Journal of Machine Learning Research 24 (2023) 1-65 Submitted 4/22; Revised 6/23; Published 11/23

The Art of BART: Minimax Optimality over Nonhomogeneous Smoothness in High Dimension

Seonghyun Jeong SJEONG@YONSEI.AC.KR
Department of Statistics and Data Science
Department of Applied Statistics
Yonsei University
Seoul 03722, Republic of Korea

Veronika Ročková VERONIKA.ROCKOVA@CHICAGOBOOTH.EDU
Booth School of Business
University of Chicago
Chicago, IL 60637, USA

Editor: Daniel Roy

Abstract

Many asymptotically minimax procedures for function estimation often rely on somewhat arbitrary and restrictive assumptions such as isotropy or spatial homogeneity. This work enhances the theoretical understanding of Bayesian additive regression trees under substantially relaxed smoothness assumptions. We provide a comprehensive study of asymptotic optimality and posterior contraction of Bayesian forests when the regression function has anisotropic smoothness that possibly varies over the function domain. The regression function can also be possibly discontinuous. We introduce a new class of sparse *piecewise heterogeneous anisotropic* Hölder functions and derive their minimax lower bound of estimation in high-dimensional scenarios under the L_2-loss. We then find that the Bayesian tree priors, coupled with a Dirichlet subset selection prior for sparse estimation in high-dimensional scenarios, adapt to unknown heterogeneous smoothness, discontinuity, and sparsity. These results show that Bayesian forests are uniquely suited for more general estimation problems that would render other default machine learning tools, such as Gaussian processes, suboptimal. Our numerical study shows that Bayesian forests often outperform other competitors such as random forests and deep neural networks, which are believed to work well for discontinuous or complicated smooth functions. Beyond nonparametric regression, we also examined posterior contraction of Bayesian forests for density estimation and binary classification using the technique developed in this study.

Keywords: Adaptive Bayesian procedure, Bayesian CART, Bayesian forests, High-dimensional inference, Posterior contraction, Sparsity priors

1. Introduction

1.1 Motivation

Many of the existing asymptotic minimaxity results for estimating regression functions are predicated on the assumption that certain smoothness conditions hold, which can be rarely satisfied/verified when confronted with real data. This creates a disconnect between theory and practice, limiting the scope of many theoretical results. For example, in nonparametric regression involving multiple predictors, the assumption of *isotropic smoothness* can be unnecessarily restrictive. A more realistic scenario is when the function exerts different degrees of smoothness in different directions and areas, with possible discontinuities that allow further flexibility. This study is motivated by the desire to evaluate the theoretical performance of Bayesian forests, one of the workhorses of Bayesian machine learning, in such broad scenarios.

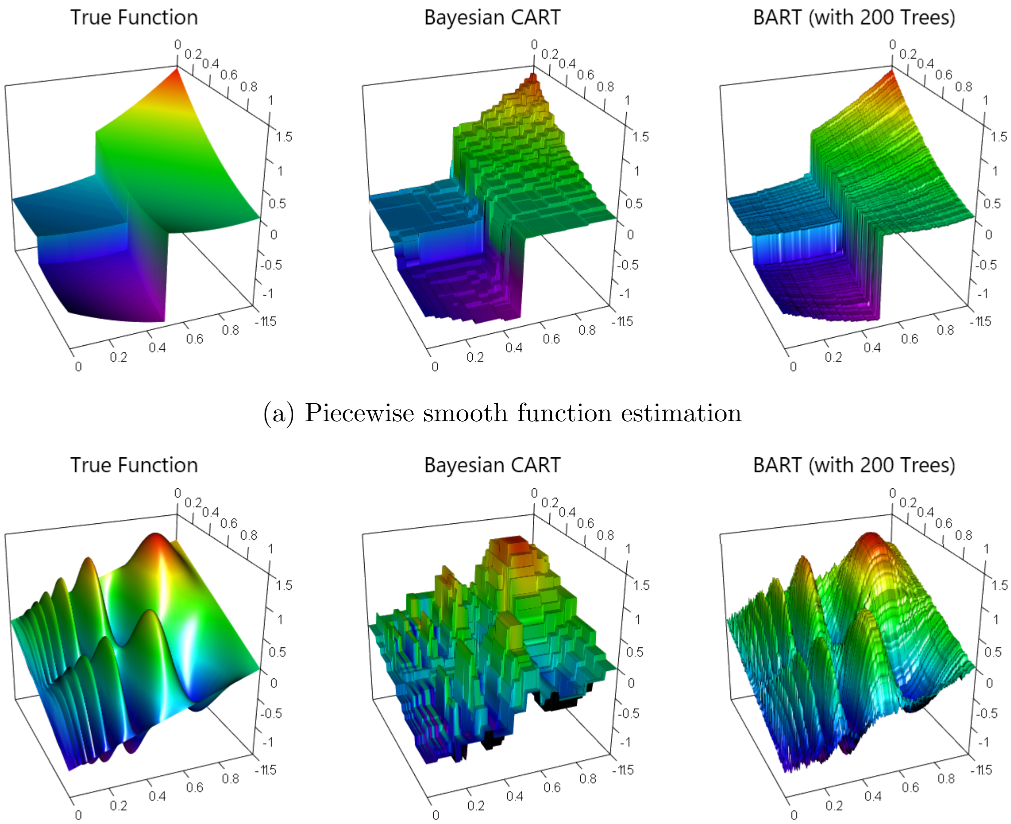
Bayesian trees and their ensembles have achieved notable empirical success in statistics and machine learning (Chipman et al., 1998; Denison et al., 1998; Chipman et al., 2010). Relative to other Bayesian machine learning alternatives, tree-based methods require comparatively less tuning and can be scaled to higher dimensions (Lakshminarayanan et al., 2013; Bleich et al., 2014; He et al., 2019). The popularity of Bayesian forests, such as Bayesian additive regression trees (BART), (Chipman et al., 2010) is growing rapidly in many areas including causal inference (Hill, 2011; Hahn et al., 2020), mean-variance function estimation (Pratola et al., 2020), smooth function estimation (Linero and Yang, 2018), variable selection (Bleich et al., 2014; Linero, 2018), interaction detection (Du and Linero, 2019), survival analysis (Sparapani et al., 2016), time series (Taddy et al., 2011), count and categorical data analysis (Murray, 2021), and density regression (Orlandi et al., 2021; Li et al., 2022). For comprehensive overviews and surveys, refer to Linero (2017), Tan and Roy (2019), and Hill et al. (2020).

Despite remarkable success in empirical studies, the theoretical properties of Bayesian forests remained unavailable until the emergence of recent literature (Ročková and van der Pas, 2020; Linero and Yang, 2018; Ročková and Saha, 2019; Castillo and Ročková, 2021). Although these pioneering findings divulge why tree-based methods perform well, they are limited to isotropic regression function surfaces, which exhibit the same level of smoothness in every direction. Isotropy is an archetypal assumption in theoretical studies, but it can be restrictive in real-world applications. This assumption is particularly unattractive in higher dimensions wherein the function can behave very poorly in certain directions.

However, empirical evidence suggests that Bayesian forests are expected to adapt to more intricate smoothness situations. For example, Figure 1 shows that BART successfully adapts to a piecewise smooth function or a Doppler-type function. The successful performance beyond isotropy is attributable to at least three reasons: (i) tree methods are based on top-down recursive partitioning, wherein splits occur more often in areas where the function is locally uneven or bumpy, making the procedure spatially adaptive; (ii) the choice of coordinates for the split is data-driven, dividing the domain more often in directions in which the function is less smooth; and (iii) tree-based learners are piecewise constant and, as such, are expected to adapt to discontinuous functions by detecting smoothness boundaries and jumps. These considerations naturally create an expectation that Bayesian forests achieve optimal estimation properties in more complex function classes *without any prior modification.*

1.2 Our Contribution

The main goal of this study is to examine optimality and posterior contraction of Bayesian forests under relaxed smoothness assumptions. We introduce a class of functions the domain of which has been cleaved into hyper-rectangles, where each rectangular piece has its



(a) Piecewise smooth function estimation

(b) Doppler-type function estimation

Figure 1: Function estimation in nonparametric regression with complicated smoothness using Bayesian CART and BART.

own anisotropic smoothness (with the same harmonic mean). We allow for possible discontinuities at the boundaries of the pieces. We call this new class of functions *piecewise heterogeneous anisotropic functions* (see Definitions 1–2 in Section 2.2). We then establish an approximation theory for this general class, which blends anisotropy with spatial inhomogeneity and which, to the best of our knowledge, has not yet been pursued in the literature. Our results complement the body of existing work on piecewise isotropic smoothness classes (e.g., Candès and Donoho, 2000, 2004; Le Pennec and Mallat, 2005; Petersen and Voigtlaender, 2018; Imaizumi and Fukumizu, 2019). Our function class subsumes the usual (homogeneous) anisotropic space for which adaptive procedures exist with optimal convergence rate guarantees, including the dyadic classification and regression trees (CART) of Donoho (1997). We refer to Barron et al. (1999), Neumann and von Sachs (1997), Hoffman and Lepski (2002), Lepski (2015), and references therein for a more complete list. There are also adaptive Bayesian procedures for anisotropic function estimation with desired asymptotic properties (e.g., Bhattacharya et al., 2014; Shen and Ghosal, 2015). There appear to be *no* theoretical properties for adaptation in the more general case of piecewise heterogeneous anisotropic smoothness. Indeed, existing theoretical studies for discontinuous piecewise smooth classes impose the isotropy assumption (e.g., Candès and Donoho, 2000, 2004; Le Pennec and Mallat, 2005; Petersen and Voigtlaender, 2018; Imaizumi and Fukumizu, 2019) and the convergence rates in spatially adaptive estimation depend on global smoothness parameters (e.g., Pintore et al., 2006; Liu and Guo, 2010; Wang et al., 2013; Tibshirani, 2014). In this respect, our study appears to be the first theoretical investigation of piecewise anisotropic function classes.

The majority of frequentist/Bayesian methods for anisotropic function estimation rely on multiple scaling (bandwidth) parameters, one for each direction. As noted by Bhattacharya et al. (2014), selecting optimal scaling parameters in a frequentist way can be computationally difficult, as adaptation in anisotropic spaces presents several challenges (Lepski and Levit, 1999). The Bayesian paradigm provides an effective remedy by assigning priors over these unknown parameters. One such example is the generalized Gaussian process priors or spline basis representations (Bhattacharya et al., 2014; Shen and Ghosal, 2015). Although these priors enjoy elegant theoretical guarantees in typical anisotropic spaces, whether they can adapt to piecewise heterogeneous anisotropic spaces without substantial modification remains unclear. Contrariwise, Bayesian forests are expected to work in these more complex scenarios without any additional scaling parameters. The approximability is controlled merely by the depth of a tree and the orientation of its branches, where no prior modifications should be required to achieve optimal performance. Moreover, computation with Gaussian processes can be quite costly (Banerjee et al., 2013; Liu et al., 2020), while Bayesian forests are more scalable and faster than their competitors.

In the context of regression or classification, Bayesian forests often rely on observed covariate values for splits in recursive partitioning (Chipman et al., 1998; Denison et al., 1998; Chipman et al., 2010). This facilitates theoretical investigation under the fixed regression design. In the context of nonparametric Gaussian regression, Ročková and van der Pas (2020) and Ročková and Saha (2019) investigated posterior contraction for BART based on this conventional manner of partitioning, whereas the dyadic CART (Donoho, 1997) splits at dyadic midpoints of the domain and can achieve optimal performance as well (Castillo and Ročková, 2021). We generalize the dyadic CART by introducing the notion of split-nets, which form a collection of candidate split-points that are not necessarily observed covariate values and/or dyadic midpoints. Our findings show that optimality can be achieved with split-nets that are sufficiently evenly distributed. By allowing the split-points to occur beyond observed values, we show that Bayesian forests enjoy the general recipe of the posterior contraction theory (Ghosal et al., 2000; Ghosal and van der Vaart, 2007), which applies to other statistical setups such as density estimation or regression/classification with random design.

Asymptotic minimaxity is often used to evaluate the optimality of statistical procedures. Yang and Tokdar (2015) derived the minimax rates of sparse function estimation in high dimensions, but their results are restricted to the isotropic cases. In fixed (low) dimensions, minimax rates over anisotropic function spaces have been extensively studied in the literature (Ibragimov and Hasminskii, 1981; Nussbaum, 1985; Birgé, 1986). If the true function only depends on a subset of coordinates, the minimax rate is improved and determined by the smoothness parameters of active coordinates (Hoffman and Lepski, 2002). However, to the best of our knowledge, there are no available studies on minimax rates over piecewise anisotropic function spaces like ours. While there exist results on piecewise isotropic classes (e.g., Imaizumi and Fukumizu, 2019), even the simpler fixed-dimensional setup without sparsity has *not* been studied for piecewise anisotropic classes. Focusing on Gaussian nonparametric regression, we derive the minimax lower bound for our piecewise heterogeneous anisotropic spaces under the high-dimensional scenario. This result verifies the finding that our obtained contraction rates for Bayesian forests are indeed minimax-optimal up to a logarithmic factor.

