The set of possible *split points* obtained with (7) 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 $(l,k) \in \mathcal{T}_{int}$ are cut in half and intervals I_{lk} for each external node $(l,k) \in \mathcal{T}_{ext}$ 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.

2.3. Tree-shaped Wavelet Priors

We now introduce tree-based wavelet shrinkage priors as a more flexible alternative to sieve priors. Traditional (linear) Haar wavelet reconstructions deploy *all* wavelet coefficients β_{lk} with resolutions l smaller than some $d > 0$. This strategy amounts to fitting the *flat tree* with d layers (as the one on 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. For $\mathcal{T} \in \mathbb{T}$, let us denote by $\mathcal{T}'_{int} = \mathcal{T}_{int} \cup \{(-1,0)\}$ the 'rooted' tree where the index $(-1,0)$ is added to $\mathcal{T}_{int}$. Note that $|\mathcal{T}'_{int}| = |\mathcal{T}_{ext}|$. Given a full binary tree $\mathcal{T} \in \mathbb{T}$ and a vector $\boldsymbol{\beta} = (\beta_{-10}, (\beta_{lk})_{0 \le l \le L-1, 0 \le k < 2^l})$, one obtains a wavelet reconstruction

$$f_{\mathcal{T},\boldsymbol{\beta}}(x) = \beta_{-10}\psi_{-10}(x) + \sum_{(l,k)\in\mathcal{T}_{int}} \beta_{lk}\psi_{lk}(x) = \sum_{(l,k)\in\mathcal{T}'_{int}} \beta_{lk}\psi_{lk}(x). \qquad (9)$$

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

$$\begin{aligned} \mathcal{T} \quad &\sim \quad \Pi_{\mathbb{T}} \\ (\beta_{lk})_{lk} \,|\, \mathcal{T} \quad &\sim \bigotimes_{(l,k)\in\mathcal{T}'_{int}} \pi(\beta_{lk}) \;\otimes \bigotimes_{(l,k)\notin\mathcal{T}'_{int}} \delta_0(\beta_{lk}), \end{aligned} \qquad (10)$$

where $\Pi_{\mathbb{T}}$ is a prior on trees as described in Section 2 and where the active wavelet coefficients β_{lk} for $(l,k) \in \mathcal{T}_{int}$ follow a distribution with a bounded and positive density $\pi(\beta_{lk})$ on $\mathbb{R}$. The prior (10) specifies the coefficients of $\boldsymbol{\beta}$ and is seen as a distribution on $\mathbb{R}^{2^L}$. We set all remaining coefficients, i.e. $\beta'_{lk}s$ for $l \ge L$, to 0, so that a distribution on the collection of all wavelet coefficients is now specified.

The prior (10) contains the so-called *sieve priors* [17] as a special case, where the sieve is with respect to the approximating spaces Vect$\{\psi_{lk}, l < d\}$ for some $d \ge 0$. 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 wavelet priors) are too rigid to enable adaptive results for more complex multiscale norms, as we demonstrate in Section 3.4.

By definition, the prior (10) 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.

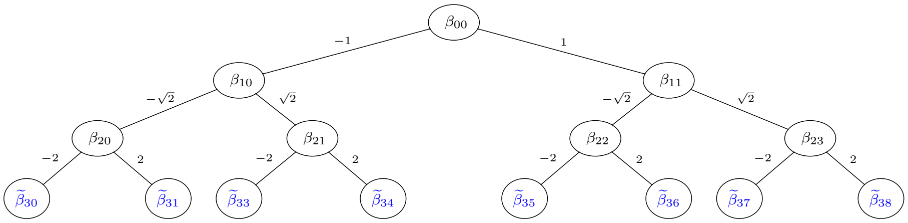


FIG 2. *Flat tree with edges weighted by the amplitude of the Haar wavelets.*

the depth of the tree) in (10) 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 (10) attain adaptive multiscale behavior. We will demonstrate in Section 3 that this optimism is indeed warranted.

2.4. Dyadic Bayesian CART priors

Each tree $\mathcal{T} = \mathcal{T}_{int} \cup \mathcal{T}_{ext}$ can be associated with two sets of coefficients: (a) *internal* coefficients β_{lk} attached to wavelets ψ_{lk} for $(l, k) \in \mathcal{T}'_{int}$ and (b) *external* coefficients $\widetilde{\beta}_{lk}$ attached to intervals I_{lk} for $(l, k) \in \mathcal{T}_{ext}$, as defined in the next paragraph. Bayesian CART priors [20, 23], as opposed to wavelet priors, assign the prior distribution externally on $\widetilde{\beta}_{lk}$.

Given a tree $\mathcal{T} \in \mathbb{T}$, itself generating a random partition via intervals I_{lk} as in Section 2.2, Bayesian CART methods yield histogram reconstructions

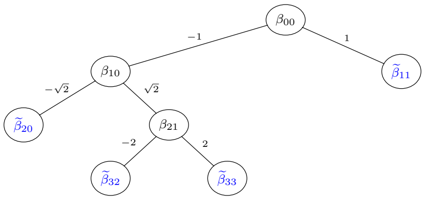
$$\widetilde{f}_{\mathcal{T},\widetilde{\boldsymbol{\beta}}}(x) = \sum_{(l,k)\in\mathcal{T}_{ext}} \widetilde{\beta}_{lk} \mathbb{I}_{I_{lk}}(x), \tag{11}$$

where $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_{lk} : (l, k) \in \mathcal{T}_{ext})'$ is a vector of reals interpreted as step heights and where I_{lk}'s are obtained from the tree $\mathcal{T}$ as in (8). We now define the *Dyadic Bayesian CART prior* using the following hierarchical model on the *external* coefficients rather than *internal* coefficients (compare with (10))

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

for $\Pi_{\mathbb{T}}$ a prior as in Section 2.1, and where the height $\widetilde{\beta}_{lk}$ at a specific $(l, k) \in \mathcal{T}_{ext}$ has a bounded and positive density $\widetilde{\pi}(\widetilde{\beta}_{lk})$ on $\mathbb{R}$. This coincides with the widely used Bayesian CART priors using a midpoint dyadic splitting rule (as we explained in Section 2.2). In practice, the density $\widetilde{\pi}$ is often chosen as centered Gaussian with some variance $\sigma^2 > 0$ [20, 23].

The histogram prior (11) can be rephrased in terms of wavelets. Indeed, linking the Haar wavelet functions ψ_{lk}'s in (3) with $\mathbb{I}_{I_{lk}}$'s via the duality $2\mathbb{I}_{I_{(l+1)2k}} =$



(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).*

$\mathbb{I}_{I_{lk}} + 2^{-l/2}\psi_{lk}$ and $2\mathbb{I}_{I_{(l+1)2k+1}} = \mathbb{I}_{I_{lk}} - 2^{-l/2}\psi_{lk}$, the indicators can be expressed in terms of the ψ_{lk}'s of smaller depths. The histogram representation (11) can then be rewritten in terms of the *internal* coefficients, i.e. $\widetilde{f}_{\mathcal{T},\widetilde{\boldsymbol{\beta}}}(x) = f_{\mathcal{T},\boldsymbol{\beta}}(x)$ as in (9), with β_{lk}'s and $\widetilde{\beta}_{lk}$'s linked through

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

where $s_k = (-1)^{k+1}$. 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 (9) through (13) provides various avenues for constructing prior distributions, and enables a re-interpretation of the Bayesian CART prior, as we now see.

2.5. *Introducing the g-prior for Trees*

For a given tree $\mathcal{T}$, let $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}'_{int})'$ 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, $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}_{ext})'$ 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 $|\mathcal{T}_{ext}| \times (|\mathcal{T}'_{int}|)$ matrix (noting $|\mathcal{T}_{ext}| = |\mathcal{T}'_{int}|$), 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

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

This results from the fact that the collection $(2^{l/2}\mathbb{I}_{lk},\ (l,k) \in \mathcal{T}_{ext})$ is an orthonormal system spanning the same space as $(\psi_{jk},\ (j,k) \in \mathcal{T}'_{int})$, 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 (10) which draws $\beta_{lk} \sim \pi$ and the practical Bayesian CART histogram prior which draws $\widetilde{\beta}_{lk} \sim \widetilde{\pi}$.

Starting *from within* the tree, one can assume standard Gaussian $\pi(\cdot)$ to obtain $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, I_{|\mathcal{T}_{ext}|})$. Through the duality (14), this implies an *independent product prior* on the external coefficients $\widetilde{\beta}_{lk}$, i.e.

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

i.e. var $\widetilde{\beta}_{lk} = 2^l$ where the variances increase with the resolution l.

The Bayesian CART prior, on the other hand, starts *from outside* the tree by assigning $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} \sim \mathcal{N}(0, g_n I_{|\mathcal{T}_{ext}|})$ for some $g_n > 0$ (given $\mathcal{T}$), ultimately setting the bottom node variances equal. This translates into the following "g-prior" on the *internal* wavelet coefficients through (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*

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

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 non-trivial, although $A'_{\mathcal{T}}A_{\mathcal{T}}$ admits some partial sparsity (we characterize basic properties of the pinball matrix in Section 7.1.1 in the Appendix). Also, comparison with (15) suggests possible choices of g_n: the independent wavelet prior makes variances of external coefficients increase as $2^l \leq 2^{L_{max}} \asymp n$, which suggests setting $g_n = n$ to cover ("undersmooth") all possible variance configurations. This choice, as well as others, is considered 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 [71]. The g-prior has been shown to have many favorable properties in terms of invariance or predictive matching [4, 3]. 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.

2.6. Connection to Spike-and-Slab Priors

The tree-priors introduced above induce *irregular* dyadic partitions, where the partitioning intervals I_{lk} are *not* necessarily of equal length. Standard wavelet thresholding [26] reconstructs regression surfaces with an inverse wavelet transform of thresholded coefficients, which *also* yields a piece-wise constant reconstruction where the pieces are not necessarily of equal size. This brings us to the following interesting connection to spike-and-slab wavelet priors.

For adaptive wavelet shrinkage, [21] 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 [42], who show adaptive minimax posterior concentration over Hölder balls for the sup-norm loss, and by [54] who subsequently quantified an adaptive BvM property in the multiscale setting using a similar prior. Independently for each wavelet coefficient β_{lk} at resolutions larger than some $l_0(n)$ (strictly increasing sequence), the prior in [54] can be written in the standard spike-and-slab form

$$\pi(\beta_{lk} \mid \gamma_{lk}) = \gamma_{lk}\pi(\beta_{lk}) + (1-\gamma_{lk})\delta_0(\beta_{lk}), \tag{18}$$

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, i.e. $\theta_l = 1$ 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". [21] suggest setting θ_l equal to the proportion of signal coefficients (at resolution l) as determined by the universal threshold value, whereas [54] specifies $n^{-a} \leq \theta_l \leq 2^{-l(1+b)}$ for some $a > 0$ and $b > 1/2$ for $l_0(n) < l \leq L_n$, where $L_n = \lfloor \log n / \log 2 \rfloor$ and $l_0(n) \asymp (\log n)^{1/(2\nu+1)}$ for some $\nu > 0$.

There are undeniable similarities between (10) and (18), in the sense that the binary inclusion indicator γ_{lk} in (18) can be regarded as the node splitting indicator γ_{lk} in (4). While the indicators γ_{lk} in (18) are *independent* under the spike-and-slab prior, they are hierarchically constrained under the CART prior, where the pattern of non-zeroes encodes the tree oligarchy. While the spike-and-slab prior has been widely regarded as the methodological ideal [42], it is not very practical in higher dimensions and is confined to one given basis. The Bayesian CART prior is widely used in practice, allows for basis selection (as will be seen in Section 4) but is not yet theoretically well understood. The seeming resemblance of the Bayesian CART prior (10) to the spike-and-slab prior (18) makes one naturally wonder whether, unlike sieve-type priors, Bayesian CART posteriors verify optimal multiscale properties.

3. Bayesian Dyadic CART is Multiscale

In this section we investigate the inference properties of tree-based posteriors, showing that they are "multiscale" in the sense that (a) they attain the minimax rate of posterior concentration in the supremum-norm sense (up to a log factor), and (b) exhibit an *adaptive* non-parametric BvM behavior in typical multiscale spaces. We complement these results by addressing the question of uncertainty quantification via the construction of adaptive confidence bands. For clarity of exposition, we focus in the main text 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 the appendix Section 7.3 for more details.

3.1. Posterior supremum-norm convergence

Let us recall the standard inequality (see e.g. (66) below), for f and f_0 two continuous (or Haar-histogram) functions with Haar-wavelet coefficients β_{lk} and β^0_{lk},

$$\|f - f_0\|_\infty \leq |\beta_{-10} - \beta^0_{-10}| + \sum_{l \geq -1} 2^{l/2} \max_{0 \leq k < 2^l} |\beta_{lk} - \beta^0_{lk}| =: \ell_\infty(f, f_0). \tag{19}$$

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 $\mathcal{T}$ 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 $\mathcal{T}'_{int}$ from Section 2.3,

$$\begin{aligned} \mathcal{T} \quad &\sim \quad \Pi_{\mathbb{T}} \\ (\beta_{lk})_{l \leq L,k} \,|\, \mathcal{T} \;\; &\sim \;\; \pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k) \notin \mathcal{T}'_{int}} \delta_0(\beta_{lk}), \end{aligned} \tag{20}$$

where $\pi(\boldsymbol{\beta}_{\mathcal{T}})$ is a distribution to be chosen on $\mathbb{R}^{|\mathcal{T}'_{int}|}$, not necessarily of product form. This is a generalization of (10), which allows for *correlated* wavelet coefficients (e.g. the g-prior). Similarly, let $\boldsymbol{X}_{\mathcal{T}}$ denote the vector of ordered responses X_{lk} in (2) for $(l, k) \in \mathcal{T}'_{int}$. From the white noise model, we have

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

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

$$\sum_{\mathcal{T} \in \mathbb{T}} \Pi[\mathcal{T} \,|\, X] \cdot \pi(\boldsymbol{\beta}_{\mathcal{T}} \,|\, X) \cdot \prod_{(l,k) \notin \mathcal{T}'_{int}} \mathbb{I}_0(\beta_{lk}), \tag{21}$$

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

$$\pi(\boldsymbol{\beta}_{\mathcal{T}} \,|\, X) = \frac{\mathrm{e}^{-\frac{n}{2}\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 + n\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}} \pi(\boldsymbol{\beta}_{\mathcal{T}})}{N_X(\mathcal{T})}, \tag{22}$$

$$\Pi[\mathcal{T} \,|\, X] = \frac{W_X(\mathcal{T})}{\sum_{\mathcal{T} \in \mathbb{T}} W_X(\mathcal{T})}, \quad \text{with} \quad W_X(\mathcal{T}) = \Pi_{\mathbb{T}}(\mathcal{T}) N_X(\mathcal{T}). \tag{23}$$

Let us note that the sum in the last display is finite, as we restrict to trees of depth at most $L = L_{max}$. While the posterior expression (22) allows for general priors $\pi(\boldsymbol{\beta}_{\mathcal{T}})$, we will focus on conditionally conjugate Gaussian priors for simplicity. Note that the classes of priors $\Pi_{\mathbb{T}}$ from Section 2 are non-conjugate: the posterior on trees is given by the somewhat intricate expression (23) and does not belong to one of the classes of $\Pi_{\mathbb{T}}$ priors.

Our first result exemplifies the potential of tree-shaped sparsity by showing that Bayesian Dyadic CART achieves the minimax rate of posterior concentration over Hölder balls in the sup-norm sense, i.e. $\varepsilon_n = (n/\log n)^{-\alpha/(2\alpha+1)}$, up to a logarithmic term. Define Hölder-type spaces of functions on $[0,1]$ as

$$\mathcal{H}(\alpha, M) \equiv \left\{ f \in \mathcal{C}[0,1] : \max_{l\geq 0,\ 0\leq k<2^l} 2^{l(\frac{1}{2}+\alpha)} |\langle f, \psi_{lk}\rangle| \vee |\langle f, \psi_{-10}\rangle| \leq M \right\}. \tag{24}$$

For balanced Haar wavelets ψ_{lk} as in (24), $\mathcal{H}(\alpha, M)$ contains the standard space of α-Hölder (resp. Lipschitz when $\alpha = 1$) functions for any $\alpha \in (0,1]$, defined as

$$\mathcal{H}_M^\alpha \equiv \left\{ f : \|f\|_\infty \leq M,\ \frac{|f(x) - f(y)|}{|x-y|^\alpha} \leq M \quad \forall x, y \in [0,1] \right\}. \tag{25}$$

Our master rate-theorem, whose proof can be found in Section 6.1, is stated below. It will be extended in various directions in the sequel.

Theorem 1. *Let $\Pi_{\mathbb{T}}$ be one of the Bayesian CART priors from Section 2.1, with parameters $\Gamma > 2e^3$ and $c > 7/4$, i.e*

(i) the Galton-Watson process prior with $p_{lk} = \Gamma^{-l}$, or
(ii) the conditionally uniform prior with $\lambda = 1/n^c$ in (5), or
(iii) the exponential prior (6) with a parameter c.

Consider the tree-shaped wavelet prior (20) with $\Pi_{\mathbb{T}}$ as above and $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$, where $\Sigma_{\mathcal{T}}$ is either $I_{|\mathcal{T}_{ext}|}$ or $g_n(A'_{\mathcal{T}} A_{\mathcal{T}})^{-1}$ with $g_n = n$. Define

$$\varepsilon_n = \left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \quad \text{for} \quad \alpha > 0. \tag{26}$$

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

$$\sup_{f_0 \in \mathcal{H}(\alpha, M)} E_{f_0} \Pi\left[f_{\mathcal{T},\boldsymbol{\beta}} : \ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}}, f_0) > M_n \varepsilon_n \,|\, X\right] \to 0. \tag{27}$$

By (19), the statement (27) also holds for the supremum loss $\|\cdot\|_\infty$.

Theorem 1 encompasses both original Bayesian CART proposals for priors on coefficients $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} \sim \mathcal{N}(0, I_{|\mathcal{T}_{ext}|})$ (the case $\Sigma_{\mathcal{T}} = g_n(A_{\mathcal{T}} A'_{\mathcal{T}})^{-1}$) as well as the mathematically slightly simpler wavelet priors $\Sigma_{\mathcal{T}} = I_{|\mathcal{T}_{ext}|}$. One may also note that the tree-shaped Bayesian CART priors occupy the middle ground between flat trees (with only a depth cutoff) and spike-and-slab priors (with general sparsity patterns). We did not fully optimize the constants in the statement; for instance, one can check that $\Gamma > 2$ for the g-prior works.

While Theorem 1 is posited 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. The following Theorem, proved in Section 7.5 in the Appendix, formalizes this intuition and generalizes Theorem 1 to (a) prior distributions over functions $f_{\mathcal{T},\boldsymbol{\beta}}$ in (9), where $\{\psi_{lk}\}$ is not necessarily the Haar basis, (b) general unstructured covariance matrices on active wavelet coefficients β_{lk}. In the next statement, $\{\psi_{lk}\}$ is an S–regular wavelet basis on $[0,1]$, such as the boundary-corrected wavelet basis of [22] (see also [36], Chapter 4), and the space $\mathcal{H}(\alpha,M)$ still defined as in (24), with $\{\psi_{lk}\}$ referring to the considered basis.

Theorem 2. *Let $\{\psi_{lk}\}_{lk}$ be an S–regular wavelet basis, $S\geq 1$, and let $\Pi_{\mathbb{T}}$ be one of the tree priors (i)-(iii) in Theorem 1, for some $\Gamma\geq\Gamma_0(S)>0$ or $c\geq c_0>0$ large enough. Let $f_0\in\mathcal{H}(\alpha,M)$ for some $M>0$ and* arbitrary *$0<\alpha\leq S$. Consider the tree-shaped wavelet prior (20) with*

$$\pi(\boldsymbol{\beta}_{\mathcal{T}})\sim\mathcal{N}(0,\Sigma_{\mathcal{T}}),\quad \text{and } \lambda_{min}(\Sigma_{\mathcal{T}})\gtrsim 1/\sqrt{\log n},\ \lambda_{max}(\Sigma_{\mathcal{T}})\lesssim n^a, \tag{28}$$

for some $a>0$. Then for any $M_n\to\infty$ and for ε_n as in (27) we have

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

Example 2. *As an example of the general covariance matrix $\Sigma_{\mathcal{T}}$, consider an autoregressive histogram prior $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}}\sim\mathcal{N}(0,\widetilde{\Sigma}_{\mathcal{T}})$ with $\widetilde{\Sigma}_{\mathcal{T}}=c_n\left(\rho^{|i-j|}\right)$ for some $0<\rho<1$ and $c_n>0$. This prior links jump sizes of neighboring histogram cells and implies $\Sigma_{\mathcal{T}}=(A'_{\mathcal{T}}A_{\mathcal{T}})^{-1}A'_{\mathcal{T}}\widetilde{\Sigma}_{\mathcal{T}}A_{\mathcal{T}}(A'_{\mathcal{T}}A_{\mathcal{T}})^{-1}$. From Proposition 2 in the Appendix (Section 7.1.1) and the Gershgorin circle theorem, the maximal eigenvalue satisfies $\lambda_{max}(\Sigma_{\mathcal{T}})\leq\lambda_{max}(\widetilde{\Sigma}_{\mathcal{T}})/\lambda_{min}(A'_{\mathcal{T}}A_{\mathcal{T}})\leq c_n(1+\frac{2}{1-\rho})$, where we have used the fact that the spectral matrix norm is sub-multiplicative, that the non-zero eigenvalues of AB and BA are the same and that $\widetilde{\Sigma}_{\mathcal{T}}$ is symmetric and positive semi-definite. Moreover, $\lambda_{min}(\Sigma_{\mathcal{T}})\geq\lambda_{min}(\widetilde{\Sigma}_{\mathcal{T}})/\lambda_{max}(A'_{\mathcal{T}}A_{\mathcal{T}})\geq 1/\sqrt{\log n}$ for large enough $c_n>0$.*

The rate ε_n in (26) coincides with the minimax rate for the supremum norm in the white noise model up to a logarithmic factor, which equals $(\log n)^{\frac{\alpha}{2\alpha+1}}$. This means that there may be a slight logarithmic term to pay (no more than $\sqrt{\log n}$ in the limit $\alpha\to\infty$ and no price in the limit $\alpha\to 0$) when using Bayesian CART. We next show that this logarithmic factor is in fact real, i.e. it is *not* an artifact from the upper-bound proof, and cannot be removed for the considered class of priors. 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 3. *Let $\Pi_{\mathbb{T}}$ be one of the Bayesian CART priors from Theorem 1. Consider the tree-shaped wavelet prior (20) with $\pi(\boldsymbol{\beta}_{\mathcal{T}})\sim\mathcal{N}(0,\Sigma_{\mathcal{T}})$, where $\Sigma_{\mathcal{T}}$ is $I_{|\mathcal{T}_{ext}|}$ and $\{\psi_{lk}\}$ an S–regular wavelet basis, $S\geq 1$. Let ε_n be the rate defined in (26) 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 $n\to\infty$,*

$$\inf_{f_0\in\mathcal{H}(\alpha,M)} E_{f_0}\Pi\left[\ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}},f_0)\leq m\,\varepsilon_n\,|\,X\right]\to 0. \tag{29}$$

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 (29) with respect to (27); we refer to [12] for more details on the notion of posterior rate lower bound). The proof of Theorem 3 can be found in Section 7.6.

