Spatial properties of Bayesian unsupervised trees

Linxi Liu LINXI_LIU@PITT.EDU
University of Pittsburgh

Li Ma LI.MA@DUKE.EDU
Duke University

Editors: Shipra Agrawal and Aaron Roth

Abstract

Tree-based methods are popular nonparametric tools for capturing spatial heterogeneity and making predictions in multivariate problems. In unsupervised learning, trees and their ensembles have also been applied to a wide range of statistical inference tasks, such as multi-resolution sketching of distributional variations, localization of high-density regions, and design of efficient data compression schemes. In this paper, we study the spatial adaptation property of Bayesian tree-based methods in the unsupervised setting, with a focus on the density estimation problem. We characterize spatial heterogeneity of the underlying density function by using anisotropic Besov spaces, region-wise anisotropic Besov spaces, and two novel function classes as their extensions. For two types of commonly used prior distributions on trees under the context of unsupervised learning—the optional Pólya tree (Wong and Ma, 2010) and the Dirichlet prior (Lu et al., 2013)—we calculate posterior concentration rates when the density function exhibits different types of heterogeneity. In specific, we show that the posterior concentration rate for trees is near minimax over the anisotropic Besov space. The rate is adaptive in the sense that to achieve such a rate we do not need any prior knowledge of the parameters of the Besov space.

Keywords: Multivariate Density Estimation, Tree-based Methods, Spatial Adaptation, Posterior Concentration, Asymptotic Minimaxity

1. Introduction

Modern applications often involve data arising from complex generative distributions supported on multivariate or even high-dimensional sample spaces. A fundamental inference objective of unsupervised learning is to identify the nature and structure of the underlying data generative distribution. However, learning a multivariate distribution in a flexible, nonparametric fashion is known to be challenging when the sample space grows beyond a handful of dimensions due to the so-called "curse-of-dimensionality", especially when the distribution involves complex structure—such as non-linear dependency, spatially varying smoothness, and local features.

Tree-based methods are quite effective for nonparametric estimation, and have achieved great success in supervised learning. Random forests (RF, Breiman 2001), tree boosting (Freund and Schapire, 1997; Friedman, 2001), Bayesian classification and regression trees (Bayesian CART, Chipman et al. 1997, 1998), and Bayesian additive regression trees (BART, Chipman et al. 2010) have been widely in use due to their computational ease and desirable predictive accuracy. In the last few years, there has been a burgeoning of interest in rigorously quantifying performance of supervised trees, especially that for RF, thereby unraveling the myth of their success (Arlot and Genuer, 2014; Biau, 2012; Scornet, 2016; Mentch and Zhou, 2020; Biau et al., 2008; Denil et al., 2014; Lin and Jeon, 2006; Scornet et al., 2015; Chi et al., 2022; Cattaneo et al., 2023; Klusowski, 2020, 2021). For unsupervised learning, many estimation or inference techniques have also been

developed on top of tree models, such as two-sample comparison through multi-resolution scanning (Soriano and Ma, 2017; Ma and Wong, 2011), mode hunting and clustering (McKinnon, 2018; Ooi, 2002), and data compression (Huffman, 1952; Poggi and Olshen, 1995). However, the theoretical analysis of tree-based methods for unsupervised problems such as density estimation is far less sophisticated compared to what has been accomplished for supervised ones.

The Bayesian tree model has gained popularity in practice, as the Bayesian framework naturally endows a stochastic search algorithm, and provides an elegant solution to accounting for model uncertainty. Quite a few recent work on Bayesian tree models contributes to deriving posterior concentration rates for trees over different function classes, as a way to investigate their spatial adaptivity. However, the existing analysis for either supervised or unsupervised trees is not adequate to provide a full picture of spatial adaptation properties of trees. For example, Liu et al. (2017, 2023) study Bayesian unsupervised trees only when the underlying true density function lies in a Besov space with *homogeneous* smoothness properties along different directions. Ročková and van der Pas (2020); Jeong and Ročková (2023); Ročková and Rousseau (forthcoming) analyze posterior concentration for Bayesian CART and BART when the regression function is isotropic Hölder, anisotropic Hölder, or region-wise anisotropic Hölder continuous. These Hölder spaces allow the smoothness of the function to vary spatially. However, they may not be a good model to describe local spikes or sharp changes as functions in these spaces are uniformly continuous.

The purpose of this paper is to provide a more complete characterization of spatial adaptation properties of tree-based methods, with a focus on the unsupervised setting. We will model spatial heterogeneity by using anisotropic Besov spaces, region-wise anisotropic Besov spaces, or more generally classes of density functions satisfying different types of sparsity condition in the spectral domain. Our theoretical analysis is under the Bayesian framework. We consider two common types of prior distributions on unsupervised trees: the optional Pólya tree prior (Wong and Ma, 2010) and the Dirichlet prior (Lu et al., 2013; Liu et al., 2023). We will provide a thorough analysis of posterior concentration properties under these two types of prior, in order to reveal spatial adaptation properties of trees from the theoretical perspective.

Rigorously justified spatial adaptivity also distinguishes tree-based methods from the other categories of nonparametric methods for unsupervised learning. The kernel method (Rosenblatt, 1956; Parzen, 1962) suffers from the curse-of-dimensionality and may have limited ability to capture inhomogeneous features in high-dimensional cases. More recently introduced generative models, such as generative adversarial networks (GANs, Goodfellow et al. 2014; Uppal et al. 2019), variational auto-encoders (VAEs, Kingma and Welling 2014), and normalizing flows (Rezende and Mohamed, 2015; Dinh et al., 2015), exhibit superior empirical performance, but spatial adaptation properties of these methods remain unknown from the theoretical perspective.

We introduce some notations at the end of the introduction. On the measurable space $(\mathbb{R}^d, \mathcal{B})$ equipped with the Lebesgue measure μ, we assume a density f_0 for the distribution of interest exists. After translation and scaling, we may assume that the sample space Ω is the unit cube in $\mathbb{R}^d$. The collection of all density functions on $(\Omega, \mathcal{B}, \mu)$ is denoted as $\mathcal{F}$. We will derive posterior concentration rates with respect to the Hellinger distance ρ, and will also use both the Kullback-Leibler (KL) divergence and the L^∞-norm, defined as $\mathrm{KL}(f, g) = \mathbb{E}_f\left(\log(f/g)(Y)\right)$ and $\|f - g\|_\infty = \sup_{y\in\Omega} |f(y) - g(y)|$ respectively. We will treat the dimension d as fixed, as it is very challenging to estimate the density in an ultra-high dimensional space.

The rest of the paper is organized as the following. In Section 2, we provide detailed characterizations of spatial heterogeneity of the density function. In Section 3, we introduce two types of

prior distributions on trees. We summarize our main theoretical results of posterior concentration rates for trees when the density exhibits different types of spatial features in Section 4, and provide an outline of the proof in Section 5.

2. Characterization of spatial heterogeneity

We will first provide a characterization of spatial heterogeneity by using anisotropic Besov spaces, or more generally, region-wise anisotropic Besov spaces. The Besov space is a rich function class that is commonly considered under the context of nonparametric modeling. Both the Hölder space and the class of functions of bounded variations can be embedded into Besov spaces (see Nikol'skii 2012 for more details). We will introduce the anisotropic Besov space based on the Besov norm, and will provide an alternative definition by using multiresolution Haar wavelets. The second definition in terms of decay of wavelet coefficients is particularly interesting, as it reflects a type of sparsity in the spectral domain. This is also the motivation for providing the two other types of characterizations of spatial heterogeneity in terms of Besov balls and weak-$\ell_{p'}$ sparsity of wavelet coefficients.

2.1. Anisotropic Besov spaces

Generally, to define the anisotropic Besov space $B^{\boldsymbol{\sigma}}_{\boldsymbol{p},q}(\mathcal{R}, L)$ on the domain $\mathcal{R} = \otimes_{l=1}^{d}(a_l, b_l)$, let $e^l = (\delta_{l1}, \ldots, \delta_{ld})^\top$ ($\delta_{ll'} = \mathbb{1}_{\{l=l'\}}$) be the l-th unit vector. The first difference of the function f along the direction of y^l is defined as $\Delta_{l,h} f(y) = f(y + he^l) - f(y)$, and the second difference is $\Delta^2_{l,h} f(y) = \Delta_{l,h}(\Delta_{l,h} f(y)) = f(y + 2he^l) - 2f(y + he^l) + f(y)$. For each l, $s_l = \lfloor \sigma_l \rfloor$ denotes the largest integer strictly less than σ_l. In the one-dimensional case, For $q < \infty$, the Besov norm along the direction of y^l is defined as

$$\|f\|_{b^{\sigma_l}_{l,p_l,q}} = \left(\int_0^\infty |h|^{(s_l - \sigma_l)q - 1} \left\| \Delta^2_{l,h} \left(\frac{\partial^{s_l}}{\partial (y^l)^{s_l}} f \right) \right\|^q_{L^{p_l}(\mathcal{R}_{l,h})} dh \right)^{1/q},$$

and for $q = \infty$,

$$\|f\|_{b^{\sigma_l}_{l,p_l,\infty}} = \sup_{h \geq 0} \left\{ |h|^{s_l - \sigma_l} \left\| \Delta^2_{l,h} \left(\frac{\partial^{s_l}}{\partial (y^l)^{s_l}} f \right) \right\|_{L^{p_l}(\mathcal{R}_{l,h})} \right\},$$

where $\mathcal{R}_{l,h} = \otimes_{l'=1}^{l-1}(a_{l'}, b_{l'}) \otimes (a_l, a_l \vee (b_l - 2h)) \otimes_{l'=l+1}^{d} (a_{l'}, b_{l'})$. For $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)^\top$ and $\boldsymbol{p} = (p_0, p_1, \ldots, p_d)^\top$, the norm associated with the anisotropic Besov space $B^{\boldsymbol{\sigma}}_{\boldsymbol{p},q}$ is

$$\|f\|_{B^{\boldsymbol{\sigma}}_{\boldsymbol{p},q}} = \|f\|_{L^{p_0}(\mathcal{R})} + \sum_{l=1}^{d} \|f\|_{b^{\sigma_l}_{l,p_l,q}}.$$

We define the anisotropic Besov space as

$$B^{\boldsymbol{\sigma}}_{\boldsymbol{p},q}(\mathcal{R}, L) = \{f : \|f\|_{B^{\boldsymbol{\sigma}}_{\boldsymbol{p},q}} \leq L\}, \quad \text{for some } L > 0.$$

According to the definition above, in the one-dimensional case, the Besov norm $\|\cdot\|_{b^{\sigma}_{p,q}}$ measures the regularity of the function by using the L^p-norm. The larger p is, the stronger the norm is. When $p = \infty$, the density function lying in the Besov space is Hölder continuous in the sense that $\|\cdot\|_{b^{\alpha}_{\infty,\infty}} < \infty$ is equivalent to α-Hölder continuity when $0 < \alpha \notin \mathbb{N}$. For smaller p, as a weak

metric is applied, the density function in the Besov space is allowed to have more local variations. This way, the space can cover those density functions with sharp changes and local spikes. Last but not the least, the anisotropic Besov space allows different smoothness levels along different directions, which is a reflection of spatial heterogeneity.

In this paper, we will focus on the case that $p_0 = \cdots = p_d = p \geq 2$, which means the same norm is considered along different directions. We require $p \geq 2$ as the Hellinger distance or equivalently the L^2-norm for bounded densities will be applied to quantify the posterior concentration rate. We also assume $\sigma_l \in (0, 1]$, as the tree-supported density function has limited approximation ability. Usually, the domain $\mathcal{R}$ is a subset of $\Omega = [0, 1]^d$.