We summarize the contribution of this study as follows.

- **Approximation theory**: The true function should be approximable by tree-based learners to establish the optimal rate of posterior contraction. Approximation theory for piecewise heterogeneous anisotropic classes is much more intricate when there are discontinuities and heterogeneity. We establish such approximation theory here under suitable regularity conditions (with smoothness up to 1 owing to the limitation of piecewise constant learners).
- **Posterior contraction**: For piecewise heterogeneous anisotropic functions, posterior contraction of Bayesian forests is established under the high-dimensional setup with a Dirichlet sparse prior. The derived rates consist of the risk of variable selection uncertainty and the risk of function estimation, similar to isotropic cases (Yang and Tokdar, 2015; Ročková and van der Pas, 2020).
- **Minimax optimality**: Minimax rates in high-dimensional spaces have been unavailable even for simple anisotropic classes. For Gaussian nonparametric regression with high-dimensional inputs, we formally derive the minimax lower bound over piecewise heterogeneous anisotropic spaces. This certifies that our obtained contraction rate for Bayesian forests is optimal up to a logarithmic factor.
- **Applications beyond regression**: Unlike the asymptotic studies of the traditional tree priors (Ročková and van der Pas, 2020; Ročková and Saha, 2019), our findings show that splits for recursive partitioning do not necessarily have to be at observed covariate values. This implies that our technique of proofs extends beyond fixed-design regression to other estimation problems such as density estimation or regression/classification with random design.

1.3 Preview and Outline of the Paper

The main results of this study begin to appear in Section 4.2 after excessive preliminary steps. Before going into the preparatory phase, here we provide a preview of our main results. Let us focus on a fixed design regression setup,

$$Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_0^2), \quad i = 1, \dots, n, \tag{1}$$

with a response $Y_i \in \mathbb{R}$ and a covariate $x_i \in [0,1]^p$, where $f_0 : [0,1]^p \to \mathbb{R}$ and $\sigma_0^2 < \infty$. Assume that f_0 depends only on d variables among p coordinates. Assume further that f_0 is a piecewise heterogeneous anisotropic function with a global smoothness harmonic mean $\bar{\alpha} \in (0,1]$ (see Definitions 1–3 for a more precise definition). Assigning the BART prior on f_0, the posterior contraction rate is obtained as $\sqrt{(d \log p)/n} + (\log n)^c n^{-\bar{\alpha}/(2\bar{\alpha}+d)}$ for some $c > 0$ (Theorem 2). This rate is minimax-optimal up to a log factor (Theorem 3). The same contraction rates are also achieved in other statistical setups (Theorems 4–6). For the additive true function, the rate has an additive form (Theorems 7).

The rest of this paper is organized as follows. In Section 2, we describe the background of function spaces and Bayesian forests. In high-dimensional scenarios, the tree priors on functions are specified in Section 3. In Section 4, we illuminate the approximation theory for our function spaces. In Section 5, we study posterior contraction of Bayesian forests and their minimax optimality in nonparametric regression with a fixed design. The section also includes a numerical study that shows the outstanding performance of BART over other methods such as random forests and deep neural networks, which are believed to work well for discontinuous or complicated smooth functions. Posterior contraction properties in other statistical models such as density estimation and binary classification are investigated in Section 6. An example of additive regression is also considered in Section 6 to emphasize a theoretical advantage of Bayesian forests over single tree models. Section 7 concludes. All technical proofs are presented in Appendix.

2. Preliminaries

2.1 Notation and Terminology

Although the main focus of this study is BART for regression in (1), we work with a general statistical experiment P_f indexed by a measurable function $f : [0,1]^p \to \mathbb{R}$ for some $p > 0$, which will be modeled by Bayesian forests. This allows us to incorporate other statistical setups, such as density estimation, into our theoretical framework. Each statistical model we are dealing with will be specified for our examples in Sections 5–6. We observe n observations with the true function denoted by f_0 and assume that p is possibly increasing with the sample size n. The notation $\mathbb{E}_0$ denotes the expectation operator under the true model with f_0.

For sequences a_n and b_n, we write $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$ equivalently) if $a_n \le C b_n$ for some constant $C > 0$, and $a_n \asymp b_n$ implies $a_n \lesssim b_n \lesssim a_n$. We also write $a_n \ll b_n$ (or $b_n \gg a_n$ equivalently) if $a_n / b_n \to 0$ as $n \to \infty$. For a subspace E of the Euclidean space, $\mathcal{C}(E)$ denotes a class of continuous functions $f : E \to \mathbb{R}$. For a given measure μ and a measurable function f, we denote by $\|f\|_{v,\mu} = (\int |f|^v d\mu)^{1/v}$ the $L_v(\mu)$-norm, $1 \le v < \infty$. We denote by $\mathcal{L}_2(\mu)$ the linear space of real valued functions equipped with inner product $\langle f, g\rangle_\mu = \int fg d\mu$ and norm $\|f\|_{2,\mu} = \langle f, f\rangle_\mu^{1/2}$. For the sake of brevity, with the Lebesgue measure on a unit hypercube, $\mathcal{L}_2$ denotes the L_2 space and $\|f\|_v$ denotes the L_v-norm. In particular, $\|f\|_\infty$ denotes the L_∞-norm of a function f defined by the essential supremum, i.e., $\|f\|_\infty = \inf\{C \ge 0 : |f(x)| \le C \text{ for almost every } x\}$.[1] The support of a measure μ is denoted by $\text{supp}(\mu)$. For a given vector u, the notations $\|u\|_v$ and $\|u\|_\infty$ represent the ℓ_v-norms, $1 \le v < \infty$, and the maximum-norm, respectively. For a semimetric space $(\mathcal{F}, \rho)$ endowed with a semimetric ρ, the expressions $D(\epsilon, \mathcal{F}, \rho)$ and $N(\epsilon, \mathcal{F}, \rho)$ are ϵ-packing and ϵ-covering numbers of $\mathcal{F}$, respectively. For a subset $S \subseteq \{1, \dots, p\}$ and $x = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$, let $x_S = (x_j, j \in S) \in \mathbb{R}^{|S|}$ be the indices chosen by S.

A q-dimensional hyper-rectangle $\Psi \subseteq [0,1]^q$ with any $q > 0$ is simply called a *box*. Precisely, a box is defined as the Cartesian product of open, closed, or semi-closed intervals;

[1] We use the L_∞-norm to measure the difference of discontinuous functions while ignoring possible disagreement at jump surfaces. The L_∞-norm is reduced to the supremum-norm for continuous functions if the domain is not a null set.

therefore, a box can be open, closed, or neither (e.g., $[a_1, b_1) \times (a_2, b_2]$) depending on the context. A partition $\mathfrak{Y} = \{\Psi_1, \dots, \Psi_J\}$ of $[0,1]^q$, consisting of J disjoint boxes $\Psi_r \subseteq [0,1]^q$, $r = 1, \dots, J$, is called a *box partition*. For the Cartesian product of q subsets of $\mathbb{R}$, i.e., $E \subseteq \mathbb{R}^q$, we denote the jth projection mapping of E by $[E]_j = \{x_j \in \mathbb{R} : (x_1, \dots, x_q)^\top \in E\}$. The length and interior of an interval $I \in \mathbb{R}$ is denoted by $\mathsf{len}(I)$ and $\mathsf{int}(I)$, respectively.

2.2 Heterogeneous Anisotropic Function Spaces with Sparsity

In this subsection, we introduce our function spaces with heterogeneous smoothness and sparsity in high dimensions. The first assumption is that the true regression function $f_0 : [0,1]^p \to \mathbb{R}$ is d-sparse, i.e., it depends on a small subset of d variables. This means that there exist a function $h_0 : [0,1]^d \to \mathbb{R}$ and a subset $S_0 \subseteq \{1, \dots, p\}$ with $|S_0| = d$, such that $f_0(x) = h_0(x_{S_0})$ for any $x \in [0,1]^p$. For example, suppose the true function is defined as $f_0(x_1, x_2) = \sin(x_1)$ on $[0,1]^2$ with $p = 2$. This function can be completely expressed by the one-dimensional function $h_0(x_1) = \sin(x_1)$ on $[0,1]$, and hence is 1-sparse by definition.

For now, we focus on the function h_0 on the low-dimensional domain $[0,1]^d$. The complete characterization of f_0 will soon be discussed. We assume that $[0,1]^d$ partitioned into many boxes and h_0 is Hölder continuous with possibly different smoothness in each box. The smoothness inside each box is anisotropic, i.e., different for each coordinate. Focusing on a single box, we first define an *anisotropic Hölder space* in the usual sense.

Definition 1 (Anisotropic Hölder space)**.** For smoothness $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in (0,1]^d$, a box $\Psi \subseteq [0,1]^d$, and a Hölder coefficient $\lambda < \infty$, we denote by $\mathcal{H}_\lambda^{\alpha,d}(\Psi)$ an anisotropic α-Hölder space on Ψ, i.e.,

$$\mathcal{H}_\lambda^{\alpha,d}(\Psi) = \left\{ h : \Psi \to \mathbb{R};\ |h(x) - h(y)| \le \lambda \sum_{j=1}^{d} |x_j - y_j|^{\alpha_j},\ x, y \in \Psi \right\}.$$

Note that the definition above imposes a restriction $\alpha \in (0,1]^d$. Although one can generalize this definition to smoother classes (e.g. Bhattacharya et al., 2014), we do not consider such extensions here, as step function estimators cannot be optimal in classes smoother than Lipschitz.

As discussed above, our targeted function class is not necessarily globally anisotropic over the entire domain $[0,1]^d$. Instead, we assume that h_0 has different anisotropic smoothness on $R \geq 1$ disjoint boxes of the domain with the same harmonic mean (the same harmonic mean is an important assumption for obtaining the minimax lower bound in Section 5.2). To be more precise, we define a set of R-tuples for smoothness parameters,

$$\mathcal{A}_{\bar\alpha}^{R,d} = \left\{ (\alpha_1, \dots, \alpha_R) : \alpha_r = (\alpha_{r1}, \dots, \alpha_{rd})^\top \in (0,1]^d,\ \bar\alpha^{-1} = d^{-1} \sum_{j=1}^{d} \alpha_{rj}^{-1},\ r = 1, \dots, R \right\}.$$

We assume that the anisotropic smoothness of h_0, the nonsparse proxy of f_0, is specified on an unknown underlying box partition $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$ of $[0,1]^d$ with $R \geq 1$ boxes. If $R = 1$, we write $\mathfrak{X}_0 = \{[0,1]^d\}$ with $\Xi_1 = [0,1]^d$. Note that each Ξ_r can be open, closed, or neither. The function space is formed by agglomerating anisotropic Hölder spaces for all

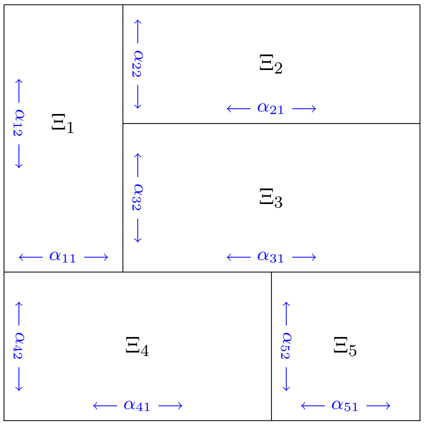


Figure 2: A graphical illustration of a piecewise heterogeneous anisotropic Hölder space with five boxes. Each piece has its own smoothness parameter, but the harmonic mean is assumed to be the same.

boxes. We emphasize that the resulting function space is not necessarily continuous, which provides a lot more flexibility relative to the conventional Hölderian class. Considering that smoothness parameters can vary across boxes and functions can be discontinuous at their boundaries, we call this new class a *piecewise heterogeneous anisotropic Hölder space*. We define these functions formally below.

Definition 2 (Piecewise heterogeneous anisotropic Hölder space)**.** Consider a smoothness parameter $A_{\bar{\alpha}} = (\alpha_r)_{r=1}^R \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ for some $\bar{\alpha} \in (0,1]$ and a box partition $\mathfrak{Y} = \{\Psi_1, \dots, \Psi_R\}$ of $[0,1]^d$ with boxes $\Psi_r \subseteq [0,1]^d$.[2] We define a piecewise heterogeneous anisotropic Hölder space as

$$\mathcal{H}_\lambda^{A_{\bar{\alpha}},d}(\mathfrak{Y}) = \left\{ h : [0,1]^d \to \mathbb{R};\ h|_{\Psi_r} \in \mathcal{H}_\lambda^{\alpha_r,d}(\Psi_r),\ r = 1, \dots, R \right\}.$$

A graphical illustration of the piecewise heterogeneous anisotropic Hölder spaces is given in Figure 2. Clearly, Definition 2 subsumes the anisotropic Hölder space in Definition 1 with $R = 1$. According to Definition 2, any $h \in \mathcal{H}_\lambda^{A_{\bar{\alpha}},d}(\mathfrak{Y})$ is anisotropic on each Ψ_r with a smoothness parameter $\alpha_r \in (0,1]^d$ and the same harmonic mean $\bar{\alpha}$ for all Ψ_r. We again emphasize that discontinuities are allowed at the boundaries of boxes Ψ_r, $r = 1, \dots, R$.