3.2. From $\|\cdot\|_\infty$ to BvM

Let us now formalize the notion of a nonparametric BvM theorem in multiscale spaces following [17] (we refer also to [16] for more background and discussion of the, different, L^2–type setting). Such spaces are defined through the speed of decay of multiscale coefficients $\beta_{lk} = \langle f, \psi_{lk}\rangle$. For a monotone increasing weighting sequence $w = (w_l)_{l=0}^\infty$, with $w_l \geq 1$ and $w_l/\sqrt{l} \to \infty$ as $l \to \infty$ (such a $w = (w_l)_{l=0}^\infty$ is called *admissible*) we define the following *multiscale sequence space*

$$\mathcal{M}(w) = \left\{ x = (x_{lk}) : \|x\|_{\mathcal{M}(w)} \equiv \sup_l \frac{\max_k |x_{lk}|}{w_l} < \infty \right\}.$$

We consider a separable closed subspace of $\mathcal{M}(w)$ defined as

$$\mathcal{M}_0(w) = \left\{ x \in \mathcal{M}(w) : \lim_{l\to\infty} \max_k \frac{|x_{lk}|}{w_l} = 0 \right\}.$$

Defining random variables $g_{lk} = \int_0^1 \psi_{lk} dW(t) \sim \mathcal{N}(0,1)$, according to Proposition 2 in [17], the Gaussian white noise $\mathbb{W} = (g_{lk})$ defines a tight Gaussian Borel measure in the space $\mathcal{M}_0(w)$ for admissible sequences w. The convergence in distribution of random variables in the multiscale space $\mathcal{M}_0(w)$ is metrised via the bounded Lipschitz metric $\beta_{\mathcal{M}_0(w)}$ defined below. For μ, η probability measures on a metric space (S, d) define

$$\beta_S(\mu,\eta) = \sup_{F:\|F\|_{BL}\leq 1} \left| \int_S F(x)(d\mu(x) - d\eta(x)) \right|,$$

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

Denote with $\mathbb{X} = \mathbb{X}^{(n)} = (X_{lk} : l \in \mathbb{N}_0, 0 \leq k < 2^l)$, where X_{lk} satisfy (2). Let $\widetilde{\Pi}_n = \Pi_n \circ \tau_{\mathbb{X}}^{-1}$ be the image measure of $\Pi(\cdot \mid X)$ under $\tau_{\mathbb{X}} : f \to \sqrt{n}(f - \mathbb{X})$. Namely, for any Borel set B we have

$$\widetilde{\Pi}_n(B) = \Pi\left(\sqrt{n}(f - \mathbb{X}) \in B \mid X\right). \tag{30}$$

The following Theorem characterizes the *adaptive* non-parametric Bernstein-von Mises behavior of posteriors under the Bayesian Dyadic CART. In the result below, one assumes that trees sampled from $\Pi_{\mathbb{T}}$ contain all nodes (j, k) for all $j \leq j_0(n) \to \infty$ slowly. Note that this constraint is easy to accommodate in the construction: for the GW process, one starts stopping splits only after depth $j_0(n)$, while for priors (5)–(6), it suffices to constrain the indicator $\mathbb{I}_{\mathcal{T}\in\mathbb{T}}$ to trees that fill all first $j_0(n)$ layers.

Theorem 4. *(Adaptive Non-parametric BvM) Let $\mathcal{M}_0 = \mathcal{M}_0(w)$ for some admissible sequence $w = (w_l)$. Assume the Bayesian CART priors $\Pi_{\mathbb{T}}$ from Theorem 1 constrained to trees that fit $j_0(n)$ layers, i.e. $\gamma_{lk} = 1$ for $l \leq j_0(n)$, for some strictly increasing sequence $j_0(n) \to \infty$ that satisfies $w_{j_0(n)} \geq c \log n$ for some $c > 0$. Consider tree-shaped priors as in Theorem 1, or using an $S-$regular wavelet basis, $S \geq 1$, as in Theorem 2. Then the posterior distribution satisfies the weak Bernstein-von Mises phenomenon in $\mathcal{M}_0$ in the sense that*

$$E_{f_0}\beta_{\mathcal{M}_0}(\widetilde{\Pi}_n, \mathcal{N}) \to 0 \quad \textit{as } n \to \infty,$$

where $\mathcal{N}$ is the law of $\mathbb{W}$ in $\mathcal{M}_0$.

This statement, proved in Section 7.7 in the Appendix, can be shown, for example, by verifying the conditions in Proposition 6 of [17]. The first condition pertains to contraction in the $\mathcal{M}_0$–space, which can be obtained from our $\|\cdot\|_\infty$ result. In order to attain BvM, we need to modify the prior to always include a few coarsest dense layers in the tree (similarly as [54]). Such trees are semi-dense, where sparsity kicks in only deeper in the tree after $j_0(n)$ dense layers have already been fitted.

While Theorem 4 will be of use in the next subsection, we now briefly mention its several implications, referring to [17] for details. First, let us consider multi-scale credible balls for f_0, which consist of functions $f_{\mathcal{T},\boldsymbol{\beta}}$ that simultaneously satisfy multi-scale linear constraints (see e.g. (5) in [17]):

$$\mathcal{B}_n = \left\{ f_{\mathcal{T},\boldsymbol{\beta}} : \ \|f_{\mathcal{T},\boldsymbol{\beta}} - \mathbb{X}\|_{\mathcal{M}(w)} \leq R_n/\sqrt{n} \right\}, \tag{31}$$

where R_n is chosen such that $\Pi[\mathcal{B}_n \mid X] = 1-\gamma$ (or the smallest radius such that $\Pi[\mathcal{B}_n \mid X] \geq 1-\gamma$), i.e. $\mathcal{B}_n$ a credible set of level $1-\gamma$. It follows from Theorem 4 and Theorem 5 in [17] that $\mathcal{B}_n$ is *also a confidence set* for f_0 in $\mathcal{M}(w)$ of level $1-\gamma$, i.e. $P_{f_0}(f_0 \in \mathcal{B}_n) \to 1-\gamma$. As a second application of Theorem 4, one directly obtains confidence bands for $F(t) := \int_0^t f(x)dx$ for $0 \leq t \leq 1$: those result from taking quantile credible bands in the following limiting distribution result. By combining Theorem 4 and Theorem 4 in [17], one indeed obtains $\beta_{C([0,1])}\left(\mathcal{L}(\sqrt{n}(F(\cdot) - \int_0^\cdot dX^{(n)} \mid X), \mathcal{L}(G)\right) \to 0$ in P_{f_0}-probability, where $(G(t) : t \in [0,1])$ is a Brownian motion.

3.3. Adaptive Honest Confidence Bands for f_0

In order for uncertainty quantification to be as informative as possible, it is desirable that the confidence sets shrink as fast as possible. When the degree of smoothness α is a priori known, one can intersect (31) with qualitative restrictions on f_0 to obtain “optimal” frequentist confidence intervals (whose diameters shrink at the sup-norm rate). For the more practical case when α is unknown, [54] obtained multiscale credible balls under the spike-and-slab prior that are adaptive and have uniform coverage over self-similar functions [53, 10, 34, 52].

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 $\varepsilon > 0$,*

$$\|K_j(f) - f\|_\infty \geq \varepsilon 2^{-j\alpha} \quad \textit{for all } j \geq j_0, \tag{32}$$

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 $\mathcal{H}_{SS}(\alpha, M, \varepsilon)$.

Following [54], 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 distribution $\Pi_{\mathbb{T}}[\cdot \mid X]$ over trees, we define the* median tree *$\mathcal{T}^*_X = \mathcal{T}^*(\Pi_{\mathbb{T}}[\cdot \mid X])$ as the following set of nodes*

$$\mathcal{T}^*_X = \{(l,k),\ l \leq L_{max},\quad \Pi[(l,k) \in \mathcal{T}_{int} \mid X] \geq 1/2\}\,. \tag{33}$$

Similarly as in the median probability model [3, 2], a node belongs to $\mathcal{T}^*_X$ if its (marginal) posterior probability to be selected by a tree estimator exceeds 1/2. Interestingly, it turns out that $\mathcal{T}^*_X$ *is an actual tree*, i.e. the nodes follow hereditary constraints (see Lemma 14 in the Appendix). We define the resulting median tree estimator as

$$\widehat{f}_T(x) = \sum_{(l,k) \in \mathcal{T}^*_X} X_{lk} \psi_{lk}(x). \tag{34}$$

Moreover, with R_n as in (31), we define the *radius* as

$$\sigma_n = \sigma_n(X) = \sup_{x \in [0,1]} \sum_{l=0}^{L_{max}} v_n \sqrt{\frac{\log n}{n}} \sum_{k=0}^{2^l - 1} \mathbb{I}_{(l,k) \in \mathcal{T}^*_X} |\psi_{lk}(x)|, \tag{35}$$

for some $v_n \to \infty$ to be chosen. Finally, we construct the credible band as

$$\mathcal{C}_n = \left\{ f_{\mathcal{T},\boldsymbol{\beta}} :\ \|f_{\mathcal{T},\boldsymbol{\beta}} - \mathbb{X}\|_{\mathcal{M}(w)} \leq R_n/\sqrt{n},\ \|f_{\mathcal{T},\boldsymbol{\beta}} - \widehat{f}_T\|_\infty \leq \sigma_n \right\}, \tag{36}$$

where $\widehat{f}_T$ as in (34) and $\sigma_n = \sigma_n(X)$ is as in (35).

The following Theorem (proved in Section 7.8) shows that uncertainty quantification with Bayesian CART is well-calibrated from a frequentist point of view, where posterior credible bands (36) have uniform coverage under self-similarity.

Theorem 5. *Let $0 < \alpha_1 \leq \alpha_2 < \infty$, $M \geq 1$, $\gamma \in (0,1)$ and $\varepsilon > 0$. Let Π be any prior as in the statement of Theorem 4 and let (w_l) be an admissible sequence such that $w_{j_0(n)}/\sqrt{\log n} \uparrow \infty$. Assume R_n as in (31), σ_n as in (35) with v_n such that $(\log n)^{1/2} = o(v_n)$ and let $\widehat{f}_T$ denote the median tree estimator (34). Then the set $\mathcal{C}_n$ defined in (36) satisfies, uniformly over $\alpha \in [\alpha_1, \alpha_2]$,*

$$\sup_{f_0 \in \mathcal{H}_{SS}(\alpha, M, \varepsilon)} |P_{f_0}(f_0 \in \mathcal{C}_n) - (1-\gamma)| \to 0,$$

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

$$|\mathcal{C}_n|_\infty = O_{P_{f_0}}((n/\log n)^{-\alpha/(2\alpha+1)} v_n), \tag{37}$$

$$\Pi[\mathcal{C}_n \,|\, X] = 1 - \gamma + o_{P_{f_0}}(1). \tag{38}$$

3.4. Flat Trees are not Multiscale

Depending on the tree topology, some tree priors may be more or less suited to derive adaptive multiscale properties. Only very few priors (actually *only* point mass spike-and-slab based priors, as discussed in the Introduction) were shown to attain adaptive sup-norm concentration rates. Theorem 1 now certifies Bayesian Dyadic CART as one of them. 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 [42]. 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 [11]. 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 trees from Section 2.1.3.

Recall that 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 $\pi(\beta_{lk}) \propto \sigma_l^{-1}\phi\left(\beta_{lk}/\sigma_l\right)$,

$$(\beta_{lk}) \,|\, d \sim \bigotimes_{l<d,k} \pi(\beta_{lk}) \;\otimes\; \bigotimes_{l\geq d,k} \delta_0(\beta_{lk}), \tag{39}$$

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, e.g. $\sigma_l = 2^{-l(\beta+1/2)}$ for some $0 < \beta \leq \alpha$. This "undersmoothing" prior requires the knowledge of (a lower bound on) α and yields a *non-adaptive* non-parametric BvM behavior [17].

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 [56, 61]. It is not hard to check that the flat-tree prior (39) 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 (18) *is ℓ_∞ adaptive* [42]. The flat tree prior (39), on the other hand, fits a few dense layers *without* internal sparsity and *is ℓ_2 adaptive* (up to a log term) [61]. However, as shown in the following Theorem proved in Section 7.9 in the Appendix, flat trees fall short of ℓ_∞ adaptation.

Theorem 6. *Assume the flat tree prior (39) with random d, where $\pi(d)$ is non-increasing and where the active wavelet coefficients β_{lk} are Gaussian iid*

$\mathcal{N}(0,1)$. *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*

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

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

Theorem 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 sub-optimal.

4. Non-dyadic Bayesian CART is Multiscale

A fundamental 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 [8, 30] 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 (non-dyadic) CART [20, 23] and its connection to Unbalanced Haar (UH) wavelet basis selection.

4.1. Unbalanced Haar Wavelets

UH wavelet basis functions [37] 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 [29]. Assume that possible values for splits are chosen from a set of $n = 2^{L_{max}}$ breakpoints $\mathcal{X} = \{x_i : x_i = i/n, 1 \leq i \leq n\}$[1]. 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 $b_{lk} = 1/2^{l+1}$, 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 $1 \leq l \leq L_{max}$ and $0 \leq k < 2^l$, set

$$\begin{aligned} l_{lk} &= l_{(l-1)\lfloor k/2 \rfloor}, \qquad r_{lk} = b_{(l-1)\lfloor k/2 \rfloor}, \quad \text{if } k \text{ is even,} \\ l_{lk} &= b_{(l-1)\lfloor k/2 \rfloor}, \qquad r_{lk} = r_{(l-1)\lfloor k/2 \rfloor}, \quad \text{if } k \text{ is odd.} \end{aligned} \tag{40}$$

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

[1] In non-parametric regression, $\mathcal{X}$ could be regarded as the set of observed covariate values.

Let A denote the set of *admissible* nodes (l,k), in that (l,k) is such that $\mathcal{X} \cap (l_{lk}, r_{lk}] \neq \emptyset$, 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 $\psi_{-10}^B = \psi_{-10} = \mathbb{I}_{(0,1)}$, one then recursively constructs wavelet functions ψ_{lk}^B from L_{lk} and R_{lk} as

$$\psi_{lk}^B(x) = \frac{1}{\sqrt{|L_{lk}|^{-1} + |R_{lk}|^{-1}}} \left(\frac{\mathbb{I}_{L_{lk}}(x)}{|L_{lk}|} - \frac{\mathbb{I}_{R_{lk}}(x)}{|R_{lk}|} \right). \tag{41}$$

The system $\Psi_A^B = \{\psi_{-10}^B, \psi_{lk}^B : (l,k) \in A\}$ is orthonormal with respect to the $L^2[0,1]$–inner product. Indeed, by construction, the functions ψ_{lk}^B have a unit L^2–norm and, for a given depth l, they have disjoint supports. Furthermore, ψ_{lk}^B integrate to 0 on their support and thereby each ψ_{lk}^B is orthogonal to $\psi_{l'k}^B$ for any $l' < l$. Similarly as in (7), where I_{lk} denotes the support of the balanced Haar wavelets ψ_{lk}, we denote with I_{lk}^B the support of ψ_{lk}^B.

We characterized Hölder functions though the speed at which the multiscale coefficients decay as a function of the resolution index l (see (24)). With UH wavelets, the decay can be expressed in terms the lengths of the right/left wavelet pieces L_{lk} and R_{lk}.

Lemma 1. *For a set A of admissible nodes (l,k) as above, let us define $\beta_{lk}^B = \langle f, \psi_{lk}^B \rangle$, where ψ_{lk}^B is the unbalanced Haar wavelet in (41) and where $f \in \mathcal{H}_M^\alpha$ was defined in (25). Then*

$$|\beta_{lk}^B| \leq M\, 2^{\alpha-1/2} \max\{|L_{lk}|, |R_{lk}|\}^{\alpha+1/2}. \tag{42}$$

For the classical Haar basis (3), one obtains (24) from (42) by noting $\max\{|L_{lk}|, |R_{lk}|\} = 2^{-(l+1)}$. [29] points out that the computational complexity of the discrete UH transform could be unnecessarily large and imposes the balancing requirement $\max\{|L_{lk}|, |R_{lk}|\} \leq E(|L_{lk}| + |R_{lk}|)\ \forall (l,k) \in A$, for some $1/2 \leq E < 1$. In order to control the combinatorial complexity of the basis system, we also require that the UH wavelets are not too imbalanced. To this end, we introduce the notion of *weakly balanced* wavelets.

Definition 5. *Consider a collection of UH wavelets $\Psi_A^B = \{\psi_{-10}^B, \psi_{lk}^B : (l,k) \in A\}$. We say that Ψ_A^B is* weakly balanced *with balancing constants $E, D \in \mathbb{N}\backslash\{0\}$ if, for any $(l,k) \in A$,*

$$\max(|L_{lk}|, |R_{lk}|) = \frac{M_{lk}}{2^{l+D}} \quad \textit{for some } M_{lk} \in \{1, \ldots, E + l\}. \tag{43}$$

Example 3. *To glean insights into the balancing condition (43), we first consider an example of UH system which is* not *weakly balanced for some given n, D, say $n = 2^4$ and $D = 2$. If we choose $b_{00} = 1/2$, $b_{10} = 1/2 - 1/n$ and $b_{11} = 3/4$, we have*

$$\begin{aligned} L_{10} &= (0, 1/2 - 1/n], & R_{10} &= (1/2 - 1/n, 1/2], \\ L_{11} &= (1/2, 3/4], & R_{11} &= (3/4, 1]. \end{aligned}$$

While the node $(1,1)$ is seen to satisfy (43) with $E = 5$, we note that $\max\{|L_{10}|, |R_{10}|\} = (n-2)/(2n) = 7/16$. *However, $7/16$ cannot be written as $M_{10}/2^{D+1} = M_{10}/8$ for* any *integer M_{10}. This is why the split points b_{00}, b_{10} and b_{11}* do not *belong to any weakly balanced UH wavelet system with balancing constant $D = 2$. Weakly balanced systems can be built by choosing splits in such a way that the "granularity" does not increase too rapidly throughout the branching process. With granularity $R(l, \Psi_A^B)$ of the l^{th} layer we mean the smallest integer $R \geq 1$ such that* $\min_{0 \leq k < 2^l} \min\{|L_{lk}|, |R_{lk}|\} = j/2^R$ *for some $j \in \{1, 2, \ldots, 2^{R-1}\}$. For instance, setting $D = 2$ and $E = 3$ one can build weakly balanced wavelets by first picking b_{00} from values $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$. If, e.g. $b_{00} = 3/4$ (i.e. $R(0, \Psi_A^B) = 2$), the next split b_{10} can be selected from $\{\frac{1}{4}, \frac{3}{8}, \frac{1}{2}\}$, while b_{11} has to be set as $7/8$.*

Our theoretical development relies in part on combinatorial properties of weakly balanced UH systems and on the speed of decay of the multiscale functionals $\beta_{lk}^B = \langle f, \psi_{lk}^B \rangle$ as the layer index $l \in \mathbb{N}$ increases. These two fundamental properties are encapsulated in the Lemma 11 which is vital to the proof of the upcoming Theorem 7.

Note that in the actual BART implementation, the splits are chosen from sample quantiles to ensure balancedness (similar to our condition (43)). Quantile splits are a natural way to generate many weakly balanced systems, providing a much increased flexibility compared to dyadic splits, which correspond to uniform quantiles.

Example 4 (Quantile Splits)**.** *Denote with G a c.d.f with a density g on $[0, 1]$ that satisfies $\|g\|_\infty \leq 2^{D-1}/(2E)$ for $E, D > 0$ chosen below and $\|1/g\|_\infty \leq C_q$ 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 $l \leq L_{max}$ and $0 \leq k < 2^l$, as

$$b_{lk} = G_{L_{max}+D}^{-1}[(2k+1)/2^{l+1}]. \tag{44}$$

The system Ψ_A^B obtained from steps (a) and (b) with splits (44) is weakly balanced for $E = 2 + 3C_q 2^{D-1}$. This is verified in Lemma 12 in the Appendix (Section 7.2.4). Moreover, Figure 4 in Section 7.2.4 illustrates the implementation of the quantile system, where splits are placed more densely in areas where $G(x)$ changes more rapidly.

4.2. Non-dyadic CART prior and multiscale properties.

The recursive construction of the weakly balanced Haar basis pertains closely to the Bayesian CART prior of [20]. Instead of confining b_{lk} to dyadic midpoints (step (i) in Section 2.1.1), such a prior draws b_{lk} from available observations. We consider a related, and more general, strategy which separates the prior on the basis $\Pi_{\mathbb{B}}$ from the prior on the trees $\Pi_{\mathbb{T}}$. We regard the *non-dyadic Bayesian CART* prior as arising from the following three steps:

- *Step 1. (Basis Generation)* Sample $B = (b_{lk})_{0\leq k<2^l-1, l\leq L}$ from $\Pi_{\mathbb{B}}$ by following the steps a)–b) around (40) subject to satisfying the *balancing condition* (43).
- *Step 2. (Tree Generation)* Independently of B, sample a binary tree $\mathcal{T}$ from one of the priors $\Pi_{\mathbb{T}}$ described in Section 2.1.1 or Section 2.1.2.
- *Step 3. (Step Heights Generation)* Given $\mathcal{T}$, we obtain the coefficients (β_{lk}^B) from the tree-shaped prior (20). 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 (9).

An example of such a prior is obtained by drawing a density at random verifying conditions as in Example 4 to generate the breakpoints and then following the construction from Section 2 for Steps 2–3. The following theorem, proved in Section 7.10, is positioned for the (non-smooth) weakly balanced UH wavelets.

Theorem 7. *Let $\Pi_{\mathbb{B}}$ be any prior on breakpoint collections that satisfy weak balancedness according to Definition 5. Let $\Pi_{\mathbb{T}}$ be one of the Bayesian CART priors discussed in Section 2.1, i.e either, for some $\Gamma > 0$, $c \geq 1$,*

(i) the Galton-Watson Process prior with $p_{lk} = \Gamma^{-l^4}$,
(ii) the conditionally uniform prior with $\pi(K) \propto \exp\left(-c\, K \log^4 n\right)$,
(iii) the exponential prior $\pi(\mathcal{T}) \propto \exp\left(-c\, |\mathcal{T}_{ext}| \log^4 n\right)$.

Moreover, consider the tree-shaped wavelet prior (20) where the conditions in (28) are replaced by $\lambda_{min}(\Sigma_{\mathcal{T}}) \gtrsim 1/\sqrt{\log^4 n}$ and $\lambda_{max}(\Sigma_{\mathcal{T}}) \lesssim n^a$ for some $a > 0$. Let $f_0 \in \mathcal{H}_\lambda^M$ as in (25) for some $M > 0$ and $0 < \alpha \leq 1$ and define

$$\varepsilon_n = (\log n)^{1+\frac{3}{2}} \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}. \tag{45}$$

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

$$E_{f_0} \Pi\left[\, \ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}}, f_0) \geq \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > M_n \varepsilon_n \,|\, X\right] \to 0. \tag{46}$$

In the context of piecewise constant priors, Theorem 7 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 a specific own prior.

Remark 1. *The log-factor in Theorem 7 depends on the amount of unbalancedness. A more general statement with a log factor* $(\log n)^{1+\frac{\delta}{2}}$ *for some* $\delta > 0$ *can be obtained under Conditions* $(B1)$ *and* $(B2)$ *in Lemma 11 and under the assumption* $\|\psi_{lk}^B\| \lesssim 2^l$. *According to Lemma 11, weakly balanced systems verify* $(B1)$ *and* $(B2)$ *with* $\delta = 3$. *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* (45) *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 are multiscale, in the sense that they attain (almost) optimal convergence rates in the supremum norm, as well as verify a fully non-parametric and adaptive Bernstein-von Mises theorem in multiscale spaces. The developed framework also allows us to construct adaptive credible bands around f_0 under self-similarity. To allow for non-dyadically organized splits, we introduced weakly balanced Haar wavelets (an elaboration on unbalanced Haar wavelets of [37]) 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, we note that the techniques of proof are non-conjugate in their key tree aspect, which opens the door to applications in many other statistical settings. The obtained results extend to fixed design regression for regular design or possibly more general designs under some conditions. A version of Bayesian CART in the model of density estimation following the ideas of the present work is currently investigated by T. Randrianarisoa as part of his PhD 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 [67]. 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. Proofs

6.1. Proof of Theorem 1

Let us first introduce the event, with $L = L_{max}$,

$$\mathcal{A} = \left\{ \max_{-1 \le l \le L,\, 0 \le k < 2^l} \varepsilon_{lk}^2 \le 2 \log\left(2^{L+1}\right) \right\}. \tag{47}$$

Assuming $\varepsilon_{lk} \sim \mathcal{N}(0,1)$, this event has large probability in the sense that $P(\mathcal{A}^c) \lesssim (\log n)^{-1}$, which follows from $P\left[\max\limits_{1\leq i\leq N} |Z_i| > \sqrt{2\log N}\right] \leq c_0/\sqrt{\log N}$ for some $c_0 > 0$ when $Z_i \sim \mathcal{N}(0,1)$ for $1 \leq i \leq N$.