2.2. Region-wise anisotropic Besov spaces

We further introduce region-wise anisotropic Besov spaces to accommodate richer spatial features of the density function. On a dyadic partition $\mathcal{A}^\star = \{\Omega_s\}_{s=1}^S$ of the domain Ω with S subregions (more details about the dyadic partition will be introduced in Section 3), we allow the density function to have different smoothness properties in different regions. More specifically, we denote the restriction of a function f on a set A as $f|_A$ and assume $f|_{\Omega_s} \in B^{\boldsymbol{\sigma}_s}_{\boldsymbol{p},q}(\Omega_s, L)$, where $\boldsymbol{\sigma}_s = (\sigma_{sl})_{1\leq l\leq d}$. The smoothness parameters $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S)$ for Besov spaces over S regions are from either of the following two sets,

$$\Sigma_\sigma = \left\{ (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) : \sigma_{sl} \in (0, 1] \text{ for all } s \text{ and } l, \sum_{l=1}^{d} \frac{1}{\sigma_{sl}} = \frac{d}{\sigma} \text{ for all } 1 \leq s \leq S \right\}, \tag{1}$$

where $\sigma \in (0, 1]$, and

$$\Sigma = \{(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) : \sigma_{sl} \in (0, 1] \text{ for all } 1 \leq s \leq S, 1 \leq l \leq d\}. \tag{2}$$

It is easy to see that $\Sigma_\sigma \subset \Sigma$. The region-wise anisotropic Besov space with parameters in Σ characterizes higher level of spatial heterogeneity.

2.3. Multiresolution wavelet basis and Besov balls

In Section 2.1 and Section 2.2, we have introduced two characterizations of spatial heterogeneity. The purpose of this subsection is to provide a multiresolution description of the Besov space and extend our characterization of spatial features to a more general case, called the Besov balls. We start with a description for the one dimensional multiresolution analysis (MRA).

In the one dimensional case, it is well known that a function ϕ can be constructed to satisfy the following properties (Meyer, 1990):

- F1. The sequence $\{\phi(\cdot - k), k \in \mathbb{Z}\}$ is an orthonormal family of $L^2(\mathbb{R})$. Let V_0 be the function space spanned by it.
- F2. For all $j \in \mathbb{Z}$, $V_j \subset V_{j+1}$ if V_j denotes the space spanned by $\{\phi_{jk}, k \in \mathbb{Z}\}$, where $\phi_{jk} = 2^{j/2}\phi(2^j \cdot -k)$.
- F3. ϕ is r times weakly differentiable, and the derivative $\phi^{(r)}$ is rapidly decreasing.

One example of such a function is the scaling function for Haar wavelet $\phi = \mathbb{1}_{[0,1)}(y)$. Then we have $\cap_{j\in\mathbb{Z}} V_j = \{0\}$, and $L^2(\mathbb{R}) = \cup_{j\in\mathbb{Z}} V_j$. Under these conditions, the space W_j is defined by $V_{j+1} = V_j \oplus W_j$. There exists a function ψ (the "wavelet") such that

M1. $\{\psi(\cdot - k), k \in \mathbb{Z}\}$ is an orthonormal basis of W_0.

M2. $\{\psi_{jk}, j \in \mathbb{Z}, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$, where $\psi_{jk} = 2^{j/2}(2^j \cdot -k)$.

M3. ψ has the same regularity property as ϕ.

Then $L^2(\mathbb{R})$ can be decomposed in the following way, for some fixed integer $j_0 \in \mathbb{Z}$:

$$L^2(\mathbb{R}) = V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots . \tag{3}$$

In this paper, we will focus on the Haar wavelet $\psi(y) = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)}$, as it is closely related to tree-based methods.

Moving to the multivariate case, to characterize the anisotropic Besov spaces, we can introduce multiresolution wavelet basis in a similar way. We will focus on the construction by Leisner (2003); Garrigós and Tabacco (2002). Let ϕ_{jk} and ψ_{jk} be scaling and translation of ϕ and ψ respectively, defined as $\phi_{jk} = 2^{j/2}\phi(2^j \cdot -k)$ and $\psi_{jk} = 2^{j/2}\psi(2^j \cdot -k)$. With smoothness parameter $\boldsymbol{\sigma}$, let $\sigma_{\min} = \min_{1\le l\le d} \sigma_l$ and $\sigma_l' = \sigma_{\min}/\sigma_l$. $\bar{\sigma}$ is the value that satisfies $(\bar{\sigma})^{-1} = \frac{1}{d}\sum_{l=1}^d \sigma_l^{-1}$. We define the multiresolution scaling function at level j as $\xi_{jk}^{\mathbf{0}}(y) = \prod_{l=1}^d \phi_{\lfloor j\sigma_l' \rfloor, k_l}$. The space $\boldsymbol{V}_j$ is

$$\boldsymbol{V}_j = \overline{\text{span}}_{L^p(\mathbb{R}^d)} \left\{ \xi_{jk}^{\mathbf{0}} : k \in \mathbb{Z}^d \right\} = \otimes_{l=1}^d V_{\lfloor j\sigma_l' \rfloor},$$

and the basis functions for $\boldsymbol{V}_j$ are naturally

$$\boldsymbol{\Phi}_{\boldsymbol{\sigma}}^{(j)} = \left\{ \xi_{jk}^{\mathbf{0}} : k \in \mathbb{Z}^d \right\}, \quad \text{where } \xi_{jk}^{\mathbf{0}}(y) = \prod_{l=1}^d \phi_{\lfloor j\sigma_l' \rfloor, k_l}.$$

In the definition, we use the subscript $\boldsymbol{\sigma}$ to emphasize that the basis functions in $\boldsymbol{\Phi}_{\boldsymbol{\sigma}}^{(j)}$ depends on the parameter $\boldsymbol{\sigma}$. Similar to the one-dimensional case, we define, for each $j \in \mathbb{Z}$, the complement space $\boldsymbol{W}_j$ to be the orthogonal complement in $\boldsymbol{V}_{j+1}$ of $\boldsymbol{V}_j$. Let $\Lambda_j = \{l \in [d] : \lfloor j\sigma_l' \rfloor < \lfloor (j+1)\sigma_l' \rfloor$ and $\gamma_{jl} = \mathbb{1}_{l \in \Lambda_j}$. We have the decomposition

$$\begin{aligned} \boldsymbol{V}_{j+1} &= \otimes_{l=1}^d (\gamma_{jl}(V_{\lfloor j\sigma_l' \rfloor} \oplus W_{\lfloor j\sigma_l' \rfloor}) \oplus (1-\gamma_{jl}) V_{\lfloor j\sigma_l' \rfloor}) \\ &= \boldsymbol{V}_j \oplus \bigoplus_{\boldsymbol{\epsilon}_j : \boldsymbol{\epsilon}_j \in \{0,1\}^d \setminus \{(0,\ldots,0)\}, \epsilon_{jl}(1-\gamma_{jl})=0} \otimes_{l=1}^d (\epsilon_{jl} W_{\lfloor j\sigma_l' \rfloor} \oplus (1-\epsilon_{jl}) V_{\lfloor j\sigma_l' \rfloor}), \end{aligned}$$

where by convention for any subspace H of $L^2(\mathbb{R})$, $1 \cdot H = H$ and $0 \cdot H = \{0\}$. The condition $\epsilon_{jl}(1-\gamma_{jl}) = 0$ is understood as, along those directions with $\lfloor j\sigma_l' \rfloor = \lfloor (j+1)\sigma_l' \rfloor$, ϵ_{jl}'s are forced to be 0. It follows that $\boldsymbol{W}_j$ consists of $2^{|\Lambda_j|} - 1$ pieces, and for each piece an orthonormal basis can be obtained by tensor product. In specific, for any $j \in \mathbb{Z}$, we define

$$\boldsymbol{\Xi}_{\boldsymbol{\sigma}}^{(j)} = \left\{ \xi_{jk}^{\boldsymbol{\epsilon}_j} : \boldsymbol{\epsilon}_j = (\epsilon^l)_{1\le l\le p} \in \{0,1\}^d \setminus \{(0,\ldots,0)\}, \epsilon_{jl}(1-\gamma_{jl}) = 0, k \in \mathbb{Z}^d \right\},$$

$$\text{where } \xi_{jk}^{\boldsymbol{\epsilon}_j}(y) = \prod_{l=1}^d \psi_{\lfloor j\sigma_l' \rfloor, k_l}^{\epsilon_{jl}}(y_l) \cdot \phi_{\lfloor j\sigma_l' \rfloor, k_l}^{1-\epsilon_{jl}}(y_l). \tag{4}$$

Then $\boldsymbol{\Phi}_{\boldsymbol{\sigma}}^{(j_0)} \cup_{j=j_0}^{\infty} \boldsymbol{\Xi}_{\boldsymbol{\sigma}}^{(j)}$ is an orthonormal basis of $L^2(\mathbb{R}^d)$, and the following decomposition holds:

$$L^2(\mathbb{R}^d) = \boldsymbol{V}_{j_0} \oplus \boldsymbol{W}_{j_0} \oplus \boldsymbol{W}_{j_0+1} \oplus \cdots .$$

This implies that for all $g \in L^2(\mathbb{R}^d)$,

$$g = \sum_{\xi^0_{j_0k} \in \mathbf{\Phi}^{(j_0)}_{\boldsymbol{\sigma}}} \langle g, \xi^0_{j_0k} \rangle \xi^0_{j_0k} + \sum_{j \geq j_0} \sum_{\xi^{(j)} \in \mathbf{\Xi}^{(j)}_{\boldsymbol{\sigma}}} \langle g, \xi^{(j)} \rangle \xi^{(j)}, \tag{5}$$

where the expansion holds under the norm of the space $L^2(\mathbb{R}^d)$.

Assume $j \in \mathbb{N}$ is an index of the resolution. With the multiresolution Haar basis introduced above, for any function $g \in L^2(\mathbb{R}^d)$, we define $P^{(j)}$ to the be projection operator onto $\boldsymbol{V}_j$ and $E^{(j)} = P^{(j+1)} - P^{(j)}$. According to Leisner (2003), for $\boldsymbol{\sigma} \in (0,1]^d$ with $\max_{1 \leq l \leq d} \sigma_l < 1/p$, $2 \leq p \leq \infty$, $1 \leq q \leq \infty$, g lies in the Besov space $B^{\boldsymbol{\sigma}}_{p,q}$ if and only if

$$\left\| P^{(j_0)} g \right\|_p + \left(\sum_{j \geq j_0} \left(2^{j \sigma_{\min}} \left\| E^{(j)} g \right\|_p \right)^q \right)^{1/q} < \infty.$$

With the multiresolution wavelet basis,

$$P^{(j_0)} g = \sum_{\xi^{\mathbf{0}}_{j_0k} \in \mathbf{\Phi}^{(j_0)}_{\boldsymbol{\sigma}}} \langle g, \xi^0_{j_0k} \rangle \xi^0_{j_0k}, \quad E^{(j)} g = \sum_{\xi^{(j)} \in \mathbf{\Xi}^{(j)}_{\boldsymbol{\sigma}}} \langle g, \xi^{(j)} \rangle \xi^{(j)}.$$

Therefore, an equivalent norm for the Besov space in terms of wavelet coefficients is

$$\left\| g \right\|_{B^{\boldsymbol{\sigma}}_{p,q}} := \left\| P^{(j_0)} g \right\|_p + \left(\sum_{j \geq j_0} \left(2^{j \sigma_{\min} + (1/2 - 1/p) \sum_{l=1}^d \lfloor j \sigma'_l \rfloor} \left\| \{ \langle g, \xi^{(j)} \rangle \}_{\xi^{(j)} \in \mathbf{\Xi}^{(j)}_{\boldsymbol{\sigma}}} \right\|_{\ell_p} \right)^q \right)^{1/q}. \tag{6}$$

For the Besov space defined on a domain $\mathcal{R}$, we can define multiresolution basis on $\mathcal{R}$, denoted as $\mathbf{\Phi}^{(j)}_{\boldsymbol{\sigma}}|_{\mathcal{R}}$ and $\mathbf{\Xi}^{(j)}_{\boldsymbol{\sigma}}|_{\mathcal{R}}$, by using those basis functions in $\mathbf{\Phi}^{(j)}_{\boldsymbol{\sigma}}$ and $\mathbf{\Xi}^{(j)}_{\boldsymbol{\sigma}}$ whose support overlap with $\mathcal{R}$, and perform multiresolution analysis in a similar way.