Definition 2 does not impose a specific structure on the partition $\mathfrak{Y}$ other than a box partition. However, we will later see that, depending on the approximation metric, our approximation theory will require $\mathfrak{X}_0$ to be a tree-based recursive structure defined in the next section (see Figure 4 below). Nonetheless, as every box partition can be extended to the required form by adding more splits, this discrepancy can be addressed, but it may harm our posterior contraction rate. We refer the reader to Section 4.1.1 for more discussion.

Remark 1. We compare Definition 2 with piecewise smooth function spaces widely investigated in the literature. Approximation rates for piecewise smooth functions with smooth

[2] For any $q > 1$, we write $\mathfrak{Y} = \{\Psi_r\}_r$ to denote an arbitrary box partition of $[0,1]^q$ with boxes $\Psi_r \subseteq [0,1]^q$, $r = 1, 2, \dots,$ and write $\Psi \subseteq [0,1]^q$ to denote an arbitrary q-dimensional box.

jump curves/surfaces have been extensively studied in two dimensions (e.g., Candès and Donoho, 2000, 2004; Guo and Labate, 2007) as well as in higher dimensions (Chandrasekaran et al., 2008; Petersen and Voigtlaender, 2018; Imaizumi and Fukumizu, 2019). All these studies deal with smooth functions with smooth jump curves/surfaces under the isotropy assumption. contrariwise, our definition deals with different anisotropic smoothness parameters for the boxes in a box partition, and hence seems to offer some flexibility. Our jump surfaces, however, are restricted to hyper-planes parallel to the coordinates.

Remark 2. We believe that our function class is not a subset of a popular one, but is originally defined in our work. For example, anisotropic and mixed smooth Besov spaces are highly flexible classes that render discontinuity and spatially varying smoothness (Suzuki, 2019; Suzuki and Nitanda, 2021), but they do not account for our piecewise heterogeneous anisotropic smoothness in Definition 2. In our construction, the axis-aligned box partition appears to be an important assumption in obtaining the optimal posterior contraction rate using our theory. Later we will see that our contraction rate depends on R, which is translated as the number of binary splits required to approximate the true $\mathfrak{X}_0^*$ (see Section 4.1.1). If the partition is not axis-aligned, infinitely many splits are needed, which will deteriorate our rate. Whether this is a fundamental limitation of BART is still unclear.

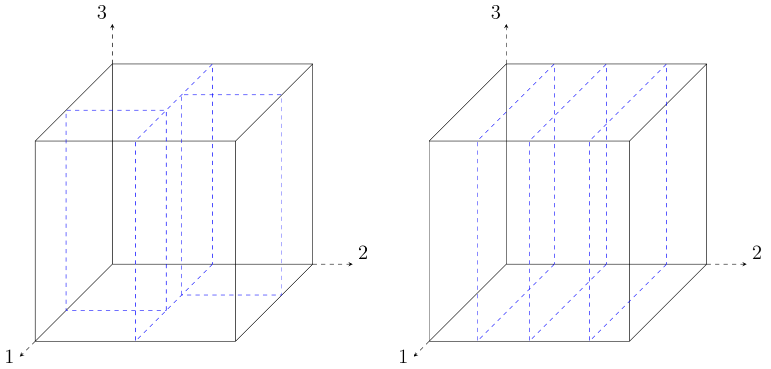
Note that Definition 2 can be used for the mapping h_0 from the lower dimensional domain $[0,1]^d$ while the true function f_0 maps the entire $[0,1]^p$ to $\mathbb{R}$. We now characterize a *sparse* elaboration of Definition 2 for the mapping $f_0 : [0,1]^p \to \mathbb{R}$. For any $S \subseteq \{1,\dots,p\}$, we denote with $W_S^p : \mathcal{C}(\mathbb{R}^{|S|}) \to \mathcal{C}(\mathbb{R}^p)$ the map that transmits $h \in \mathcal{C}(\mathbb{R}^{|S|})$ onto $W_S^p h : x \mapsto h(x_S)$. Similar to Yang and Tokdar (2015) for the isotropic cases, we now formalize d-sparse function spaces as follows.

Definition 3 (Sparse function space)**.** For the space $\mathcal{H}_\lambda^{A_{\bar{\alpha}},d}(\mathfrak{Y})$ in Definition 2, we define a d-sparse piecewise heterogeneous anisotropic Hölder space as

$$\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{Y}) = \bigcup_{S\subseteq\{1,\dots,p\}:|S|=d} W_S^p\big(\mathcal{H}_\lambda^{A_{\bar{\alpha}},d}(\mathfrak{Y})\big).$$

That is, $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{Y})$ is read as the collection of p-dimensional d-sparse functions over $\mathfrak{Y}$ with piecewise anisotropic $\bar{\alpha}$ smoothness and a Lipschitz constant λ. For an unknown smoothness parameter $A_{\bar{\alpha}} = (\alpha_r)_{r=1}^R \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ (with possibly decreasing $\bar{\alpha}$) and model components R, d, p, and λ (which are possibly increasing with n), the true function f_0 is assumed to belong to the class $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$ which allows for discontinuities, or to its continuous variant $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)\cap\mathcal{C}([0,1]^p)$. This means that there exists a function $h_0 : [0,1]^d \to \mathbb{R}$ and a subset $S_0 \subseteq \{1,\dots,p\}$ with $|S_0| = d$ such that $f_0 = W_{S_0}^p h_0$. The continuous variant $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$ achieves approximability under more relaxed assumptions (see Theorem 1 in Section 4.2). The two spaces are identical if $R = 1$.

Note that the true underlying $\mathfrak{X}_0$ is the box partition of the d-dimensional cube $[0,1]^d$. Considering the domain $[0,1]^p$ of f_0, it will be convenient to extend $\mathfrak{X}_0$ to the corresponding box partition of the p-dimensional cube $[0,1]^p$. To this end, we extend each Ξ_r to the p-dimensional box $\Xi_r^* = \{x \in [0,1]^p : x_{S_0} \in \Xi_r, x_{S_0^c} \in [0,1]^{p-d}\} \subseteq [0,1]^p$ using the true



(a) $\{1,2\}$-chopped partition (b) $\{2\}$-chopped partition

Figure 3: Examples of sparse partitions in three dimensions.

sparsity index S_0; that is, Ξ_r is the projection of Ξ_r^* onto the coordinates in S_0. The boxes Ξ_r^* then constitute the box partition $\mathfrak{X}_0^* = \{\Xi_1^*, \dots, \Xi_R^*\}$ of $[0,1]^p$.[3] We emphasize that $\mathfrak{X}_0^*$ is determined by the unknown sparsity index S_0 of the true function f_0. Observe also that our definition gives rise to $\mathfrak{X}_0^* = \{[0,1]^p\}$ with $\Xi_1^* = [0,1]^p$ if $R = 1$.

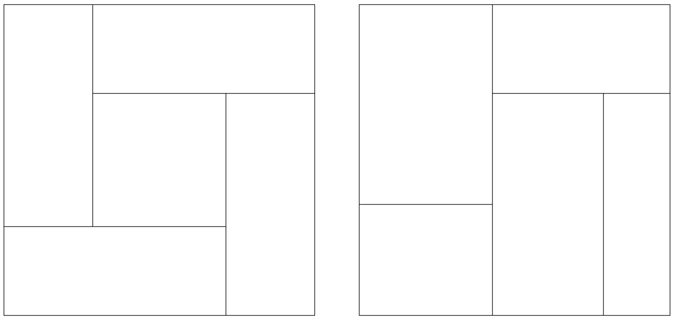
Apart from the notion of sparsity for functions, we also introduce sparsity of box partitions as follows.

Definition 4 (Sparse partition)**.** Consider a box partition $\mathfrak{Y} = \{\Psi_1, \dots, \Psi_J\}$ of $[0,1]^p$ with boxes $\Psi_r \subseteq [0,1]^p$, $r = 1, \dots, J$. For a subset $S \subseteq \{1, \dots, p\}$, the partition $\mathfrak{Y}$ is called S-chopped if $\max_{j \in S} \mathsf{len}([\Psi_r]_j) < 1$ and $\min_{j \notin S} \mathsf{len}([\Psi_r]_j) = 1$ for every $r = 1, \dots, J$.

A graphical illustration of sparse partitions is provided in Figure 3. According to Definition 4, the extended box partition $\mathfrak{X}_0^*$ is S-chopped for some $S \subseteq S_0$. Observe that $\mathfrak{X}_0^*$ is not always S_0-chopped, since $\mathfrak{X}_0$ may not have been cleaved in some coordinates. For example, if $f_0(x_1, x_2, x_3) = h_0(x_1, x_3) = \sin(x_1)\cos(x_3)\mathbb{1}(0 \le x_1 \le 0.5)\mathbb{1}(0 \le x_2 \le 1)$ with $p = 3$ and $d = 2$, then $S_0 = \{1, 3\}$, but $\mathfrak{X}_0^* = \{[0, 0.5] \times [0,1]^2, (0.5, 1] \times [0,1]^2\}$ is $\{1\}$-chopped. In particular, $\mathfrak{X}_0^*$ is $\varnothing$-chopped if $R = 1$ irrespective of what S_0 is. It is then clear that sparsity of $\mathfrak{X}_0^*$ is not the same as sparsity of f_0. In what follows, we write $S_0^* \subseteq S_0$ to denote sparsity of $\mathfrak{X}_0^*$; that is, $\mathfrak{X}_0^*$ is S_0^*-chopped.

Remark 3. Throughout the study, the model parameters $\bar{\alpha}$, R, d, p, and λ are treated as positive sequences of n, which can vary at appropriate rates so that our target posterior contraction rate in (7) changes. Accordingly, the model objects related to these sequences, e.g., $\mathfrak{X}_0$, $\mathfrak{X}_0^*$, and $A_\alpha \in \mathcal{A}_\alpha^{R,d}$, can also vary with n. The only exception is the minimax study in Section 5.2, where a fixed d provides a correct interpretation of the obtained minimax lower bound (see the lower bound in Theorem 3). With a slight abuse of notation, we usually suppress the dependency on n for the sake of notational simplicity.

[3] The notations $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$ and $\mathfrak{X}_0^* = \{\Xi_1^*, \dots, \Xi_R^*\}$ are used only to denote the true underlying box partition for the anisotropic smoothness of h_0 and its extension to the p-dimensional space for f_0, respectively.



(a) A non-tree box partition (b) A tree partition.

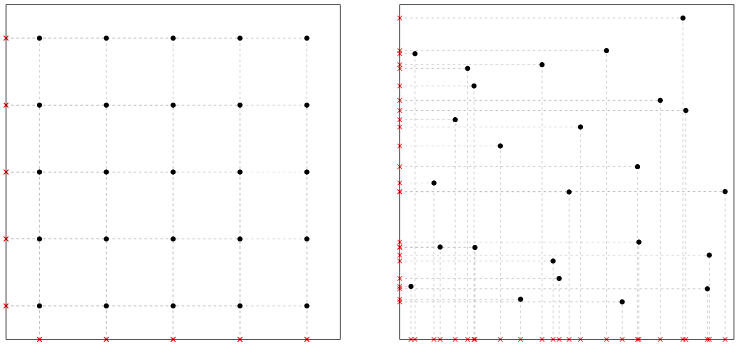
Figure 4: Examples of non-tree box partitions and tree partitions

2.3 Tree-Based Partitions

In this work, for estimators of the true function f_0, we focus on piecewise constant learners, i.e., step functions that are constant on each piece of a box partition of $[0,1]^p$. A precise description of piecewise constant learners requires an underlying partitioning rule that produces a partition for these step functions. In tree-structured models, the idea is based on recursively applying binary splitting rules to split the domain $[0,1]^p$. Here we shed light on this mechanism to construct tree-based partitions, while deferring a complete description of the induced step functions to Section 2.4.

For a given box $\Psi \subseteq [0,1]^p$, choose a *splitting coordinate* $j \in \{1,\dots,p\}$ and a *split-point* $\tau_j \in \mathsf{int}([\Psi]_j)$. The pair (j,τ_j) then dichotomizes Ψ along the jth coordinate into two boxes: $\{x \in \Psi : x_j \le \tau_j\}$ and $\{x \in \Psi : x_j > \tau_j\}$, where x_j is jth entry of x. Starting from the root node $[0,1]^p$, the procedure is iterated $K-1$ times in a top-down manner by picking one box for a split each time. This generates K disjoint boxes $\Psi_1,\dots,\Psi_K$, called *terminal nodes*, which constitute a tree-shaped partition of $[0,1]^p$, called a *tree partition*. We call this iterative procedure the *binary tree partitioning*. We will further refer to the resulting tree partitions as *flexible tree partitions* to emphasize that splits can occur everywhere in the domain $[0,1]^p$ (not necessarily at dyadic midpoints or observed covariate values). According to Definition 4, we say that a flexible tree partition is S-chopped if splitting coordinates j are restricted to a subset $S \subseteq \{1,\dots,p\}$. Note that while flexible tree partitions are always box partitions, the reverse is not generally true; see Figure 4.