6.1.1. Posterior Probability of Deep Trees

The first step is to show that, on the event $\mathcal{A}$, the posterior concentrates on reasonably small trees (i.e. with small depth $d(\mathcal{T})$). Let us define the cutoff $\mathcal{L}_c = \mathcal{L}_c(\alpha, M)$ as

$$\mathcal{L}_c = \left\lceil \log_2\left((8M)^{\frac{1}{\alpha+1/2}} \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}} \right) \right\rceil. \tag{48}$$

Lemma 2. *Under the assumptions of Theorem 1, on the event $\mathcal{A}$,*

$$\Pi[d(\mathcal{T}) > \mathcal{L}_c \,|\, X] \to 0 \quad (n \to \infty). \tag{49}$$

Proof. We first show (49) for the GW-process prior as in Section 2.1.1. Consider one tree $\mathcal{T} \in \mathbb{T}$ such that $d(\mathcal{T}) \geq 1$ 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 $\mathcal{T}^- = \mathcal{T}^-_{int} \cup \mathcal{T}^-_{ext}$, where

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

Note that $\mathcal{T}^-$ is a full binary tree, where 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 (23),

$$\Pi[\mathbb{T}_d \,|\, X] = \frac{\sum_{\mathcal{T}\in\mathbb{T}_d} W_X(\mathcal{T})}{\sum_{\mathcal{T}\in\mathbb{T}} W_X(\mathcal{T})} = \frac{\sum_{\mathcal{T}\in\mathbb{T}_d} \frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} W_X(\mathcal{T}^-)}{\sum_{\mathcal{T}\in\mathbb{T}} W_X(\mathcal{T})}, \tag{50}$$

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

Let $\boldsymbol{X}_{\mathcal{T}} = (X_{lk} : (l,k) \in \mathcal{T}'_{int})'$ and $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}'_{int})'$ 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 $K = |\mathcal{T}_{ext}| = |\mathcal{T}_{int}| + 1$,

$$\begin{aligned}\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} &= \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \sqrt{\frac{|\Sigma_{\mathcal{T}^-}|}{2\pi|\Sigma_{\mathcal{T}}|}} \frac{\int \mathrm{e}^{n\boldsymbol{X}'_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}} - \boldsymbol{\beta}'_{\mathcal{T}}[nI_K + \Sigma^{-1}_{\mathcal{T}}]\boldsymbol{\beta}_{\mathcal{T}}/2} d\boldsymbol{\beta}_{\mathcal{T}}}{\int \mathrm{e}^{n\boldsymbol{X}'_{\mathcal{T}^-}\boldsymbol{\beta}_{\mathcal{T}^-} - \boldsymbol{\beta}'_{\mathcal{T}^-}[nI_{K-1} + \Sigma^{-1}_{\mathcal{T}^-}]\boldsymbol{\beta}_{\mathcal{T}^-}/2} d\boldsymbol{\beta}_{\mathcal{T}^-}} \\ &= \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \sqrt{\frac{|\Sigma_{\mathcal{T}^-}|}{|\Sigma_{\mathcal{T}}|}} \sqrt{\frac{|nI_{K-1} + \Sigma^{-1}_{\mathcal{T}^-}|}{|nI_K + \Sigma^{-1}_{\mathcal{T}}|}} \frac{\mathrm{e}^{n^2\boldsymbol{X}'_{\mathcal{T}}(nI_K + \Sigma^{-1}_{\mathcal{T}})^{-1}\boldsymbol{X}_{\mathcal{T}}/2}}{\mathrm{e}^{n^2\boldsymbol{X}'_{\mathcal{T}^-}(nI_{K-1} + \Sigma^{-1}_{\mathcal{T}^-})^{-1}\boldsymbol{X}_{\mathcal{T}^-}/2}}.\end{aligned} \tag{51}$$

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

The Independent Prior. We first consider the independent prior $\Sigma_{\mathcal{T}} = I_K$. Using the expression (51) and since (l_1,k_1) is the deepest rightmost internal node in $\mathcal{T}$ of depth $d = d(\mathcal{T}) = l_1+1$, using the definition of the GW prior,

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \prod_{(l,k)\in\mathcal{T}\backslash\mathcal{T}^-} \frac{\mathrm{e}^{\frac{n^2}{2(n+1)}X_{lk}^2}}{\sqrt{n+1}} = \frac{p_{d-1}(1-p_d)^2}{1-p_{d-1}} \frac{\mathrm{e}^{\frac{n^2}{2(n+1)}X_{l_1k_1}^2}}{\sqrt{n+1}}.$$

From the Hölder continuity (24), one gets $8|\beta_{l_1k_1}| \leq \sqrt{\log n/n}$ for $l_1 \geq \mathcal{L}_c$, where $\mathcal{L}_c$ is as in (48). Conditionally on the event (47), we can then write

$$|X_{l_1k_1}| \leq \frac{1}{\sqrt{n}}\left[\frac{1}{8}\sqrt{\log n} + \sqrt{2\log n + \log 4}\right] \tag{52}$$

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 $d > \mathcal{L}_c$,

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

Going back to the ratio (50), 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 set of all possible trees $\mathcal{T}^-$ that correspond to some $\mathcal{T}\in\mathbb{T}_d$. Using this bound one deduces that, on the event $\mathcal{A}$,

$$\Pi[d(\mathcal{T}) > \mathcal{L}_c\,|\,X] = \sum_{d=\mathcal{L}_c+1}^{L} \Pi[\mathbb{T}_d\,|\,X] \leq 4\,n^{3/4}\sum_{d=\mathcal{L}_c+1}^{L} 2^{d-1}p_{d-1} \tag{53}$$

$$< 4\,n^{3/4}\,L\exp\left[-\mathcal{L}_c\log(\Gamma/2)\right]. \tag{54}$$

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

The g-prior. With $\Sigma_{\mathcal{T}} = g_n\,(A_{\mathcal{T}}'A_{\mathcal{T}})^{-1}$, Proposition 1 implies

$$nI_K + \Sigma_{\mathcal{T}}^{-1} = \begin{pmatrix} nI_{K-1} + \Sigma_{\mathcal{T}^-}^{-1} + \frac{1}{g_n}\boldsymbol{v}\boldsymbol{v}' & \boldsymbol{0} \\ \boldsymbol{0}' & n + \frac{2^{l_1+1}}{g_n} \end{pmatrix}. \tag{55}$$

Using the formula $|A + \boldsymbol{u}\boldsymbol{u}'| = |A|(1+\boldsymbol{u}'A^{-1}\boldsymbol{u})$ for A invertible, and setting $M = (nI_{K-1} + \Sigma_{\mathcal{T}^-}^{-1})^{-1}$, one gets

$$\frac{|nI_{K-1} + \Sigma_{\mathcal{T}^-}^{-1}|}{|nI_K + \Sigma_{\mathcal{T}}^{-1}|} = \frac{1}{(n + 2^{l_1+1}/g_n)\,[1 + \boldsymbol{v}'M\boldsymbol{v}/\,g_n]} < \frac{1}{(n + 2^{l_1+1}/g_n)}.$$

Because $||\boldsymbol{v}||_2^2 = 1+\sum_{l=0}^{l_1-1} 2^l = 2^{l_1}$ and $\lambda_{max}(A'_{\mathcal{T}^-}A_{\mathcal{T}^-})^{-1} = \lambda_{min}(A'_{\mathcal{T}^-}A_{\mathcal{T}^-})^{-1}$ is at most 1, using that the smallest eigenvalue of $A'_{\mathcal{T}}A_{\mathcal{T}}$ equals 2^{d+1} where d is the depth of the most shallow parent of a leaf node in $\mathcal{T}$, we can write, again with the help of the determinant formula above,

$$\frac{|\Sigma_{\mathcal{T}^-}|}{|\Sigma_{\mathcal{T}}|} = \frac{2^{l_1+1}}{g_n}\left(1+\boldsymbol{v}'(A'_{\mathcal{T}^-}A_{\mathcal{T}^-})^{-1}\boldsymbol{v}\right) \leq \frac{2^{l_1+1}}{g_n}\left(1+2^{l_1}\right).$$

Let us set $D := \boldsymbol{X}'_{\mathcal{T}}(nI+\Sigma_{\mathcal{T}}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}} - \boldsymbol{X}'_{\mathcal{T}^-}(nI+\Sigma_{\mathcal{T}^-}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}^-}$. Combining with (55), it follows from a variant of the Sherman–Morrison's matrix inversion formula (Lemma 6) that

$$(nI_K+\Sigma_{\mathcal{T}}^{-1})^{-1} = \begin{pmatrix} M - \frac{M\boldsymbol{v}\boldsymbol{v}'M}{g_n+\boldsymbol{v}'M\boldsymbol{v}} & \boldsymbol{0} \\ \boldsymbol{0}' & 1/(n+2^{l_1+1}/g_n) \end{pmatrix},$$

from which one deduces that

$$D = \frac{X_{l_1k_1}^2}{n+2^{l_1+1}/g_n} - \frac{\boldsymbol{X}'_{\mathcal{T}^-}M\boldsymbol{v}\boldsymbol{v}'M\boldsymbol{X}_{\mathcal{T}^-}}{g_n+\boldsymbol{v}'M\boldsymbol{v}} < \frac{X_{l_1k_1}^2}{n+2^{l_1+1}/g_n}. \tag{56}$$

Since for $l_1 > \mathcal{L}_c$ we have $2X_{l_1k_1}^2 \leq 5\log n/n$, we can write

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)}\frac{\Pi_{\mathbb{T}}(\mathcal{T}^-)}{\Pi_{\mathbb{T}}(\mathcal{T})} < \sqrt{\frac{2^{l_1+1}\left(1+2^{l_1}\right)}{g_n(n+2^{l_1+1}/g_n)}}\,\mathrm{e}^{\frac{X_{l_1k_1}^2 n^2}{2(n+2^{l_1+1}/g_n)}} < \sqrt{\frac{2^{2(l_1+1)}}{ng_n}}n^{5/4}.$$

For $g_n = n$, and for $\mathcal{T}\in\mathbb{T}_d$, so that $2^{l_1}\lesssim 2^d$, the last display is bounded by a constant times $n^{-1/4}2^d p_d$, and the argument can be completed in similar vein as before, with now $\Pi[d(\mathcal{T}) > \mathcal{L}_c \,|\, X] = o_P(1)$ if $\Gamma > 2$.

Other Tree Priors $\Pi_{\mathbb{T}}$. The only modification needed to carry over the proof to the other two priors is the bound for the ratio $\Pi_{\mathbb{T}}(\mathcal{T})/\Pi_{\mathbb{T}}(\mathcal{T}^-)$. Consider the prior from Section 2.1.2. Denoting $K = |\mathcal{T}_{ext}|$ and $\mathbb{C}_K$ the number of full binary trees with $K+1$ leaves, we have

$$\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} = \frac{\pi(K)}{\pi(K-1)}\frac{\mathbb{C}_{K-1}}{\mathbb{C}_{K-2}},$$

and Lemma 7 now implies, for a universal constant $C>0$,

$$\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \lesssim \frac{\lambda}{K}\frac{4^{K-1}(K-2)^{3/2}}{4^{K-2}(\{K-1\}\vee 1)^{3/2}} \leq C\lambda/K.$$

Choosing $\lambda = 1/n^c$ for some $c > 7/4$, it follows that $\Pi[d(\mathcal{T}) > \mathcal{L}_c \,|\, X] \leq 4\,\lambda\, n^{3/4}\sum_{d=\mathcal{L}_c+1}^{L} 2^d/K \leq 4\,\lambda\, n^{3/4}2^L \to 0$. Finally, for the prior (6), one has $\Pi_{\mathbb{T}}(\mathcal{T})/\Pi_{\mathbb{T}}(\mathcal{T}^-) = 1/n^c$, so one can argue similarly. □

6.1.2. Posterior Probability of Missing Signal

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

Lemma 3. *Let us denote, for $A > 0$ to be chosen suitably large,*

$$S(f_0; A) = \left\{ (l,k) : \ |\beta^0_{lk}| \geq A \frac{\log n}{\sqrt{n}} \right\}. \tag{57}$$

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

$$\Pi\left[\{\mathcal{T} : S(f_0; A) \not\subseteq \mathcal{T}\} \mid X\right] \to 0 \qquad (n \to \infty). \tag{58}$$

Proof. As before, we start the proof with the GW prior from Section 2.1.1. 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 $l_S \leq \mathcal{L}_c$ (for n large enough). Let $\mathbb{T}_{\backslash(l_S,k_S)} = \{\mathcal{T} \in \mathbb{T} : (l_S, k_S) \notin \mathcal{T}_{int}\}$ 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 $\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}$ 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 $\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}$ as follows. Denote by $(l_0, k_0) \in \mathcal{T}_{ext} \cap [(0,0) \leftrightarrow (l_S, k_S)]$ the external node which is *closest* to (l_S, k_S) on a route from the root to (l_S, k_S) in a flat tree. 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 $\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}$. Now $\Pi\left[\mathbb{T}_{\backslash(l_S,k_S)} \,|\, X\right]$ equals

$$\frac{\sum_{\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}} W_X(\mathcal{T})}{\sum_{\mathcal{T} \in \mathbb{T}} W_X(\mathcal{T})} \leq \frac{\sum_{\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}} \frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} W_X(\mathcal{T}^+)}{\sum_{\mathcal{T} \in \mathbb{T}_{(l_S,k_S)}} W_X(\mathcal{T})}. \tag{59}$$

Let us denote by $\mathcal{T}^{(j)}$ for $j = -1, \ldots, S$ 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

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} \prod_{s=0}^{S} \frac{N_X(\mathcal{T}^{(s-1)})}{N_X(\mathcal{T}^{(s)})}. \tag{60}$$

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

$$\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} = \frac{1 - p_{l_0}}{1 - p_{l_S+1}} \left(\prod_{l=l_0}^{l_S} \frac{1}{p_l} \right) \left(\prod_{l=l_0+1}^{l_S+1} \frac{1}{1 - p_l} \right), \tag{61}$$

which is bounded by $2^{l_S - l_0 + 2} \Gamma^{(l_0 + l_S)(l_S - l_0 + 1)/2} < 4\Gamma^{2 l_S^2}$.

The Independent Prior. Assuming the independent prior $\Sigma_{\mathcal{T}} = I_K$, we can write for any $\mathcal{T}$ in $\mathbb{T}_{\backslash(l_S,k_S)}$

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} \prod_{(l,k) \in \mathcal{T}^+ \backslash \mathcal{T}} \frac{\sqrt{n+1}}{e^{\frac{n^2}{2(n+1)} X_{lk}^2}}. \tag{62}$$

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

$$\exp\left\{-\frac{n^2}{2(n+1)} X^2_{l_S k_S}\right\} \leq \exp\left\{-\frac{n^2(\beta^0_{l_S k_S})^2}{4(n+1)} + \frac{n \log 2(\log_2 n + 1)}{n+1}\right\},$$

The ratio in (62) can be thus bounded, for any $\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}$, by

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

where we bounded the exponential term in (62) from below by 1. We now continue to bound the ratio (59). For each given $\mathcal{T}^+$, there is *at most* l_S trees $\widetilde{\mathcal{T}} \in \mathbb{T}_{\backslash(l_S,k_S)}$ which have the same extended tree $\widetilde{\mathcal{T}}^+ = \mathcal{T}^+$. This is because $\mathcal{T}^+$ is obtained by extending one branch by adding no more than l_S nodes. Using this fact and the definition of $b(n, l_S)$ above,

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

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

$$\Pi\left[(l_S, k_S) \notin \mathcal{T}_{int} \,|\, X\right] \leq \mathrm{e}^{(3/2 + 3\log\Gamma)(\log_2 n + 1)^2 - \frac{A^2}{8}\log^2 n} \leq \mathrm{e}^{-\frac{A^2}{16}\log^2 n}.$$

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

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

The g-prior. We now modify the proof for the g-prior obtained with $\Sigma_{\mathcal{T}} = g_n(A'_{\mathcal{T}} A_{\mathcal{T}})^{-1}$. Denoting with $K_s = |\mathcal{T}^s_{ext}|$ and because $\mathcal{T}^{(s-1)}$ is obtained from $\mathcal{T}^{(s)}$ by removing two children nodes, we can apply Proposition 1 to obtain the following upper bound for $N_X(\mathcal{T}^{(s-1)})/N_X(\mathcal{T}^{(s)})$. Namely, invoking again the matrix determinant lemma $|A + \boldsymbol{u}\boldsymbol{u}'| = |A|(1 + \boldsymbol{u}'A^{-1}\boldsymbol{u})$ and the matrix inversion lemma (Lemma 6), one obtains the bound

$$\sqrt{\frac{(g_n n + 2^{l_s+1})(g_n + \boldsymbol{v}'M\boldsymbol{v})}{g_n 2^{l_s+1}}} \exp\left\{-\frac{n^2\, X^2_{l_s k_s}}{2(n + 2^{l_s+1}/g_n)} + n^2 \frac{\boldsymbol{X}'_{\mathcal{T}^{(s-1)}} M\boldsymbol{v}\boldsymbol{v}'M\boldsymbol{X}_{\mathcal{T}^{(s-1)}}}{2(g_n + \boldsymbol{v}'M\boldsymbol{v})}\right\},$$

where $M = (nI_{K_{s-1}} + \Sigma^{-1}_{\mathcal{T}^{(s-1)}})^{-1}$ for some $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{T}^{(s-1)}|}$. Next, for $C > 0$ a large enough constant,

$$\|\boldsymbol{X}_{\mathcal{T}^{(s-1)}}\|_2^2 = X^2_{-10} + \sum_{l=0}^{l_{s-1}} \sum_{k=0}^{2^l-1} X^2_{lk} \leq C\{1 + \sum_{l=0}^{l_{s-1}} (2^{-2\,l\,\alpha} + \frac{2^l}{n}\log n)\}, \qquad (63)$$

which is uniformly bounded. Moreover, since

$$\begin{aligned}\boldsymbol{X}'_{\mathcal{T}^{(s-1)}} M\boldsymbol{v}\boldsymbol{v}' M \boldsymbol{X}_{\mathcal{T}^{(s-1)}} &\leq \|\boldsymbol{X}_{\mathcal{T}}^{(s-1)}\|_2^2 \lambda_{max}(M\boldsymbol{v}\boldsymbol{v}'M) \leq \|\boldsymbol{X}_{\mathcal{T}}^{(s-1)}\|_2^2 \mathrm{tr}(M\boldsymbol{v}\boldsymbol{v}'M) \\ &\leq \|\boldsymbol{X}_{\mathcal{T}}^{(s-1)}\|^2 \boldsymbol{v}'MM\boldsymbol{v} \leq \|\boldsymbol{X}_{\mathcal{T}}^{(s-1)}\|_2^2 \lambda_{max}(M)\boldsymbol{v}'M\boldsymbol{v}\end{aligned}$$

and $\lambda_{\max}(M)^{-1} = n + \lambda_{min}(A_{\mathcal{T}}^{(s-1)\prime} A_{\mathcal{T}}^{(s-1)})/g_n > n$, one can write

$$n^2 \frac{\boldsymbol{X}'_{\mathcal{T}^{(s-1)}} M\boldsymbol{v}\boldsymbol{v}' M \boldsymbol{X}_{\mathcal{T}^{(s-1)}}}{2(g_n + \boldsymbol{v}'M\boldsymbol{v})} < \frac{n^2 \|\boldsymbol{X}_{\mathcal{T}^{(s-1)}}\|_2^2 \lambda_{max}(M)}{2g_n/(\boldsymbol{v}'M\boldsymbol{v}) + 2} \leq C_4 \frac{2^{l_s}}{2g_n},$$

where we used the fact that $\|\boldsymbol{v}\|_2^2 = 2^{l_s}$. Finally, because $X_{l_S k_S}^2 \geq C_5 A^2 \log^2 n/n$ for some $C_5 > 0$ we have

$$\begin{aligned}\prod_{s=0}^{S} \frac{N_X(\mathcal{T}^{(s-1)})}{N_X(\mathcal{T}^{(s)})} &< \left(\frac{n(g_n+2)}{g_n}\right)^{S+1} \exp\left\{C_4 \sum_{s=0}^{S} \frac{2^{l_s - 1}}{g_n} - \frac{nC_5 A^2 \log^2 n}{2(n + 2^{l_s+1}/g_n)}\right\} \\ &< \exp\left\{(S+1)\log(3n) + C_4 \frac{(S+1)2^{l_S-1}}{g_n} - C_5 A^2 \log^2 n/4\right\}.\end{aligned}$$

With $g_n = n$, the exponent is dominated by the last term. One then proceeds with (60) as above.

Other Tree Priors $\pi(\mathcal{T})$. As before, the only modification needed is the bound for $\Pi_{\mathbb{T}}(\mathcal{T})/\Pi_{\mathbb{T}}(\mathcal{T}^+)$. Denote by $K^+ = |\mathcal{T}_{ext}^+|$ and $K = |\mathcal{T}_{ext}|$ and note that $K^+ - K = l_S - l_0$. For the prior from Section 2.1.2, we then have

$$\begin{aligned}\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)} = \lambda^{-(l_S-l_0)} \frac{K^+!\mathbb{C}_{K^+-1}}{K!\mathbb{C}_{K-1}} &\lesssim \left(\frac{\lambda}{4}\right)^{-(l_S-l_0)} \frac{(K^+)!}{K!} \\ &\lesssim \left(\frac{\lambda}{4}\right)^{-(l_S-l_0)} \mathrm{e}^{(l_S-l_0)\{1+\log[K+(l_S-l_0)]\}}.\end{aligned}$$

With $\lambda = n^{-c}$ and $K \leq 2^{\mathcal{L}_c+1}$, this is bounded from above by $C\mathrm{e}^{C\,\log^2 n}$ for some $C > 0$ and the proof is completed as before. For the prior (6), one similarly uses $\Pi_{\mathbb{T}}(\mathcal{T})/\Pi_{\mathbb{T}}(\mathcal{T}^+) = \mathrm{e}^{c(l_S-l_0)\log n} \leq \mathrm{e}^{c\log^2 n}$. □

6.1.3. Posterior Concentration Around Signals

Let us now show that the posterior does not distort large signals too much.

Lemma 4. *Let us denote, for $\mathcal{L}_c$ as in (48) and $S(f_0; A)$ as in (57),*

$$\mathsf{T} = \{\mathcal{T} : d(\mathcal{T}) \leq \mathcal{L}_c,\ S(f_0; A) \subset \mathcal{T}\}. \tag{64}$$

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

$$\int \max_{(l,k)\in\mathcal{T}'_{int}} |\beta_{lk} - \beta^0_{lk}| d\Pi[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}] < C'\sqrt{\frac{\log n}{n}}, \tag{65}$$

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

Proof. A sketch of the proof for the independent prior $\Sigma_{\mathcal{T}} = I$ is as follows: under the conditional posterior, β_{lk} is normally distributed, centered very close to X_{lk} (and hence to β_{lk}^0 on the event $\mathcal{A}$), and the expectation of the maximum of n Gaussians of variance of order $1/n$ is of order $\sqrt{\log n/n}$. A detailed proof including the g-prior case can be found in the Appendix. □

6.2. *Supremum norm Convergence Rate*

Let us write $f_0 = f_0^{\mathcal{L}_c} + f_0^{\backslash\mathcal{L}_c}$, where $f_0^{\mathcal{L}_c}$ the L^2–projection of f_0 onto the first $\mathcal{L}_c$ layers of wavelet coefficients. Under the Hölder condition the equality holds also pointwise and $\|f_0^{\backslash\mathcal{L}_c}\|_\infty \leq \sum_{l>\mathcal{L}_c} 2^{l/2}2^{-l(1/2+\alpha)} \lesssim (\log n/n)^{\alpha/(2\alpha+1)}$.

The following inequality bounds the supremum norm by the ℓ_∞–norm,

$$
\begin{aligned}
\|f - f_0\|_\infty &\leq \sum_{l\geq -1} \max_{0\leq k<2^l} |\beta_{lk} - \beta_{lk}^0| \sum_{0\leq k<2^{-l}} \|\psi_{lk}\|_\infty \\
&\leq |\langle f - f_0, \varphi\rangle| + \sum_{l\geq 0} 2^{l/2} \max_{0\leq k<2^l} |\beta_{lk} - \beta_{lk}^0| = \ell_\infty(f, f_0). \qquad (66)
\end{aligned}
$$

We use the notation $S(f_0; A), \mathsf{T}$ as in (57)–(64) and

$$
\mathcal{E} = \{f_{\mathcal{T},\boldsymbol{\beta}} : \mathcal{T} \in \mathsf{T}\}. \qquad (67)
$$

Using the definition of the event $\mathcal{A}$ from (47), one can write

$$
\begin{aligned}
E_{f_0}\Pi[f_{\mathcal{T},\boldsymbol{\beta}} :\ \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X] &\leq P_{f_0}[\mathcal{A}^c] + E_{f_0}\Pi[\mathcal{E}^c \,|\, X] \\
&+ E_{f_0}\left\{\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{E} :\ \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X]\mathbb{I}_{\mathcal{A}}\right\}. \qquad (68)
\end{aligned}
$$

By Markov's inequality and the previous bound (66),

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

With T as above, the integral in the last display is bounded by, for $l \leq \mathcal{L}_c$,

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

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 $|\beta_{lk}^0| \leq A\log n/\sqrt{n}$.

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 $f_0 \in \mathcal{H}(\alpha, M)$,

$$\sum_{l\leq\mathcal{L}_c} 2^{\frac{l}{2}} \min\left(\max_{0\leq k<2^l} |\beta^0_{lk}|, A\frac{\log n}{\sqrt{n}}\right) \leq \sum_{l\leq L^*} 2^{\frac{l}{2}} A\frac{\log n}{\sqrt{n}} + \sum_{L^*<l\leq\mathcal{L}_c} 2^{\frac{l}{2}} M 2^{-l(\frac{1}{2}+\alpha)}$$
$$\leq C2^{L^*/2}A\frac{\log n}{\sqrt{n}} + C2^{-L^*\alpha} \leq \widetilde{C}2^{-L^*\alpha} \leq c\left(n^{-1}\log^2 n\right)^{\frac{\alpha}{2\alpha+1}}. \quad (69)$$

Using $P_{f_0}[\mathcal{A}^c] + E_{f_0}\Pi[\mathcal{E}^c \mid X] = o(1)$ and Lemma 4, one obtains

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

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

References

[1] Arbel, J., G. Gayraud, and J. Rousseau (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics 40*, 549–570.