For $\boldsymbol{\sigma} \in (0, 1/p)^d$, it has been shown by Leisner (2003) that the definition above matches that in the literature (Nikol'skii, 2012), and is equivalent to the definition of Besov spaces based on the Besov norm in Section 2.1 and Section 2.2. For $\boldsymbol{\sigma}$ with $\sigma_l \geq 1/p$, although it may not be consistent with the standard description of the Besov space, we can still introduce a space of functions with fast decaying wavelet coefficients, called an anisotropic Besov ball, in the following way. The anisotropic Besov ball $\widetilde{B}^{\boldsymbol{\sigma}}_{p,q}, \sigma_l > 0, 2 \leq p \leq \infty$ and $1 \leq q \leq \infty$ with radius L is defined as

$$\widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\mathcal{R}, L) = \left\{ g : \|g\|_{B^{\boldsymbol{\sigma}}_{p,q}} \leq L, \text{ where the norm } \|\cdot\|_{B^{\boldsymbol{\sigma}}_{p,q}} \text{ is defined in (6)} \right\}.$$

For anisotropic Besov balls, we would like to allow some entries of the smoothness parameter equal to ∞. According to our definition of multiresolution basis, it means that for the function along the direction l with $\sigma_l = \infty$, one does not perform MRA at all.

We can also introduce region-wise anisotropic Besov balls. For a fixed dyadic partition $\mathcal{A}^\star = \{\Omega_s\}_{s=1}^S$, f lies in anisotropic Besov balls region-wise if $f|_{\Omega_s} \in \widetilde{B}^{\boldsymbol{\sigma}_s}_{p,q}(\Omega_s, L)$ and smoothness parameters $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S)$ are from either of the following two sets

$$\widetilde{\Sigma}_\sigma = \left\{ (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) : \sigma_{sl} > 0, \sum_{l=1}^d \frac{1}{\sigma_{sl}} = \frac{d}{\sigma} \text{ for all } 1 \leq s \leq S \right\}, \tag{7}$$

for some $\sigma \in (0, 1]$, and

$$\widetilde{\Sigma} = \{(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) : \sigma_{sl} > 0 \text{ for all } 1 \le s \le S, 1 \le l \le d\}. \tag{8}$$

The anisotropic Besov ball and its region-wise representation are a more general characterization of spatial heterogeneity. They can cover the anisotropic Besov space as a special case. The intuition behind this characterization is that with respect to a sequence of carefully designed multiresolution Haar basis, high resolution wavelet coefficients decay very fast. An extreme case is that with respect to the Haar wavelet a piecewise constant function defined on a dyadic partition $\mathcal{A}_0$ lies in a Besov ball with arbitrarily large σ_l's.

2.4. Weak-$\ell_{p'}$ sparsity in the spectral domain

Besov balls are collections of functions with fast decaying wavelet coefficients. This is essentially a reflection of sparsity of wavelet coefficients. Now we further generalize our definition of spatial heterogeneity by imposing a weaker sparsity condition on Haar coefficients, namely, the weak-$\ell_{p'}$ condition.

With respect to a mutiresolution Haar basis $\boldsymbol{\Phi}_{\boldsymbol{\sigma}}^{(j_0)}|_\Omega \cup (\cup_{j=j_0}^{\infty} \boldsymbol{\Xi}_{\boldsymbol{\sigma}}^{(j)}|_\Omega)$ as defined in Section 2.3, for any function $f \in L^2(\Omega)$, the expansion (5) holds. We can rearrange wavelet coefficients in expansion (5) according to their size: $|\langle g, \xi_{(1)}\rangle| \ge |\langle g, \xi_{(2)}\rangle| \ge \cdots \ge |\langle g, \xi_{(k)}\rangle| \ge \cdots$. Then the sparsity condition imposed on f is that the decay of wavelet coefficients follows a power law,

$$|\langle g, \xi_{(k)}\rangle| \le C k^{-1/p'} \text{ for all } k \in \mathbb{N} \text{ and } 0 < p' < 2, \tag{9}$$

where C is a constant. The condition (9) is called the weak-$\ell_{p'}$ condition in the literature (Abramovich et al., 2006). It has been commonly imposed to characterize the sparsity of images (Candès and Tao, 2006; DeVore et al., 1992), as well as the sparsity of signals in a Gaussian sequence model (Abramovich et al., 2006). Compared to Besov balls, the weak-$\ell_{p'}$ condition allows more local spikes, as wavelet coefficients in the Besov ball are organized according to resolutions and exhibit a fast decay from low resolutions to high ones, while for those satisfying the weak-$\ell_{p'}$ condition, a high-resolution coefficient is allowed to have a larger size. A similar condition has been introduced by Liu et al. (2023) to describe regularity of density functions. In our paper, the condition (9) can cover more different types of spatial heterogeneity, as we allow scaling of the basis function to rely on a parameter $\boldsymbol{\sigma}$.

In order to avoid extremely local peaks that cannot be efficiently captured by trees, we further impose a constraint on how local or equivalently high resolution a leading coefficient can be.

Condition 2.1 (Moderate Resolution) *Let $\xi_{(k)}$ be the multiresolution wavelet basis corresponding to the k-th largest coefficient in size in expansion (5), we assume that there exist constants $c_v, c_\mu > 0$, such that the volume of the supporting rectangle for $\xi_{(k)}$ satisfies*

$$\mu\left(supp(\xi_{(k)})\right) \ge c_\mu k^{-c_v}. \tag{10}$$

The Condition 2.1 is quite mild in the sense that it does allow a high resolution wavelet coefficient to have a larger size than a low resolution one.

We define the weak-$\ell_{p'}$ ball as

$$\begin{aligned} \mathrm{wl}_{\boldsymbol{\sigma}}^{p'}(\Omega, C) \quad = \quad & \{f \in L^2(\Omega) : f \text{ satisfies (9) and (10) with respect to} \\ & \text{the multiresolution Haar basis } \boldsymbol{\Phi}_{\boldsymbol{\sigma}}^{(j_0)}|_\Omega \cup (\cup_{j=j_0}^{\infty} \boldsymbol{\Xi}_{\boldsymbol{\sigma}}^{(j)}|_\Omega)\}. \end{aligned}$$

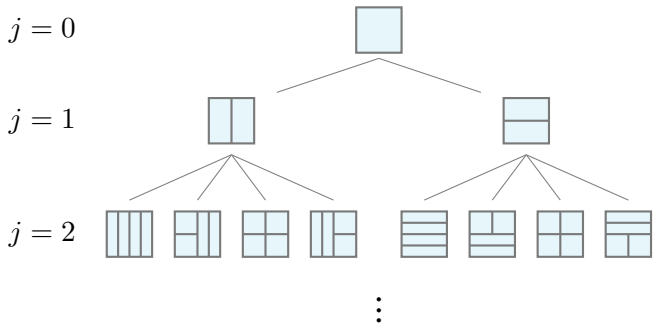

Figure 1: Recursive partitions of $[0,1]^2$ up to level 2 without stopping. The recursion is with respect to tree depth.

Between characterizations of spatial heterogeneity in terms of anisotropic Besov balls and weak-$\ell_{p'}$ sparsity in the spectral domain, the latter one is more general, in the sense that the former function class can be embedded into a weak-$\ell_{p'}$ ball with appropriately chosen p' and constant C. In specific, for $\boldsymbol{\sigma} = (\sigma_l)_{1\le l\le d}$, we have $\widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L) \subset \mathrm{wl}^{p'}_{\boldsymbol{\sigma}}(\Omega, C)$ with $p' = 1/(\bar{\sigma}/d + 1/2)$ and $C = 2^d L$.

3. Prior distributions on trees

In this section, we introduce partitions or tree topologies of the sample space, and present two types of priors on density functions defined on trees.

3.1. Optional Pólya tree

Under the Bayesian framework, a natural conjugate prior for probability measures is the optional Pólya tree prior (OPT, Wong and Ma 2010), which is an extension of the Pólya tree (Ferguson, 1974). By introducing a stopping rule, the prior can flexibly adjust the probability mass on trees of different depths, which plays a key role in variance reduction.

When constructing a tree, we consider a recursive procedure with respect to the tree depth. To simplify the theoretical analysis, we assume that the partition of the sample space is *dyadic*. The tree growth starts from splitting the root Ω into two equally sized disjoint rectangles along the *midpoint* of a randomly selected dimension: $\Omega = \Omega_{\kappa,0} \cup \Omega_{\kappa,1}$, $\kappa \in [d]$, where each $\Omega_{\kappa,\epsilon}$, $\epsilon \in \{0,1\}$ is called a level-1 elementary region, and can in turn be divided into level-2 elementary regions, while Ω itself can also be viewed as a level-0 elementary region. In general, for any level-j elementary region A, the volume is 2^{-j}, and there are always d ways to partition it; that is,

$$A = A_{\kappa,0} \cup A_{\kappa,1}, \quad \text{where } \kappa \in [d]. \tag{11}$$

The collection of all possible level-j elementary regions is denoted as $\mathcal{R}^{(j)}$. Figure 1 displays all possible partitions of Ω up to level 2 *without stopping* in the two-dimensional case.

To grow an optional Pólya tree, a recursive procedure is employed to generate a random recursive partition of Ω and a random probability measure Q. Suppose after j steps of the recursion, we have obtained a random recursive partition $\{\Omega_k\}_{k=1}^t$, $1 \le t \le 2^j$. In addition, we have also obtained a random probability measure $Q^{(j)}$ on Ω which is uniformly distributed within each region Ω_k, $1 \le k \le t$. In the $(j+1)$-th step, the random recursive partition and the random probability measure $Q^{(j+1)}$ are obtained by further partitioning non-stopped level-j regions in $\{\Omega_k\}_{k=1}^t$ and randomly assigning probability mass to child regions. In specific, for each non-stopped level-j

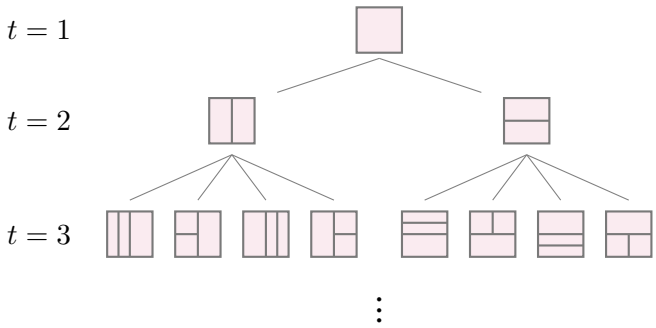

Figure 2: All recursive partitions of $[0,1]^2$ of size up to 3. The recursion is with respect to the number of terminal nodes.

region A, we introduce a *stopping rule* as an independent Bernoulli random variable,

$$S(A) \sim \text{Bernoulli}\left(\tau(A)\right), \quad \text{stopping rule for a level-}j\text{ region } A,$$

where the parameter $\tau(A)$ can be region specific. If $S(A) = 0$, stop further partitioning of A. Otherwise, select a dimension uniformly from $\{1, \ldots, d\}$ and split A along the midpoint of the selected dimension κ as described in (11). Note that once a region is stopped, it will remain intact in later steps. To generate the random probability $Q^{(j+1)}$, on any stopped region, $Q^{(j+1)}$ is the same as $Q^{(j)}$. If non-stopped level-j region A is split into $A_{\kappa,0}$ and $A_{\kappa,1}$ in the $(j+1)$-th step, then we assign probability mass randomly according to an independent beta random variable:

$$Q^{(j+1)}(A_{\kappa,0}) = Q^{(j)}(A) \cdot \omega(A), \quad Q^{(j+1)}(A_{\kappa,1}) = Q^{(j)}(A) \cdot (1 - \omega(A)),$$
$$\text{where} \quad \omega(A) \sim \text{Beta}(a(A,\kappa), b(A,\kappa)).$$

The parameters of the beta random variable may depend on the region A and the selected way of partitioning.