Although the binary tree partitioning allows splits to occur anywhere in the domain, Bayesian tree models usually take advantage of priors that choose split-points from a predetermined discrete set. For example, in regression with continuous covariates, observed covariate values are typically used for split-points (Chipman et al., 1998; Denison et al., 1998; Chipman et al., 2010). Following this manner, Ročková and van der Pas (2020) and Ročková and Saha (2019) investigated posterior contraction of BART in Gaussian nonparametric regression with fixed covariates. Here, we relax this restriction while keeping split-points chosen from a discrete set. To this end, we define a discrete collection of locations where splits can occur, which we call a split-net.



(a) Regular grid system (b) Split-net without duplication

Figure 5: Examples of the split-net with $b_n = 25$ in two dimensions. For the regular grid in (a), one can easily see that $b_j(\mathcal{Z}) = 5$, $j = 1, 2$; hence, initial splits eliminate the possibility of other splits. The split-candidates of the split-net in (b) are unique in every coordinate, so $b_j(\mathcal{Z}) = b_n$, $j = 1, 2$.

Definition 5 (split-net)**.** For an integer sequence b_n, a split-net $\mathcal{Z} = \{z_i \in [0,1]^p,\ i = 1, \ldots, b_n\}$ is a set of b_n points $z_i = (z_{i1}, \ldots, z_{ip})^\top \in [0,1]^p$ at which possible splits occur along coordinates.

For a given split-net $\mathcal{Z}$, we call each point $z_i = (z_{i1}, \ldots, z_{ip})^\top$ a *split-candidate*. For a given splitting coordinate j and a split-net $\mathcal{Z}$, a split-point will be chosen from $[\mathcal{Z}]_j \cap \mathsf{int}([\Psi]_j)$ to dichotomize a box Ψ. Note that $[\mathcal{Z}]_j = \{z_{ij} \in [0,1], i = 1, \ldots, b_n\}$ may have fewer elements than $\mathcal{Z}$ owing to duplication. We denote by $b_j(\mathcal{Z})$ the cardinality of $[\mathcal{Z}]_j$, i.e., the number of unique values in the b_n-tuple $(z_{1j}, \ldots, z_{b_n j})$. We then obtain $\max_{1\le j \le p} b_j(\mathcal{Z}) \le b_n$ by definition. For example, consider a regular (equidistant) grid system illustrated in Figure 5a, wherein $b_j(\mathcal{Z}) = b_n^{1/p} < b_n$, $j = 1, \ldots, p$. This simplest split-net will be further discussed in Section 4.3.1. It is also possible to construct a split-net such that $b_j(\mathcal{Z}) = b_n$, $j = 1, \ldots, p$, as shown in Figure 5b. As noted above, another typical example of $\mathcal{Z}$ is the observed covariate values in fixed-design nonparametric regression with $b_n = n$ (supposing that all x_i are different). This specific example will be discussed in Section 4.3.2. Our definition of split-nets yields additional flexibility in situations when no deterministic covariate values are available, such as density estimation or in the analysis of nonparametric regression with random covariates. A subset of the observed covariate values can also be used in a fixed-design regression setup.

In assigning a prior over tree partitions, we will assume that splits in the binary partitioning rule occur only at the points in $\mathcal{Z}$; that is, for every splitting box $\Psi \subseteq [0,1]^p$ with a splitting coordinate j, a split-point τ_j is chosen such that $\tau_j \in [\mathcal{Z}]_j \cap \mathsf{int}([\Psi]_j)$. As a split is restricted to the interior of a given interval, some split-candidates may have already been eliminated in the previous steps of the splitting procedure (see Figure 5a). Clearly, a tree partition constructed by $\mathcal{Z}$ is an instance of flexible tree partitions, but the reverse is not the case. To distinguish between the two more clearly, we make the following definition.

Definition 6 ($\mathcal{Z}$-tree partition)**.** For a given split-net $\mathcal{Z}$, a flexible tree partition $\mathcal{T} = \{\Omega_1, \ldots, \Omega_K\}$ of $[0,1]^p$ with boxes $\Omega_k \subseteq [0,1]^p$, $k = 1, \ldots, K$, is called a $\mathcal{Z}$-tree partition if every split occurs at points $z_i \in \mathcal{Z}$.[4]

In summary, we obtain the following relationship among the three types of partitions: $\{\mathcal{Z}\text{-tree partitions}\} \subseteq \{\text{Flexible tree partitions}\} \subseteq \{\text{Box partitions}\}$. Similar to flexible tree partitions, $\mathcal{Z}$-tree partitions can be S-chopped for a subset $S \subseteq \{1, \ldots, p\}$ irrespective of what $\mathcal{Z}$ is employed. As we aim to do sparse estimation in high-dimensional setups, we are primarily interested in S-chopped $\mathcal{Z}$-tree partitions for some low-dimensional S. In what follows, we denote by $\mathscr{T}_{S,K,\mathcal{Z}}$ the set of all S-chopped $\mathcal{Z}$-tree partitions with K boxes.

Remark 4. The definition of a $\mathcal{Z}$-tree partition is introduced to restrict possible splits to a discrete set. This means that we assign a discrete prior on the tree topologies (see Section 3). One may instead assign a prior on the topology of flexible tree partitions, in which case a split-net $\mathcal{Z}$ is not needed. For regression problems, most of the recent BART procedures deploy a discrete set of split-candidates in their prior constructions using the observed covariate values. We aim to generalize this conventional idea while incorporating it into our framework. A discrete prior has an advantage in that it is invariant to a transformation of predictor variables (Chipman et al., 1998). We only consider placing a discrete tree prior using a given split-net $\mathcal{Z}$, and a continuous prior on flexible tree partitions is not considered.

2.4 Bayesian Trees and Forests

We now describe our piecewise constant learners using $\mathcal{Z}$-tree partitions. While single tree learners have received some attention (Chipman et al., 1998; Denison et al., 1998), it is widely accepted that additive aggregations of small trees are much more effective for prediction (Chipman et al., 2010). Noting that single trees are a special case of tree ensembles (forests), we will focus on forests throughout the rest of the paper.

We consider a fixed number T of trees. For a given split-net $\mathcal{Z}$ and for each $t \leq T$, we denote with $\mathcal{T}^t = \{\Omega_1^t, \ldots, \Omega_{K^t}^t\}$ a $\mathcal{Z}$-tree partition of size K^t and with $\beta^t = (\beta_1^t, \ldots, \beta_{K^t}^t)^\top \in \mathbb{R}^{K^t}$ the heights of the step function, called the *step-heights*. An additive tree-based learner is then fully described by a tree ensemble $\mathcal{E} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ and terminal node parameters $B = (\beta^{1\top}, \ldots, \beta^{T\top})^\top \in \mathbb{R}^{\sum_{t=1}^T K^t}$ through

$$f_{\mathcal{E},B}(x) = \sum_{t=1}^{T} \sum_{k=1}^{K^t} \beta_k^t \mathbb{1}(x \in \Omega_k^t). \tag{2}$$

That is, $f_{\mathcal{E},B}$ is constant on the boxes constructed by overlapping $\mathcal{Z}$-tree partitions $\mathcal{T}^1, \ldots, \mathcal{T}^T$. Chipman et al. (2010) recommends the choice $T = 200$, which was seen to provide good empirical results. For a given ensemble $\mathcal{E}$, we henceforth define $\mathcal{F}_{\mathcal{E}} = \{f_{\mathcal{E},B} : B \in \mathbb{R}^{\sum_{t=1}^T K^t}\}$ the set of functions in (2). If $\mathcal{E}$ consists of a single tree $\mathcal{T}$, we instead write $\mathcal{F}_{\mathcal{T}}$ to denote $\mathcal{F}_{\mathcal{E}}$.

[4] The notation $\mathcal{T} = \{\Omega_k\}_k$ is used only for the $\mathcal{Z}$-tree partitions with a split-net $\mathcal{Z}$, with some suitable superscript and/or superscript if required. We denote flexible tree partitions by $\mathfrak{Y} = \{\Psi_k\}_k$ as general box partitions.

Our objective is to characterize the posterior asymptotic properties of the tree learners in (2) in estimating the true function f_0 belonging to $\Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$ or $\Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)\cap\mathcal{C}([0,1]^p)$. This goal requires two nice attributes of the procedure. First, appropriate prior distributions should be assigned to the tree learners $f_{\mathcal{E},B}$ in (2) so that the induced posterior can achieve the desired asymptotic properties. Second, there should exist a piecewise tree learner approximating f_0 with a suitable approximation error matched to our target rate. In the following two sections, we elucidate these in detail.

3. Tree and Forest Priors in High Dimensions

3.1 Priors over Tree Topologies with Sparsity

Conventional tree priors (Chipman et al., 1998; Denison et al., 1998) are not designed for high-dimensional data with a sparse underlying structure. Prior modifications are thus required for trees to meet demands of high-dimensional applications (Linero, 2018; Linero and Yang, 2018; Ročková and van der Pas, 2020). Ročková and van der Pas (2020) adopted a spike-and-slab prior for BART to achieve adaptability to unknown sparsity levels, but the computation of the posterior distribution is much more challenging than the original BART algorithm owing to the nature of a point mass prior. Linero (2018) and Linero and Yang (2018) considered a sparse Dirichlet prior on splitting coordinates for a computationally feasible algorithm, while achieving the theoretical optimality in the high-dimensional scenario. We deploy the sparse Dirichlet prior developed by Linero (2018) for ease of computation for the posterior distribution.

Unlike the original tree priors, the BART model with the sparse Dirichlet prior chooses a splitting coordinate j is from a proportion vector $\eta=(\eta_1,\dots,\eta_p)^\top$ belonging to the p-dimensional simplex $\mathbb{S}^p=\{(x_1,\dots,x_p)^\top\in\mathbb{R}^p:\sum_{j=1}^p x_j=1, x_j\ge 0, j=1,\dots,p\}$. A proportion vector η has a Dirichlet prior with $\zeta>0$ and $\xi>1$,

$$\eta=(\eta_1,\dots,\eta_p)^\top\sim\mathrm{Dir}(\zeta/p^\xi,\dots,\zeta/p^\xi). \tag{3}$$

The requirement $\xi>1$ is needed for technical reasons. The prior imposes a sparsity into splitting variables (we refer the reader to Figure 2 of Linero (2018)). Given a proportion vector η, the BART prior is assigned, as in Chipman et al. (2010), with a minor modification. Assuming an independent product prior for $\mathcal{E}$, i.e., $\Pi(\mathcal{E})=\prod_{t=1}^T\Pi(\mathcal{T}^t)$, a Bayesian CART prior (Chipman et al., 1998) is assigned to each $\mathcal{T}^t$. The procedure begins with the root node $[0,1]^p$ of depth $\ell=0$, where the depth of a node means the number of nodes along the path from the root node down to that node. For each $\ell=0,1,2,\dots,$ each node at depth ℓ is split with prior probability $\nu^{\ell+1}$ for $\nu\in(0,1/2)$. If a node corresponding to a box Ω is split, a splitting coordinate j is drawn from the proportion vector η and a split-point τ_j will be chosen randomly from $[\mathcal{Z}]_j\cap\mathsf{int}([\Omega]_j)$ for a given $\mathcal{Z}$. The procedure repeats until all nodes are terminal.

The original CART prior proposed by Chipman et al. (1998) uses a splitting probability that decays polynomially. Ročková and Saha (2019) showed that this decay may not be fast enough, and suggested using an exponentially decaying probability as ours. This modification gives rise to the desirable exponential tail property of tree sizes. Linero and Yang (2018) handled this issue by assigning a prior on the number T of trees. As we want to fix T as in the practical implementation of BART, we use the exponentially decaying prior probability for splits.

3.2 Prior on Step-Heights

To complete the prior on the sparse function space, what remains to be specified is the prior on step-heights B in (2). Given $K^1, \ldots, K^T$ induced by $\mathcal{E}$, Chipman et al. (2010) suggests using a Gaussian prior on B (after shifting and rescaling the responses):

$$d\Pi(B|K^1, \ldots, K^T) = \prod_{t=1}^{T} \prod_{k=1}^{K^t} \phi(\beta_k^t; 0, c_\beta/T),$$

where $c_\beta > 0$ is a constant and $\phi(\,\cdot\,; \mu, \tau^2)$ is the Gaussian density with mean μ and variance τ^2. The variance c_β/T shrinks step-heights toward zero, limiting the effect of individual components by keeping them small enough for large T. This choice is preferred in view of the practical performance, but any zero-mean multivariate Gaussian prior on B gives rise to the same optimal properties as soon as the eigenvalues of the covariance matrix are bounded below and above. Throughout the paper, we place a Gaussian prior on the step-heights B in most cases. From the computational point of view, this choice is certainly appealing in Gaussian nonparametric regression owing to its semi-conjugacy. For theoretical purposes, a prior with exponentially decaying thicker tails, such as a Laplace distribution, can easily replace a Gaussian prior for the same optimality under relaxed conditions. Although such a prior may loosen a restriction on $\|f_0\|_\infty$ (Ročková, 2020; Jeong and Ghosal, 2021a), we primarily consider normal priors throughout the paper, even for non-Gaussian models for the sake of simplicity. We consider non-Gaussian priors only when required for theoretical purposes; see, for example, a truncated prior for regression with random design in Section 6.