[2] Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *The Annals of Statistics 32*, 870–897.

[3] Barbieri, M. M., J. O. Berger, E. I. George, and V. Ročková (2018). The median probability model and correlated variables. arXiv preprint 1807.08336.

[4] Bayarri, M. J., J. O. Berger, A. Forte, and G. García-Donato (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics 40*, 1550–1577.

[5] Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research 13*(1), 1063–1095.

[6] Blanchard, G., C. Schäfer, and Y. Rozenholc (2004). Oracle bounds and exact algorithm for dyadic classification trees. In *Learning theory*, Volume 3120 of *Lecture Notes in Comput. Sci.*, pp. 378–392. Springer, Berlin.

[7] Bleich, J., A. Kapelner, E. I. George, and S. Jensen (2014). Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics 8*(3), 1750–1781.

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

[9] Brown, L. D. and M. G. Low (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics 3*, 2384–2398.

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

[11] Cai, T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *Journal of Multivariate Analysis 99*, 421–436.

[12] Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics 2*, 1281–1299.

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

[14] Castillo, I. (2017). Pólya tree posterior distributions on densities. *Ann. Inst. Henri Poincaré Probab. Stat. 53*(4), 2074–2102.

[15] Castillo, I. and R. Mismer (2019). Spike and slab adaptive Pólya tree posterior densities. Manuscript to be submitted, Chapter 3 of R. Mismer PhD thesis.

[16] Castillo, I. and R. Nickl (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics 41*, 1999–2028.

[17] Castillo, I. and R. Nickl (2014). On the Bernstein–von Mises theorem for nonparametric Bayes proceduress. *The Annals of Statistics 42*, 1941–1969.

[18] Chipman, H., E. George, R. McCulloch, and T. Shively (2016). High-dimensional nonparametric monotone function estimation using BART. arXiv preprint 1612.01619.

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

[20] Chipman, H., E. I. George, and R. E. McCulloch (1997). Bayesian CART model search. *Journal of the American Statistical Association 93*, 935–960.

[21] Chipman, H., E. D. Kolaczyk, and R. McCulloch (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association 92*, 1413–1421.

[22] Cohen, A., I. Daubechies, and P. Vial (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal. 1*(1), 54–81.

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

[24] Devroye, L., A. Mehrabian, and T. Reddad (2018). The total variation distance between high-dimensional Gaussians. arXiv preprint 1810.08693.

[25] Donoho, D. (1997). CART and best-ortho-basis: a connection. *Annals of Statistics 25*, 1870–1911.

[26] Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*, 1200–1224.

[27] Engel, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis 49*, 242–254.

[28] Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Mathematical Statistics 27*, 1119–1140.

[29] Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association 480*, 1318–

1327.

[30] Gey, S. and E. Nédélec (2005). Model selection for CART regression trees. *IEEE Trans. Inf. Th. 51*, 658–670.

[31] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis 74*, 49–68.

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

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

[34] Giné, E. and R. Nickl (2010). Confidence bands in density estimation. *Ann. Statist. 38*(2), 1122–1170.

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

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

[37] Girard, M. and W. Sweldens (1997). A new class of unbalanced Haar wavelets that form an unconditional basis for l_p on general measure spaces. *Journal of Fourier Analysis and Applications 3*, 457–474.

[38] Golub, G. and C. van Loan (1996). *Matrix Computations*. The John Hopkins University Press.

[39] Hahn, P. R., J. S. Murray, and C. Carvalho (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. arXiv preprint 1706.09523.

[40] Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*, 217–240.

[41] Hoffmann, M. and R. Nickl (2011). On adaptive inference and confidence bands. *Annals of Statistics 39*, 2383–2409.

[42] Hoffmann, M., J. Rousseau, and J. Schmidt-Hieber (2015). On adaptive posterior concentration rates. *The Annals of Statistics 43*, 2259–2295.

[43] Hooker, G. and L. Mentch (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research 17*, 841–881.

[44] Leahu, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electronic Journal of Statistics 4*, 373–404.

[45] Linero, A. (2016). Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association 113*, 626–636.

[46] Linero, A. and Y. Yang (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Association 80*, 1087–1110.

[47] Liu, Y., V. Ročková, and Y. Wang (2018). ABC variable selection with Bayesian forests. arXiv preprint 1806.02304.

[48] Murray, J. (2017). Log-linear Bayesian additive regression trees for cate-

gorical and count responses. arXiv preprint 1706.09523.

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

[50] Nickl, R. and K. Ray (2019). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *The Annals of Statistics*. to appear.

[51] Nickl, R. and J. Söhl (2019). Bernstein–von Mises theorems for statistical inverse problems II: compound Poisson processes. *Electronic Journal of Statistics 13*, 3513–3571.

[52] Nickl, R. and B. Szabó (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications 126*, 3913–3934.

[53] Picard, D. and K. Tribouley (2000). Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics 28*, 298–335.

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

[55] Ročková, V. and E. Saha (2019). On theory for BART. *Proceedings of Machine Learning Research:* 22^{nd} *International Conference on Artificial Intelligence and Statistics 89*, 2839–2848.

[56] Ročková, V. and S. van der Pas (2017). Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 1–40. to appear.

[57] Scornet, E., G. Biau, and J. Vert (2015). Consistency of random forests. *Annals of Statistics 43*, 1716–1741.

[58] Scott, C. and R. D. Nowak (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inform. Theory 52*(4), 1335–1353.

[59] Scricciolo, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal. 9*(2), 475–520.

[60] Stanley, R. P. *Enumerative combinatorics. Vol. 2*, Volume 62 of *Cambridge Studies in Advanced Mathematics.*

[61] van der Pas, S. and V. Ročková (2017). Bayesian dyadic trees and histograms for regression. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2086–2096.

[62] van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge: Cambridge University Press.

[63] Varah, J. M. (1975). A lower bound for the smallest singular value of a matrix. *Linear Algebra and Its Applications 11*, 3–5.

[64] Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*, 1228–1242.

[65] Wager, S. and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests. arXiv preprint 1503.06388.

[66] Willett, R. and R. Nowak (2007). Multiscale Poisson intensity and density estimation. *IEEE Transactions on Information Theory 53*, 3171–3187.

[67] Wong, W. H. and L. Ma (2010). Optional Pólya tree and Bayesian inference. *The Annals of Statist. 38*(3), 1433–1459.

[68] Yoo, W., R. Rivoirard, and J. Rousseau (2017). Adaptive supremum norm posterior contraction: wavelet spike-and-slab and anisotropic Besov spaces. arXiv preprint 1708.01909.

[69] Yoo, W. and A. van der Vaart (2017). The Bayes Lepski's method and credible bands through volume of tubular neighborhoods. arXiv preprint 1711.06926.

[70] Yoo, W. W. and S. Ghosal (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics 44*(3), 1069–1102.

[71] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti 6*, 233–243.

7. Appendix

7.1. *Basic Lemmata*

7.1.1. Properties of the pinball matrix (14)

While $A'_{\mathcal{T}} A_{\mathcal{T}}$ is not proportional to an identity matrix (for trees other than flat trees), it *does* have a nested sparse structure which will be exploited in our analysis.

Proposition 1. *Denote with (l_1, k_1) the deepest rightmost internal node in the tree $\mathcal{T}$, i.e. the node $(l,k) \in \mathcal{T}_{int}$ with the highest index $2^l + k$. Let $\mathcal{T}^-$ be a tree obtained from $\mathcal{T}$ by turning (l_1, k_1) into a terminal node. Then*

$$A'_{\mathcal{T}} A_{\mathcal{T}} = \begin{pmatrix} A'_{\mathcal{T}^-} A_{\mathcal{T}^-} + \boldsymbol{v}\boldsymbol{v}' & \mathbf{0} \\ \mathbf{0}' & 2^{l_1+1} \end{pmatrix} \tag{70}$$

for a vector of zeros $\mathbf{0} \in \mathbb{R}^{|\mathcal{T}_{ext}|-1}$ and a vector $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{T}_{ext}|-1}$ obtained from $A_{\mathcal{T}}$ by first deleting its last column and then transposing the last row of this reduced matrix.

Proof. The index (l_1, k_1), by definition, corresponds to the last entry in the vector $\boldsymbol{\beta}_{\mathcal{T}}$. We note that $\mathcal{T}^-_{int} = \mathcal{T}_{int} \backslash \{(l_1, k_1)\}$ and $\mathcal{T}^-_{ext} = \mathcal{T}_{ext} \backslash \{(l_1+1, 2k_1), (l_1+1, 2k_1+1)\} \cup \{(l_1, k_1)\}$. The matrix $A_{\mathcal{T}^-}$ can be obtained from $A_{\mathcal{T}}$ by deleting the last column of $A_{\mathcal{T}}$ and then deleting the last row, further denoted with $\boldsymbol{v}'$. The desired statement (70) is obtained by noting that the last column of $A_{\mathcal{T}}$ (associated with β_{l_1,k_1}) is orthogonal to all the other columns. This is true because (a) this column has only two nonzero entries that correspond to the last two siblings $\{(l_1+1, 2k_1), (l_1+1, 2k_1+1)\}$, (b) the last two rows of $A_{\mathcal{T}}$ differ only in the sign of the last entry because $\{(l_1+1, 2k_1), (l_1+1, 2k_1+1)\}$ are siblings and share the same ancestry with the same weights up to the sign of their immediate parent. Finally, the entry 2^{l_1+1} follows from (13). □

Corollary 1. *Under the prior (17), the coefficient β_{lk} of any internal node (l,k) which has terminal descendants is independent of all the remaining internal coefficients.*

Proof. Follows directly from Prop. 1 after reordering the nodes. □

The following proposition characterizes the eigenspectrum of $A'_{\mathcal{T}} A_{\mathcal{T}}$ which will be exploited in our proofs.

Proposition 2. *The eigenspectrum of $A'_{\mathcal{T}} A_{\mathcal{T}}$ consists of the diagonal entries of $\boldsymbol{D} = \mathrm{diag}(\widetilde{d}_{lk,lk}) = A_{\mathcal{T}} A'_{\mathcal{T}}$ in (15). Moreover, the diagonal entries $\mathrm{diag}(A'_{\mathcal{T}} A_{\mathcal{T}}) = \{d_{lk,lk}\}_{lk \in \mathcal{T}_{int}}$ satisfy $d_{-10,-10} = |\mathcal{T}_{ext}|$ and $d_{lk,lk} = \sum_{j=l+1}^{d(\mathcal{T})} 2^j \sum_{m=0}^{2^{j-1}} \mathbb{I}[\beta_{lk} \in [(0,0) \leftrightarrow (j,m)]_{\mathcal{T}}]$ with $[(0,0) \leftrightarrow (l,k)]_{\mathcal{T}} \equiv \{(0,0), (1, \lfloor k/2^{l-1} \rfloor), \ldots, (l-1, \lfloor k/2 \rfloor)\}$.*

Proof. The first statement follows from (15) and the fact that $A'_{\mathcal{T}} A_{\mathcal{T}}$ and $A_{\mathcal{T}} A'_{\mathcal{T}}$ have the same spectrum, and the second statement from (13). □

7.1.2. Other Lemmata

Lemma 5. *Assume that a square matrix A is diagonally dominant by rows (i.e., $a_{kk} > \sum_{j\neq k}|a_{kj}|$). Then*

$$\|A\|_\infty < \frac{1}{\min_k(|a_{kk}| - \sum_{j\neq k}|a_{kj}|)}.$$

Proof. Theorem 1 in Varah [63]. □

Lemma 6. *For an invertible matrix $M \in \mathbb{R}^{p\times p}$ and $\boldsymbol{v} \in \mathbb{R}^p$ we have*

$$(M^{-1} + \boldsymbol{v}\boldsymbol{v}'/g_n)^{-1} = M - \frac{M\boldsymbol{v}\boldsymbol{v}'M}{g_n + \boldsymbol{v}'M\boldsymbol{v}} \quad \textit{for } g_n > 0.$$

Proof. Follows immediately by direct computation. □

Lemma 7. *Let $\mathbb{C}_K$ denote the number of full binary trees with $K+1$ leaves. Then*

$$\mathbb{C}_K = \frac{(2K)!}{(K+1)!K!} \asymp 4^K/K^{3/2}.$$

Proof. The number $\mathbb{C}_K$ is the Catalan number (see e.g. [60]), which verifies the identity. The second assertion follows from Stirling's formula. □

Lemma 8. *Let $\boldsymbol{Y} \sim \mathcal{N}_K(\boldsymbol{\mu}, \Sigma)$ be a Gaussian random vector. Denote with $\{\sigma_i\}_{i=1}^K = \mathrm{diag}(\Sigma)$, with $\bar{\mu} = \max_{1\leq i\leq K}\mu_i$ and with $\bar{\sigma}^2 = \max_{1\leq i\leq K}\sigma_i^2$ the maximal mean and variance. Then*

$$\mathbb{E}\left[\max_{1\leq i\leq K}|Y_i|\right] \leq \bar{\mu} + \sqrt{2\,\bar{\sigma}^2\log K} + 2\sqrt{2\pi\bar{\sigma}^2}. \tag{71}$$

Proof. We start by noting that $|Y_i| \leq \bar{\mu} + |Y_i - \mu_i|$. Next, one can use the formula, valid for any real μ_i, $c > 0$ and real random variables Y_i,

$$\mathbb{E}\,[\max_{1\leq i\leq K}|Y_i - \mu_i|] \leq c + \sum_{i=1}^K\int_c^\infty \mathbb{P}(|Y_i - \mu_i| > x)dx. \tag{72}$$

Assuming the Gaussian distribution, the integral is of order $\int_c^\infty 2\mathrm{e}^{-x^2/2\sigma_i^2}dx \leq \sqrt{2\pi\sigma_i^2}\,\mathrm{e}^{-c^2/2\sigma_i^2}$. Then (71) follows from (72) by choosing $c = \sqrt{2\bar{\sigma}^2\log K}$. □

Lemma 9 (see, e.g., [24])**.** *For a positive integer d, let $\mu, \mu_1, \mu_2 \in \mathbb{R}^d$ and let $\Sigma, \Sigma_1, \Sigma_2$ be positive definite $d\times d$ matrices. Then the exist universal constants $C_1, C_2 > 0$ such that, for TV the total variation distance,*

$$TV\left(\mathcal{N}(\mu,\Sigma_1), \mathcal{N}(\mu,\Sigma_2)\right) \leq C_1\|\Sigma_1^{-1}\Sigma_2 - I_d\|_F$$

$$TV\left(\mathcal{N}(\mu_1,\Sigma), \mathcal{N}(\mu_2,\Sigma)\right) \leq C_2\frac{\|\mu_1 - \mu_2\|^2}{\sqrt{(\mu_1-\mu_2)'\Sigma(\mu_1-\mu_2)}},$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof. The first inequality follows from Theorem 1.1 in [24] and the second by Theorem 1.2 in [24] (by setting $\Sigma = \Sigma_1 = \Sigma_2$ in their statement). □

7.2. Properties of Unbalanced systems

7.2.1. Proof of Lemma 1

Proof. Let us denote by $\bar{C} = 1/\sqrt{|L_{lk}|^{-1} + |R_{lk}|^{-1}}$. We have

$$
\begin{aligned}
|\beta_{lk}^B| &= \left|\bar{C}\left\{\int_{L_{lk}} \frac{f(x)}{|L_{lk}|}dx - \int_{R_{lk}} \frac{f(x)}{|R_{lk}|}dx\right\}\right| \\
&\leq \frac{\bar{C}}{|L_{lk}|}\int_0^{|L_{lk}|}\left|f(x+l_{lk}) - f\left(b_{lk} + x\frac{|R_{lk}|}{|L_{lk}|}\right)\right|dx.
\end{aligned}
$$

Next, from α-Hölder continuity (25), we have

$$
\left|f(x+l_{lk}) - f\left(b_{lk} + x\frac{|R_{lk}|}{|L_{lk}|}\right)\right| \leq M\left||L_{lk}| + x\left(\frac{|R_{lk}|}{|L_{lk}|} - 1\right)\right|^{\alpha}.
$$

Suppose that $|R_{lk}| > |L_{lk}|$, then for $x \in (0, |L_{lk}|)$ the above can be bounded with $|R_{lk}|^{\alpha}$. When the contrary is true, the bound is $2^{\alpha}|L_{lk}|^{\alpha}$. □

7.2.2. Granularity lemma

Lemma 10. *Denote with* Ψ_A^B *a weakly balanced UH system according to Definition 5. Then for any* $(l,k) \in A$,

$$
\min\{|L_{lk}|, |R_{lk}|\} = \frac{m_{lk}}{2^{l+D}} \qquad \textit{for some} \quad m_{lk} \in \{1, \ldots, C+l\}.
$$

Proof. We prove the statement by induction. First, from the definition of weak balancedness, we have $\min\{|L_{00}|, |R_{00}|\} = 1 - M_{00}/2^D = j/2^D$ (for $j = 2^D - M_{00}$) and by definition this is less than $M_{00}/2^D \leq C/2^D$, so $j \leq C$. Assume now that the statement holds for $l - 1 \geq 0$ and consider a node $(l,k) \in A$ for some $0 \leq k < 2^l$. The union $L_{lk} \cup R_{lk}$ is either $L_{l-1\,\lfloor k/2 \rfloor}$ or $R_{l-1\,\lfloor k/2 \rfloor}$; without loss of generality, suppose it is $R_{l-1\,\lfloor k/2 \rfloor}$. Then, from weak balancedness, we find

$$
\min\{|L_{lk}|, |R_{lk}|\} = |R_{l-1\,\lfloor k/2 \rfloor}| - M_{lk}/2^{l+D}. \tag{73}
$$

If $|R_{l-1\,\lfloor k/2 \rfloor}| \leq |L_{l-1\,\lfloor k/2 \rfloor}|$, we use induction to find $|R_{l-1\,\lfloor k/2 \rfloor}| = j_1/2^{l-1+D}$ for some $j_1 \in \{1, \ldots, C+l-1\}$ and thereby (73) equals $j/2^{l+D}$ for $j = 2j_1 - M_{lk}$. As this is at most $M_{lk}/2^{l+D} = \max\{|L_{lk}|, |R_{lk}|\}$, one deduces $M_{lk} \geq j_1$ and then $j/2^{l+D} \leq j_1/2^{l+D}$ with $j_1 \leq C + l - 1 \leq C + l$. If $|R_{l-1\,\lfloor k/2 \rfloor}| > |L_{l-1\,\lfloor k/2 \rfloor}|$, we again use weak balancedness to write (73) as $j/2^{l+D}$ with $j = 2M_{l-1\,\lfloor k/2 \rfloor} - M_{lk} \leq M_{lk}$, using again $M_{lk}/2^{l+D} = \max\{|L_{lk}|, |R_{lk}|\}$, so that j is again less than $C + l$. The result follows by induction. □

7.2.3. Complexity lemma

Lemma 11. *Consider a* weakly balanced *UH wavelet system* $\Psi_A^B = \{\Psi_{lk}^B : (l,k) \in A\}$ *according to* (43) *and let* $f \in \mathcal{H}_M^\alpha$*. Then the following conditions hold for* $\delta = 3$*, with constants independent of* B*: for any* $(l,k) \in A$

(B1) the basis function ψ_{lk}^B *can be expressed as a linear combination of at most* $C_0 l^\delta$ *Haar functions* ψ_{jk} *for* $j \le l + D$ *and some* $C_0 > 0$*, and*

(B2) there exists $C_1 > 0$ *(depending only on* E, D *from* (43)*) such that* $|\beta_{lk}^B| \le C_1 M l^{\delta/2} 2^{-l(\alpha+1/2)}$.

Proof. First, the function ψ_{lk}^B belongs to $\mathrm{Vect}\{\mathbb{I}_{I_{(l+D)m}} : 0 \le m < 2^{l+D}\}$ and the support of ψ_{lk}^B spans at most $2(E+l)$ of the indicators $\mathbb{I}_{I_{(l+D)m}}$. These indicators can be expressed in terms of at most $l+D$ of ψ_{lk}'s (one per level above $l+D$), which yields an upper bound $2(E+l)(l+D) \asymp l^2$ and thereby (B1) with $\delta = 2$. Second, the balancing condition (43) gives $\max\{|L_{lk}|, |R_{lk}|\} \le (E+l)2^{-l-D}$ which, combined with Lemma 1 implies

$$|\beta_{lk}^B| \le M 2^{\alpha-1}(E+l)^{\alpha+1/2} 2^{-(l+D)(\alpha+1/2)} \le C_1 M l^{3/2} 2^{-l(\alpha+1/2)},$$

by taking the worst case $\alpha = 1$, which proves (B2) with $\delta = 3$. □

7.2.4. The Quantile Example

Lemma 12. *The quantile system* Ψ_A^B *from Example* 4 *is weakly balanced in the sense of Definition* 5 *for balancing constants satisfying* $E = 2 + 3\,C_q 2^{D-1}$*, where* $\|1/g\|_\infty \le C_q$ *and* $\|g\|_\infty < 2^{D-1}/(2E)$.