The collection of hyper-parameters for the OPT prior is

$$\Phi = \{\tau(A), a(A,\kappa), b(A,\kappa)\}_{A \in \cup_{j \geq 0} \mathcal{R}^{(j)}, \kappa \in [d]}.$$

In this paper, we consider the following specification of hyper parameters. For any level-j elementary region A, we set the stopping parameter $\tau(A) = c_\tau 2^{-\nu j}$ for some constants $0 < c_\tau \leq 1/2$ and $\nu \geq 1$, and the parameters for the beta distribution $a(A,\kappa) = b(A,\kappa) = \alpha_0$, $0 < \alpha_0 < 1$ for all A and κ. This way, we impose a penalty on complexity of tree topology by penalizing the depth: the larger j is, the more unlikely a level-j region will be further split. This specific type of OPT prior is denoted as $\pi_{\text{OPT}}(\cdot)$.

For the rest, we will use the simplified notation $\mathcal{A} = \{\Omega_k\}_{k=1}^t$ to denote a tree structure. A tree $\mathcal{A}$ is of depth $\mathfrak{D}$ if all leaf nodes Ω_k's are elementary regions of level at most $\mathfrak{D}$ and at least one leaf node is exactly level-$\mathfrak{D}$. Intuitively, the depth roughly characterizes the complexity of tree topology.

3.2. The Dirichlet prior

As opposed to imposing a penalty on tree depth, the Dirichlet prior (Lu et al., 2013; Liu et al., 2023) directly penalizes the number of terminal nodes in a tree. When growing the tree, we still only allow *dyadic* partitions and employ a recursive partitioning procedure. But for the Dirichlet prior

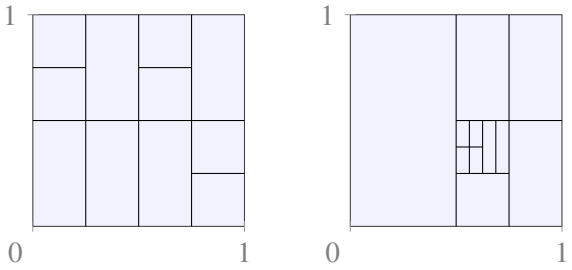

Figure 3: The two tree topologies are of same size, but different depths. Both trees are of size 11. The left tree is of depth 4, while the right one is of depth 7.

we assume the recursion is with respect to the number of terminal nodes, instead of the depth: at every step, *one* region or node will be randomly selected and divided into two along the midpoint of a randomly chosen dimension. We refer the number of terminal nodes in a partion as the size. Figure 2 presents all recursive partitions of size up to 3 in the two-dimensional case.

To impose a prior on tree-supported densities, we first assign a prior to different tree topologies; and given a tree structure, we sample probability weights associated with leaf nodes from a Dirichlet distribution. More specifically, we assume all tree topologies of the same size share the same prior, and the prior probability on a size t tree $\mathcal{A}_t$ is proportional to $\exp(-\lambda t \log t)$ for some $\lambda > 0$. Given the partition $\mathcal{A}_t$, the prior on leaf node probabilities is the Dirilet distribution with parameters $\alpha_1 = \cdots \alpha_t = \alpha_0$ for some $\alpha_0 \in (0, 1)$. We denote this prior as $\pi_{\text{Dir}}(\cdot)$.

3.3. A comparison between the two types of priors

We provide a simple example to illustrate the difference between the two types of prior distributions, namely π_{OPT} and π_{Dir}. In summary, the major difference is how the prior imposes a penalty on tree complexity: the OPT penalizes the depth, while the Dirichlet prior penalizes the size. In Figure 3, we show two tree topologies of the same size, but different depths. The Dirichlet prior assigns equal probability to the two tree structures, while the OPT is in favor of the left one. This toy example also suggests the Dirichlet prior may have a stronger ability to capture local features. We will further explain this property when we present the results on posterior concentration rates.

4. Spatial adaptation properties

Assume we have an i.i.d. sample $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ from the unknown distribution f_0. Under a prior distribution π on the space of densities $\mathcal{F}$, the posterior distribution on $\mathcal{F}$ is the random measure

$$\Pi(B|\mathcal{Y}_n) = \frac{\int_B \prod_{i=1}^n f(Y_i) d\pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n f(Y_i) d\pi(f)},$$

for any measurable set $B \subset \mathcal{F}$. We will derive the posterior concentration rate, denoted as $\epsilon_n \to 0$, such that for a sequence $\{M_n\}_{n \geq 1}$ that increases at most as polynomials of $\log n$,

$$\Pi(f : \rho(f, f_0) \geq M_n \epsilon_n | \mathcal{Y}_n) \to 0, \quad \text{in } \mathbb{P}^n_{f_0}\text{-probability}.$$

In order to guarantee that on average there are enough number of data points in each leaf node, we assume that for the OPT prior the probability of making a further split is 0 beyond depth $\overline{\mathfrak{D}} =$

$\lfloor \log_2(n/\log n) \rfloor$, and for the Dirichlet prior the probability vanishes on tree topologies with size larger than $\bar{t} = \lfloor n/\log n \rfloor$.

For the Besov space $B^{\boldsymbol{\sigma}}_{p,q}(\mathcal{R}, L)$ or the Besov ball $\widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\mathcal{R}, L)$ with smoothness parameter $\boldsymbol{\sigma}$, $\bar{\sigma}$ is the value that satisfies $(\bar{\sigma})^{-1} = (\sum_{l=1}^{d} \sigma_l^{-1})/d$. In the following, we always assume that $2 \leq p \leq \infty$, $1 \leq q \leq \infty$. In order to avoid an assumption that restricts f_0 to be bounded away from zero, or obtaining a density function taking negative values after truncating an L^2-expansion of form (5), we choose to describe spatial features of $g_0 = \sqrt{f_0}$ instead of f_0. We first present our results when the heterogeneity is modeled by the anisotropic Besov space or the anisotropic Besov ball.

Theorem 1 *Assume $g_0 \in B^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ with $\boldsymbol{\sigma} \in (0,1]^d$ or $g_0 \in \widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ with $\boldsymbol{\sigma} \in (0,\infty]^d$. Under both the OPT prior and the Dirichlet prior, the posterior concentration rate for unsupervised trees is $\epsilon_n = (Ld)^{\frac{d}{2\bar{\sigma}+d}} n^{-\frac{\bar{\sigma}}{2\bar{\sigma}+d}}$.*

For the anisotropic Besov space, the rate is minimax up to a logarithmic term (near minimax). The theorem suggests that unsupervised trees under both types of priors can well adapt to this type of spatial heterogeneity. The statement in Theorem 1 for anisotropic Besov balls also covers the variable screening property by trees as a special case. As we explained before, a smoothness parameter $\sigma_l = \infty$ means there is no need to perform MRA along that direction, and in the density estimation problem this is equivalent to saying that, the density function does not show any variations along that direction. In this case, if we denote $d' = |\{l \in [d] : \sigma_l < \infty\}|$ and $(\bar{\sigma}')^{-1} = (\sum_{l:\sigma_l<\infty} \sigma_l^{-1})/d'$, then the rate in this special case is $\epsilon_n = (Ld)^{\frac{d'}{2\bar{\sigma}'+d'}} n^{-\frac{\bar{\sigma}'}{2\bar{\sigma}'+d'}}$, implying that tree-based methods can successfully identify effective dimensions, and achieve a rate mainly determined by d' instead of the full dimension d. The rate in Theorem 1 is also adaptive in the sense that to achieve such a rate, one does not need any prior knowledge of $\boldsymbol{\sigma}$.

Next, we consider region-wise anisotropic Besov spaces and region-wise anisotropic Besov balls. For a tuple of smoothness parameters $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S)$, we define $\bar{\sigma}_{\min} = \min_{1 \leq s \leq S} \bar{\sigma}_s$, where each $\bar{\sigma}_s$ is calculated based on $\boldsymbol{\sigma}_s$.

Theorem 2 *Given a fixed dyadic partition $\mathcal{A}^\star = \{\Omega_s\}_{s=1}^S$ of Ω, assume $g_0|_{\Omega_s} \in B^{\boldsymbol{\sigma}_s}_{p,q}(\Omega_s, L)$ or $g_0 \in \widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ with unknown $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S)$. When the tuple of smoothness parameters $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) \in \Sigma_\sigma$ for region-wise anisotropic Besov spaces or lies in $\widetilde{\Sigma}_\sigma$ for region-wise anisotropic Besov balls, under both the OPT prior and the Dirichlet prior the posterior concentration rate for unsupervised trees is $\epsilon_n = (Ld)^{\frac{d}{2\sigma+d}} S^{\frac{\sigma+d}{2\sigma+d}} n^{-\frac{\sigma}{2\sigma+d}}$. When $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) \in \Sigma$ for anisotropic Besov spaces or in the set $\widetilde{\Sigma}$ respectively for anisotropic Besov balls, under the Dirichlet prior the posterior concentration rate for unsupervised trees is at least $\epsilon_n = (Ld)^{\frac{d}{2\bar{\sigma}_{\min}+d}} S^{\frac{\bar{\sigma}_{\min}+d}{2\bar{\sigma}_{\min}+d}} n^{-\frac{\bar{\sigma}_{\min}}{2\bar{\sigma}_{\min}+d}}$.*

Theorem 2 reveals how strong spatial adaptivity tree-based methods have. This is especially the case when spatial features are described by region-wise anisotropic Besov balls. In this case, we only require that in each region Ω_s one can find a system of multiresolution Haar basis functions whose scaling and translation at different levels are determined by some unknown parameter $\boldsymbol{\sigma}_s$, such that the density function has fast decaying coefficients with respect to those basis functions as the resolution increases. Extremely flexible designs of multiresolution basis suggest the wide range of spatial heterogeneity that can be characterized by this condition. When applying tree-based methods, we do not need *any* information of the specific design of the basis, nor the decay

rate of the wavelet coefficients. The methods will learn a partition that best adapt to the unknown spatial features. Note that given the embedding between the Hölder space and the Besov one, our results can cover adaption to anisotropic Hölder classes or region-wise anisotropic Hölder ones as a special case.

Our last result will be spatial adaptation by trees under the weak-$\ell_{p'}$ characterization of the spatial heterogeneity.