4. Approximating the True Function

Recall that tree learners $f_{\mathcal{E},B}$ in (2) are piecewise constant, whereas the true function f_0 does not have to be. This will not be an issue as long as there exists a tree learner that can approximate f_0 sufficiently well. In this section, we establish the approximation theory for tree ensembles in the context of our targeted function spaces.

For isotropic classes, balanced k-d trees (Bentley, 1979) are known to give rise to rate-optimal approximations under mild regularity conditions (Ročková and van der Pas, 2020). This is not necessarily the case for our general setup where smoothness may vary over the domain and where cycling repeatedly through the coordinates (as is done in the k-d tree) may not be enough to capture localized features of f_0. We thus generalize the notion of k-d trees and show that there exists a good partitioning scheme for piecewise heterogeneous anisotropic classes. Although our primary interest lies in additive tree aggregations in (2), we show that a single deep tree can approximate well. We thereby consider only single trees $\mathcal{T}$ and suppress the superscript t throughout this section.

4.1 Split-Nets for Approximation

Approximation properties of tree-based estimators are driven by the granularity and fineness of a chosen split-net. Roughly speaking, a good approximation requires that a split-net have

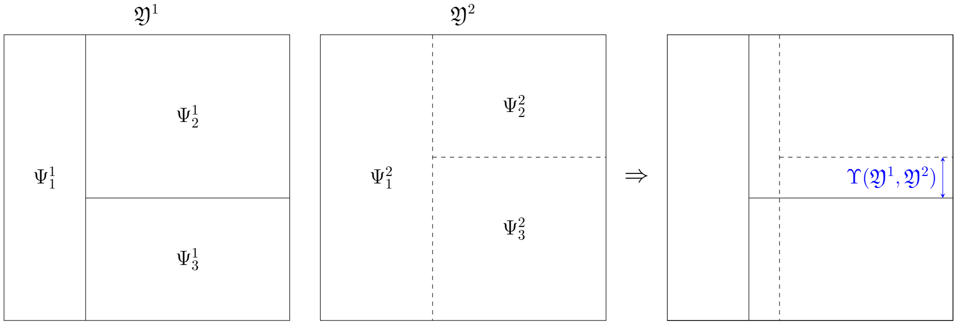


Figure 6: A two-dimensional example of the Hausdorff-type divergence in Definition 7. The divergence is the maximum dependency of the boxes in the partitions.

two properties: (i) it should be dense enough so that the boundaries of the box partition $\mathfrak{X}_0^* = \{\Xi_1^*, \dots, \Xi_R^*\}$, extended from $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$, can be detected by a $\mathcal{Z}$-tree partition with a minimal error; and (ii) it should be regular enough so that there exists a $\mathcal{Z}$-tree partition that captures local/global features of f_0 on each Ξ_r^*. We elucidate these two properties.

4.1.1 Dense Split-Nets: Global Approximability

Recall that the underlying partition $\mathfrak{X}_0^* = \{\Xi_1^*, \dots, \Xi_R^*\}$ for the true function is unknown. From the sheer flexibility of binary tree partitioning, we expect that the boundaries can be detected well enough by a $\mathcal{Z}$-tree partition if $\mathfrak{X}_0^*$ is a flexible tree partition. If the prior rewards partitions that are sufficiently close to $\mathfrak{X}_0^*$, Bayesian CART (BART) is expected to adapt to unknown $\mathfrak{X}_0^*$ without much loss of efficiency. We examine when this adaptivity can be achieved in more detail below.

The ability to detect $\mathfrak{X}_0^*$ is thus closely tied to the density of the split-net $\mathcal{Z}$; it should be dense enough so that a $\mathcal{Z}$-tree partition can be constructed that is sufficiently close to $\mathfrak{X}_0^*$. Therefore, we need a gadget to measure the closeness between two partitions. To this end, we introduce a Hausdorff-type divergence; see Figure 6 for an illustration.

Definition 7 (Hausdorff-type divergence)**.** For any two box partitions $\mathfrak{Y}^1 = \{\Psi_1^1, \dots, \Psi_J^1\}$ and $\mathfrak{Y}^2 = \{\Psi_1^2, \dots, \Psi_J^2\}$ with the same number J of boxes, we define a divergence between $\mathfrak{Y}^1$ and $\mathfrak{Y}^2$ as

$$\Upsilon(\mathfrak{Y}^1, \mathfrak{Y}^2) = \min_{(\pi(1)\dots\pi(J))\in P_\pi[J]} \max_{1\le r\le J} \mathrm{Haus}(\Psi_r^1, \Psi_{\pi(r)}^2),$$

where $P_\pi[J]$ denotes the set of all permutations $(\pi(1)\dots\pi(J))$ of $\{1,\dots,J\}$ and $\mathrm{Haus}(\cdot,\cdot)$ is the Hausdorff distance.

The permutation in Definition 7 makes the specification immune to the ordering of boxes. We want the split-net $\mathcal{Z}$ to produce a $\mathcal{Z}$-tree partition $\mathcal{T}$ such that $\Upsilon(\mathfrak{X}_0^*, \mathcal{T})$ is smaller than some threshold. Section 4.2 establishes how small these thresholds should be so that the tree learner is close to f_0 (for various approximation metrics). The following definition will be useful in characterizing the details.

Definition 8 (Dense split-net)**.** For a given subset $S \subseteq \{1,\dots,p\}$ and an integer $J \geq 1$, consider an S-chopped partition $\mathfrak{Y} = \{\Psi_1,\dots,\Psi_J\}$ of $[0,1]^p$ with boxes $\Psi_r \subseteq [0,1]^p$, $r = 1,\dots,J$. For any given $c_n \geq 0$, a split-net $\mathcal{Z} = \{z_i \in [0,1]^p, i = 1,\dots,b_n\}$ is said to be $(\mathfrak{Y}, c_n)$-dense if there exists an S-chopped $\mathcal{Z}$-tree partition $\mathcal{T} = \{\Omega_1,\dots,\Omega_J\}$ of $[0,1]^p$ such that $\Upsilon(\mathfrak{Y},\mathcal{T}) \leq c_n$.

In Section 4.2, the approximation theory will require that $\mathcal{Z}$ be $(\mathfrak{X}_0^*, c_n)$-dense for some suitable $c_n \geq 0$. Note that the ideal case $c_n = 0$ can be achieved only when $\mathfrak{X}_0^*$ is a $\mathcal{Z}$-tree partition. This condition, while obviously satisfied in the case $R = 1$, is very restrictive in the most situations. This is because, if $J = 1$, i.e., $\mathfrak{Y} = \{[0,1]^p\}$, we obtain $\Upsilon(\mathfrak{Y},\mathcal{T}) = 0$ for $\mathcal{T} = \{[0,1]^p\}$. Hence, every split-net $\mathcal{Z}$ is $(([0,1]^p), 0)$-dense. However, we will see in Theorem 1 that, in many cases, it is sufficient that c_n tends to zero at a suitable rate. This means that $\mathfrak{X}_0^*$ should be at least a flexible tree partition, but not necessarily a $\mathcal{Z}$-tree partition. If $\mathfrak{X}_0^*$ is a box partition but not a flexible tree partition, we can redefine $\mathfrak{X}_0^*$ by adding more splits to make it a flexible tree partition. For example, the non-tree box partition in Figure 4 can be extended to a tree partition with a single extra split. However, this approach increases R and hence may deteriorate the result (observe that our rate in (7) is dependent on R). In particular, if $\mathfrak{X}_0^*$ is not a box partition (e.g., jumps are not axis-parallel), the redefined R increases to infinity. For our theory to be valid, $\mathfrak{X}_0^*$ must be at least a box partition. In Section 4.3, we present some examples of dense split-nets.

Dense split-nets have nested properties. That is, a $(\mathfrak{Y}, c_n)$-dense split-net is also $(\mathfrak{Y}, \tilde{c}_n)$-dense for every $\tilde{c}_n \geq c_n$. We are interested in the smallest possible c_n. In particular, every split-net $\mathcal{Z}$ is $(\mathfrak{Y}, 1)$-dense for any box partition $\mathfrak{Y}$.

4.1.2 Regular Split-Nets: Local Approximability

Beyond closely tracking smoothness boundaries, good tree partitions should be able to capture local/global smoothness features of f_0. In other words, there should exist a $\mathcal{Z}$-tree partition that achieves an optimal approximation error determined by our target rate. In Section 4.1.1, we focused on *global* approximability of underlying partitions, which requires split-nets to be suitably dense. Now, we focus on *local* approximability.

Assume that $\mathfrak{X}_0^*$ can be approximated well (as discussed in the previous section) by an S_0^*-chopped $\mathcal{Z}$-tree partition $\mathcal{T}^* = \{\Omega_1^*,\dots,\Omega_R^*\}$,[5] which is formally written as

$$\mathcal{T}^* = \operatorname*{arg\,min}_{\mathcal{T} \in \mathscr{T}_{S_0^*,R,\mathcal{Z}}} \Upsilon(\mathfrak{X}_0^*, \mathcal{T}). \tag{4}$$

We now focus on local approximability inside each box Ω_r^*. Ideally, one would want to construct a sub-tree partition of this local box that balances out approximation errors in all coordinates. Therefore, we first need to devise a splitting scheme to achieve this balancing condition. The regularity of split-nets can then be spelled out based on such a law.

We now zoom onto a single box Ω_r^*. Recall that the true function f_0 has anisotropic smoothness on each of Ξ_r^*. Intuitively, denser subdivisions are required for less smooth coordinates to capture the local features. Allowing splits to occur more often in certain directions, we define the *anisotropic k-d tree*, which achieves the desired approximation error

[5] The notation $\mathcal{T}^* = \{\Omega_1^*,\dots,\Omega_R^*\}$ with an asterisk is only used to denote an S_0^*-chopped $\mathcal{Z}$-tree partition approximating $\mathfrak{X}_0^* = \{\Xi_1^*,\dots,\Xi_R^*\}$.

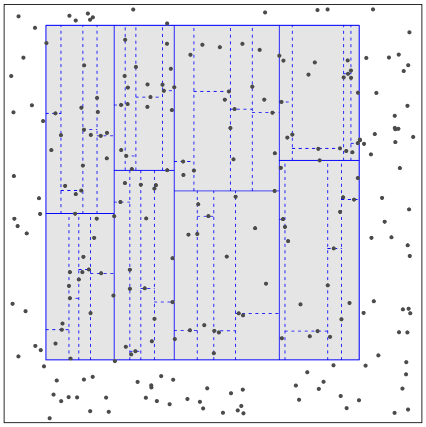


Figure 7: A realization of the anisotropic k-d tree with smoothness parameters $\alpha_1 = 0.25$ (for the horizontal axis) and $\alpha_2 = 0.5$ (for the vertical axis), and a box Ψ (the shaded box) that is a subspace of $[0,1]^2$ (the outer square). Because $2\alpha_1 = \alpha_2$, the subset Ψ splits twice as often in the vertical direction than in the horizontal direction.

for anisotropic smoothness. The definition requires the notion of *midpoint-splits* defined as follows. For a given box Ψ and a splitting coordinate j, a midpoint-split picks up the $\lceil \tilde{b}_j(\mathcal{Z},\Psi)/2 \rceil$th split-candidate in $[\mathcal{Z}]_j \cap \mathsf{int}([\Psi]_j)$ as a split-point τ_j, where $\tilde{b}_j(\mathcal{Z},\Psi)$ is the cardinality of $[\mathcal{Z}]_j \cap \mathsf{int}([\Psi]_j)$.

Definition 9 (Anisotropic k-d tree)**.** Consider a smoothness vector $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in (0,1]^d$, a box $\Psi \subseteq [0,1]^p$, a split-net $\mathcal{Z} = \{z_i \in [0,1]^p,\ i = 1, \dots, b_n\}$, an integer $L > 0$, and an index set $S = \{s_1, \dots, s_d\} \subseteq \{1, \dots, p\}$ with $|S| = d$. We define the *anisotropic k-d tree* $\mathsf{Akd}(\Psi; \mathcal{Z}, \alpha, L, S)$ as the iterative splitting procedure that partitions Ψ into disjoint boxes as follows.