Proof. Let us start by writing explicitly the intervals L_{lk}, R_{lk}. Assuming without loss of generality that k is odd, i.e. (l,k) is the right child node,

$$|L_{lk}| = b_{lk} - b_{(l-1)\lfloor k/2 \rfloor} = G^{-1}_{L_{max}+D}[(2k+1)/2^{l+1}] - G^{-1}_{L_{max}+D}[(2\lfloor k/2 \rfloor + 1)/2^l],$$
$$|R_{lk}| = b_{(l-2)\lfloor k/4 \rfloor} - b_{lk} = G^{-1}_{L_{max}+D}[(2\lfloor k/4 \rfloor + 1)/2^{l-1}] - G^{-1}_{L_{max}+D}[(2k+1)/2^{l+1}].$$

We first show by contradiction that $\max\{|L_{lk}|, |R_{lk}|\} \ge E/2^{l+D}$ for $E \ge 1$. Let us denote $y_1 = G^{-1}[(2k+1)/2^{l+1}]$, $y_2 = G^{-1}[(2\lfloor k/2 \rfloor + 1)/2^l]$ and $y_3 = G^{-1}[(2\lfloor k/4 \rfloor + 1)/2^{l-1}]$. Assuming $|L_{lk}| < E/2^{l+D}$, one obtains

$$\lfloor 2^{L_{max}+D} y_1 \rfloor - \lfloor 2^{L_{max}+D} y_2 \rfloor < E 2^{L_{max}-l},$$

and thereby $y_1 - y_2 < E 2^{-l-D+1}$. Next, using the fact that k is odd,

$$\begin{aligned} \frac{1}{2^{l+1}} &= |(2k+1)/2^{l+1} - (2\lfloor k/2 \rfloor + 1)/2^l| \\ &= |G(y_1) - G(y_2)| \le \|g\|_\infty |y_1 - y_2| \le \|g\|_\infty 2E 2^{-l-D}, \end{aligned}$$

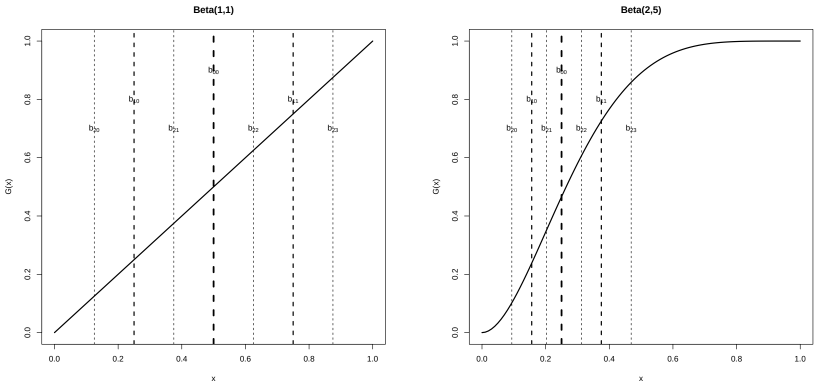
which yields a contradiction when $\|g\|_\infty < 2^{D-1}/(2E)$. Similarly, when $|R_{lk}| < E/2^{l+D}$, one obtains

$$\begin{aligned} 1/2^{l+1} &< |(2k+1)/2^{l+1} - (2\lfloor k/4 \rfloor + 1)/2^{l-1}| \\ &= |G(y_3) - G(y_1)| \le \|g\|_\infty |y_3 - y_1| < \|g\|_\infty 2E\,3\,2^{-l-D}. \end{aligned}$$

Next, we note that for $\|1/g\|_\infty \leq C_q$ and $E \equiv \left(2 + 3\, C_q 2^{D-1}\right)$,

$$\begin{aligned} |R_{lk}| &= \frac{1}{2^{L_{max}+D}} \left[\lfloor 2^{L_{max}+D} y_1 \rfloor - \lfloor 2^{L_{max}+D} y_3 \rfloor \right] \\ &\leq \frac{2}{2^{L_{max}+D}} + \left\| \frac{1}{g} \right\|_\infty \frac{3}{2^{l+1}} \leq \frac{E}{2^{l+D}}. \end{aligned}$$

Similarly, one obtains $|L_{lk}| < E/2^{l+D}$, which concludes that the quantile system is weakly balanced. □



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

FIG 4. *Example of quantile splits for a uniform density $g(x)$ and a non-uniform beta density $g(x)$ using $L_{max} = 6$.*

7.3. Multi-dimensional extensions

Our tree-shaped wavelet reconstruction generalizes to the multivariable case, where a fixed number $d \geq 1$ of covariate directions are available for split. We outline one such generalization using the tensor product of Haar basis functions ψ_{lk} from (3) defined as

$$\Psi_{l\boldsymbol{k}}(\boldsymbol{x}) := \psi_{lk_1}(x_1) \cdots \psi_{lk_d}(x_d)$$

for $l \geq 0$ and $\boldsymbol{k} = (k_1, \ldots, k_d)'$ with $0 \leq k_i \leq 2^l - 1$ for $i = 1, \ldots, d$, where $\Psi_{-1\boldsymbol{0}}(\boldsymbol{x}) = \mathbb{I}_{(0,1]^d}(\boldsymbol{x})$. These wavelet tensor products can be associated with d-ary trees (as opposed to binary trees), where each internal node has 2^d children. The nodes in a d-ary tree satisfy a hierarchical constraint: $(l, \boldsymbol{k}) \in \mathcal{T}, l \geq 1 \Rightarrow (l-1, \lfloor \boldsymbol{k}/2 \rfloor) \in \mathcal{T}$, where the floor operation is applied element-wise. This intuition can be gleaned from Figure 5 which organizes tensor wavelets with $l = 0, 1$ and $d = 2$ in a flat 4-ary tree. We assume that f_0 belongs to α-Hölder

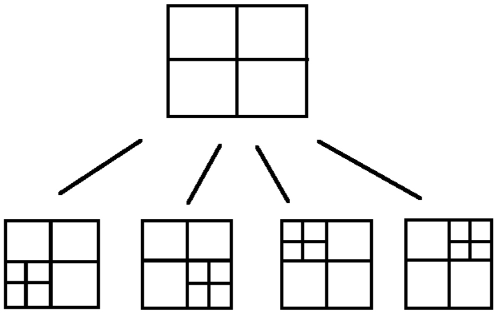


Fig 5. *A plot of tensor Haar wavelets. The top figure plots $\Psi_{0\,(0,0)'}$ and the bottom figures are $\Psi_{1\,(0,0)'}, \Psi_{0\,(1,0)'}, \Psi_{0\,(0,1)'}, \Psi_{0\,(1,1)'}$ (from left to right).*

functions on $[0,1]^d$ for $0 < \alpha \leq 1$ defined as

$$\mathcal{H}_M^{\alpha,d} \equiv \left\{ f \in \mathcal{C}([0,1]^d) : \|f\|_\infty + \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|^\alpha} \leq M \right\}. \tag{74}$$

The multiscale coefficients $\beta_{-1\mathbf{0}} = \langle f_0, \Psi_{-1\mathbf{0}} \rangle$ and

$$\beta_{l\boldsymbol{k}} = \langle f_0, \Psi_{l\boldsymbol{k}} \rangle = \int_{[0,1]^d} f_0(\boldsymbol{x}) \Psi_{l\boldsymbol{k}}(\boldsymbol{x})\, d\boldsymbol{x}.$$

can be verified to satisfy, for some universal constant $C > 0$,

$$|\beta_{l\boldsymbol{k}}| \leq C 2^{-l(\frac{1}{2}+\alpha)d}. \tag{75}$$

Similarly as in Section 2.3, denoting with $\mathcal{T}'_{int}$ the collection of internal nodes $(l, \boldsymbol{k})$ in a d-ary tree (including the node $(-1, \mathbf{0})$), one then obtains a wavelet reconstruction $f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{(l,\boldsymbol{k}) \in \mathcal{T}'_{int}} \beta_{l\boldsymbol{k}} \Psi_{l\boldsymbol{k}}(\boldsymbol{x})$, where coefficients $\beta_{l\boldsymbol{k}}$ can be assigned, for instance, a Gaussian independent product prior. There are several options for defining the d–dimensional version of the prior $\Pi_{\mathbb{T}}$. Restricting to Galton-Watson type priors, the most direct extension, for each node $(l, \boldsymbol{k})$ to be potentially split, either does not split it with probability $1 - \Gamma^{-l}$, or splits it into 2^d children, leading to a full 2^d–ary tree. Another, more flexible option, is to split $(l, \boldsymbol{k})$ into a random number of children inbetween 0 and 2^d, where a split in each specific direction occurs with probability Γ^{-l}, for Γ a large enough constant.

Assuming that d is fixed as $n \to \infty$, the general proving strategy of Theorem 1 can still be applied to conclude ℓ_∞–posterior convergence at the rate $\varepsilon_n = (\log n/n)^{\alpha/(2\alpha+d)} \log^\delta n$ for some $\delta > 0$. The proof requires the threshold $\mathcal{L}_c$ in (48) to be modified as satisfying $2^{\mathcal{L}_c} \asymp (n/\log n)^{1/(2\alpha+d)}$.

7.4. Proof of Lemma 4

Proof. For a given tree $\mathcal{T}$ with $K = |\mathcal{T}_{ext}|$ leaves, we denote by $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}'_{int})'$ 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. We have seen in (22) that, given $\boldsymbol{X}_{\mathcal{T}}$ (so for fixed ε_{lk}) and $\mathcal{T}$, the vector $\boldsymbol{\beta}_{\mathcal{T}}$ has a Gaussian distribution

$$\boldsymbol{\beta}_{\mathcal{T}} \,|\, \boldsymbol{X}_{\mathcal{T}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{T}}, \widetilde{\Sigma}_{\mathcal{T}}),$$

where $\widetilde{\Sigma}_{\mathcal{T}} = (nI_K + \Sigma_{\mathcal{T}}^{-1})^{-1}$ and

$$\boldsymbol{\mu}_{\mathcal{T}} = n\widetilde{\Sigma}_{\mathcal{T}} \left(\boldsymbol{\beta}^0_{\mathcal{T}} + \frac{1}{\sqrt{n}} \boldsymbol{\varepsilon}_{\mathcal{T}} \right).$$

Next, using Lemma 8, we have

$$\mathbb{E}\left[\|\boldsymbol{\beta}_{\mathcal{T}} - \boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty \,|\, \boldsymbol{X}_{\mathcal{T}} \right] \leq \|\boldsymbol{\mu}_{\mathcal{T}} - \boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty + \sqrt{2\,\bar{\sigma}^2 \log K} + 2\sqrt{2\pi\bar{\sigma}^2}, \tag{76}$$

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

$$\|\boldsymbol{\mu}_{\mathcal{T}} - \boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty \leq \sqrt{n}\|\widetilde{\Sigma}_{\mathcal{T}} \boldsymbol{\varepsilon}_{\mathcal{T}}\|_\infty + \|(n\widetilde{\Sigma}_{\mathcal{T}} - I_K)\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty. \tag{77}$$

Using the fact $(I+B)^{-1} = I - (I + B^{-1})^{-1}$, we obtain $n\widetilde{\Sigma}_{\mathcal{T}} - I_K = -(I_K + n\Sigma_{\mathcal{T}})^{-1}$. For the g-prior, we have $\lambda_{max}(\widetilde{\Sigma}_{\mathcal{T}}) < 1/n$ and $\lambda_{max}(I_K + n\Sigma_{\mathcal{T}})^{-1} < \lambda_{max}(A'_{\mathcal{T}}A_{\mathcal{T}})/(ng_n) < 1/g_n$. These inequalities hold trivially also for the independent prior. Next, we note that $\|Bx\|_\infty \leq \|B\|_\infty \|x\|_\infty \leq \sqrt{K}\lambda_{max}(B)\|x\|_\infty$, where K is the number of rows (columns) of a symmetric matrix B and where $\|B\|_\infty$ denotes the matrix sup norm (maximum absolute row sum of the matrix). Assuming $g_n = n$ we can thus write

$$\|(n\widetilde{\Sigma}_{\mathcal{T}} - I_K)\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty \leq \frac{\|\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty \sqrt{K}}{1 + n\lambda_{min}(\Sigma_{\mathcal{T}})} \leq \frac{C\sqrt{K}\lambda_{max}(A'_{\mathcal{T}}A_{\mathcal{T}})}{n\, g_n} \leq C/\sqrt{n}. \tag{78}$$

Next, we note that $\widetilde{\Sigma}^{-1}_{\mathcal{T}}$ is strictly diagonally dominant. This is trivially true for the independent prior and holds also for the g-prior with $g_n = n$ in which case $\widetilde{\Sigma}^{-1}_{\mathcal{T}} = nI_K + \frac{1}{g_n}A'_{\mathcal{T}}A_{\mathcal{T}}$ and $\frac{1}{g_n}\|A'_{\mathcal{T}}A_{\mathcal{T}}\|_\infty \leq \frac{\sqrt{K}}{g_n}\lambda_{max}(A'_{\mathcal{T}}A_{\mathcal{T}}) < \sqrt{n}$. Writing $A'_{\mathcal{T}}A_{\mathcal{T}} = (a_{ij})_{i,j}^{K,K}$, it then follows from Lemma 5 (Theorem 1 in [63]) that

$$\|\widetilde{\Sigma}_{\mathcal{T}}\|_\infty \leq \frac{1}{n + \frac{1}{g_n} \min\limits_{1 \leq k \leq K} \Delta_k}, \quad \text{where} \quad \Delta_k = |a_{kk}| - \sum_{j \neq k} |a_{kj}|. \tag{79}$$

Since $\Delta_k/g_n > -\frac{1}{g_n}\|A'_{\mathcal{T}}A_{\mathcal{T}}\|_\infty > -\sqrt{n}$ and using the fact that $\|\boldsymbol{\varepsilon}_{\mathcal{T}}\|_\infty \lesssim \sqrt{\log n}$ on the event $\mathcal{A}$, we obtain

$$\sqrt{n}\|\widetilde{\Sigma}_{\mathcal{T}}\boldsymbol{\varepsilon}_{\mathcal{T}}\|_\infty \lesssim \sqrt{\frac{\log n}{n}}. \tag{80}$$

The sum of the remaining two terms in (76) can be bounded by a multiple of $\sqrt{\log n/n}$ by noting that $\bar{\sigma}^2 \leq \|\widetilde{\Sigma}_{\mathcal{T}}\|_\infty \lesssim 1/n$. The statement (65) then follows from (76). □

7.5. Proof of Theorem 2

The strategy of the proof of Theorem 1 can be directly applied for an S–regular wavelet basis, also noting that the corresponding $\mathcal{H}(\alpha, M)$ space in (24) contains the usual Hölder–space for $\alpha \le S$, using the properties of S–regular wavelets (see e.g. Section 2.2 in [13]).

We now show how the proof can be modified by assuming a general covariance matrix $\Sigma_{\mathcal{T}}$ on the internal wavelet coefficients. Recall the ratio (51) from Section 6.1.1

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} = \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \frac{\sqrt{|I + n\Sigma_{\mathcal{T}^-}|}}{\sqrt{|I + n\Sigma_{\mathcal{T}}|}} \frac{\mathrm{e}^{n^2 \boldsymbol{X}'_{\mathcal{T}}(nI+\Sigma_{\mathcal{T}}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}}/2}}{\mathrm{e}^{n^2 \boldsymbol{X}'_{\mathcal{T}^-}(nI+\Sigma_{\mathcal{T}^-}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}^-}/2}}. \tag{81}$$

Using the fact that eigenvalues of a principal submatrix interlace the eigenvalues of the original matrix (Theorem 8.1.7 of [38]), we can write

$$\frac{|I + n\Sigma_{\mathcal{T}^-}|}{|I + n\Sigma_{\mathcal{T}}|} \le \frac{1}{1 + n\lambda_{min}(\Sigma_{\mathcal{T}})}.$$

Using the matrix inversion formula $(I + B)^{-1} = I - (I + B^{-1})^{-1}$, we get

$$(nI + \Sigma_{\mathcal{T}}^{-1})^{-1} = \frac{1}{n}\left[I - (I + n\Sigma_{\mathcal{T}})^{-1}\right]$$

and thereby

$$\boldsymbol{X}'_{\mathcal{T}}(nI + \Sigma_{\mathcal{T}}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}} = \frac{1}{n}\|\boldsymbol{X}_{\mathcal{T}}\|_2^2 - \frac{1}{n}\boldsymbol{X}'_{\mathcal{T}}(I + n\Sigma_{\mathcal{T}})^{-1}\boldsymbol{X}_{\mathcal{T}}.$$

Writing $\boldsymbol{X}_{\mathcal{T}} = (\boldsymbol{X}_{\mathcal{T}^-}, X_{l_1k_1})'$, where (l_1, k_1) it the deepest rightmost internal node in $\mathcal{T}$ (as in Section 6.1.1), and $Z \equiv \boldsymbol{X}'_{\mathcal{T}}(nI + \Sigma_{\mathcal{T}}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}} - \boldsymbol{X}'_{\mathcal{T}^-}(nI + \Sigma_{\mathcal{T}^-}^{-1})^{-1}\boldsymbol{X}_{\mathcal{T}^-}$, we have

$$\begin{aligned} Z &= X_{l_1k_1}^2/n - \frac{1}{n}\left[\boldsymbol{X}'_{\mathcal{T}}(I + n\Sigma_{\mathcal{T}})^{-1}\boldsymbol{X}_{\mathcal{T}} - \boldsymbol{X}'_{\mathcal{T}^-}(I + n\Sigma_{\mathcal{T}^-})^{-1}\boldsymbol{X}_{\mathcal{T}^-}\right] \\ &< \frac{X_{l_1k_1}^2}{n}\left(1 - \frac{1}{1 + n\lambda_{max}(\Sigma_{\mathcal{T}})}\right) \\ &\qquad + \frac{\|\boldsymbol{X}_{\mathcal{T}^-}\|_2^2}{n}\left(\frac{1}{1 + n\lambda_{min}(\Sigma_{\mathcal{T}^-})} - \frac{1}{1 + n\lambda_{max}(\Sigma_{\mathcal{T}})}\right). \end{aligned}$$

It follows from the proof of Lemma 2 that $X_{l_1k_1}^2 \lesssim \log n/n$ and $\|\boldsymbol{X}_{\mathcal{T}^-}\|_2^2 \le C_1$ (as was shown in (63)). Moreover, from our assumption (28) we have $\lambda_{min}(\Sigma_{\mathcal{T}^-}) \ge 1/\sqrt{\log n}$ and thereby

$$\begin{aligned} \frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^-)} &< \sqrt{\frac{\log n}{n}} \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \exp\left\{\frac{nX_{l_1k_1}^2}{2} + \frac{n\|\boldsymbol{X}_{\mathcal{T}^-}\|_2^2}{2(1 + n\lambda_{min}(\Sigma_{\mathcal{T}^-}))}\right\} \\ &< \frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^-)} \mathrm{e}^{(C+C_1)\log n}. \end{aligned}$$

Proceeding as in the proof of Lemma 2, one can show (49).

Regarding missing the signal, we use the same strategy as in the proof of Lemma 3. We deploy the interlacing eigenvalue theorem in (60) to obtain the following upper bound for $\frac{N_X(\mathcal{T}^{(s-1)})}{N_X(\mathcal{T}^{(s)})}$ (using matrix determinant and inversion lemmata as before)

$$\sqrt{1+n\lambda_{\max}(\Sigma_{\mathcal{T}^{(s)}})}\exp\left\{-\frac{nX_{l_sk_s}^2}{2}\frac{n\lambda_{min}(\Sigma_{\mathcal{T}^{(s)}})}{1+n\lambda_{min}(\Sigma_{\mathcal{T}^{(s)}})}+\frac{n\|\boldsymbol{X}_{\mathcal{T}^{(s-1)}}\|_2^2}{2(1+n\lambda_{min}(\Sigma_{\mathcal{T}^{(s)}}))}\right\}.$$

Using the expression (60) and assumptions $\lambda_{max}(\Sigma_{\mathcal{T}})\lesssim n^a$ for some $a\geq 1$ and $\lambda_{min}(\Sigma_{\mathcal{T}^{(s)}})\geq 1/\sqrt{\log n}$, we obtain for $C_2,C_3>0$

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^+)}<\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^+)}\exp\left\{C_2(S+1)\sqrt{\log n}-C_3A^2\log^2 n\right\}.$$

Using this bound, one can proceed as in the proof of Lemma 3 and show (112).

For the posterior concentration around signals, we modify the proof of Lemma 4. Similarly as in (78), we find that when $\lambda_{min}(\Sigma_{\mathcal{T}})\geq 1/\sqrt{\log n}$ we have $\|(n\widetilde{\Sigma}_{\mathcal{T}}-I_K)\boldsymbol{\beta}^0_{\mathcal{T}}\|_\infty\leq C\sqrt{\log n/n}$. Next, because

$$\|\Sigma_{\mathcal{T}}^{-1}\|_\infty\leq\sqrt{K}\lambda_{max}(\Sigma_{\mathcal{T}}^{-1})\leq\sqrt{K\log n}<\sqrt{n\log n}$$

the matrix $\widetilde{\Sigma}_{\mathcal{T}}=(nI_K+\Sigma_{\mathcal{T}}^{-1})^{-1}$ is diagonally dominant and thereby (using Lemma 5)

$$\|\widetilde{\Sigma}_{\mathcal{T}}\|_\infty\leq\frac{1}{n+\min\limits_{1\leq k\leq K}\Delta_k}\quad\text{where}\quad\Delta_k=|\sigma_{kk}|-\sum_{j\neq k}|\sigma_{kj}|$$

and where $\Sigma_{\mathcal{T}}^{-1}=(\sigma_{jk})_{j,k=1}^{K,K}$. Since $\Delta_k>-\sqrt{n\log n}$ for all $k=1,\dots,K$, the inequalities (80) and (65) hold. The rest of the proof can be completed using similar arguments as in Section 6.2.

7.6. Proof of Theorem 3

Define a sequence

$$L^*=\left\lceil\log_2\left[M^{1/(\alpha+1/2)}\left(n/\log^2 n\right)^{1/(2\alpha+1)}\right]\right\rceil,$$

so that $2^{L^*}\asymp(n/\log^2 n)^{1/(2\alpha+1)}$. Define the sequence of functions f_0^n (below we write simply f_0 for simplicity) by its sequence of wavelet coefficients as follows: set all coefficients β_{lk}^0 to 0 except for $\beta^0_{L^*0}=M2^{-L^*(1/2+\alpha)}$. By definition, f_0 belongs to $\mathcal{H}(\alpha,M)$. Let us also note that if $(L^*,0)$ does not belong to the tree $\mathcal{T}$, one can bound from below

$$\ell_\infty(f_{\mathcal{T},\boldsymbol{\beta}},f_0)\geq 2^{L^*/2}|\beta_{L^*0}|=M2^{-L^*\alpha}\geq C'\varepsilon_n.$$

So, to prove the result, it is enough to show that $\Pi[(L^*,0)\notin\mathcal{T}_{int}\,|\,X]\to 1$, i.e. the node $(L^*,0)$ does not belong to a tree sampled from the posterior with probability going to 1, or equivalently, if $\mathbb{T}_{L^*0}$ denotes the set of all full binary trees (of depth at most L_{max}) that contain $(L^*,0)$ as an internal node, that $\Pi[\mathbb{T}_{L^*0}\,|\,X]=o_P(1)$. To prove this, let us consider a given tree $\mathcal{T}\in\mathbb{T}_{L^*0}$. As it contains the node $(L^*,0)$, it must also contain all nodes $(\lambda,0)$ with $0\le\lambda\le L^*$, in particular $(L_1,0)$, where $L_1=\lceil L^*/2\rceil$, say. We note that $L^*\asymp L^*-L_1\asymp\log n$. Let τ^* be the maximal subtree of $\mathcal{T}$ that has $(L_1,0)$ as its root. Next, let $\mathcal{T}^*_-$ denote the remainder tree built from $\mathcal{T}$ by erasing all of τ^* except for the node $(L_1,0)$ (so that $\mathcal{T}^*_-$ still has a full-binary tree structure). So, $\mathcal{T}^*_-$ and τ^* have only the node $(L_1,0)$ in common, and the union of their nodes gives the original tree $\mathcal{T}$. Let us now write

$$\Pi[\mathbb{T}_{L^*0}\,|\,X]=\frac{\sum_{\mathcal{T}\in\mathbb{T}_{L^*0}}W_X(\mathcal{T})}{\sum_{\mathcal{T}\in\mathbb{T}}W_X(\mathcal{T})}=\frac{\sum_{\mathcal{T}\in\mathbb{T}_{L^*0}}\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^*_-)}W_X(\mathcal{T}^*_-)}{\sum_{\mathcal{T}\in\mathbb{T}}W_X(\mathcal{T})}.$$

Let $q=q(\tau^*)$ denote the number of internal nodes τ^*_{int} of the subtree τ^*. From the Galton-Watson prior, we obtain

$$\frac{\Pi_{\mathbb{T}}[\mathcal{T}]}{\Pi_{\mathbb{T}}[\mathcal{T}^*_-]}=\prod_{(l,k)\in\tau^*_{int}}p_l\prod_{(l,k)\in\tau^*_{ext}}(1-p_l)\frac{1}{1-p_{L_1}}\le 2\prod_{(l,k)\in\tau^*_{int}}\Gamma^{-l}. \tag{82}$$

Then, by definition of $\mathcal{T}^*_-$ and τ^*,

$$\begin{aligned}\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}^*_-)}&=\frac{\Pi_{\mathbb{T}}(\mathcal{T})}{\Pi_{\mathbb{T}}(\mathcal{T}^*_-)}\prod_{(l,k)\in\tau^*_{int}}\frac{\exp\{(n+1)X_{lk}^2/2\}}{\sqrt{n+1}}\\&\le 2(n+1)^{-q/2}\cdot\prod_{(l,k)\in\tau^*_{int}}\Gamma^{-l}\exp\{(n+1)X_{lk}^2/2\}.\end{aligned} \tag{83}$$

We bound the data-dependent part in the previous line by using $(a+b)^2\le 2a^2+2b^2$. Furthermore, noting that the noise variables $|\varepsilon_{lk}|$ are uniformly bounded for $l+1\le L_{max}, 0\le k<2^l$, by $2\log n$ on an event of overwhelming probability, we can upper-bound $(n+1)\sum_{(l,k)\in\tau^*_{int}}X_{lk}^2/2$ by

$$\begin{aligned}&(n+1)\sum_{(l,k)\in\tau^*_{int}}\left[(\beta^0_{L^*0})^2\mathbb{I}_{(l,k)=(L^*,0)}+\frac{1}{n}\max_{l+1\le L_{max},k}\varepsilon_{lk}^2\right]\\&\le(n+1)(\beta^0_{L^*0})^2+2\frac{n+1}{n}(\log n)q\le\frac{n+1}{n}\log^2 n+2\frac{n+1}{n}(\log n)q.\end{aligned}$$

Now, using that for $(l,k)\in\tau^*_{int}$, we have $l\ge L_1\ge\frac{1}{2\alpha+1}\log_2\left(\frac{M^2n}{\log n}\right)\equiv c(\alpha,M,n)$, one notes that

$$\sum_{(l,k)\in\tau^*_{int}}l\ge\max\left(c(\alpha,M,n)q,\sum_{l=L_1}^{L^*}l\right)\ge c(\alpha,M,n)\max\left(q,\frac{3}{2}c(\alpha,M,n)\right),$$

which is bounded from below by $c(\alpha, M, n)q$, where we have used that $q \geq L^* - L_1 + 1 \geq L_1 \geq c(\alpha, M, n)$ and $L_1 + L^* > 3c(\alpha, M, n)$. One then deduces that the product of terms Γ^{-l} dominates (83), as long as $\log(\Gamma)$ is large enough (noting also that $1/(2\alpha + 1) \geq 1/(2S + 1)$), in the control of $W_X(\mathcal{T})/W_X(\mathcal{T}_-^*)$. That is, for some constant $C > 0$,

$$\frac{W_X(\mathcal{T})}{W_X(\mathcal{T}_-^*)} \leq \exp\{-C(\log n)q\} =: b_q,$$

where the last bound only depends on the number of internal nodes q of τ^*. By coming back to the above bound on the posterior $\Pi[\mathbb{T}_{L^*0} \,|\, X]$, let us split the sum on the numerator as follows. Let $\mathbb{T}_{L^*0}^q$ denote the set of trees $\mathcal{T} = \mathcal{T}_-^* \cup \tau^*$ in $\mathbb{T}_{L^*0}$ such that $|\tau_{int}^*| = q$. Then $\Pi[\mathbb{T}_{L^*0} \,|\, X]$ is bounded by

$$\sum_{q \geq 1} \frac{\sum_{\mathcal{T} \in \mathbb{T}_{L^*0}^q} b_q W_X(\mathcal{T}_-^*)}{\sum_{\mathcal{T} \in \mathbb{T}} W_X(\mathcal{T})} \leq \sum_{q \geq 1} \frac{\sum_{\mathcal{T}_1 \in \mathbb{T}_-^{*q}} a_q b_q W_X(\mathcal{T}_1)}{\sum_{\mathcal{T}_1 \in \mathbb{T}_-^{*q}} W_X(\mathcal{T}_1)},$$

where $\mathbb{T}_-^{*q}$ denotes the set of all possible $\mathcal{T}_-^*$ that can be obtained from $\mathcal{T} \in \mathbb{T}_{L^*0}^q$ and where a_q denotes the number of different possible trees τ^* such that $|\tau_{int}^*| = q$. To obtain the last bound, we also used that each $\mathcal{T} \in \mathbb{T}_{L^*0}$ is uniquely caracterised by a pair $(\mathcal{T}_-^*, \tau^*)$, so that the sum over $\mathcal{T}$ can be rewritten as a double sum over $\mathcal{T}_-^*$ and τ^*. One deduces that, as q cannot be larger than 2^L,

$$\Pi[\mathbb{T}_{L^*0} \,|\, X] \leq \sum_{q=1}^{2^L} a_q b_q.$$

As a_q is less (because of the restriction $|\mathcal{T}| \leq L$) or equal to the number of full binary trees with q internal nodes, i.e. with $2q + 1$ nodes in total, we have $a_q \leq \mathbb{C}_{2q}$, which is bounded from above by 4^{2q} by Lemma 7. We conclude that $\Pi[\mathbb{T}_{L^*0} \,|\, X]$ is bounded above by $\exp(-C \log^2 n)$ for some $C > 0$, on an event of overwhelming probability, which concludes the proof for the Galton-Watson prior. Similarly, for the exponential prior, we replace (82) directly with

$$\frac{\Pi_{\mathbb{T}}[\mathcal{T}]}{\Pi_{\mathbb{T}}[\mathcal{T}_-^*]} \propto \mathrm{e}^{-c(|\mathcal{T}_{ext}| - |\mathcal{T}_{-ext}^*|) \log n} = \mathrm{e}^{-c\, q \log n},$$

where we used the fact that $|\mathcal{T}_{ext}| - |\mathcal{T}_{-ext}^*| = |\tau_{ext}^*| - 1 = |\tau_{int}^*| = q$. For the conditionally uniform prior, we have for $\lambda = 1/n^c$ for some $c > 0$

$$\frac{\Pi_{\mathbb{T}}[\mathcal{T}]}{\Pi_{\mathbb{T}}[\mathcal{T}_-^*]} = \frac{\pi(|\mathcal{T}_{ext}|)}{\pi(|\mathcal{T}_{-ext}^*|)} \frac{\mathbb{C}_{|\mathcal{T}_{-int}^*|}}{\mathbb{C}_{|\mathcal{T}_{int}|}} \lesssim \lambda^q \mathrm{e}^{-q \log |\mathcal{T}_{ext}|} \lesssim \mathrm{e}^{-c\, q \log n},$$

and the end of the proof is then the same as for the GW prior. □

7.7. *Proof of Theorem 4*

This BvM statement can be shown, for example, by verifying the conditions in Proposition 6 of [17] or by proceeding as in the proof of Theorem 3.5 of [54]. The first requirement is the "tightness condition" (Proposition 6 of [17]) summarized by the following lemma.