Theorem 3 *With respect to a system of multiresolution Haar basis on Ω, denoted by $\mathbf{\Phi}_{\boldsymbol{\sigma}}^{j_0}|_\Omega \cup (\cup_{j\geq j_0}\mathbf{\Xi}_{\boldsymbol{\sigma}}^{(j)}|_\Omega)$, assume that $g_0 \in wl_{\boldsymbol{\sigma}}^{p'}(\Omega, C)$. Under the Dirichlet prior, the posterior concentration rate for unsupervised trees is $\epsilon_n = C^{p'/2} n^{-\frac{1-p'/2}{2}}$.*

Except for the region-wise characterization, the weak-$\ell_{p'}$ sparsity is the most general characterization of the spatial heterogeneity, in the sense that for $\boldsymbol{\sigma} \in (0, 1/p)^d$, $B_{p,q}^{\boldsymbol{\sigma}}(\Omega, L) \subset \widetilde{B}_{p,q}^{\boldsymbol{\sigma}}(\Omega, L)$ and for $\boldsymbol{\sigma} \in (0, \infty]^d$ $\widetilde{B}_{p,q}^{\boldsymbol{\sigma}}(\Omega, L) \subset \mathrm{wl}_{\boldsymbol{\sigma}}^{p'}(\Omega, C)$ with $p' = 1/(\bar{\sigma}/d + 1/2)$. The embedding also shows that under the Dirichlet prior tree-based methods have stronger ability to adapt to spatial features in the sense that they can achieve fast convergence over a larger function class. This is mainly due to the different way of imposing penalty on tree complexity. Similar to Theorem 2, the result in Theorem 3 is adaptive in the sense that trees can successfully capture spatial heterogeneity described by a system of multiresolution basis with unknown parameter $\boldsymbol{\sigma}$. For Besov balls, region-wise anisotropic Besov balls and weak-$\ell_{p'}$ balls, the rate can be near the parametric rate, provided organized wavelet coefficients show fast decay. This is different from results obtained under the Hölder-continuity based characterization of spatial heterogeneity, where the rate usually takes the form $(Ld)^{\frac{d'}{2\alpha+d'}}(S/n)^{\frac{\alpha}{2\alpha+d'}}$, with α representing the smoothness level and taking value at most 1, and d' being interpreted as the effective dimension (Jeong and Ročková, 2023). By using the other characterization, the rate cannot be very close to $\sqrt{n}$. Empirically, we do observe tree-based methods can perform quite well, especially when the underlying density has a simple structure. From the semiparametric efficiency perspective (van der Vaart, 2000), it should be possible to achieve near parametric rate by trees. For example, this would be the case if the underlying density can be parametrized by finite number of wavelet basis functions. Therefore, we think the characterization of spatial heterogeneity by decay of wavelet coefficients can better reveal spatial adaptation properties of trees.

5. Outline of the proof

The machinery for analyzing posterior concentration rates has been introduced in the landmark works by Ghosal et al. (2000) and by Shen and Wasserman (2001). In specific, we will apply Theorem 2.1 in the paper by Ghosal et al. (2000) here. Assume Π_n is a prior distribution on the infinite dimensional space $\mathcal{F}$ of density functions, where the subscript n indicates that the specification of the prior can be data dependent, which is the case when we apply a truncation to both π_{OPT} and π_{Dir}. Suppose that for a sequence of ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, a constant $D > 0$ and a sieve $\mathcal{F}_n \subset \mathcal{F}$, we have

$$\log N(\epsilon_n/2, \mathcal{F}_n, \rho) \leq n\epsilon_n^2, \tag{12}$$

$$\Pi_n(\mathcal{F} \setminus \mathcal{F}_n) \leq \exp(-n\epsilon_n^2(D+4)), \tag{13}$$

$$\Pi_n\left(f : \mathrm{KL}(f_0, f) \le \epsilon_n^2, \int f_0(\log f_0/f)^2 \le \epsilon_n^2\right) \ge \exp(-Dn\epsilon_n^2). \tag{14}$$

Then for a sequence of M_n that increases at most as polynomials of $\log n$, we have that $\Pi_n(f : \rho(f_0, f) \ge M_n\epsilon_n|\mathcal{Y}_n) \to 0$ in $\mathbb{P}^n_{f_0}$-probability. The condition (12) is a bound of the covering number for the sieve $\mathcal{F}_n$'s, where $N(\epsilon, \mathcal{F}, \rho)$ is the minumum number of ϵ-balls under the metric ρ that are needed to cover the space $\mathcal{F}$. Condition (13) requires that the tree prior should decay fast enough as the tree complexity increases, while the last condition is to guarantee that the prior puts enough probability mass around the true density and is also called the condition on prior thickness. We will first show the following proposition for the OPT prior in Appendix A by checking these three conditions.

Proposition 4 *Assume that for any function f lying in a function class $\mathcal{G}$, we can find a sequence of depth-$\mathfrak{D}$ tree-supported densities $f_{\mathfrak{D}}$'s, such that $\rho(f, f_{\mathfrak{D}}) \le c_{\text{approx}}\mathfrak{D}2^{-r\mathfrak{D}}$ for all $\mathfrak{D}$, where $c_{\text{approx}} > 0$ and $r > 0$ are some constants. Then for $f_0 \in \mathcal{G}$, the posterior concentration rate of Bayesian unsupervised trees under the OPT prior is $\epsilon_n = (c_{\text{approx}})^{\frac{1}{2r+1}} n^{-\frac{r}{2r+1}}$ up to a logarithmic term.*

Similarly, if f in a function class can by approximated by a size-t tree-supported density at a rate $c_{\text{approx}}(t/\log t)^{-r}$ under the Hellinger distance, then the posterior concentration rate under the Dirichlet prior is also $\epsilon_n = (c_{\text{approx}})^{\frac{1}{2r+1}} n^{-\frac{1}{2r+1}}$. This is exactly Theorem 4 in the paper by Liu et al. (2023). These two results imply, as long as we can obtain the approximation rate for different function spaces, we can directly obtain the posterior concentration rate. The approximation results will be presented in Appendix B.

6. Discussion

In this paper, we introduce multiple ways to characterize spatial features of a data distribution, and show how tree-based methods can effectively adapt to the unknown structure. Our analysis is under the Bayesian framework, and we mainly examine two types of priors on trees. Our characterization of spatial heterogeneity can also be extended to the supervised learning setting, to describe spatial properties of the conditional mean in either a regression or a classification problem. It would be interesting to perform the analysis for supervised trees in future work, which can help better understand this class of widely used learning methods.

Acknowledgments

This research is partly supported by NSF grants DMS-1749789 and DMS-2152999.

References

Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint: arXiv:1407.3939*, 2014.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13: 1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, 2008.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Emmanuel. J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

Matias D. Cattaneo, Jason M. Klusowski, and William G. Underwood. Inference with Mondrian random forests. *arXiv preprint: arXiv2310.09702*, 2023.

Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.

Hugh Chipman, Edward I. George, and Robert E. McCulloch. A Bayesian approach to CART. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, volume R1 of *Proceedings of Machine Learning Research*, pages 91–102. Fort Lauderdale, FL, USA, 1997.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, pages 665–673. Beijing, China, 2014.

Ronald A. DeVore, B.D. Jawerth, and B.J. Lucier. Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38(2):719–746, 1992.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, ICLR, Workshop Track Proceedings*. San Diego, CA, USA, 2015.

Thomas S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2:615–629, 1974.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Gustavo Garrigós and Anita Tabacco. Wavelet decompositions of anisotropic besov spaces. *Mathematische Nachrichten*, 239-240(1):80–102, 2002.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Montréal, Canada, 2014.

David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

Seonghyun Jeong and Veronika Ročková. The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension. *Journal of Machine Learning Research*, 24(337):1–65, 2023.

Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*. ICLR, 2014.

Jason Klusowski. Sparse learning with CART. In *Advances in Neural Information Processing Systems*, volume 33, pages 11612–11622. 2020.

Jason Klusowski. Sharp analysis of a simple model for random forests. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 757–765. 2021.

Christopher Leisner. Nonlinear wavelet approximation in anisotropic besov spaces. *Indiana University Mathematics Journal*, 52(2):437–455, 2003.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Linxi Liu, Dangna Li, and Wing Hung Wong. Convergence rates of a partition based Bayesian multivariate density estimation method. In *Advances in Neural Information Processing Systems 30*, pages 4738–4746. Long Beach, California, 2017.

Linxi Liu, Dangna Li, and Wing Hung Wong. Convergence rates of a class of multivariate density estimation methods based on adaptive partitioning. *Journal of Machine Learning Research*, 24 (50):1–64, 2023.

Luo Lu, Hui Jiang, and Wing H. Wong. Multivariate density estimation by Bayesian sequential partitioning. *Journal of the American Statistical Association*, 108(504):1402–1410, 2013.

Li Ma and Wing Hung Wong. Coupling optional Pólya trees and the two sample problem. *Journal of the American Statistical Association*, 106(496):1553–1565, 2011.

Katherine M. McKinnon. Flow cytometry: An overview. *Current Protocols in Immunology*, 120 (1):5.1.1–5.1.11, 2018.

Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020.

Y. Meyer. *Ondelettes et opérateurs*. Number v. 1 in Actualités mathématiques. Hermann, 1990.

Sergei Mihailovic Nikol'skii. *Approximation of Functions of Several Variables and Imbedding Theorems*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 2012.

Hong Ooi. Density visualization and mode hunting using trees. *Journal of Computational and Graphical Statistics*, 11(2):328–347, 2002.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

G. Poggi and R. A. Olshen. Pruned tree-structured vector quantization of medical images with segmentation and improved prediction. *IEEE Transactions on Image Processing*, 4(6):734–742, 1995.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530–1538. Lille, France, 2015.

Veronika Ročková and Judith Rousseau. Ideal bayesian spatial adaptation. *Journal of the American Statistical Association*, forthcoming.

Veronika Ročková and Stéphanie van der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131, 2020.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62 (3):1485–1500, 2016.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741, 2015.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Jacopo Soriano and Li Ma. Probabilistic multi-resolution scanning for two-sample differences. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):547–572, 2017.

Ananya Uppal, Shashank Singh, and Barnabás Poczós. Nonparametric density estimation and convergence rates for GANs under Besov IPM losses. In *Advances in Neural Information Processing Systems 32*, pages 9086–9097. Vancouver, Canada, 2019.

A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

Wing Hung Wong and Li Ma. Optional Pólya tree and Bayesian inference. *The Annals of Statistics*, 38(3):1433–1459, 2010.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.

Appendix A. Proof of Proposition 4

The proof is finished by checking conditions (12), (13) and (14). We use Θ_t to denote the collection of size-t tree supported density functions. The sieve $\mathcal{F}_n$ is set to be Θ_{t_n} with $t_n = \lfloor (c_{\text{approx}}^2 n / \log n)^{\frac{1}{2r+1}} \rfloor$.

A.1. An upper bound for the covering number

The purpose of the subsection is to obtain an upper bound for the covering number of the tree space, and verify the condition (12).

Lemma 5 *Take ρ to be the Hellinger distance. Let $\Theta_t^{\mathcal{A}} = \{f : f \text{ is a density defined on the partition } \mathcal{A} = \{\Omega_k\}_{k=1}^t\}$. Then,*

$$\log N(u, \Theta_t^{\mathcal{A}}, \rho) \leq \frac{t}{2} \log t + t \log \frac{2}{u} + b_{entropy}, \tag{15}$$

where $b_{entropy}$ is a constant not dependent on the recursive partition.