1. Start from the root node by setting $\Omega_1^\circ = \Psi$ and set $l_j = 0$, $j = 1, \dots, d$.
2. For splits at iteration $1 + \sum_{j=1}^d l_j$, choose j corresponding to the smallest $l_j \alpha_j$. If the smallest $l_j \alpha_j$ is duplicated with multiple js, choose the smallest j among such j's.
3. For all boxes Ω_k°, $k = 1, \dots, 2^{\sum_{j=1}^d l_j}$, at the current iteration, do the midpoint-splits with the given $\mathcal{Z}$ and the splitting coordinate s_j chosen by j. Relabel the generated new boxes as Ω_k°, $k = 1, \dots, 2^{1+\sum_{j=1}^d l_j}$, and then increase l_j by one for chosen j.
4. Repeat 2–3 until either $\sum_{j=1}^d l_j = L$ or the midpoint-split is no longer available. Return $(l_1, \dots, l_d)^\top$ and $\mathcal{T}^\circ = \{\Omega_1^\circ, \dots, \Omega_{2^{L^\circ}}^\circ\}$, where $L^\circ = \sum_{j=1}^d l_j$.

Note that the anisotropic k-d tree construction depends on the smoothness that is unknown. Rather than a practical estimator, we use this to show that there exists a good tree approximator in the technical proof. One possible realization of the anisotropic k-d tree generating process is given in Figure 7. Observe that $\mathsf{Akd}(\Psi; \mathcal{Z}, \alpha, L, S)$ returns a tree partition $\mathcal{T}^\circ = \{\Omega_1^\circ, \dots, \Omega_{2^{L^\circ}}^\circ\}$ of Ψ and a vector $(l_1, \dots, l_d)^\top$ such that $L^\circ = \sum_{j=1}^d l_j \le L$.[6]

[6] The notation $\mathcal{T}^\circ = \{\Omega_k^\circ\}_k$ with a circle is used only for tree partitions of some box $\Psi \subseteq [0,1]^p$, returned by the anisotropic k-d trees, with some suitable subscript if required.

Although these returned items clearly depend on the inputs of the anisotropic k-d tree procedure (i.e., Ψ, $\mathcal{Z}$, α, L, and S), we suppress them throughout the paper. Each l_j is a counter of how many times the jth coordinate has been used. The procedure is designed so that every l_j is approximately proportional to α_j^{-1} after enough iterations. The total number of splits for the jth coordinate is thus close to $2^{C/\alpha_j}$ for every j with some $C > 0$. In the proof of Theorem 1, this matching is indeed clearly optimal and minimizes the induced bias.

To play a role as a 'sieve' for approximation, Ψ needs to be sufficiently finely subdivided to capture the global/local behavior of a function. The threshold L determines the resolution of the returned tree partition $\mathcal{T}^\circ = \{\Omega_1^\circ, \dots, \Omega_{2^{L^\circ}}^\circ\}$. For a good approximation, we are particularly interested in the situation when $L^\circ = L$, i.e., the resulting tree has the desired depth. If $L^\circ < L$ owing to insufficient split-candidates, the resolution may not be good enough.

Now, we can define the regularity of a split-net on $\Psi \subseteq [0,1]^p$ using $\mathcal{T}^\circ$. The desirable situation is when all the splits occur nearly at the center of boxes such that, for any given $j \in S$, all $\mathsf{len}([\Omega_k^\circ]_j)$, $k = 1, \dots, 2^L$, are balanced well. The evenness of the returned partition is solely determined by the regularity of a split-net $\mathcal{Z}$. Intuitively, the split-net should be regularly distributed to give rise to an appropriate partition, in which we say a split-net is regular. We make the definition technically precise below, which will be used as a basis for approximating the function classes. See Verma et al. (2009) for a related regularity condition.

Definition 10 (Regular split-net)**.** For a given box $\Psi \subseteq [0,1]^p$, an integer $L > 0$, and an index set $S = \{s_1, \dots, s_d\} \subseteq \{1, \dots, p\}$, we say that a split-net $\mathcal{Z}$ is (Ψ, α, L, S)-regular if $\mathcal{T}^\circ = \{\Omega_1^\circ, \dots, \Omega_{2^{L^\circ}}^\circ\}$ and $(l_1, \dots, l_d)^\top$, returned by $\mathsf{Akd}(\Psi; \mathcal{Z}, \alpha, L, S)$, satisfy $L^\circ = L$ and $\max_k \mathsf{len}([\Omega_k^\circ]_{s_j}) \lesssim \mathsf{len}([\Psi]_{s_j}) 2^{-l_j}$ for every $j = 1, \dots, d$.

The condition $\max_k \mathsf{len}([\Omega_k^\circ]_{s_j}) \lesssim \mathsf{len}([\Psi]_{s_j}) 2^{-l_j}$ is the key to obtaining optimal approximation results. In the ideal case that all the splits occur exactly at the center, this condition is trivially satisfied as $\max_k \mathsf{len}([\Omega_k^\circ]_{s_j}) = \mathsf{len}([\Psi]_{s_j}) 2^{-l_j}$. The inequality provides a lot more flexibility where the condition can be satisfied in most cases except for very extreme situations. See Section 4.3 for examples of regular split-nets.

Similar to dense split-nets, regular split-nets also have nested properties. If a split-net $\mathcal{Z}$ is (Ψ, α, L, S)-regular for some Ψ, α, L, and S, then it is also $(\Psi, \alpha, \tilde{L}, S)$-regular for any $\tilde{L} \le L$. This can be easily shown by noting that the latter is determined only by a pruned tree of the full-blown tree for the former. We are particularly interested in the largest possible L.

Remark 5. As regular split-nets require the desired depth, i.e., $L^\circ = L$, it is of interest to see which L achieves this precondition. Consider a box $\Psi \subseteq [0,1]^p$ and a split-net $\mathcal{Z} = \{z_i \in [0,1]^p,\ i = 1, \dots, b_n\}$. If there are no ties in $\mathcal{Z}$ for any coordinate, i.e., $b_j(\mathcal{Z}) = b_n$, $j = 1, \dots, p$, it can be easily checked that any integer $L \le \lfloor \log_2(\tilde{b}_j(\mathcal{Z}; \Psi) + 1) \rfloor$ gives rise to $L^\circ = L$ with the anisotropic k-d tree. (Observe that all $\tilde{b}_j(\mathcal{Z}; \Psi)$ are identical in this case.) If there are ties, L may need to be much smaller to achieve $L^\circ = L$, but a tight upper bound may not be obtained for the general case.

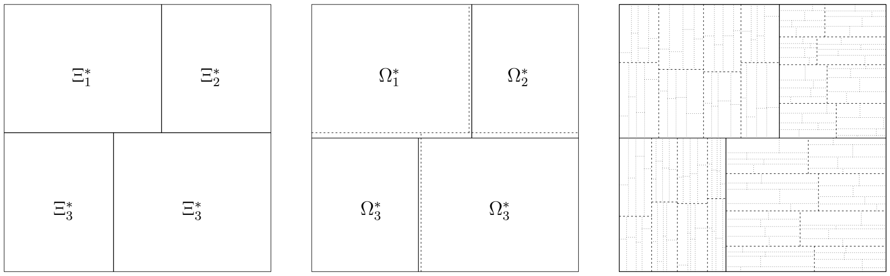


Figure 8: An example of constructing $\widehat{\mathcal{T}}$. First, $\mathfrak{X}_0^* = \{\Xi_1^*, \dots, \Xi_4^*\}$ is approximated by $\mathcal{T}^* = \{\Omega_1^*, \dots, \Omega_4^*\}$. Then, each Ω_r^* is subdivided by the anisotropic k-d tree, producing $\mathcal{T}_r^\circ$ a constituent of $\widehat{\mathcal{T}}$ displayed on the rightmost panel.

4.2 Approximation Theory

Our goal is to establish the contraction rate of the posterior distribution. The construction requires that tree learners be able to approximate functions in the spaces $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$ and $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$ appropriately. Here, we investigate the approximation properties for these sparse function spaces.

Recall that a split-net $\mathcal{Z}$ is required to be suitably dense and regular. First, a split-net $\mathcal{Z}$ should be $(\mathfrak{X}_0^*, c_n)$-dense for some appropriate c_n, so that the boundaries of $\mathfrak{X}_0^* = \{\Xi_1^*, \dots, \Xi_R^*\}$ can be detected well by the binary tree partitioning rule. As $\mathfrak{X}_0^*$ is approximated by a $\mathcal{Z}$-tree partition with a given $\mathcal{Z}$, the underlying partition $\mathfrak{X}_0^*$ should be at least a flexible tree partition, but a stronger result is obtained if $\mathfrak{X}_0^*$ is a $\mathcal{Z}$-tree partition (see Theorem 1 below). Denoting by $\mathcal{T}^* = \{\Omega_1^*, \dots, \Omega_R^*\}$ the S_0^*-chopped $\mathcal{Z}$-tree partition in (4), each box Ω_r^* should be appropriately subdivided to capture the local/global nature of the true function on Ξ_r^*. (If $R = 1$, we write $\mathcal{T}^* = \mathfrak{X}_0^* = \{[0,1]^p\}$ with $\Omega_1^* = [0,1]^p$.) Hence, for a smoothness parameter $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ and some suitably chosen $L > 0$, $\mathcal{Z}$ should also be $(\Xi_r^*, \alpha_r, L, S_0)$-regular, $r = 1, \dots, R$. The integer sequence L will eventually be chosen such that the approximation error is balanced with our target rate (see L_0 in Theorem 1). Let $\mathcal{T}_r^\circ = \{\Omega_{r1}^\circ, \dots, \Omega_{r2^L}^\circ\}$ be the tree partition of Ω_r^* returned by $\mathsf{Akd}(\Omega_r^*; \mathcal{Z}, \alpha_r, L, S_0)$, $r = 1, \dots, R$. Then, the approximating partition $\widehat{\mathcal{T}}$ is formed by agglomerating all sub-tree partitions $\mathcal{T}_r^\circ$, leading to an S_0-chopped $\mathcal{Z}$-tree partition

$$\widehat{\mathcal{T}} = \left\{\Omega_{11}^\circ, \dots, \Omega_{12^L}^\circ, \dots, \Omega_{R1}^\circ, \dots, \Omega_{R2^L}^\circ\right\}. \tag{5}$$

(Note that each $\mathcal{T}_r^\circ$ is S_0-chopped, not S_0^*-chopped.) A graphical illustration of constructing $\widehat{\mathcal{T}}$ is given in Figure 8.

The strongest approximation results relative to the L_∞-norm for $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$ are of particular interest. Owing to the possible discontinuity or heterogeneity at the unknown boundaries of $\mathfrak{X}_0^*$, however, such results are not practically obtained except for the case $R = 1$. As the following theorem shows, the conditions can be relaxed if we opt for weaker metrics, which often suffice in many statistical setups. For example, in our examples of Gaussian nonparametric regression in Section 5.1, we only need an approximation rate in L_2- or empirical L_2-sense. The approximation results for the continuous variant $\Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$ require even milder conditions.

Theorem 1 (Approximation theory)**.** *For $c_n \geq 0$ specified below, assume that a split-net $\mathcal{Z}$ is $(\mathfrak{X}_0^*, c_n)$-dense. For a smoothness parameter $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ and an integer $L > 0$, assume that $\mathcal{Z}$ is $(\Xi_r^*, \alpha_r, L, S_0)$-regular for every $r = 1, \dots, R$. Let $\tilde{\epsilon}_n$ be a sequence satisfying $\tilde{\epsilon}_n \gtrsim \lambda d 2^{-\bar{\alpha}L/d}$ and construct $\widehat{\mathcal{T}}$ as in (5) (through $\mathcal{T}^*$ in (4)). Then, for any $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$, there exists $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ such that*

(i) $\|f_0 - \hat{f}_0\|_\infty \lesssim \tilde{\epsilon}_n$ *if* $c_n = 0$*;*
(ii) $\|f_0 - \hat{f}_0\|_v \lesssim \tilde{\epsilon}_n$ *if* $c_n \lesssim (\tilde{\epsilon}_n/\|f_0\|_\infty)^v \min_{r,j} \mathsf{len}([\Xi_r^*]_j)/|S_0^*|$ *for any* $v \geq 1$*;*
(iii) $\|f_0 - \hat{f}_0\|_{v,P_{\mathcal{Z}}} \lesssim \tilde{\epsilon}_n$ *for any* $v \geq 1$*, where* $P_{\mathcal{Z}}(\cdot) = b_n^{-1}\sum_{i=1}^{b_n} \delta_{z_i}(\cdot)$*.*

Further, for any $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$, there exists $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ such that

(iv) $\|f_0 - \hat{f}_0\|_\infty \lesssim \tilde{\epsilon}_n$ *if* $c_n^{\min_{r,j}\alpha_{rj}} \lesssim \tilde{\epsilon}_n/(\lambda|S_0^*|)$*;*
(v) $\|f_0 - \hat{f}_0\|_v \lesssim \tilde{\epsilon}_n$ *if* $c_n^{1+v\min_{r,j}\alpha_{rj}} \lesssim (\tilde{\epsilon}_n/\lambda)^v \min_{r,j} \mathsf{len}([\Xi_r^*]_j)/|S_0^*|^{v+1}$ *for any* $v \geq 1$*.*

In particular, if we choose $L = L_0$ such that $2^{L_0} \asymp (n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)}$, then the above assertions hold for $\tilde{\epsilon}_n = \bar{\epsilon}_n := (\lambda d)^{d/(2\bar{\alpha}+d)}((R\log n)/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}$.