Lemma 13. *Under the assumptions of Theorem 4, we have*

$$E_{f_0}\Pi(\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_{\mathcal{M}(w)} \geq M_n n^{-1/2} \mid X) \to 0.$$

Proof. Similarly as in Section 6.2, for $j \in \mathbb{N}$ and $f \in L^2[0,1]$ we denote with f^j the L^2 projection onto the first j layers of wavelet coefficients and write $f = f^j + f^{\setminus j}$. Similarly as in the proof of Theorem 1, we denote with $\mathcal{A}$ the event (47) and with $S(f_0; A)$ the set (57). Recall also the notation

$$\mathsf{T} = \{\mathcal{T} : d(\mathcal{T}) \leq \mathcal{L}_c, S(f_0; A) \subset \mathcal{T}\} \quad \text{and} \quad \mathcal{E} = \{f_{\mathcal{T},\boldsymbol{\beta}} : \mathcal{T} \in \mathsf{T}\}$$

from (67), where $\mathcal{E}$ is the subset of tree-based functions $f_{\mathcal{T},\boldsymbol{\beta}}$ with up to $\mathcal{L}_c$ leaves that do not miss any signal. Similarly as in the proof of Theorem 1, we will condition on the event $\mathcal{A}$ and focus on the set $\mathcal{E}$ (as in (68)). For simplicity, we will write $j_0 = j_0(n)$. Following [54], one can write for some suitably chosen $D = D(\eta) > 0$, where $\eta > 0$ is a fixed small constant.

$$\begin{aligned}
&E_{f_0}\left\{\Pi(f_{\mathcal{T},\boldsymbol{\beta}} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_{\mathcal{M}_{(w)}} \geq M_n n^{-1/2} \mid X)\right\} \leq o(1) \\
&\quad + E_{f_0}\left\{\Pi(f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{E} : \|f^{j_0}_{\mathcal{T},\boldsymbol{\beta}} - f^{j_0}_0\|_{\mathcal{M}_{(w)}} \geq D\, n^{-1/2} \mid X)\mathbb{I}_{\mathcal{A}}\right\} && (84)\\
&\quad + E_{f_0}\left\{\Pi(f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{E} : \|f^{\setminus j_0}_{\mathcal{T},\boldsymbol{\beta}} - f^{\setminus j_0}_0\|_{\mathcal{M}_{(w)}} \geq \widetilde{M}_n\, n^{-1/2} \mid X)\mathbb{I}_{\mathcal{A}}\right\}, && (85)
\end{aligned}$$

where $\widetilde{M}_n = M_n - D \to \infty$ as $n \to \infty$. We have for $f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{E}$

$$\|f^{\setminus j_0}_{\mathcal{T},\boldsymbol{\beta}} - f^{\setminus j_0}_0\|_{\mathcal{M}_{(w)}} \leq \sup_{j_0 < l \leq \mathcal{L}_c} \frac{\max_k |\beta_{lk} - \beta^0_{lk}|}{w_l} + \|f_0^{\setminus \mathcal{L}_c}\|_{\mathcal{M}(w)}. \tag{86}$$

From the Hölder property (24) and the definition of $\mathcal{L}_c$ in (48) we have

$$\|f_0^{\setminus \mathcal{L}_c}\|_{\mathcal{M}(w)} = \max_{l > \mathcal{L}_c} \frac{\max_k |\beta^0_{lk}|}{w_l} \lesssim \frac{2^{-\mathcal{L}_c(\alpha+0.5)}}{\sqrt{\mathcal{L}_c}} \leq C/\sqrt{n}, \tag{87}$$

where we used the fact that $\{w_l\}$ is *admissible* in the sense that $w_l/\sqrt{l} \to \infty$ as $l \to \infty$. Using Markov's inequality and bounds (86) and (87), the term (85) can be bounded with

$$E_{f_0}\left\{\frac{\sqrt{n}}{\widetilde{M}_n\, w_{j_0}} \int_{\mathcal{E}} \max_{j_0 < l \leq \mathcal{L}_c} \max_{0 \leq k \leq 2^l} |\beta_{lk} - \beta^0_{lk}| d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \mid X]\mathbb{I}_{\mathcal{A}}\right\} + C/\widetilde{M}_n. \tag{88}$$

Using similar arguments as in Section 6.2 and using Lemma 4, we can upper bound the integral above on the event $\mathcal{A}$ by

$$\begin{aligned}
&\sum_{\mathcal{T} \in \mathsf{T}} \pi[\mathcal{T} \mid X] \int \max_{j_0 < l \leq \mathcal{L}_c} \max_{0 \leq k \leq 2^l} |\beta_{lk} - \beta^0_{lk}| d\Pi[\boldsymbol{\beta}_{\mathcal{T}} \mid X] \\
&\qquad \leq \left(A\frac{\log n}{\sqrt{n}} + C'\sqrt{\frac{\log n}{n}}\right) \lesssim \frac{\log n}{\sqrt{n}}.
\end{aligned}$$

For $w_{j_0} \geq c \log n$ for some $c > 0$, the term (88) goes to zero. Now we focus on the first term (84). By Markov's inequality and using the notation $\mathbb{W} = (g_{lk})$ for the white noise from Section 1 and $\mathbb{X} = (X_{lk})$ for the observation sequence, we find the following upper bound

$$\frac{\sqrt{n}}{D}\left\{E_{f_0}\int_{\mathcal{E}} \|f^{j_0}_{\mathcal{T},\boldsymbol{\beta}} - f^{j_0}_0\|_{\mathcal{M}(w)} d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \mid X]\mathbb{I}_{\mathcal{A}}\right\} \leq \frac{E_{f_0}\left\{\|\mathbb{W}^{j_0}\|_{\mathcal{M}(w)}\mathbb{I}_{\mathcal{A}}\right\}}{D} \tag{89}$$

$$+ \frac{\sqrt{n}}{D} E_{f_0}\left\{\int_{\mathcal{E}} \|\mathbb{X}^{j_0} - f^{j_0}_{\mathcal{T},\boldsymbol{\beta}}\|_{\mathcal{M}(w)} d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \mid X]\mathbb{I}_{\mathcal{A}}\right\}. \tag{90}$$

We can write the second term as

$$\sum_{\mathcal{T}\in\mathsf{T}} \pi[\mathcal{T} \mid X] E_{f_0} \int_{\mathcal{E}} \left(\sup_{l \leq j_0} l^{-1/2} \max_{0\leq k<2^l} \frac{\sqrt{n}}{D}|X_{lk} - \beta_{lk}|\right) d\Pi[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}]. \tag{91}$$

Note that all trees $\mathcal{T} \in \mathsf{T}$ fit j_0 layers and under both the g-prior and the independent prior, the coefficients β_{lk} for $0 \leq l \leq j_0$ are a-priori (and a-posteriori) independent given $\mathcal{T}$. Similarly as in the proof of Theorem 2 in [17], we can show that the term (91) is bounded by a constant by first showing that for each $l \leq j_0$ and $0 \leq k < 2^l$

$$E_{f_0}\left\{\int \mathrm{e}^{t\sqrt{n}(\beta_{lk} - X_{lk})} d\Pi[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}]\mathbb{I}_{\mathcal{A}}\right\} \leq C\mathrm{e}^{t^2/2}. \tag{92}$$

This follows from [17]. The second term $E_{f_0}\left\{\|\mathbb{W}^{j_0}\|_{\mathcal{M}(w)}\mathbb{I}_{\mathcal{A}}\right\}$ is also bounded by $C^\star/D$ for some $C^\star > 0$. Choosing $D = D(\eta) > 0$ large enough, the term on the left side of (89) can be made smaller than $\eta/2$. □

The second step in the proof of Theorem 4 is showing convergence of finite-dimensional distributions (as in Proposition 6 of [17]). Similarly as in the proof of Theorem 2 of [17], convergence of the finite-dimensional distributions can be established by showing BvM for the projected posterior distribution onto $V_j = \mathrm{Vect}\{\psi_{lk}, l \leq j\}$ for any fixed $j \in \mathbb{N}$. Denote with $\boldsymbol{\beta}_j = (\beta_{-10}, \beta_{00}, \ldots, \beta_{j\,2^j-1})'$ the Haar wavelet coefficients up to the level j. The prior on $\boldsymbol{\beta}_j$ consists of $\boldsymbol{\beta}_j \sim \mathcal{N}(0, \Sigma_j)$, where Σ_j is the submatrix of Σ that corresponds to coefficients up to level j.

Let us first consider the case of the independent prior $\Sigma_{\mathcal{T}} = I_K$. Because $j_0(n) \to \infty$, for large enough n we have an *independent product prior* on $\boldsymbol{\beta}_j$ when $\Sigma_j = I$. Then one is exactly in the setting of Theorem 7 of [16] which derives finite-dimensional BvM for product priors (see the paragraph below the statement of Theorem 7 in [16] for two different arguments).

The case of the g–prior $\Sigma_{\mathcal{T}} = g_n(A'_{\mathcal{T}}A_{\mathcal{T}})^{-1}$ is more involved, as the induced prior distribution on the first coordinates is not of product form. Nevertheless, one can express the posterior distribution on coefficients as a mixture over trees (all containing the first $j_0(n)$ layers) of certain $\mathcal{T}$–dependent Gaussian distributions (complemented by zeroes for the coefficients outside the tree $\mathcal{T}$), and study each individual mixture component separately. Let P_{V_j} be the $n \times n$ projection matrix onto V_j and $P_{V_j}^{\mathcal{T}}$ the $|\mathcal{T}_{ext}| \times |\mathcal{T}_{ext}|$ projection matrix onto V_j, projecting the coordinates corresponding to nodes in $\mathcal{T}$ only (recalling that by definition of the prior, all nodes of V_j are in trees $\mathcal{T}$ sampled from the prior). We also denote by I_{V_j} the identity matrix on V_j. It is enough for our needs to show, if $\mathrm{TV}(P, Q)$ denotes the total variation distance between the probability distributions P and Q, that

$$\mathrm{TV}\left(\Pi[\cdot \,|\, X] \circ P_{V_j}^{-1}, R_j^X\right) = o_P(1), \tag{93}$$

where $R_j^X := \mathcal{N}(P_{V_j} X, I_{V_j}/n)$. From the expression of the posterior (22),

$$\beta_{\mathcal{T}} \,|\, \mathcal{T}, X \sim \mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}) =: Q_{\mathcal{T}}^X, \tag{94}$$

where $\mu_{\mathcal{T}}(X) := n\widetilde{\Sigma}_{\mathcal{T}} X_{\mathcal{T}}$ and $\widetilde{\Sigma}_{\mathcal{T}} = (nI_K + \Sigma_{\mathcal{T}})^{-1}$. Further, the coefficients β_{lk} for $(l,k) \notin \mathcal{T}'_{int}$ are zero, which together gives a prior on $\boldsymbol{\beta}_{L-1} \in \mathbb{R}^{2^L} = \mathbb{R}^n$. By definition of the prior distribution, all trees $\mathcal{T}$ sampled from the prior contain the nodes $(l,k), l \leq j_0(n)$, in particular all nodes corresponding to V_j, so (identifying in slight abuse of notation a matrix with its corresponding linear map) the projected posterior $\Pi[\cdot \,|\, X, \mathcal{T}] \circ P_{V_j}^{-1}$ coincides with $\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}) \circ P_{V_j}^{\mathcal{T}\ -1} =: Q_{\mathcal{T},j}^X$. Then

$$\begin{aligned}
&\mathrm{TV}\left(\Pi[\cdot \,|\, X] \circ P_{V_j}^{-1}, R_j^X\right) = \mathrm{TV}\left(\sum_{\mathcal{T}} \Pi[\mathcal{T} \,|\, X] Q_{\mathcal{T},j}^X \,,\, R_j^X\right) \\
&= \mathrm{TV}\left(\sum_{\mathcal{T}} \Pi[\mathcal{T} \,|\, X] Q_{\mathcal{T},j}^X \,,\, \sum_{\mathcal{T}} \Pi[\mathcal{T} \,|\, X] R_j^X\right) \leq \sum_{\mathcal{T}} \Pi[\mathcal{T} \,|\, X] \mathrm{TV}\left(Q_{\mathcal{T},j}^X \,,\, R_j^X\right) \\
&\leq \max_{\mathcal{T}} \mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}) \circ P_{V_j}^{\mathcal{T}\ -1}, \mathcal{N}(X_{\mathcal{T}}, I_K/n) \circ P_{V_j}^{\mathcal{T}\ -1}\right) \\
&\leq \max_{\mathcal{T}} \mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}), \mathcal{N}(X_{\mathcal{T}}, I_K/n)\right),
\end{aligned}$$

where sums and maxima in the last display span over trees that fill in the first $j_0(n)$ layers of nodes, and where the last line uses that the total variation distance can only decrease after projecting onto V_j (one restricts to marginal probabilities in the definition of the t.v. distance). In order to obtain (93), one now needs to bound individual distances given the tree $\mathcal{T}$, in a uniform way with respect to $\mathcal{T}$. By the triangle inequality, for any $\mathcal{T}$ as above,

$$\begin{aligned}
&\mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}), \mathcal{N}(X_{\mathcal{T}}, I_K/n)\right) \\
&\leq \mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}), \mathcal{N}(\mu_{\mathcal{T}}(X), \frac{I_K}{n})\right) + \mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \frac{I_K}{n}), \mathcal{N}(X_{\mathcal{T}}, \frac{I_K}{n})\right).
\end{aligned}$$

Both terms on the right hand side of the last display can be bounded using Lemma 9, where one sets $d = K = |\mathcal{T}_{ext}|$, $\mu = \mu_{\mathcal{T}}(X) = \mu_1$, $\mu_2 = X_{\mathcal{T}}$, and $\Sigma = I_K/n = \Sigma_1$, $\Sigma_2 = \widetilde{\Sigma}_{\mathcal{T}}$. Then $\Sigma_1^{-1}\Sigma_2 - I_K = n(nI_K + \Sigma_{\mathcal{T}}^{-1})^{-1} - I_K = -(I_K + n\Sigma_{\mathcal{T}})^{-1}$, using the formula $(I+B)^{-1} = I - (I+B^{-1})^{-1}$ for B invertible.

Setting $M_{\mathcal{T}} := (I_K + n\Sigma_{\mathcal{T}})^{-1}$, the first and second inequalities of Lemma 9 lead to

$$\mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}}), \mathcal{N}(\mu_{\mathcal{T}}(X), \frac{I_K}{n})\right) \lesssim \|M_{\mathcal{T}}\|_F$$

$$\mathrm{TV}\left(\mathcal{N}(\mu_{\mathcal{T}}(X), \frac{I_K}{n}), \mathcal{N}(X_{\mathcal{T}}, \frac{I_K}{n})\right) \lesssim \frac{\|M_{\mathcal{T}}X_{\mathcal{T}}\|^2}{\sqrt{\frac{1}{n}\|M_{\mathcal{T}}X_{\mathcal{T}}\|^2}} = \sqrt{n}\|M_{\mathcal{T}}X_{\mathcal{T}}\|.$$

One now notes $\|M_{\mathcal{T}}\|_F \leq \sqrt{K}\lambda_{max}(M_{\mathcal{T}}) = \sqrt{K}/\lambda_{min}(I_K + n\Sigma_{\mathcal{T}})$. By Proposition 2, we have $\lambda_{min}((A'_{\mathcal{T}}A_{\mathcal{T}})^{-1})$ is at least $2^{-L} \geq 1/n$, and one deduces that $\lambda_{min}(I_K + n\Sigma_{\mathcal{T}}) \gtrsim 1 + ng_n/n \geq 1 + g_n$, so that $\|M_{\mathcal{T}}\|_F \lesssim \sqrt{K}/g_n \lesssim 2^{L/2}/g_n \lesssim 1/\sqrt{n} = o(1)$, uniformly over $\mathcal{T}$. On the other hand, we have, as $\lambda_{max}(M_{\mathcal{T}}) \leq 1/(1+g_n)$ as seen above and $X_{\mathcal{T}} = \beta^0_{\mathcal{T}} + \varepsilon_{\mathcal{T}}/\sqrt{n}$, so that, working on the event $\mathcal{A}$ from (47),

$$\begin{aligned}\|M_{\mathcal{T}}X_{\mathcal{T}}\|^2 &\leq \lambda_{max}(M_{\mathcal{T}})^2\|X_{\mathcal{T}}\|^2 \lesssim g_n^{-2}(\|\beta^0_{\mathcal{T}}\|^2 + \|\varepsilon_{\mathcal{T}}\|^2/n) \\ &\lesssim g_n^{-2}(1 + n(\log n)/n) \lesssim (\log n)/g_n^2,\end{aligned}$$

where we have used that $\|\beta^0\|^2 = \|f^0\|^2$ is bounded and $\|\varepsilon_{\mathcal{T}}\|^2 \lesssim n\log n$ on the event $\mathcal{A}$. The previous bounds together imply that the total variation distance between $\mathcal{N}(\mu_{\mathcal{T}}(X), \widetilde{\Sigma}_{\mathcal{T}})$ and $\mathcal{N}(X_{\mathcal{T}}, I_K/n)$ goes to 0 uniformly in $\mathcal{T}$ on the event $\mathcal{A}$. As $P[\mathcal{A}^c]$ vanishes, this proves (93).

7.8. *Proof of Theorem 5*

In the proof, we repeatedly use the properties of the median tree $\mathcal{T}^*_X$ established in Lemma 15. We denote by $\mathcal{E}$ the event from Lemma 15. We first show the diameter statement (37). The depth of the median tree estimator $\widehat{f}_T$ verifies condition (i) of Lemma 15 on the event $\mathcal{E}$. For any $f, g \in \mathcal{C}_n$, by definition of $\mathcal{C}_n$, we then have, for $C(\psi)$ a constant depending on the wavelet basis only,

$$\begin{aligned}\|f - g\|_\infty &\leq \|f - \widehat{f}_T\|_\infty + \|\widehat{f}_T - g\|_\infty \\ &\leq 2\sup_{x\in[0,1]} \sum_{l=0}^{L_{max}} v_n\sqrt{\frac{\log n}{n}} \sum_{k=0}^{2^l-1} \mathbb{I}_{(l,k)\in\mathcal{T}^*_X}|\psi_{lk}(x)| \\ &\leq 2v_nC(\psi)\sqrt{\frac{\log n}{n}} \sum_{l:2^l\leq C_1 2^{\mathcal{L}_c}} 2^{l/2} \leq C' v_n\sqrt{\frac{\log n}{n}2^{\mathcal{L}_c}}.\end{aligned}$$

We now turn to the confidence statement. First, one shows that the median estimator (34) is (nearly) rate optimal. Denote with $\widehat{f}_{T,lk} = \langle \widehat{f}_T, \psi_{lk}\rangle$ and let $\mathcal{S} = \{(l,k) :\ |\beta^0_{lk}| \geq A\log n/\sqrt{n}\}$. Let us consider the event

$$B_n = \{\widehat{f}_{T,lk} \neq 0,\ \forall\,(l,k) \in \mathcal{S}\} \cap \{\widehat{f}_{T,lk} = 0,\ \forall\,(l,k) :\ 2^l \geq C_1\,2^{\mathcal{L}_c}\} \cap \mathcal{A}, \quad (95)$$

where the noise-event $\mathcal{A}$ is defined in (47). Lemma 15 together with $P_{f_0}(\mathcal{A}) = 1 + o(1)$ imply that $P_{f_0}(B_n) = 1 + o(1)$. On the event B_n, we have

$$
\begin{aligned}
\|\widehat{f}_T - f_0\|_\infty \le & \sum_{l:\, 2^l \le C_1 2^{\mathcal{L}_c}} 2^{l/2} \max\left(\max_{0\le k<2^l:\, (l,k)\in\mathcal{S}} |X_{lk} - \beta^0_{lk}|, \max_{0\le k<2^l:\, (l,k)\notin\mathcal{S}} \{|\beta^0_{lk}|\} \right) \\
& + \sum_{l:\, 2^l > C_1 2^{\mathcal{L}_c}} 2^{l/2} \max_{0\le k<2^l} |\beta^0_{lk}| \\
& \lesssim 2^{\mathcal{L}_c/2}\sqrt{\frac{\log n}{n}} + \sum_{l:\, 2^l \le C_1 2^{\mathcal{L}_c}} 2^{l/2} \min\left(\max_{0\le k<2^l} |\beta^0_{lk}|, A\frac{\log n}{\sqrt{n}} \right) + 2^{-\alpha\mathcal{L}_c},
\end{aligned}
$$

where we have used the definition of $\mathcal{S}$, that $\widehat{f}_T$ equals 0 or X_{lk}, that f_0 belongs to $\mathcal{H}(\alpha, M)$ and $\max(a,b) \le a+b$ (note also that the term with the minimum in the last display is an upper bound of the maximum over $(l,k) \notin \mathcal{S}$ on the first line of the display). This shows that the median tree estimator is rate-optimal up to a logarithmic factor, in probability under P_{f_0}. In particular, on B_n, we have for some $C > 0$

$$
\|\widehat{f}_T - f_0\|_\infty \le C(\log^2 n/n)^{\alpha/(2\alpha+1)}, \tag{96}
$$

where we used the inequality in (69) in the case of smooth wavelets. Second, we now show that σ_n is appropriately large. By the proof of Proposition 3 of [41], we have for $f_0 \in \mathcal{H}_{SS}(\alpha, M, \varepsilon)$, for $l_n \ge j_0$ suitable sequence chosen later

$$
\sup_{(l,k):\ l\ge l_n} |\beta^0_{lk}| \ge C(M,\psi,\alpha,\varepsilon) 2^{-l_n(\alpha+1/2)},
$$

for some constant $C(M,\psi,\alpha,\varepsilon)$ depending on α, M, the wavelet basis and ε (as in (2.12) of [41]). Let $\Lambda_n(\alpha)$ be defined by, for $\eta > 0$ to be chosen below,

$$
\eta(n/\log^2 n)^{1/(2\alpha+1)} \le 2^{\Lambda_n(\alpha)} \le 2\eta(n/\log^2 n)^{1/(2\alpha+1)}
$$

Combining the previous two displays leads to, for $f_0 \in \mathcal{H}_{SS}(\alpha, M, \varepsilon)$,

$$
\sup_{(l,k):\ l\ge \Lambda_n(\alpha)} |\beta^0_{lk}| \ge C(M,\psi,\alpha,\varepsilon)\eta^{-\alpha-1/2}\frac{\log n}{\sqrt{n}}.
$$

By taking η small enough, one obtains that there exists (λ, κ) with $\lambda \ge \Lambda_n(\alpha)$ verifying $|\beta^0_{\lambda\kappa}| \ge A\log n/\sqrt{n}$ and thus, in turn, $\widehat{f}_{T,\lambda\kappa} \ne 0$, by the second part of Lemma 15. One deduces that the term $(l,k) = (\lambda,\kappa)$ in the sum defining σ_n is nonzero on the event B_n, so that

$$
\sigma_n \ge v_n c(\psi)\sqrt{\frac{\log n}{n}}\|\psi_{\lambda\kappa}\|_\infty \ge v_n c(\psi)\sqrt{\frac{\log n}{n}} 2^{\Lambda_n(\alpha)/2}.
$$

This leads to

$$
\sigma_n \ge c' \frac{v_n}{(\log n)^{1/2}} \left(\frac{\log^2 n}{n}\right)^{\alpha/(2\alpha+1)}. \tag{97}
$$

The ratio in the last display goes to infinity for v_n of larger order than $\log^{1/2} n$. Now, on the event B_n, one can thus write $\|\widehat{f}_T - f_0\|_\infty \leq \sigma_n/2$ for large enough n, uniformly over $f_0 \in \mathcal{H}_{SS}(\alpha, M, \varepsilon)$, implying that $B_n \subset \{\|\widehat{f}_T - f_0\|_\infty \leq \sigma_n\}$. This implies the desired coverage property, since

$$P_{f_0}(f_0 \in C_n) = P_{f_0}\left(\{\|f_0 - \mathbb{X}\|_{\mathcal{M}(w)} \leq R_n/\sqrt{n}\} \cap B_n\right) + o_{P_{f_0}}(1) = 1-\gamma+o_{P_{f_0}}(1),$$

where we used Theorem 5 of [17].