The lemma is a direct consequence of Lemma 14 in Liu et al. (2023), as the covering number is always bounded by the covering number with bracketing. Based on this bound, we can obtain an upper bound for the covering number of Θ_t.

Lemma 6 *For the space of piecewise constant densities Θ_t,*

$$\log N(u, \Theta_t, \rho) \leq (b_{path} + \frac{1}{2}) t \log t + t \log d + t log(2/u) + b_{entropy}, \tag{16}$$

where $b_{path} > 0$ is a constant.

Proof As $\Theta_t = \cup_{\mathcal{A}:\mathcal{A} \text{ is a size-}t \text{ partition}} \Theta_t^{\mathcal{A}}$, the covering number for Θ_t is bounded by

$$N(u, \Theta_t, \rho) \leq \sum_{\mathcal{A}:\mathcal{A} \text{ is a size-}t \text{ partition}} N(u, \Theta_t^{\mathcal{A}}, \rho).$$

Different number of partitions of size t is bounded by $d^t t!$. In combination with the bound (15), we obtain (16). ■

By applying the upper bound (15), after some simple calculation, we can verify that the condition (12) is satisfied when $\epsilon_n = \eta n^{\frac{r}{2r+1}}$, where $\eta > 1$ is allowed to increase at most as polynomials of $\log n$.

A.2. Prior thickness

The purpose of this subsection is to verify the condition (14). We will first provide some analysis of the OPT prior.

A.2.1. PRIOR ON A NON-STOPPED TREE TOPOLOGY

We consider a special type of tree topologies of depth exactly equal to $\mathfrak{D}$, generated under the OPT prior by growing the tree without stopping any intermediate nodes at all lower levels $j < \mathfrak{D}$, and stopping all level-$\mathfrak{D}$ regions. This implies that the volume of all terminal nodes is $2^{-\mathfrak{D}}$. We denote such a depth-$\mathfrak{D}$ tree as $\mathcal{T}$, and assume it is generated through non-stopped intermediate trees $\mathcal{T}^{(0)}, \mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(\mathfrak{D}-1)}$ of depth $0, 1, \ldots, \mathfrak{D}-1$ respectively ($\mathcal{T}^{(0)}$ is simply Ω). The terminal depth-$\mathfrak{D}$ tree can also be denoted as $\mathcal{T}^{(\mathfrak{D})}$ with a convention $\mathcal{T} = \mathcal{T}^{(\mathfrak{D})}$. The size of $\mathcal{T}^{(j)}$ is equal to 2^j as there is no stopping at lower levels. We use $\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}$ to denote the collection of density functions defined on $\mathcal{T}$ grown as $\mathcal{T}^{(0)}, \mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(\mathfrak{D})}$. We have the following lemma for prior probability mass on $\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}$.

Lemma 7 *Under the OPT prior, the prior on a non-stopped tree topology defined as above satisfies*

$$(\tau_{\mathfrak{D}}/d)^{2^{\mathfrak{D}}} \cdot \pi_{OPT}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}) \le \pi_{OPT}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}+1\}}) \le (2\tau_{\mathfrak{D}}/d)^{2^{\mathfrak{D}}} \cdot \pi_{OPT}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}),$$

and this holds for all $\mathfrak{D} \ge 0$.

Proof First, we can show the following recursion for the prior on non-stopped tree structures

$$\pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}+1\}}) = \left(\frac{\tau_{\mathfrak{D}}(1-\tau_{\mathfrak{D}+1})^2}{d(1-\tau_{\mathfrak{D}})}\right)^{2^{\mathfrak{D}}} \cdot \pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}), \quad \text{for } \mathfrak{D} \ge 0,$$

where $\tau_{\mathfrak{D}}$ is the probability of making a further split at level $\mathfrak{D}$. According to our specification of hyper-parameters, the stopping rule for all level-j regions is solely a function of j. The proof is based on the observation that the level-$(\mathfrak{D}+1)$ non-stopped tree is obtained by making a further split of every terminal node in the level-$\mathfrak{D}$ tree. When making the further split, we also need to randomly choose a dimension for splitting at each node.

As $2\tau_{\mathfrak{D}+1} \le \tau_{\mathfrak{D}}$ ($\nu \ge 1$) and $1-\tau_{\mathfrak{D}} \ge 1/2$ ($c_\tau \le 1/2$), it is not hard to see that $2 \ge \frac{(1-\tau_{\mathfrak{D}+1})^2}{1-\tau_{\mathfrak{D}}} = \frac{1-2\tau_{\mathfrak{D}+1}+\tau_{\mathfrak{D}+1}^2}{1-\tau_{\mathfrak{D}}} \ge 1$. The desired result follows. ■

Lemma 8 *Under the OPT prior, we have the following lower and upper bounds for the prior on a non-stopped tree,*

$$\log \pi_{OPT}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}) \ge \log(1-c_\tau) + (2^{\mathfrak{D}}-2)\log(c_\tau/d) - \nu \log 2\left((\mathfrak{D}-1)2^{\mathfrak{D}} - 2^{\mathfrak{D}} + 2\right),$$

$$\log \pi_{OPT}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}) \le \log(1-c_\tau) + (2^{\mathfrak{D}}-2)\log(2c_\tau/d) - \nu \log 2\left((\mathfrak{D}-1)2^{\mathfrak{D}} - 2^{\mathfrak{D}} + 2\right).$$

Proof We will show the upper bound, and the proof for the lower bound is similar. According to the result of Lemma 7 and the fact that $\tau_j = c_\tau 2^{-\nu j}$, for all $\mathfrak{D} \ge 0$,

$$\log \pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}+1\}}) \le 2^{\mathfrak{D}} \log(2/d) - (\nu\mathfrak{D})2^{\mathfrak{D}} \log 2 + 2^{\mathfrak{D}} \log c_\tau + \log \pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}).$$

Since $\pi_{\text{OPT}}(\mathcal{F}_{\mathcal{T}^{(0)}}) = 1 - c_\tau$, based on simple calculations we have

$$\begin{aligned}
\log \pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}+1\}}) &\le \log(1-c_\tau) + \sum_{j=1}^{\mathfrak{D}} \left(2^j \log(2c_\tau/d) - \nu j \cdot 2^j \log 2\right) \\
&= \log(1-c_\tau) + (2^{\mathfrak{D}+1}-2)\log(2c_\tau/d) \\
&\quad -\nu \log 2\left(\mathfrak{D}2^{\mathfrak{D}+1} - 2^{\mathfrak{D}+1} + 2\right).
\end{aligned}$$

This finished the proof. ■

A.2.2. A LOWER BOUND FOR THE OPT PRIOR

Assume $\mathcal{T}$ is a non-stopped tree topology as described in the previous subsection. It is grown as $\mathcal{T}^{(0)}, \mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(\mathfrak{D})}$. Let $g^{(\mathfrak{D})}$ be the L^2-projection of an arbitrary density g onto $\mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}}$. In the following, we consider an L^∞-ball around $g^{(\mathfrak{D})}$, and derive a lower bound for the prior probability on this ball. For a density function f and a set A, we denote $\mu_f(A) = \int_A f d\mu$, where μ is the Lebesgue measure. For some constant $\eta \geq 1$, we define a ball around $g^{(\mathfrak{D})}$ to be the collection of tree-supported densities that are uniformly close to $g^{(\mathfrak{D})}$ on all leaf nodes, and are bounded away from 0:

$$\begin{aligned} B_{\mathfrak{D}}(g) \;\coloneqq\; \Big\{ f \in \mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}} : |\mu_f(A) - \mu_g(A)| \leq \eta \mathfrak{D}/n \\ \text{and } \mu_f(A) \geq \left(\frac{1}{2}\eta/n\right)^{\mathfrak{D}} \text{ for all } A \in \mathcal{T}^{(\mathfrak{D})} \Big\}. \end{aligned}$$

As $\mathcal{T}$ is non-stopped with each leaf node of volume $2^{-\mathfrak{D}}$, the set $B_{\mathfrak{D}}(g)$ can be equivalently written as

$$B_{\mathfrak{D}}(g) \coloneqq \left\{ f \in \mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}} : \|f - g^{(\mathfrak{D})}\|_\infty \leq \eta \mathfrak{D} 2^{\mathfrak{D}}/n, \text{ and } f \geq \left(\frac{1}{2}\eta/n\right)^{\mathfrak{D}} 2^{\mathfrak{D}} \right\}.$$

Then, we will show the following lemma for the prior probability on $B_{\mathfrak{D}}(g)$.

Lemma 9 *Under the OPT prior, for the set $B_{\mathfrak{D}}(g)$ defined above,*

$$\log \pi_{\mathit{OPT}}(B_{\mathfrak{D}}(g)) \geq \log \pi_{\mathit{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}}) - 2^{\mathfrak{D}} \log n + 2^{\mathfrak{D}} \log((\eta/2) \cdot \Gamma(2\alpha)/(2\Gamma(\alpha))).$$

Proof Under the OPT prior, sampling of a density function can be finished by the following two steps: first, sample a tree topology from the prior; second, conditional on the tree topology, split probabilities on each node according to a beta random variable. Under this view, it suffices to obtain a lower bound for $\pi_{\text{OPT}}(B_{\mathfrak{D}}(g)|\mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}})$.

We claim that a sufficient condition for obtaining a density function in $B_{\mathfrak{D}}(g)$ is to split probability mass at all intermediate nodes $A \in \mathcal{T}^{(j)}$, $0 \leq k < \mathfrak{D}$ in a way that is close to how $g^{(\mathfrak{D})}$ splits. More specifically, for an intermediate node A^{int} being split into left child A_0^{int} and right child A_1^{int}, to split the probability mass, $g^{(\mathfrak{D})}$ will split according to $\mu_g(A_0^{\text{int}})/\mu_g(A^{\text{int}})$, while the OPT splits according to an independent beta random variable $\omega(A^{\text{int}}) \sim \text{Beta}(a, a)$. If $|\omega(A^{\text{int}}) - \mu_g(A_0^{\text{int}})/\mu_g(A^{\text{int}})| \leq \eta/n$ holds for all intermediate nodes, then we that claim the sampled density f satisfies $|\mu_f(A) - \mu_g(A)| \leq \eta\mathfrak{D}/n$ for all terminal nodes A. This is because under the OPT prior, $\mu_f(A)$ can always be written as a product of $\mathfrak{D}$ weights. Without loss of generality, we may assume a path from the root Ω to A is $\{A^{(j)}, 0 \leq j < \mathfrak{D}\}$, where $A^{(0)} = \Omega$. Let $\epsilon_j = 1$ if $A^{(j+1)}$ is the left child of $A^{(j)}$ and to be 0 otherwise. Then

$$\mu_f(A) = \prod_{j=0}^{\mathfrak{D}-1} \omega(A^{(j)})^{\epsilon_j} (1 - \omega(A^{(j)}))^{1-\epsilon_j}.$$

Simple calculation shows

$$|\mu_f(A) - \mu_g(A)| \leq \sum_{j=0}^{\mathfrak{D}-1} \left| \omega(A^{(j)}) - \mu_g(A^{(j+1)})/\mu_g(A^{(j)}) \right|.$$

This shows our claim. $\mu_f(A)$ is a product of $\mathfrak{D}$ beta random variables. If each of them is bounded from below by $\eta/(2n)$, then the lower bound for $\mu_f(A)$ holds.