Proof. See Section A.1 in Appendix. ■

Although Theorem 1 holds for any $\tilde{\epsilon}_n \gtrsim \lambda d 2^{-\bar{\alpha}L/d}$, the results are particularly useful for our purposes when combined with L_0 and $\bar{\epsilon}_n$, motivated by our target rate ϵ_n in (7). The assertion in (i) gives the strongest result with the L_∞-norm. However, the condition $c_n = 0$ requires that the boundaries of the pieces be correctly detectable by the binary tree partitioning rule with a given split-net $\mathcal{Z}$; that is, $\mathfrak{X}_0^*$ should be a $\mathcal{Z}$-tree partition. Except for the case $R = 1$, this limitation is too restrictive and impractical, as the locations of the boundaries are unknown (every split-net $\mathcal{Z}$ is $(\mathfrak{X}_0^*, 0)$-dense if $R = 1$). The assertion in (iv) relaxes this limitation by means of the continuity restriction. We will use (i) and (iv) for a density estimation problem in Section 6.2.

The assertions in (ii) and (v) are with respect to the L_v-norm, $v \geq 1$, which is useful in many statistical setups. We note that, despite the continuity restriction, the condition for c_n of (v) is not always milder than that of (ii). Indeed, the former is milder than the latter only if $\lambda|S_0^*|c_n^{\min_{r,j}\alpha_{rj}} \lesssim \|f_0\|_\infty$, which is often satisfied, as the left-hand side is prone to be decreasing with a suitably chosen c_n. We will use the results in (ii) and (v) for nonparametric regression and binary classification with random design in Sections 6.1 and 6.3.

The assertion in (iii) is particularly useful in regression setups with $\mathcal{Z}$ chosen by fixed covariates; see Sections 4.3.2 and 6.4. Note that (iii) only explicitly requires the regularity of a split-net $\mathcal{Z}$, and an upper bound for c_n is not specified. This is because the closeness between f_0 and $\hat{f}_0$ is measured only at points in $\mathcal{Z}$, and the boundary detection needs to be performed much loosely compared with the other metrics. Although not explicitly stated, (iii) still requires a dense split-net in an implicit way. Indeed, every assertion in Theorem 1 necessitates a condition on $\mathcal{T}^*$ imposed implicitly by the regularity with Ξ_r^*; for $\mathcal{Z}$ to be regular for every Ξ_r^*, it must be sufficiently evenly distributed and hence suitably dense.

As stated above, if $R = 1$, i.e., the global anisotropic case, we always obtain the strongest result in (i) as soon as a split-net is suitably regular. If $R > 1$, a split-net should also be suitably dense except for the case of the empirical $\|\cdot\|_{v,P_{\mathcal{Z}}}$-norm in (iii). As the conditions on c_n depend on unknown model specification, e.g., $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$, λ, and $|S_0^*|$, more practical conditions can be obtained by plugging in reasonable bounds of the unknown components. For example, we cannot hope for better than $\bar{\epsilon}_n \gtrsim (\lambda d R(\log n)/n)^{1/3}$ owing to the fundamental limitation of piecewise constant learners. We can also assume that $\min_{r,j} \mathsf{len}([\Xi_r^*]_j)$ is bounded away from zero or decreases at most polynomially. To establish the posterior contraction rate, we will eventually assume $\|f_0\|_\infty \lesssim \sqrt{\log n}$ (see (A3) below). Because the necessary conditions $d/\bar{\alpha} \ll \log n$ and $\lambda^{\bar{\alpha}/d} R \ll n$ are required for consistent estimation (see the rate in (7) below), making mild assumptions on d and λ is not prohibitive (note that $|S_0^*| \le d$). Putting everything together, the conditions on c_n can be easily satisfied if c_n is a decreasing polynomial in n with a suitable exponent. The results are formalized in the following corollary.

Corollary 1 (Approximation with L_∞ and L_v when $R > 1$)**.** *Under the setup of Theorem 1 with $L = L_0$, suppose that $R > 1$ and $d \lesssim \log n$. Then, the following assertions hold.*

(i) *Suppose that $\min_{r,j} \alpha_{rj} \ge a_1$ and $\lambda \lesssim n^{a_2}$ for some constants $a_1 > 0$ and $a_2 \ge 0$. If $c_n \lesssim n^{-(1+2a_2)/(3a_1)}(\log n)^{-1/(3a_1)}$, then for every $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$, there exists $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ such that $\|f_0 - \hat{f}_0\|_\infty \lesssim \bar{\epsilon}_n$.*

(ii) *Suppose that $\|f_0\|_\infty \lesssim \sqrt{\log n}$ and $\min_{r,j} \mathsf{len}([\Xi_r^*]_j) \gtrsim n^{-a_3}$ for some constant $a_3 \ge 0$. Fix any $v \ge 1$. If $c_n \lesssim n^{-(v/3+a_3)}(\log n)^{-(\max\{0,1-v/3\}+v/6)}$, then for every $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$, there exists $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ such that $\|f_0 - \hat{f}_0\|_v \lesssim \bar{\epsilon}_n$.*

Proof. See Section A.1 in Appendix. ■

Corollary 1 implies that the target approximation error is attained with both the L_∞- and L_v-norms as soon as c_n decreases polynomially. The assertion in (i) provides the stronger result with the aid of the continuous restriction. It also requires a constant lower bound of the minimum smoothness parameter $\min_{r,j} \alpha_{rj}$, causing $\bar{\alpha}$ to be bounded away from zero. In contrast, (ii) removes such a restriction at the expense of a tighter upper bound. In general, the conditions for (ii) are much milder, yielding a relatively weaker but still useful result in many statistical setups.

Remark 6. No upper bounds for b_n and $b_j(\mathcal{Z})$ are made for Theorem 1; the approximation results are more easily achieved with larger values of $b_j(\mathcal{Z})$, $j = 1, \dots, p$. However, values increasing too fast may harm the contraction rate as they escalate the model complexity. In Section 5, we will see that our main results on the optimal posterior contraction require that $\max_{1 \le j \le p} \log b_j(\mathcal{Z}) \lesssim \log n$. We are ultimately interested in well-balanced split-nets.

Remark 7. Our approximation theory is presented with the error $\bar{\epsilon}_n$ motivated by our target rate ϵ_n in (7). However, what we really need is the weaker approximation error ϵ_n, which is identical to the posterior contraction rate (see Sections 5–6). Although the latter slightly relaxes the required conditions, we stick to the approximation result with $\bar{\epsilon}_n$ because such generalization complicates the technical details too much for a small gain.

Remark 8. The assertion in (iii) requires $\mathcal{Z}$ to be regular over $[0,1]^p$. Because the assertion is with respect to $L_v(\mathcal{P}_{\mathcal{Z}})$-norm, one may anticipate the regularity over $[0,1]^p$ to be relaxed into a smaller subset. Indeed, we can restrict our attention to a subset of $[0,1]^p$ and the technical details require the regularity $\mathcal{Z}$ only over such a smaller subset. We do not consider such an extension so that the $L_v(\mathcal{P}_{\mathcal{Z}})$-consistency can be interpreted as an approximate result for the L_v-norm, which is more appealing in the usual sense.

4.3 Examples of Split-Nets for Approximation

Although the notion of dense and regular split-nets is crucial in characterizing the approximation theory in Section 4.2, how to obtain such a good split-net in practice remains unsolved. Clearly, a split-net attains the suitable density and regularity more easily with larger b_n. As mentioned in Remark 6, however, we will see that a split-net must satisfy $\max_{1\le j\le p}\log b_j(\mathcal{Z}) \lesssim \log n$ to establish the optimal posterior contraction rate. Accordingly, our primary concern is examining split-nets that are suitably dense and regular under the restriction on $\log b_j(\mathcal{Z})$. In this subsection, we show that the two split-nets described in Section 2.3 are dense and regular as required, and hence fulfill the requirements of Theorem 1 and Corollary 1.

4.3.1 Regular Grid

We first consider a regular grid $\mathcal{Z} = \{(i-1/2)/b_n^{1/p}, i=1,\dots,b_n^{1/p}\}^p$ for b_n such that $b_n^{1/p}$ is an integer. This simplest example is a split-net according to Definition 5. We will see that a regular grid can be useful for density estimation, binary classification, and nonparametric regression with random design, but it also has the potential to be used for many other statistical models. A two-dimensional example is illustrated in Figure 5a. The following lemma shows that, with an appropriately chosen b_n, a regular grid is suitably dense and regular under mild conditions.

Lemma 1 (Regular grid)**.** *Consider a regular grid $\mathcal{Z}$ with $b_n = n^{cp}$ for a constant $c \ge 1$. If $\min_{r,j}\mathsf{len}([\Xi_r^*]_j) \ge n^{-c}$ and $\lambda d/\min_{r,j}\mathsf{len}([\Xi_r^*]_j)^{\bar\alpha/d+1/2} \lesssim n^{c\bar\alpha/d+(c-1)/2}\sqrt{R\log n}$, then $\mathcal{Z}$ is $(\mathfrak{X}_0^*, c_n)$-dense and $(\Xi_r^*, \alpha_r, L_0, S_0)$-regular for $r=1,\dots,R$, where $c_n = n^{-c}\mathbb{1}(R>1)$.*

Proof. See Section A.2 in Appendix. ■

The second condition is replaced by $\lambda d/\sqrt{\min_{r,j}\mathsf{len}([\Xi_r^*]_j)} \lesssim n^{(c-1)/2}\sqrt{R\log n}$ if we consider the worst-case scenario $\bar\alpha \to 0$ with the upper bound $\bar\alpha/d \le 1$. Combined with the necessary conditions $d/\bar\alpha \ll \log n$ and $\lambda^{\bar\alpha/d}R \ll n$ for consistent estimation (see (7)), the conditions are very mild as soon as c is suitably large. The choice $c=1$ may even be sufficient with stronger boundedness conditions, i.e., $\lambda \lesssim 1$, $d \lesssim \sqrt{\log n}$, and $\min_{r,j}\mathsf{len}([\Xi_r^*]_j) \gtrsim 1$. In particular, the first condition is trivially satisfied if $R=1$, i.e., $\mathfrak{X}_0^* = \{[0,1]^p\}$. In this case, we obtain the strongest result in (i) of Theorem 1 as soon as the second condition is satisfied (recall that $\Gamma_\lambda^{A_{\bar\alpha},d,p}(\mathfrak{X}_0) = \Gamma_\lambda^{A_{\bar\alpha},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$ if $R=1$). If $R>1$, c_n is a decreasing polynomial in n with our choice of b_n. This concludes that, with a suitably large c, the assertions in (ii) and (iv) of Theorem 1 (or the assertions in (i) and (ii) of Corollary 1) hold. Note that (iii) of Theorem 1 also holds trivially with this $\mathcal{Z}$.

As $\max_{1\le j\le p}\log b_j(\mathcal{Z}) = p^{-1}\log b_n \lesssim \log n$, a regular grid satisfies the condition for the optimal posterior contraction specified in Section 5 (see Remark 6). This makes a regular grid very appealing for practical use given its simplicity, and there is little benefit of considering more complicated split-nets. The only exception is a set of fixed design points commonly used in the literature of BART (Chipman et al., 2010; Ročková and van der Pas, 2020).

A regular grid can easily be extended to an irregular rectangular grid with boxes of different sizes. If every mesh-size of an irregular checkerboard is asymptotically proportional to $1/b_n^{1/p}$, the above results still hold with minor modification. This extension is particularly interesting in a regression setup where the distribution of covariates is explicitly available. For example, it allows us to use the quantiles for grid points, which is a natural way to generate a weakly balanced system (Castillo and Ročková, 2021).

Remark 9. Lemma 1 indicates that a large value of c is preferred in the sense of making the required conditions mild. Furthermore, a large c does not harm the posterior contraction rate, as the boundedness condition $\max_{1\le j\le p}\log b_j(\mathcal{Z}) \lesssim \log n$ is satisfied for any $c > 0$. Nonetheless, the empirical performance is affected by the size of c; an extremely large c produces unnecessarily many split-candidates, making the algorithm inefficient. Consequently, we want to choose a suitable but not extremely large c. A good choice of c is model-specific. In Section 6.2, we will see that density estimation requires approximation with respect to the L_∞-norm, which can be fulfilled by (i) of Corollary 1 with the continuity assumption on f_0. If $\lambda \lesssim 1$, $d \lesssim \sqrt{\log n}$, $\min_{r,j} \mathsf{len}([\Xi_r^*]_j) \gtrsim 1$, and $\min_{r,j} \alpha_{rj} > 1/3$, then $c = 1$ and the corresponding c_n satisfy the requirements for (i) of Corollary 1 and Lemma 1. The most disappointing assumption is the lower bound for the minimum smoothness parameter, $\min_{r,j} \alpha_{rj} > 1/3$. Although we recommend $c = 1$ as the default choice by assuming such requirements, increasing c is recommended if the density function is thought to be less smooth.[7] In contrast, nonparametric regression with random design and binary classification require approximation with respect to the L_2-norm (see Sections 6.1 and 6.3), which is obtained by (ii) of Corollary 1. One can easily verify that, if $\lambda \lesssim 1$, $d \lesssim \sqrt{\log n}$, and $\min_{r,j} \mathsf{len}([\Xi_r^*]_j) \gtrsim 1$, then $c = 1$ and the corresponding c_n satisfy the conditions for (ii) of Corollary 1 and Lemma 1, and hence $c = 1$ is the default choice.