For the credibility statement, we only need to show that the second constraint in $\mathcal{C}_n$ is satisfied with posterior probability tending to one, since $\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - \mathbb{X}\|_{\mathcal{M}(w)} \leq R_n/\sqrt{n} \mid X] = 1-\gamma$ by definition of R_n. In addition, we note that the posterior distribution (from Theorem 2 part a)) and the median estimator $\widehat{f}_T$ (from (96)) converge at a rate strictly faster than σ_n on the event B_n, using again the lower bound on σ_n in (97). In particular, because $B_n \subset \{\|\widehat{f}_T - f_0\|_\infty \leq \sigma_n\}$ we can write

$$E_{f_0}\left(\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - \widehat{f}_T\|_\infty \leq \sigma_n \mid X]\right) \geq E_{f_0}\left(\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty \leq \sigma_n/2 \mid X]\mathbb{I}_{B_n}\right) + o_{P_{f_0}}(1).$$

The right side converges to 1 in P_{f_0}-probability, which concludes the proof of the theorem.

Lemma 14. *The set of nodes $\mathcal{T}_X^*$ in (33) P_{f_0}-almost surely defines a binary tree.*

Proof. Let us recall that $\mathbb{T}$ is the set of all admissible trees that can be obtained by sampling from the prior $\Pi_{\mathbb{T}}$ and with depth at most L_{max}. For any given node (l_1, k_1) with $0 \leq l_1 \leq L_{max}$, one can write

$$\begin{aligned}\Pi[(l_1,k_1) \in \mathcal{T} \mid X] &= \sum_{\mathcal{T}_1 \in \mathbb{T}} \Pi_{\mathbb{T}}[\mathcal{T}_1 \mid X] \times \Pi[(l_1,k_1) \in \mathcal{T}_1 \mid X, \mathcal{T} = \mathcal{T}_1] \\ &= \sum_{\mathcal{T}_1 \in \mathbb{T}:\ (l_1,k_1) \in \mathcal{T}_1} \Pi_{\mathbb{T}}[\mathcal{T}_1 \mid X].\end{aligned}$$

Let (l_1-1, k_1^-) denote the parent node of (l_1,k_1) in $\mathcal{T}_1$, where $k_1^- = \lfloor k_1/2 \rfloor$. Any (full-)binary tree that contains (l_1,k_1) must also contain (l_1-1,k_1^-), so that, using the formula in the last display, $\Pi[(l_1,k_1) \in \mathcal{T} \mid X] \leq \Pi[(l_1-1,k_1^-) \in \mathcal{T} \mid X]$. This implies, by definition of $\mathcal{T}_X^*$, that if a given node (l_1,k_1) belongs to $\mathcal{T}_X^*$, so does the node (l_1-1,k_1^-). Therefore $\mathcal{T}_X^*$ is a tree. □

Lemma 15. *Consider a prior distribution Π as in Theorem 1. There exists an event $\mathcal{E}$ such that $P_{f_0}[\mathcal{E}] = 1+o(1)$ on which the tree $\mathcal{T}_X^*$ defined in (33) has the following properties: there exists a constant $C_1 > 0$ such that*

(i) the depth of the tree satisfies $2^{d(\mathcal{T}_X^*)} \leq C_1 2^{\mathcal{L}_c} \asymp (n/\log n)^{\alpha/(2\alpha+1)}$, where $\mathcal{L}_c$ is as in (48),

(ii) the tree contains as interior nodes all nodes (l,k) that satisfy $|\beta^0_{lk}| \geq A \log n/\sqrt{n}$, for some $A > 0$.

Proof. We focus on the GW-prior, the proof for the other two priors $\Pi_{\mathbb{T}}$ being similar. Let $\mathbb{T}^{(1)}$, respectively $\mathbb{T}^{(2)}$, denote the set of binary trees that satisfy condition (i), respectively (ii), in the statement of the lemma. By the proof of Theorem 1, $\Pi[\mathbb{T}^{(1)} \,|\, X]$ and $\Pi[\mathbb{T}^{(2)} \,|\, X]$ both tend to 1 in probability under P_{f_0}, hence also $\Pi[\mathbb{T}^{(1)} \cap \mathbb{T}^{(2)} \,|\, X]$. In fact, it also follows from the proof of Theorem 1 that, for $\mathbb{T}^{(1)}$, we also have the stronger estimate $\Pi[d(\mathcal{T}) > d \,|\, X] \leq 2^{-c_1 d \log \Gamma}$ for some $c_1 > 0$ under the GW process prior, uniformly over $\mathcal{L}_c < d \leq L_{max}$, on an event $\mathcal{A}$ of P_{f_0}-probability going to 1. The latter probability is $o(2^{-d})$ provided Γ is chosen large enough, which will be used below. Defining $\mathcal{E} = \{\Pi[\mathbb{T}^{(1)} \cap \mathbb{T}^{(2)} \,|\, X] \geq 3/4\}$, we have $P_{f_0}[\mathcal{E}] \to 1$ as $n \to \infty$. For any node (l_2, k_2) such that $|\beta^0_{l_2 k_2}| \geq C \log n/\sqrt{n}$, we have

$$\Pi[(l_2, k_2) \in \mathcal{T}_{int} \,|\, X] = \sum_{\mathcal{T}_2 \in \mathbb{T}:\ (l_2,k_2) \in \mathcal{T}_{2\,int}} \Pi[\mathcal{T}_2 \,|\, X] \geq \Pi[\mathbb{T}^{(2)} \,|\, X],$$

where we used that any tree in $\mathbb{T}^{(2)}$ must, by definition, contain (l_2, k_2). As $\Pi[\mathbb{T}^{(2)} \,|\, X] \geq 3/4 > 1/2$, we deduce that (l_2, k_2) belongs to $\mathcal{T}^*_X$ on the event $\mathcal{E}$. In other words, $\mathcal{T}^*_X$ verifies the second property (ii) of the lemma on $\mathcal{E}$. To conclude the proof of the lemma, one observes that on $\mathcal{E}$, for a given node (l_3, k_3) with $2^{l_3} > C_1 2^{\mathcal{L}_c}$,

$$\Pi[(l_3, k_3) \in \mathcal{T}_{int} \,|\, X] \leq \Pi[d(\mathcal{T}) > l_3 \,|\, X],$$

Recall that $\Pi[d(\mathcal{T}) > l_3 \,|\, X] \leq C 2^{-c_1 l_3 \log \Gamma}$ on $\mathcal{A}$ (which holds uniformly over $l_3 \in [\mathcal{L}_c, L_{max}]$). Then, on the event $\mathcal{A}$, we can write

$$\begin{aligned} P_{f_0}[\{\mathcal{T}^*_X \notin \mathbb{T}^{(1)}\} \cap \mathcal{A}] &\leq P_{f_0}[\{\exists\, (l_3, k_3):\ 2^{l_3} > C_1 2^{\mathcal{L}_c},\ (l_3, k_3) \in \mathcal{T}^*_X\} \cap \mathcal{A}] \\ &\leq \sum_{l_3:\ 2^{l_3} > C_1 2^{\mathcal{L}_c}}^{L_{max}} \sum_{k_3=0}^{2^{l_3}-1} P_{f_0}\left[\{\Pi[(l_3, k_3) \in \mathcal{T}_{int} \,|\, X] \geq 1/2\} \cap \mathcal{A}\right] \\ &\leq \sum_{l_3:\ 2^{l_3} > C_1 2^{\mathcal{L}_c}}^{L_{max}} 2^{l_3+1} E_{f_0}\left[\Pi[d(\mathcal{T}) > l_3 \,|\, X] \mathbb{I}_{\mathcal{A}}\right] = o(1). \end{aligned}$$

Using that $P_{f_0}[\mathcal{A}]$ goes to 1, one obtains $P_{f_0}[\mathcal{T}^*_X \notin \mathbb{T}^{(1)}\}] = o(1)$, which concludes the proof. □

7.9. Proof of Theorem 6

Denote with $\mathcal{T}^F_D$ the flat tree of depth $D+1$ (i.e. all β_{lk}'s for $l \leq D$ are active). The formula (23) gives

$$\begin{aligned} \Pi[\mathcal{T}^F_D \,|\, X] \propto W_X(\mathcal{T}^F_D) &= \Pi_{\mathbb{T}}(\mathcal{T}^F_D) \prod_{(l,k) \in \mathcal{T}^{F\,\prime}_{D\,int}} \frac{\mathrm{e}^{\frac{n^2}{2(n+1)} X^2_{lk}}}{\sqrt{n+1}} \\ &\propto \exp\left\{ -\log \Pi_{\mathbb{T}}(\mathcal{T}^F_D) - 2^{\mathcal{D}} \log(n+1) + \frac{n^2}{2(n+1)} \|\boldsymbol{X}^{(D)}\|_2^2 \right\}, \end{aligned}$$

where $\|\boldsymbol{X}^{(D)}\|_2^2 = \sum_{l\leq D,k} X_{lk}^2$ is the squared L^2–norm of the signal, truncated at the level D. Next, we have

$$\begin{aligned}\|\boldsymbol{X}^{(D)}\|_2^2 &= \sum_{l\leq D,k}(\beta_{lk}^0)^2 + \sum_{l\leq D,k}\frac{2}{\sqrt{n}}\varepsilon_{lk}\beta_{lk}^0 + \sum_{l\leq D,k}\frac{1}{n}\varepsilon_{lk}^2,\\ &= C_n - \sum_{D<l\leq L_{max},k}(\beta_{lk}^0)^2 - \sum_{D<l\leq L_{max},k}\frac{2}{\sqrt{n}}\varepsilon_{lk}\beta_{lk}^0 + \sum_{l\leq D,k}\frac{1}{n}\varepsilon_{lk}^2,\end{aligned}$$

where $C_n = C(n,\{\varepsilon_{lk}\},f_0)$ does not depend on D. We can also write

$$\|\boldsymbol{X}^{(D)}\|^2 = C_n - \sum_{D<l\leq L_{max},k}(\beta_{lk}^0)^2 + \frac{2^{D+1}}{n} - \frac{2}{\sqrt{n}}Z(D) + \frac{1}{n}Q(D), \qquad (98)$$

where we have used $\sum_{l\leq D,k} 1 = 2^{D+1}$ and have set

$$Z(D) = \sum_{D<l\leq L_{max},k}\beta_{lk}^0\varepsilon_{lk}, \qquad Q(D) = \sum_{l\leq D,k}(\varepsilon_{lk}^2 - 1).$$

Let D^* be an integer defined as, for f_0 to be chosen below,

$$D^* = \underset{0\leq D\leq n}{\operatorname{argmin}}\left[2^D\log(n+1) + \frac{n}{2}\sum_{l=D+1}^{L_{max}}\sum_k(\beta_{lk}^0)^2\right].$$

Consider the following true signal $f_0 = f_0^*$ which belongs to $\mathcal{H}(\alpha, M)$ with $M = 1$ (which we can assume without loss of generality) and which is characterized by the following wavelet coefficients

$$\beta_{lk}^{0*} = \begin{cases} 2^{-l(\frac{1}{2}+\alpha)} & \text{if } k = 0,\\ 0 & \text{otherwise.}\end{cases} \qquad (99)$$

For such a signal, D^* above has the following behavior

$$2^{D^*} \asymp \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+2}}. \qquad (100)$$

With the maximum-type norm ℓ_∞ defined in (66), we use the decomposition $\ell_\infty(f,f_0) = \ell_\infty(f,f_0^D) + \ell_\infty(f_0^D,f_0)$, where f_0^D is the L^2–projection of f_0 onto the first D levels of wavelet coefficients. Moreover, using $\ell_\infty(f_0^D,f_0) = c\,2^{-D\alpha}$ for some $c > 0$, we can write, for $\rho_n = (\log n/n)^{\alpha/(2\alpha+2)}$

$$\begin{aligned}\Pi[\ell_\infty(f,f_0) < \mu\rho_n \,|\, X] &\leq \Pi[\ell_\infty(f_0,f_0^D) < \mu\rho_n \,|\, X]\\ &= \Pi[c2^{-D\alpha} < \mu\rho_n \,|\, X] = \Pi[2^D > (c\mu^{-1}\rho_n^{-1})^{1/\alpha} \,|\, X]\\ &\leq \Pi[2^D > (c\mu^{-1})^{1/\alpha}2^{D^*} \,|\, X].\end{aligned}$$

To conclude, it is enough to show that for $B = \{2^D > (c\mu^{-1})^{1/\alpha}2^{D^*}\}$, where $\mu > 0$ is a small enough constant, we have $\Pi[B \,|\, X] = o(1)$ or, equivalently,

$\Pi[B \mid X] = o(\Pi[B^c \mid X])$ (possibly on an event of vanishing probability). Rewriting $B = \{D : \; D > cD^*\}$ for $c = c(\mu) \geq 1$ (up to taking μ small enough), and using the above expression of $\Pi[\mathcal{T}_D^F \mid X]$, one obtains

$$\frac{\Pi[B \mid X]}{\Pi[B^c \mid X]} = \frac{\sum_{D>cD^*} \exp\left\{-\log \Pi(\mathcal{T}_D^F) - 2^D \log(n+1) + \frac{n^2}{2(n+1)} \|\boldsymbol{X}^{(D)}\|_2^2\right\}}{\sum_{D \leq cD^*} \exp\left\{-\log \Pi(\mathcal{T}_D^F) - 2^D \log(n+1) + \frac{n^2}{2(n+1)} \|\boldsymbol{X}^{(D)}\|_2^2\right\}}$$
$$\leq \frac{\sum_{D>cD^*} \exp\left\{-\log \Pi(\mathcal{T}_D^F) - 2^D \log(n+1) + \frac{n^2}{2(n+1)} \|\boldsymbol{X}^{(D)}\|_2^2\right\}}{\exp\left\{-\log \Pi(\mathcal{T}_{D^*}^F) - 2^{D^*} \log(n+1) + \frac{n^2}{2(n+1)} \|\boldsymbol{X}^{(D^*)}\|_2^2\right\}}.$$

Since $c \geq 1$ we have $D \geq D^* + 1$ for any $D > cD^*$ and from the monotonicity assumption on the prior we obtain $\log \Pi(\mathcal{T}_{D^*}^F) - \log \Pi(\mathcal{T}_D) \leq 0$ on B. In addition, note that $2^{D^*} - 2^D \leq -2^D/2$ on B, which implies

$$(2^{D^*} - 2^D) \log(n+1) \leq -\frac{1}{2} 2^D \log(n+1).$$

Going further, using the decomposition of $\|\boldsymbol{X}^{(D)}\|_2^2$ in (98) we have for $Z = \|\boldsymbol{X}^{(D)}\|_2^2 - \|\boldsymbol{X}^{(D^*)}\|_2^2$ the following

$$Z = \sum_{D^* < l \leq D, k} (\beta_{lk}^0)^2 + \frac{1}{n}(2^{D+1} - 2^{D^*+1}) - \frac{2}{\sqrt{n}}(Z(D) - Z(D^*)) + \frac{1}{n}(Q(D) - Q(D^*))$$
$$\leq \sum_{D^* < l \leq L_{max}, k} (\beta_{lk}^0)^2 + \frac{2^{D+1}}{n} + \frac{2}{\sqrt{n}}(|Z(D)| + |Z(D^*)|) + \frac{1}{n}(|Q(D)| + |Q(D^*)|).$$

We now provide bounds for the stochastic terms Z and Q. First, for any $D > D^*$, denoting $\sigma_D^2 := \sum_{D < l \leq L_{max}, k} (\beta_{lk}^0)^2$, we have

$$|Z(D)| \leq \sigma_D \max_{D^* \leq D \leq L_{max}} \sigma_D^{-1} |Z(D)|.$$

The variables $Z(D)/\sigma_D$ are standard normal, which implies that, on some event A_1 such that $P_{f_0}(A_1^c) = o(1)$, we have, uniformly in $D \in B$,

$$|Z(D)| \leq \sigma_D \sqrt{2 \log L_{max}}.$$

To bound the term $Q(D)$, one can use the following standard concentration bound for chi-square distributions. Namely, for ξ_q standard normal variables and any $t > 0$, we can write

$$\mathbb{P}\left[\sum_{q=1}^{Q} (\xi_q^2 - 1) \geq t\right] \leq \exp\left\{-\frac{t^2}{4(Q+t)}\right\}.$$

Applying this bound for the noise variables ε_{lk} and choosing $t = t_D := (D2^D)^{1/2}$ leads to

$$\mathbb{P}\left[\sum_{l \leq D, k} (\varepsilon_{lk}^2 - 1) > t_D\right] \leq \exp\left\{-\frac{D2^D}{4(2^{D+1} + t_D)}\right\}.$$

For $D \geq D^*$, one has $t_D \leq 2^{D+1}$ so the last display is bounded from above by $\exp\{-C_1 D\}$. Let us consider the event, with t_D as above,

$$A_2 = \bigcap_{D=D^*}^{L_{max}} \left\{ \sum_{l \leq D} \sum_{k=0}^{2^l-1} (\varepsilon_{lk}^2 - 1) \leq t_D \right\}.$$

A union bound gives $P_{f_0}[A_2^c] \leq C\exp(-c_1 D^*)$, which is a $o(1)$ using the previous bound. Now let us choose μ small enough in such a way that $C_2 2^{D^*} \leq 2^D/2$ for any D in the set B defined above (this is possible by definition of B) and thereby

$$\frac{n}{2} \sum_{D^* < l \leq L_{max}, k} (\beta_{lk}^0)^2 \leq \frac{2^D}{4} \log(n+1)$$

for any D in B. This in particular implies that $\sigma_D \leq (2^D \log(n+1)/n)^{1/2}$. Now, on the event $A_1 \cap A_2$, we have

$$\begin{aligned}
\frac{\Pi[B \mid X]}{\Pi[B^c \mid X]} \leq \sum_{D > cD^*} \exp\Big\{ & -\frac{1}{2} 2^D \log(n+1) + \frac{2^D}{4} \log(n+1) + 2^D \\
& + 2\sqrt{n}\sigma_D \sqrt{2 \log L_{max}} + (D2^D)^{1/2} \Big\} \\
\leq \sum_{D > cD^*} \exp\Big\{ & -\frac{1}{8} 2^D \log(n+1) \\
& + \Big[2^D + 2\sqrt{2^D \log(n+1) 2 \log L_{max}} + (D2^D)^{1/2} - \frac{1}{8} 2^D \log(n+1) \Big] \Big\} \\
\leq \sum_{D > cD^*} \exp\Big\{ & -\frac{1}{8} 2^D \log(n+1) \Big\} \leq \exp\Big\{ -C2^{D^*} \log(n+1) \Big\},
\end{aligned}$$

where we have used that the term under brackets in the second inequality is negative for large enough n, as $2^D \gtrsim 2^{D^*}$ goes to infinity. This shows that the last display goes to 0, which concludes the proof.

7.10. *Proof of Theorem 7*

As the breakpoints verify the balancing condition (43), they verify the properties (B1)–(B2) in the complexity Lemma 11 for $\delta = 3$. The Gaussian white noise model projects onto the Haar system $\Psi_A^B = \{\psi_{-10}^B, \psi_{lk}^B : (l,k) \in A)\}$ as follows:

$$X_{lk}^B = \beta_{lk}^{0B} + \frac{1}{\sqrt{n}} \varepsilon_{lk}^B, \tag{101}$$

where $X_{lk}^B = \langle X, \psi_{lk}^B \rangle$, $\beta_{lk}^{0\,B} = \langle f_0, \psi_{lk}^B \rangle$ and $\varepsilon_{lk}^B = \langle W, \psi_{lk}^B \rangle$. As the functions ψ_{lk}^B form an orthonormal system, the variables ε_{lk}^B are iid standard Gaussian given B. The observations here are viewed as the collection of X_{lk}^B variables which depend on B. We regard the breakpoints B as one extra "variable" in the model. Given the breakpoints B, we use the same notation $\boldsymbol{X}_{\mathcal{T}}^{B}$ and $\boldsymbol{\varepsilon}_{\mathcal{T}}^{B}$ for the ordered responses and noise variables (as in the proof of Theorem 1). Similarly, $\boldsymbol{\beta}_{\mathcal{T}} = \boldsymbol{\beta}_{\mathcal{T}}^{B}$ are the ordered internal wavelet coefficients.

As the priors on breakpoints B and trees $\mathcal{T}$ are *independent*, the tree posterior remains relatively tractable where the amount of signal at each location (l,k) now depends on B, which requires a separate "*uniform* in B" treatment.