Generally, for a random variable W following the beta distribution with parameter $(\alpha, \alpha)(0 < \alpha < 1)$, for any $w \in [0,1]$ and n large enough

$$\begin{aligned}
&\mathbb{P}(|W - w| \leq \eta/n, W \geq \frac{1}{2}\eta/n, (1-W) \geq \frac{1}{2}\eta/n) \\
&= \int_{\max\{\frac{1}{2}\eta/n, w-\eta/n\}}^{\min\{1-\frac{1}{2}\eta/n, w+\eta/n\}} \frac{\Gamma(2\alpha)}{2\Gamma(\alpha)} u^{\alpha-1}(1-u)^{\alpha-1} du \\
&\geq \frac{\Gamma(2\alpha)}{2\Gamma(\alpha)} \left(\frac{1}{2}\eta/n\right).
\end{aligned}$$

The inequality is based on the observation that for $\alpha < 1$ the integrand is larger than 1 on an interval of length at least $(\eta/2n)$. On the non-stopped tree $\mathcal{T}$ grown as $\{\mathcal{T}^{(j)} : 0 \leq j \leq \mathfrak{D}\}$, there are $2^{\mathfrak{D}} - 1$ such beta random variables. This implies that

$$\pi_{\text{OPT}}(B_{\mathfrak{D}}(g) | \mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}}) \geq \left(\frac{\Gamma(2\alpha)}{2\Gamma(\alpha)} \cdot \frac{\eta}{2n}\right)^{2^{\mathfrak{D}}-1}.$$

In combination with the lower bound for $\pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\leq j\leq \mathfrak{D}\}})$, we obtain the lower bound for $\pi_{\text{OPT}}(B_{\mathfrak{D}}(g))$. ■

The following lemma show that the density function in $B_{\mathfrak{D}}(g)$ is close to f_0 under the Hellinger distance, as long as $g^{(\mathfrak{D})}$ is close.

Lemma 10 *For any $f \in B_{\mathfrak{D}}(g)$, $\rho(f_0, f) \leq \rho(f_0, g^{(\mathfrak{D})}) + \left(4\eta\mathfrak{D}2^{\mathfrak{D}}/n\right)^{1/2}$.*

Proof First, we provide a bound for the Hellinger distance between f and $g^{(\mathfrak{D})}$. Let $\epsilon^2 = \eta\mathfrak{D}2^{\mathfrak{D}}/n$. For any $f \in B_{\mathfrak{D}}(g)$,

$$\begin{aligned}
\rho^2(g^{(\mathfrak{D})}, f) &= \int_{f:f\leq\epsilon^2/2} (\sqrt{g^{(\mathfrak{D})}} - \sqrt{f})^2 + \int_{f:f>\epsilon^2/2} (\sqrt{g^{(\mathfrak{D})}} - \sqrt{f})^2 \\
&\leq \epsilon^2/2 + \epsilon^2/2 + \epsilon^2 + \int_{f:f>\epsilon^2/2} (g^{(\mathfrak{D})} - f)^2/(\sqrt{g^{(\mathfrak{D})}} + \sqrt{f})^2 \\
&\leq 2\epsilon^2 + 2\epsilon^2.
\end{aligned}$$

Therefore, the desired result holds by the triangle inequality of the Hellinger distance. ■

In next step, we show that for any $f \in B_{\mathfrak{D}}(g)$, both the KL divergence $\text{KL}(f_0, f)$ and $\int f_0(\log(f_0/f))^2$ can be bounded by the squared Hellinger distance up to a logarithmic factor. This is a direct result of Theorem 5 in the paper by Wong and Shen (1995).

Lemma 11 *Let f, f_0 be two densities, $\rho^2(f_0, f) \leq \epsilon^2$. Suppose that $M_\delta^2 = \int_{\{f_0/f \geq e^{1/\delta}\}} f_0 (f_0/f)^\delta < \infty$ for some $\delta \in (0, 1]$. Then for all $\epsilon^2 \leq \frac{1}{2}(1 - e^{-1})^2$, we have*

$$\int f_0 \log (f_0/f) \leq \left(6 + \frac{2 \log 2}{(1 - e^{-1})^2} + (8/\delta) \max \{1, \log (M_\delta/\epsilon)\}\right) \epsilon^2,$$

$$\int f_0 \left(\log (f_0/f)\right)^2 \leq 5\epsilon^2 \left(\frac{1}{\delta} \max \{1, \log (M_\delta/\epsilon)\}\right)^2.$$

Proof See the paper by Wong and Shen (1995), Theorem 5 in Section 6. ■

In our case, to apply Lemma 11 for $f \in B_{\mathfrak{D}}(g)$ to bound $\mathrm{KL}(f_0, f)$ and $\int f_0 (\log(f_0/f))^2$, we may set $\delta = 1$. Note that any $f \in B_{\mathfrak{D}}(g)$ is bounded from below. When $\mathfrak{D}$ is at the order $\log n$, $\log(M_\delta)$ is at the order $(\log n)^2$.

To verify the condition (14), as in our paper the spatial heterogeneity assumption is imposed on $\sqrt{f_0}$, the L^2-approximation to $\sqrt{f_0}$ is exactly approximation to f_0 under the Hellinger distance. If there is a depth-$\mathfrak{D}$ tree supported density $f_{\mathfrak{D}}$ such that $\rho(f_0, f_{\mathfrak{D}}) \leq c_{\text{approx}} 2^{-r\mathfrak{D}}$, then we can always find a non-stopped depth-$\mathfrak{D}$ tree, satisfying that the $f_{\mathfrak{D}}^{(\mathfrak{D})}$ achieves the same approximation error. If we set $\mathfrak{D} = \mathfrak{D}_n = \lfloor \log_2 (c_{\text{approx}}^2 n / \log n)^{\frac{r}{2r+1}} \rfloor$, then for η increases at most as polynomials of $\log n$,

$$B_{\mathfrak{D}_n}(f_{\mathfrak{D}_n}) \subset \{f : \mathrm{KL}(f_0, f) \leq (\eta\epsilon_n)^2, \int f_0 (\log(f_0/f))^2 \leq (\eta\epsilon_n)^2\}.$$

This is a result of Lemmas 10 and 11. Then by applying Lemma 8 and Lemma 9, we can obtain the desired lower bound for the prior probability on $B_{\mathfrak{D}_n}(f_{\mathfrak{D}_n})$.

A.3. Decay of the OPT prior

The purpose of this subsection is to verify the condition (13).

First, we would like to obtain an upper bound in the following form,

$$\Pi(\Theta_t) \leq \exp(-\lambda t \log t) \quad \text{for some } \lambda > 0. \tag{17}$$

Assume the tree topology $\mathcal{A}_t$ of size t is generated by a specific path $\mathcal{A}_t^{(1)}, \mathcal{A}_t^{(2)}, \ldots, \mathcal{A}_t^{(\mathfrak{D}_a)}$, where each $\mathcal{A}_t^{(j)}, 1 \leq j \leq \mathfrak{D}_a$ is a partition obtained after j steps of recursion with respect to the depth as we grow an OPT. The random measures supported by $\mathcal{A}_t$ grown in this specific way is denoted as $\mathcal{F}_{\{\mathcal{A}_t^{(j)}: 0 \leq j \leq \mathfrak{D}_a\}}$.

A.3.1. A SHALLOW TREE IS PREFERRED

Our first lemma is to rigorously justify the intuition we described in Section 3.3.

Lemma 12 *For any two equal sized partitions $\mathcal{A}_t$ and $\mathcal{B}_t$ of depths $\mathfrak{D}_a$ and $\mathfrak{D}_b$ respectively,*

$$\pi_{OPT}\left(\mathcal{F}_{\{\mathcal{A}_t^{(j)}: 0 \leq j \leq \mathfrak{D}_a\}}\right) \geq \pi_{OPT}\left(\mathcal{F}_{\{\mathcal{B}_t^{(j)}: 0 \leq j \leq \mathfrak{D}_b\}}\right) \quad \text{if} \quad \mathfrak{D}_a < \mathfrak{D}_b. \tag{18}$$

Proof The proof is based on the observation that when we prune a pair of higher level leaves in $\mathcal{B}_t$, and make a further split of a lower level node, the prior on the modified tree will never decrease. We call such an operation as *prune and split*. As the size of $\mathcal{A}_t$ and $\mathcal{B}_t$ are the same, we can always find a pair of leave nodes in $\mathcal{B}_t$ of level $\mathfrak{D}_b$, denoted as $B_{1,\text{left}}^{(\mathfrak{D}_b)}$ and $B_{1,\text{right}}^{(\mathfrak{D}_b)}$, as two children of an intermediate node $B_1^{(\mathfrak{D}_b-1)}$. We can find another terminal node leave node $B_2^{(j)}$in $\mathcal{B}_t$ of level j no larger than $\mathfrak{D}_b - 1$. Assume $B_{1,\text{left}}^{(\mathfrak{D}_b)}$ and $B_{1,\text{right}}^{(\mathfrak{D}_b)}$ are pruned and $B_2^{(j)}$ is further split into two subregions, leading to a new partition $\widetilde{\mathcal{B}}_t$ with the path $\widetilde{\mathcal{B}}_t^{(1)}, \ldots, \widetilde{\mathcal{B}}_t^{(\widetilde{\mathfrak{D}}_b)}$, where $\widetilde{\mathfrak{D}}_b = \mathfrak{D}_b$ or $\mathfrak{D}_b - 1$. Then

$$\pi_{\text{OPT}}(\mathcal{F}_{\{\widetilde{\mathcal{B}}_t^{(j)}:1\le j\le\widetilde{\mathfrak{D}}_b\}}) = \pi_{\text{OPT}}(\mathcal{F}_{\{\mathcal{B}_t^{(j)}:0\le j\le\mathfrak{D}_b\}}) \cdot \frac{1-\tau_{\mathfrak{D}_b-1}}{(1-\tau_{\mathfrak{D}_b})^2\tau_{\mathfrak{D}_b-1}} \cdot \frac{(1-\tau_{j+1})^2\tau_j}{1-\tau_j},$$

since the assumption $\tau_j = c_\tau 2^{-\nu j}$ for some $\nu > 1$, it is easy to show

$$\frac{(1-\tau_{\mathfrak{D}_b-1})(1-\tau_{j+1})^2}{(1-\tau_j)(1-\tau_{\mathfrak{D}_b})^2} \ge 1 \quad \text{and} \quad \frac{\tau_j}{\tau_{\mathfrak{D}_b-1}} \ge 1.$$

Another useful observation is that if for one split or several consecutive splits, different dimensions are chosen for making the split, the prior on the set of random measures supported by the modified partition remains the same. We call this type of operation as *reshaping*.