The Multiscale Posterior Distribution. To determine the posterior distribution on f, it is enough to consider the posterior on wavelet coefficients (β_{lk}), which then induces a posterior on f via

$$f_{\mathcal{T},\widetilde{\boldsymbol{\beta}}}^{B}(x) = \sum_{(l,k)\in\mathcal{T}_{ext}} \widetilde{\beta}_{lk}^{B} I_{lk}^{B}(x) = \sum_{(l,k)\in\mathcal{T}'_{int}} \beta_{lk}^{B}\psi_{lk}^{B}(x). \tag{102}$$

Again, the *internal unbalanced* Haar wavelet coefficients $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk}^{B} : (l,k) \in \mathcal{T}'_{int})$ are linked to the *external* histogram coefficients $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} = (\widetilde{\beta}_{lk}^{B} : (l,k) \in \mathcal{T}_{ext})$ through $\widetilde{\boldsymbol{\beta}}_{\mathcal{T}} = A_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}$ for some sparse matrix $A_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}_{ext}|\times|\mathcal{T}_{ext}|}$ (a generalization of (14)). This section describes the posterior distribution over coefficients (β_{lk}) driven by the prior distribution

$$\begin{aligned} (B,\mathcal{T}) &\sim \Pi_{\mathbb{B}} \otimes \Pi_{\mathbb{T}} \\ (\beta_{lk})_{l\le L,k} \,|\, B,\mathcal{T} &\sim \pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k)\notin\mathcal{T}'_{int}} \delta_0(\beta_{lk}), \end{aligned} \tag{103}$$

where $L = L_{max} = \lfloor \log_2 n \rfloor$. From the white noise model, we have, given B,

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

The joint density of $(B,\mathcal{T},(\beta_{lk})_{l\le L,k},X)$ arising from the above distributions equals

$$\begin{aligned} &\Pi_{\mathbb{T}}(\mathcal{T})\Pi_{\mathbb{B}}(B)\pi(\boldsymbol{\beta}_{\mathcal{T}}) \left[\prod_{(l,k)\in\mathcal{T}'_{int}} \phi_{\frac{1}{\sqrt{n}}}(X_{lk}^{B} - \beta_{lk})\right] \left[\prod_{(l,k)\notin\mathcal{T}'_{int}} \phi_{\frac{1}{\sqrt{n}}}(X_{lk}^{B} - \beta_{lk})\mathbb{I}_0(\beta_{lk})\right] \\ &= \Pi_{\mathbb{T}}(\mathcal{T})\Pi_{\mathbb{B}}(B) \left[\prod_{l\le L,k} \phi_{\frac{1}{\sqrt{n}}}(X_{lk}^{B})\right] \left[\prod_{(l,k)\notin\mathcal{T}'_{int}} \mathbb{I}_0(\beta_{lk})\right] \mathrm{e}^{-\frac{n}{2}\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 + n\boldsymbol{X}_{\mathcal{T}}^{B\prime}\boldsymbol{\beta}_{\mathcal{T}}}\pi(\boldsymbol{\beta}_{\mathcal{T}}). \end{aligned}$$

Integrating out (β_{lk}), one obtains the marginal density of $(B,\mathcal{T},X)$ as

$$\left[\Pi_{\mathbb{B}}(B)\prod_{l\le L,k} \phi_{\frac{1}{\sqrt{n}}}(X_{lk}^{B})\right] \Pi_{\mathbb{T}}(\mathcal{T})N_X^B(\mathcal{T}), \tag{104}$$

where

$$N_X^B(\mathcal{T}) = \int \prod_{(l,k)\in\mathcal{T}'_{int}} \mathrm{e}^{nX_{lk}^{B}\beta_{lk} - n\beta_{lk}^2/2} d\pi(\boldsymbol{\beta}_{\mathcal{T}}).$$

The first bracket in (104) only depends on B and X, from which one deduces that the posterior distribution of $\mathcal{T}$, given B and X, satisfies

$$\Pi[\mathcal{T} \mid B, X] = \frac{W_X^B(\mathcal{T})}{\sum_{\mathcal{T}\in\mathbb{T}_L} W_X^B(\mathcal{T})}, \quad \text{with} \quad W_X^B(\mathcal{T}) = \Pi_{\mathbb{T}}(\mathcal{T}) N_X^B(\mathcal{T}).$$

Next, the posterior distribution on B, given X, is given by

$$\Pi[B \mid X] \propto \Pi_{\mathbb{B}}(B) \prod_{l\leq L,k} \phi_{\frac{1}{\sqrt{n}}}(X_{lk}^B) \left\{ \sum_{\mathcal{T}\in\mathbb{T}} W_X^B(\mathcal{T}) \right\}.$$

Also, we have

$$(\beta_{lk})_{l\leq L,k} \mid (X_{lk})_{l\leq L,k}, \mathcal{T}, B \ \sim \ \pi(\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}^B) \otimes \bigotimes_{(l,k)\notin\mathcal{T}'_{int}} \delta_0(\beta_{lk}), \tag{105}$$

where the posterior density on the selected coefficients on $\mathcal{T}$ is (in slight abuse of notation writing in the same way the distribution and its density)

$$\pi(\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{X}_{\mathcal{T}}^B) = \frac{\mathrm{e}^{-\frac{n}{2}\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 + n\boldsymbol{X}_{\mathcal{T}}^{B\prime}\boldsymbol{\beta}_{\mathcal{T}}} \pi(\boldsymbol{\beta}_{\mathcal{T}})}{N_X^B(\mathcal{T})}. \tag{106}$$

Controlling the Noise. Similarly as in the proof of Theorem 1, we will condition on a set of large probability, where the noise level is relatively small. Denote with $\mathbb{B}$ the set of *all* breakpoints B that can be obtained by performing steps (a) and (b) in Section 4.1 and that yield a system Ψ_B^A satisfying conditions (B1)–(B2) from Lemma 11. Recall $L = L_{max} = \lfloor \log_2 n \rfloor$ and $\varepsilon_{lk}^B = \int_0^1 \psi_{lk}^B(u)dW(u)$, and let δ be as in $(B2)$. We define

$$\mathcal{A}_{\mathbb{B}} = \left\{ \max_{B\in\mathbb{B}} \max_{l\in[0,L],k\in[0,2^l-1]} (\varepsilon_{lk}^B)^2 \leq D_1 \log^{1+\delta} n \right\} \tag{107}$$

for some $D_1 > 0$. Using assumption $(B1)$, one can express every single ψ_{lk}^B for $l \leq L$ in terms of a number $C_0 l^\delta$ of ψ_{jm}'s for $j \leq l + D$, where ψ_{jm} are the regular Haar wavelet functions from (3). That is, with $\mathcal{X}_{lk}^B$ the set of such pairs (j, m) and $\mathrm{Card}(\mathcal{X}_{lk}^B) \leq C_0 l^\delta$, we have

$$\psi_{lk}^B = \sum_{(j,m)\in\mathcal{X}_{lk}^B} p_{jm}^B \psi_{jm},$$

for some real numbers p_{jm}^B that satisfy $\sum_{(j,m)\in\mathcal{X}_{lk}^B} (p_{jm}^B)^2 = 1$ (since ψ_{lk}^B has a unit L^2–norm and ψ_{lk}'s are orthonormal in $L^2[0, 1]$). Next, we have

$$\varepsilon_{lk}^B = \sum_{(j,m)\in\mathcal{X}_{lk}^B} p_{jm}^B \varepsilon_{jm}.$$

This itself implies the following, by the Cauchy-Schwarz inequality,

$$|\varepsilon^B_{lk}| \le \max_{l\le L,k}\left\{\mathrm{Card}(\mathcal{X}^B_{lk})\max_{l\le L+D,k}\varepsilon^2_{lk}\right\}^{1/2} \le C_0^{1/2}L^{\delta/2}\max_{l\le L+D,k}|\varepsilon_{lk}|.$$

Using $L \le \log_2 n$ and denoting

$$\mathcal{A} \equiv \left\{\max_{l\in[0,L+D],k\in[0,2^l-1]}\varepsilon^2_{lk} \le 2\log(2^{L+D+1})\right\},$$

one obtains the inclusion $\mathcal{A} \subset \mathcal{A}_{\mathbb{B}}$, provided that D_1 is chosen larger than a universal constant (in particular it is independent of B). This implies $P_{f_0}(\mathcal{A}^c_{\mathbb{B}}) \le P_{f_0}(\mathcal{A}^c) \le c_0/\sqrt{\log(2^{L+D+1})}$. Next, we follow the structure of the proof of Theorem 1.

Posterior Probability of Too Deep Trees. For a given tree $\mathcal{T}$, we again denote with $\mathcal{T}^-$ the pruned subtree obtained by turning the deepest rightmost internal node $(l_1,k_1) \in \mathcal{T}_{int}$ into a leaf. Given $B \in \mathbb{B}$, we proceed as in the proof of Lemma 2 and evaluate the ratio $W^B_X(\mathcal{T})/W^B_X(\mathcal{T}^-)$. When $l_1 > \mathcal{L}_c$, with $\mathcal{L}_c$ as in (48), Lemma 11 leads to $(B2)$, that is $|\beta^B_{l_1k_1}| \lesssim (\log n)^{\delta/2}\sqrt{\log n/n}$ for large enough n. Similarly as in (52), we can write for $l_1 > \mathcal{L}_c$ and some $C_2 > 0$ (depending on E, D only), on the set $\mathcal{A}_{\mathbb{B}}$ from (107),

$$(X^B_{l_1k_1})^2 \le \frac{C_2}{n}(\log n)^{1+\delta}.$$

Under the Galton-Watson process prior from Section 2.1.1 with $p_l \le 1/2$ and the independent prior with $\Sigma_{\mathcal{T}} = I_{|\mathcal{T}_{ext}|}$ this gives *for all* $B \in \mathbb{B}$ and $d \ge \mathcal{L}_c$,

$$\frac{W^B_X(\mathcal{T})}{W^B_X(\mathcal{T}^-)} \le 2\,p_{d-1}\mathrm{e}^{(C_2/2)(\log n)^{1+\delta}},$$

from which one deduces that

$$\Pi[d(\mathcal{T}) > \mathcal{L}_c \,|\, B, X] \le 4\,\mathrm{e}^{(C_2/2)(\log n)^{1+\delta}}\sum_{d=\mathcal{L}_c+1}^{L_{max}} 2^{d-1}p_{d-1}. \qquad (108)$$

The right side goes to 0 at rate $\mathrm{e}^{-C(\log n)^{1+\delta}}$ if, e.g., p_d is of the order $(1/\Gamma)^{d^{1+\delta}}$ for some large enough $\Gamma > 0$. This also holds for a variant of the tree prior (6) using instead $\pi(\mathcal{T}) \propto \mathrm{e}^{-c|\mathcal{T}_{ext}|\log^{1+\delta} n}$ and the conditionally uniform prior from Section 2.1.2 using $\pi(K) \propto \mathrm{e}^{-cK\log^{1+\delta} n}$ where $K = |\mathcal{T}_{ext}|$. Using a similar strategy as in the proof of Theorem 2, a statement similar to (108) can also be obtained for the general prior $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0,\Sigma_{\mathcal{T}})$ where $\lambda_{min}(\Sigma_{\mathcal{T}}) > \sqrt{1/(\log n)^{1+\delta}}$.

Posterior Probability of Missing a Significant Node. We show a variant of Lemma 4 assuming instead that a signal node (l_S,k_S) satisfies

$$l_S \le \mathcal{L}_c, \qquad |\beta^{0B}_{l_Sk_S}| \ge \frac{A(\log n)^{1+\frac{\delta}{2}}}{\sqrt{n}}, \qquad (109)$$

for some $A > 0$ to be chosen below. As before, for a tree $\mathcal{T} \in \mathbb{T}_{\backslash(l_S,k_S)}$ that does not have a cut at (l_S, k_S), we denote with $\mathcal{T}^+$ the smallest full binary tree (in terms of number of nodes) that contains $\mathcal{T}$ and cuts at (l_S, k_S). Using similar arguments as in the proof of Lemma 4, we use the fact $(X^B_{l_S k_S})^2 \geq (\beta^{0B}_{l_S k_S})^2/2 - (\varepsilon^B_{l_S k_S})^2/n$ to find that on the event $\mathcal{A}_{\mathbb{B}}$ and for $A > 0$ large enough,

$$\frac{W^B_X(\mathcal{T})}{W^B_X(\mathcal{T}^+)} \leq \Gamma^{l_S^{1+\delta}(l_S+1)} \mathrm{e}^{\frac{3}{2}D_1(l_S+1)\log^{1+\delta} n - \frac{A^2}{8}\log^{2+\delta} n} \leq \mathrm{e}^{-\frac{A^2}{16}\log^{2+\delta} n} \tag{110}$$

under the independent Gaussian prior on $\boldsymbol{\beta}_{\mathcal{T}}$ and the Galton-Watson process prior from Section 2.1.1 with $p_l \asymp (1/\Gamma)^{l^{1+\delta}}$. Following the steps in the proof of Lemma 3, one can show similarly that $\Pi\left[(l_S, k_S) \notin \mathcal{T}_{int} \,|\, X, B\right] \to 0$ for each $B \in \mathbb{B}$ sufficiently quickly. More precisely, if

$$S^B(f_0; A) = \left\{ (l,k) : \ |\beta^{0B}_{lk}| \geq A \frac{(\log n)^{1+\frac{\delta}{2}}}{\sqrt{n}} \right\}, \tag{111}$$

where $\mathcal{L}_c$ is defined in (48), we have, on the event $\mathcal{A}_{\mathbb{B}}$ and for A large enough,

$$\Pi\left[\left\{\mathcal{T} : \ S^B(f_0; A) \not\subseteq \mathcal{T}\right\} \,|\, X\right] \leq \mathrm{e}^{-C(\log n)^{1+\delta}}. \tag{112}$$

uniformly in $B \in \mathbb{B}$. This statement can be obtained also for the general prior $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$ with $\lambda_{max}(\Sigma_{\mathcal{T}}) \lesssim n^a$ for some $a \geq 1$ and for other tree priors from Section 2.1.2.

Putting Pieces Together. Let us also set

$$\mathsf{T}^B = \{\mathcal{T} : d(\mathcal{T}) \leq \mathcal{L}_c, \, S^B(f_0; A) \subset \mathcal{T}\}, \qquad \mathcal{E}^B = \{f_{\mathcal{T},\boldsymbol{\beta}} : \mathcal{T} \in \mathsf{T}^B\}. \tag{113}$$

From the two previous subsections one obtains that for some constant $C > 0$

$$\Pi[\mathcal{T} \notin \mathsf{T}^B \,|\, X, B] \leq \mathrm{e}^{-C(\log n)^{1+\delta}},$$

for any possible set of breakpoints $B \in \mathbb{B}$ (that satisfy the balancing conditions). The uniformity in B is essential in the next bounds.

Using the definition of the event $\mathcal{A}_{\mathbb{B}}$ from (107), one can bound

$$E_{f_0}\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X] \leq P_{f_0}[\mathcal{A}^c_{\mathbb{B}}] + E_{f_0}\left\{\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X]\mathbb{I}_{\mathcal{A}_{\mathbb{B}}}\right\}.$$

By decomposing the posterior along B and $\mathcal{T}$ and using Markov's inequality one obtains, on the event $\mathcal{A}_{\mathbb{B}}$,

$$\begin{aligned}
&\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X] = \sum_B \Pi[B \,|\, X] \sum_{\mathcal{T}} \Pi[\mathcal{T} \,|\, X, B]\, \Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X, \mathcal{T}, B] \\
&\leq \sum_B \Pi[B \,|\, X]\Pi[\mathcal{T} \notin \mathsf{T}^B \,|\, X, B] + \sum_B \Pi[B \,|\, X] \sum_{\mathcal{T} \in \mathsf{T}^B} \Pi[\mathcal{T} \,|\, B, X]\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X, \mathcal{T}, B] \\
&\leq \mathrm{e}^{-C(\log n)^{1+\delta}} + \sum_B \Pi[B \,|\, X] \sum_{\mathcal{T} \in \mathsf{T}^B} \Pi[\mathcal{T} \,|\, B, X]\varepsilon_n^{-1} \int \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \,|\, X, \mathcal{T}, B].
\end{aligned}$$

Let us now turn to bounding $\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty$. First, note that unlike the traditional Haar basis, the UH basis system is never built up until $L = \infty$ because, by construction, we stop splitting when there are no x_i are available (i.e. we do not split nodes that are not *admissible*). In result, the very high frequencies are not covered by the system, which might induce some unwanted bias. This is, however, *not an issue* with our *weakly balanced* UH wavelets. The following Lemma shows that in weakly balanced UH systems, all nodes at levels $l \leq \Lambda := \lfloor \log_2(n/\log^c n) \rfloor$ for any $c > 0$ are admissible.

Lemma 16. *Consider a weakly balanced UH wavelet system Ψ_A^B, where A is the set of admissible nodes (l,k) in the sense that $\mathcal{X} \cap (l_{lk}, r_{lk}] \neq \emptyset$ with $\mathcal{X} = \{x_i : x_i = 1/n, 1 \leq i \leq n\}$. Let $c > 0$, then for $\Lambda = \Lambda(c) = \lfloor \log_2(n/\log^c n) \rfloor$, we have*

$$A \supset \{(l,k) : l \leq \Lambda\}.$$

Proof. The proof follows from the fact that the granularity of weakly balanced UH systems is very close to l. In Example 3 we defined the granularity $R(l, \Psi_A^B)$ of the l^{th} layer as the smallest integer $R \geq 1$ such that $\min_{0 \leq k < 2^l} \min\{|L_{lk}|, |R_{lk}|\} = j/2^R$ for some $j \in \{1, 2, \dots, 2^{R-1}\}$. From Lemma 10, the granularity of weakly balanced systems Ψ_A^B is no larger than $l + D$. This means that for $l < \Lambda$, $0 \leq k < 2^l$, any $c > 0$ and n large enough

$$\min\{|L_{lk}|, |R_{lk}|\} \geq 1/2^{l+D} > \frac{\log^c n}{2^D n} > 1/n.$$

This implies that $\mathcal{X} \cap (l_{lk}, r_{lk}] \neq \emptyset$ for any (l,k) with $l \leq \Lambda$, where we used the fact that $(l_{lk}, r_{lk}]$ is either $R_{l-1\,\lfloor k/2 \rfloor}$ (for when (l,k) is the right child) or $L_{l-1\,\lfloor k/2 \rfloor}$ (for when (l,k) is the left child). □

Next, we show that the weakly balanced UH systems are indeed rich enough to approximate f_0 well.

Lemma 17. *Consider the weakly balanced UH system Ψ_A^B. Let f_0^Λ denote the L^2–projection of $f_0 \in \mathcal{H}_M^\alpha$ onto Vect$\{\psi_{lk}^B : l \leq \Lambda\}$ for $\Lambda = \lfloor \log_2(n/\log^c n) \rfloor$ with some $c > 0$. Then*

$$\|f_0 - f_0^\Lambda\|_\infty \lesssim |\Lambda 2^{-\Lambda}|^\alpha \lesssim (\log^{c+1} n/n)^\alpha.$$

Proof. The L^2–projection is a step function $f_0^\Lambda = \sum_m \mathbb{I}_{\Omega_m} \widetilde{\beta}_m$ supported on the pieces $\Omega_m \in \{L_{\Lambda k}, R_{\Lambda k} : 0 \leq k < 2^\Lambda\}$ where the jump sizes equal $\widetilde{\beta}_m = |\Omega_m|^{-1} \int_{\Omega_m} f_0(x)dx$. From the Hölder continuity in (25) we have $|f_0(x) - f_0^\Lambda(x)| \leq M|\Omega_m|^\alpha$ for $x \in \Omega_m$. From the definition of weakly balanced UH systems, we have $\max_m |\Omega_m| \leq \frac{C+\Lambda}{2^{\Lambda+D}}$. The rest follows from the definition of Λ. □

We can now write the following decomposition. For $f_{\mathcal{T},\boldsymbol{\beta}}$ in $\mathcal{E}^B$, we have

$$\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty \lesssim \sum_{l\leq \mathcal{L}_c} 2^{l/2} \max_{0\leq k<2^l} |\beta_{lk}^B - \beta_{lk}^{0B}| \\ + \sum_{\mathcal{L}_c \leq l \leq \Lambda} 2^{l/2} \max_{0\leq k<2^l} |\beta_{lk}^{0B}| + \|f_0 - f_0^\Lambda\|_\infty. \qquad (114)$$

In the last display, we have used the fact that for weakly balanced UH systems one has

$$\max_{0\leq k<2^l} \|\psi_{lk}^B\|_\infty < \max_{0\leq k<2^l} \left[\left(\frac{1}{|L_{lk}|} \vee \frac{1}{|R_{lk}|}\right) \frac{1}{\sqrt{|L_{lk}|^{-1} + |R_{lk}|^{-1}}}\right] < 2^{(l+D)/2}.$$

The second term in (114) can be upper-bounded by $(\log n)^{1+\delta/2}(\log n/n)^{\alpha/(2\alpha+1)}$ by using $(B2)$ and the definition of $\mathcal{L}_c$. Using Lemma 17, the term $\|f_0 - f_0^\Lambda\|_\infty$ is always of smaller order than the previous one (as the bound decreases as $n^{-\alpha}$ up to a logarithmic factor).

Regarding the first term, one obtains for $\mathcal{T} \in \mathsf{T}^B$

$$\int \max_{0\leq k<2^l} |\beta_{lk}^B - \beta_{lk}^{0B}| d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \,|\, X, \mathcal{T}, B]$$
$$= \int \max\left(\max_{0\leq k<2^l,\,(l,k)\notin \mathcal{T}_{int}} |\beta_{lk}^{0B}|, \max_{0\leq k<2^l,\,(l,k)\in \mathcal{T}_{int}} |\beta_{lk}^B - \beta_{lk}^{0B}| \right) d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \,|\, X, \mathcal{T}, B]$$
$$\leq A\frac{(\log n)^{1+\frac{\delta}{2}}}{\sqrt{n}} + \int \max_{0\leq k<2^l,\,(l,k)\in \mathcal{T}_{int}} |\beta_{lk}^B - \beta_{lk}^{0B}| d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \,|\, X, \mathcal{T}, B],$$

where we have used that on the set $\mathcal{E}^B$, selected trees cannot miss any true signal larger than $A(\log n)^{1+\delta/2}/\sqrt{n}$. This means that any node (l,k) that is not in a selected tree must satisfy $|\beta_{lk}^{0B}| \leq A(\log n)^{1+\delta/2}/\sqrt{n}$.

We now focus on the independent prior $\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, I_{|\mathcal{T}_{ext}|})$. We have seen above that, given X, B (so for fixed ε_{lk}^B) and $\mathcal{T}$, if (l,k) belongs to $\mathcal{T}_{int}$, the difference $\beta_{lk}^B - \beta_{lk}^{0B}$ has a Gaussian distribution Q_{lk} given by

$$Q_{lk} \overset{\mathcal{L}}{=} X_{lk} - \beta_{lk}^{0B} + \mathcal{N}\left(0, \frac{1}{n+1}\right) = -\frac{\beta_{lk}^{0B}}{n+1} + \frac{\sqrt{n}\varepsilon_{lk}^B}{n+1} + \mathcal{N}\left(0, \frac{1}{n+1}\right).$$

If Z_{lk} are arbitrary random variables distributed according to Q_{lk}, and $\mathcal{Z}_{lk}$ arbitrary $\mathcal{N}(0,1)$ random variables,

$$\mathbb{E}\left[\max_{0\leq k<2^l} |Z_{lk}|\right] \leq \max_{0\leq k<2^l} \frac{|\beta_{lk}^{0B}|}{n} + \max_{0\leq k<2^l} \frac{|\varepsilon_{lk}^B|}{\sqrt{n}} + \frac{1}{\sqrt{n}}\mathbb{E}\left[\max_{0\leq k<2^l} |\mathcal{Z}_{lk}|\right].$$

On the event $\mathcal{A}_{\mathbb{B}}$ from (107), the sum of the first two terms on the last display is bounded by $M/n + C\sqrt{(\log n)^{1+\delta}/n}$ while the last expectation is at most $C\sqrt{l/(n+1)}$ by Lemma 8. This implies

$$\int \max_{0\leq k<2^l,\,(l,k)\in \mathcal{T}_{int}} |\beta_{lk}^B - \beta_{lk}^{0B}| d\Pi[f_{\mathcal{T},\boldsymbol{\beta}} \,|\, X, \mathcal{T}, B] \leq C'\sqrt{\frac{(\log n)^{1+\delta}}{n}}$$

uniformly over B and T^B, where we have used $l \leq C \log n$. Putting the various pieces together and using the fact that $P_{f_0}[\mathcal{A}_{\mathbb{B}}^c] = o(1)$, we obtain

$$
\begin{aligned}
&E_{f_0}\Pi[\|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_\infty > \varepsilon_n \,|\, X] \\
&\leq o(1) + \varepsilon_n^{-1}\left\{\sum_{l \leq \mathcal{L}_c} 2^{l/2}\left[A\frac{(\log n)^{1+\delta/2}}{\sqrt{n}} + C'\sqrt{\frac{\log^{1+\delta} n}{n}}\right] + 2(\log n)^{1+\frac{\delta}{2}}\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right\} \\
&\leq o(1) + \varepsilon_n^{-1}\left\{\left[A(\log n)^{\frac{1+\delta}{2}} + C'(\log n)^{\frac{\delta}{2}}\right]2\sqrt{\frac{2^{\mathcal{L}_c}\log n}{n}} + 2(\log n)^{1+\frac{\delta}{2}}\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right\} \\
&\leq o(1) + \varepsilon_n^{-1}2C'\left[A(\log n)^{\frac{1+\delta}{2}} + 4(\log n)^{1+\frac{\delta}{2}}\right]\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}.
\end{aligned}
$$

This means that one can set

$$\varepsilon_n = (\log n)^{1+\delta/2}\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$$

and this is the obtained posterior rate in terms of the supremum norm. A similar conclusion can be obtained for the general prior $\pi(\boldsymbol{\beta}_{\mathcal{T}}) \sim \mathcal{N}(0, \Sigma_{\mathcal{T}})$ using Lemma 4 under the assumption $\lambda_{min}(\Sigma_{\mathcal{T}}) \gtrsim \sqrt{1/\log^{1+\delta} n}$.