Note that the partition $\mathcal{A}_t$ under the path $\mathcal{A}_t^{(0)}, \mathcal{A}_t^{(1)}, \ldots, \mathcal{A}_t^{(\mathfrak{D}_a)}$ can be obtained by applying a number of "prune and split" and "reshaping" operations to $\mathcal{B}_t$, and these operations will *not* lower the prior. ■

A.3.2. AN UPPER BOUND FOR Θ_t

Generally, given a size t, with $2^j \le t < 2^{j+1}$ for some $j \in \mathbb{N}$, we can define a partition $\mathcal{A}_t$ that is close to a non-stopped one as the following: first, the tree grows without stopping for the first j levels as $\mathcal{T}_{j+1}^{(0)}, \mathcal{T}_{j+1}^{(1)}, \ldots, \mathcal{T}_{j+1}^{(j)}$; then at level j, the split for $t - 2^j$ level-j regions will continue for one more step, while the remaining $2^{j+1} - t$ ones being stopped. The complete path for obtaining $\mathcal{A}_t$ is still denoted as $\mathcal{A}_t^{(0)}, \mathcal{A}_t^{(1)}, \ldots, \mathcal{A}_t^{(j+1)}$. It is easy to see,

$$\pi_{\text{OPT}}\left(\mathcal{F}_{\{\mathcal{A}_t^{(l)}:0\le l\le k+1\}}\right) \le \pi_{\text{OPT}}\left(\mathcal{F}_{\{\mathcal{T}^{(j)}:0\le j\le\mathfrak{D}\}}\right).$$

Applying Lemma 8, we have

$$\pi_{\text{OPT}}\left(\mathcal{F}_{\{\mathcal{A}_t^{(l)}:0\le l\le k+1\}}\right) \lesssim \exp\left(-\frac{\nu t}{2}\log t + \nu t\log 2 - t\log d\right).$$

As shown before, such a non-stopped partition of shallow depth is preferred in the sense that (18) holds. Within the papameter space Θ_t, there are at most $d^t t!$ such paths. This is because we may consider obtaining a size t partition as the following: at each step a subregions is randomly selected for partitioning, and for the chosen subregions the split is along the midpoint of a randomly chosen dimension. This type of recursive partitioning is repeated for $t-1$ steps. If we denote the decisions made at these $t-1$ steps as $(J_1, \ldots, J_{t-1})$, then each sequence of decisions can uniquely

determine a partition with a specific path. For each path, there is at least one sequence of decisions corresponding to it. Therefore, the number of paths is bounded by the number of different decisions, which is in turn bounded by $d^t t!$.

Combining the bounds together, we have

$$\begin{aligned} \pi_{\text{OPT}}(\Theta_t) &\leq d^t t! \cdot \exp\left(-\frac{\nu t}{2}\log t + \nu t \log 2 - t\log d\right) \\ &\leq \exp\left(-(\nu/2-1)t\log t + \nu t \log 2\right). \end{aligned} \tag{19}$$

Then we can set $t = t_n$ and verify condition (13) by using (19).

Appendix B. Approximation results

In this section, we provide approximation results to different function classes.

B.1. Anisotropic Besov spaces and balls

The anisotropic Besov space has been rigorously studied. Researchers have derived approximation results by using wavelet basis, see for example book by Nikol'skii (2012). We summarize the existing results here.

Lemma 13 *For any $f \in B^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ or $f \in \widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$, we can find a depth-$\mathfrak{D}$ tree-supported function $f_{\mathfrak{D}}$, such that $\rho(f, f_{\mathfrak{D}}) \leq c_{\text{approx}} 2^{-\bar{\sigma}/d\cdot\mathfrak{D}}$, where c_{approx} can be chosen as $2^{d+1}Ld$.*

Proof In our case, as we have assume $p \geq 2$, and for both Besov spaces and Besov balls the following embedding holds for any $p' \geq p$ and $1 \leq q \leq \infty$: $B^{\boldsymbol{\sigma}}_{p',q}(\Omega, L) \subset B^{\boldsymbol{\sigma}}_{p,q}(\Omega, L) \subset B^{\boldsymbol{\sigma}}_{p,\infty}(\Omega, L)$. It suffices to obtain a bound for $f \in B^{\boldsymbol{\sigma}}_{2,\infty}(\Omega, L)$. At depth $\mathfrak{D}$, we can define the wavelet basis at level $J_{\mathfrak{D}} = (\lfloor \mathfrak{D}\bar{\sigma}/\sigma_1 \rfloor, \ldots, \lfloor \mathfrak{D}\bar{\sigma}/\sigma_d \rfloor)$ as

$$\boldsymbol{\Xi}^{(J_{\mathfrak{D}})} = \Big\{ \prod_{l=1}^{d} \phi_{j_l,k_l}, j_l = \lfloor \mathfrak{D}\bar{\sigma}/\sigma_l \rfloor, k_l \in \mathbb{Z} \Big\}.$$

Then, for any $f \in B^{\boldsymbol{\sigma}}_{2,\infty}(\Omega, L)$, there exists an approximation to $f_{\mathfrak{D}}$ in the space spanned by $\boldsymbol{\Xi}^{(J_{\mathfrak{D}})}$, denoted as $P^{(J_{\mathfrak{D}})} f = \sum_{\xi^{(J_{\mathfrak{D}})} \in \Xi^{(J_{\mathfrak{D}})}} \langle f, \xi^{(J_{\mathfrak{D}})} \rangle \xi^{(J_{\mathfrak{D}})}$, such that

$$\|f - P^{(J_{\mathfrak{D}})} f\|_2 \leq \left(\sum_{l=1}^{d} 2^{-\sigma_l j_l} \right) \|f\|_{B^{\boldsymbol{\sigma}}_{2,\infty}} \leq 2^d Ld \cdot 2^{-\bar{\sigma}/d\cdot\mathfrak{D}}.$$

Set $f = g_0$ and $g_{\mathfrak{D}} = P^{(J_{\mathfrak{D}})} g_0$, then we obtain a bound for L^2-distance between them. $g^2_{\mathfrak{D}}$ may not be an appropriate density function, we scale $g^2_{\mathfrak{D}}$ to $\tilde{g}^2_{\mathfrak{D}} = g^2_{\mathfrak{D}} / (\int g^2_{\mathfrak{D}})$, then

$$\rho(f_0, \tilde{g}^2_{\mathfrak{D}}) \leq 2^{d+1} Ld \cdot 2^{-\bar{\sigma}/d\cdot\mathfrak{D}}.$$

This finishes the proof for anisotropic Besov space. For anisotropic Besov balls, the proof is similar. ■

If we can obtain depth-$\mathfrak{D}$ approximation, then it is also a size-t approximation by setting $t = 2^{\mathfrak{D}}$. Therefore, we have the following corollary immediately.

Corollary 14 *For any $f \in B^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ or $f \in \widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$, we can find a size-$t$ tree-supported function f_t, such that $\rho(f, f_t) \leq c_{\text{approx}} \mathfrak{D} t^{-\bar{\sigma}/d}$, where c_{approx} can be chosen as $2^{d+2} Ld$.*

Theorem 1 can be obtained by Proposition 4 and Theorem 4 in Liu et al. (2023) in combination with Lemma 13 and Corollary 14 respectively.

B.2. Regionwise anisotropic Besov spaces and balls

In this subsection, we derive approximation rate to functions lying in a region-wise anisotropic Besov space or region-wise anisotropic Besov ball.

Lemma 15 *Given a fixed dyadic partition $\{\Omega_s\}_{s=1}^S$ of Ω, assume $g_0|_{\Omega_s} \in B^{\boldsymbol{\sigma}_s}_{p,q}(\Omega_s, L)$ or $g_0 \in \widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ with unknown $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S)$. When the tuple of smoothness parameters $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) \in \Sigma_\sigma$ or lies in $\widetilde{\Sigma}_\sigma$ in the case of region-wise anisotropic Besov ball, there exists a sequence of depth-$\mathfrak{D}$ approximations $f_{\mathfrak{D}}$, such that $\rho(f_0, f_{\mathfrak{D}}) \leq c_{\text{approx}} S \cdot 2^{-\sigma/d \cdot (\mathfrak{D} - \lfloor \log_2 S \rfloor)}$.*

The lemma can be shown by applying Lemma 13 to each region Ω_s separately.

For the case with unequal "average" smoothness parameters, we derive an upper bound for the approximation error.

Lemma 16 *Given a fixed dyadic partition $\{\Omega_s\}_{s=1}^S$ of Ω, assume $g_0|_{\Omega_s} \in B^{\boldsymbol{\sigma}_s}_{p,q}(\Omega_s, L)$ or $g_0 \in \widetilde{B}^{\boldsymbol{\sigma}}_{p,q}(\Omega, L)$ with unknown $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S)$. Let $\bar{\sigma}_{\min} = \min_{1 \leq s \leq S} \bar{\sigma}_s$. When $(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_S) \in \Sigma$ or in the set $\widetilde{\Sigma}$ respectively for Besov balls, there exists a sequence of size-t approximations f_t, such that $\rho(f_0, f_{\mathfrak{D}}) \leq c_{\text{approx}} S \cdot (t/S)^{-\bar{\sigma}_{\min}/d}$.*

Proof The approximation result can be obtained by allocating t/S terminal nodes to each region Ω_s. ■

Theorem 2 can be obtained by Proposition 4 and Theorem 4 in Liu et al. (2023) in combination with Lemma 15 and Lemma 16 respectively.

B.3. Weak -$\ell_{p'}$ balls

Lemma 17 *Assume for a d-dimensional density function f_0, $g_0 = \sqrt{f_0} \in wl^{p'}_{\boldsymbol{\sigma}}(\Omega, C)$. Then there exists a sequence of approximations $f_t \in \Theta_t$, such that $\rho(f_0, f_t) \lesssim c_{\text{approx}} t^{-(1/p' - 1/2)} (\log t)^{1/p' - 1/2}$, where c_{approx} can be chosen as $(2^d + 1 + \log c_\mu + c_v)^{1/p' - 1/2} C \sqrt{2/(2/p' - 1)}$. c_μ and c_v are constants from Condition 2.1.*

Proof Let $g_K = \sum_{k=1}^K \langle g_0, \xi_{(k)} \rangle \xi_{(k)}$. From condition (9) we have

$$\begin{aligned} \|g_0 - g_K\|_2^2 &= \Big\| \sum_{k=K+1}^{+\infty} \langle g_0, \xi_{(k)} \rangle \xi_{(k)} \Big\|_2^2 = \sum_{k=K+1}^{+\infty} \langle g_0, \xi_{(k)} \rangle^2 \\ &\leq C^2 \sum_{k=K+1}^{+\infty} k^{-2/p'} \leq \frac{C^2}{2/p' - 1} K^{-(2/p' - 1)}. \end{aligned}$$

g_K^2 make not be an appropriate density function. We consider normalizing g_K^2 to $\tilde{g}_K^2$ as $\tilde{g}_K^2 = g_K^2/(\int g_K^2)$. Then,

$$\begin{aligned}\rho^2(f_0, \tilde{g}_K^2) &= \|g_0 - \tilde{g}_K\|_2^2 = \|g_0 - g_K\|_2^2 + \Big(1 - \frac{1}{\|g_K\|_2}\Big)^2 \|g_K\|_2^2 \\ &\leq \|g_0 - g_K\|_2^2 + 1 - \|g_K\|_2^2 \\ &\leq 2\|g_0 - g_K\|_2^2 \leq \frac{2C^2}{2/p' - 1} K^{-(2/p'-1)}. \end{aligned} \tag{20}$$

Clearly, $\tilde{g}_K$ is a tree-supported density function. We would like to derive a bound for the size of the tree. Note that for a multivariate Haar basis function, the positive and negative parts can further divide its supporting rectangle into smaller subregions, and the total number of such subregions is upper bounded by 2^d. Based on the moderate resolution assumption (Condition 2.1), the basis function $\xi_{(k)}$ can be viewed as function define on a tree of size at most $2^d + \lceil c_v \log_2 k + \log c_\mu \rceil$. Moreover, as basis function are defined in a specific way according to a vector $\boldsymbol{\sigma}$, supporting rectangles for all basis functions are about of the same shape. Introducing more individual basis functions will only increase the size of the partition in a linear way. Therefore, $(2^d+1+\log c_\mu)K + c_v K \log K$ is the largest possible sized partition on which the density function $\tilde{g}_K$ is piecewise constant. Replacing K in (20) by $(2^d + 1 + \log c_\mu + c_v)^{-1} t / \log t$, we get the desired result of the approximation rate. ■

Theorem 3 can be obtained by combining Lemma 17 and Theorem 4 in Liu et al. (2023).