4.3.2 Fixed Design Points

Now we focus on a fixed design regression setup, where observed covariate values are readily available. In this case, using fixed design points is particularly appealing in that (iii) of Theorem 1 (coupled with this split-net) gives an approximation error relative to the empirical probability measure as soon as it is suitably regular (the assertion does not require a further bound on c_n). The strategy is conventional in the literature of Bayesian CART and BART (Chipman et al., 1998; Denison et al., 1998; Chipman et al., 2010).

Suppose that a split-net $\mathcal{Z} = \{z_i \in [0,1]^p,\ i = 1,\dots,n\}$ consists of the observed covariate values in a regression setup. We need to assume that the design points are sufficiently evenly distributed in S_0. The required assumption is formalized as follows.

(F) For every $\alpha \in (0,1]^d$ and every box $\Psi \subseteq [0,1]^p$ with $nP_{\mathcal{Z}}(\Psi) \gg 1$, $\mathcal{Z}$ is (Ψ, α, L, S_0)-regular with $L = \lfloor \log_2(cnP_{\mathcal{Z}}(\Psi)) \rfloor$ for some constant $c > 0$.

[7] A careful examination of the proof indicates that the isotropy assumption eliminates the condition $\min_{r,j} \alpha_{rj} > 1/3$, so $c = 1$ works for all smoothness levels. This is because isotropy causes $\min_{r,j} \alpha_{rj} = \bar{\alpha}$, and there is enough cancellation in simplifying (iv) of Theorem 1. To maintain anisotropy throughout the paper, we do not investigate such a particular situation in greater detail.

Although assumption (F) may appear nontrivial, it is actually not restrictive. As $P_{\mathcal{Z}}$ is defined as $P_{\mathcal{Z}}(\cdot) = b_n^{-1}\sum_{i=1}^{b_n}\delta_{z_i}(\cdot)$, for $\mathcal{Z}$ chosen above, $nP_{\mathcal{Z}}(\Psi)$ denotes the number of split-candidates contained in Ψ. Hence, the condition $nP_{\mathcal{Z}}(\Psi) \gg 1$ implies that the number of design points in Ψ increases with n, which is a certainly mild assumption. As noted in Remark 5, if $\mathcal{Z}$ is balanced very well in S_0 and there are no ties so that splits can occur $nP_{\mathcal{Z}}(\Psi)$ times, then $\mathcal{Z}$ is (Ψ, α, L, S_0)-regular for $L = \lfloor \log_2(nP_{\mathcal{Z}}(\Psi)+1)\rfloor$. Our requirement in (F) is milder with the aid of the constant c.

Lemma 2 (Fixed design points)**.** *Consider fixed design points $\mathcal{Z} = \{z_i, i = 1, \dots, n\}$ satisfying assumption* (F)*. If $\lambda d \lesssim (n/R)^{\bar{\alpha}/d}\sqrt{\log n}$, $\min_r P_{\mathcal{Z}}(\Xi_r^*) \gtrsim R^{-1}$, and $R \ll n$, then $\mathcal{Z}$ is $(\Xi_r^*, \alpha_r, L_0, S_0)$-regular for $r = 1, \dots, R$.*

Proof. See Section A.2 in Appendix. ■

As $nP_{\mathcal{Z}}(\Xi_r^*)$ is the number of split-candidates in Ξ_r^*, the condition $\min_r P_{\mathcal{Z}}(\Xi_r^*) \gtrsim R^{-1}$ implies that the number of split-candidates should be balanced well among the R boxes. Our condition $\lambda d \lesssim (n/R)^{\bar{\alpha}/d}\sqrt{\log n}$ slightly relaxes the condition $\lambda d \lesssim \sqrt{\log n}$ of Theorem 4.1 in Ročková and van der Pas (2020) (for the case of global isotropy). The latter is obtained if we consider the worst-case scenario $\bar{\alpha} \to 0$. We see that (iii) of Theorem 1 directly follows from this lemma. As the design points are used as $\mathcal{Z}$, the term $\|f_0 - \hat{f}_0\|_{v,P_{\mathcal{Z}}}$ is translated into the approximation error relative to the empirical probability measure. In regression setups, this fact makes fixed design points much more attractive than other split-nets in the previous sections. We also note that the requirement $\max_{1\le j\le p}\log b_j(\mathcal{Z}) \lesssim \log n$ for the optimal posterior contraction is trivially satisfied.

5. BART in Nonparametric Regression

5.1 Posterior Contraction Rates

BART is an archetypal example of Bayesian forests (Chipman et al., 1998; Denison et al., 1998; Chipman et al., 2010). For a fixed design Gaussian nonparametric regression, Ročková and van der Pas (2020) and Ročková and Saha (2019) established L_2 rate-optimal posterior contraction of BART for high-dimensional isotropic regression functions. Our investigation goes beyond these studies in three aspects: (i) we treat the variance parameter σ^2 as unknown with a prior; (ii) we consider both fixed and random regression design; and, most importantly, (iii) the true function is assumed to be in the piecewise heterogeneous anisotropic space introduced earlier. The last point significantly enlarges the optimality scope of BART.

We separately deal with fixed and random designs. This section is focused on the fixed design case, while the random design case will be considered in Section 6.1. The fixed design regression model writes as

$$Y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_0^2), \quad i = 1, \dots, n, \tag{6}$$

where $x_i = (x_{i1}, \dots, x_{ip})^\top \in [0,1]^p$, $i = 1, \dots, n$, are fixed. The model is independent but not identically distributed, and hence the asymptotic studies are established under the product measure for the n observations. The general theory of posterior contraction requires an exponentially powerful test function of a semimetric under this product measure (Ghosal and van der Vaart, 2017). In nonparametric regression with fixed design, such a good test function can be directly constructed for the empirical L_2-distance even when the noise error is unknown (Ning et al., 2020; Jeong and Ghosal, 2021b; Lim and Jeong, 2023). The general theory also requires desirable properties of the prior. We show that the tree priors in Section 3 satisfy those conditions.

We impose the following assumptions on the true parameters f_0 and σ_0^2.

(A1) For $d > 0$, $\lambda > 0$, $R > 0$, $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$, and $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ with $\bar{\alpha} \in (0, 1]$, the true function satisfies $f_0 \in \Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$ or $f_0 \in \Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0, 1]^p)$.

(A2) It is assumed that d, p, λ, R, and $\bar{\alpha}$ satisfy $\epsilon_n \ll 1$, where

$$\epsilon_n = \sqrt{\frac{d \log p}{n}} + (\lambda d)^{d/(2\bar{\alpha}+d)} \left(\frac{R \log n}{n}\right)^{\bar{\alpha}/(2\bar{\alpha}+d)}. \tag{7}$$

(A3) The true function satisfies $\|f_0\|_\infty \lesssim \sqrt{\log n}$.

(A4) The true variance parameter satisfies $\sigma_0^2 \in [C_0^{-1}, C_0]$ for a sufficiently large $C_0 > 1$.

Assumption (A1) means that the true regression function f_0 lies on a sparse piecewise heterogeneous anisotropic space. If the continuity assumption is further imposed, the approximation results in Theorem 1 are obtained under milder conditions. Assumption (A2) is required to make our target rate ϵ_n tend zero. The boundedness condition in (A3) is made to guarantee a sufficient prior concentration under the normal prior on the step-heights specified in (P2) below. Although the Gaussian prior can be replaced by a thick-tailed prior (e.g., Ročková, 2020), we only consider the Gaussian prior to leverage its semi-conjugacy. Assumption (A4) allows one to assign a standard prior to σ^2, e.g., an inverse gamma distribution.

It is also important to choose a suitable split-net so that Theorem 1 can be deployed. For regression with fixed design, we need an approximation result with respect to the empirical L_2-norm $\|\cdot\|_n$ defined as $\|f\|_n^2 = n^{-1} \sum_{i=1}^n |f(x_i)|^2$. We make the following assumptions on the split-net $\mathcal{Z}$. The notation dep means the depth of a node, the number of nodes along the path from the root node down to that node.

(A5) The split-net $\mathcal{Z}$ satisfies $\max_{1 \le j \le p} \log b_j(\mathcal{Z}) \lesssim \log n$.

(A6) The split-net $\mathcal{Z}$ is suitably dense and regular to construct a $\mathcal{Z}$-tree partition $\widehat{\mathcal{T}}$ such that there exists $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ satisfying $\|f_0 - \hat{f}_0\|_n \lesssim \bar{\epsilon}_n$ by Theorem 1.

(A7) The $\mathcal{Z}$-tree partition $\mathcal{T}^* = \{\Omega_1^*, \dots, \Omega_R^*\}$ approximating $\mathfrak{X}_0^*$ satisfies $\max_r \mathsf{dep}(\Omega_r^*) \lesssim \log n$.

Assumption (A5) is required for a suitable bound of the entropy and a good prior concentration (see Lemma 4). Assumption (A6) provides the desired approximation error with respect to the $\|\cdot\|_n$-distance. Owing to (iii) of Theorem 1 and Lemma 2, using fixed design points as $\mathcal{Z}$ is of particular interest, as $\|\cdot\|_{2,P_{\mathcal{Z}}}$ is equivalent to the empirical L_2-norm $\|\cdot\|_n$ in this case. Assumption (A7) is a technical requirement which is certainly mild. This condition is trivially satisfied if R is bounded.

Lastly, careful prior specification is required to obtain the optimal posterior contraction. We consider the following prior distributions discussed in Section 3.

(P1) For a fixed $T > 0$, each tree $\mathcal{T}^t$, $t = 1, \dots, T$, is independently assigned a tree prior with Dirichlet sparsity.

(P2) The step-heights B are assigned a normal prior with a zero-mean and a covariance matrix whose eigenvalues are bounded below and above.

(P3) The variance parameter σ^2 is assigned an inverse gamma prior.

Under the above assumptions and priors, the following theorem formalizes the posterior contraction rate of model (6).

Theorem 2 (Nonparametric regression, fixed design)**.** *Consider model* (6) *with Assumptions* (A1)–(A7) *and the prior assigned through* (P1)–(P3)*. Then, there exists a constant* $M > 0$ *such that for* ϵ_n *in* (7)*,*

$$\mathbb{E}_0 \Pi \Big\{ (f, \sigma^2) : \|f - f_0\|_n + |\sigma^2 - \sigma_0^2| > M\epsilon_n \,\Big|\, Y_1, \dots, Y_n \Big\} \to 0.$$

Proof. See Section A.3 in Appendix. ■

Intuitively, the rate in (7) resembles a near-minimax rate of estimation of high-dimensional anisotropic functions. The first part in (7) is the near-minimax risk of the penalty for not knowing the subset S_0 (Raskutti et al., 2011). The second part in (7) is incurred by anisotropic regression function estimation. Although λ and R can be a polynomial in n with a suitably small power to satisfy $\epsilon_n \to 0$, a particularly interesting case is when both are at most $\log^c n$ for some $c > 0$. The second term then corresponds to the near-minimax rate of anisotropic function estimation (Hoffman and Lepski, 2002). Whether the rate in (7) is in fact the actual (near) minimax rate remains to be established. The answer to this question is provided in the following subsection, where we formally derive the minimax lower bound with respect to the L_2-risk.

Remark 10. In isotropic regression using BART, Ročková and van der Pas (2020) assumed that the first part of the rate in (7) is dominated by the second part, whereby the resulting rate is simplified such that it only depends on the risk of function estimation. As this restriction is not required, we keep the rate in the form of (7).

5.2 Minimax Lower Bound

In Section 5.1, we established the posterior contraction rate of BART under relaxed smoothness assumptions. Although the rate in (7) consists of two logical components (a penalty for variable selection uncertainty and a rate of anisotropic function estimation), it is not guaranteed that the *whole rate* is (nearly) minimax optimal. While the minimax rates in high-dimensional *isotropic* function estimation were studied exhaustively in Yang and Tokdar (2015), extensions to (piecewise) *anisotropic* functions *have not* been obtained in the literature. We fill this gap by deriving a minimax lower bound in our general smoothness setup. These results will certify that the rates obtained in Section 5.1 are indeed minimax optimal (with respect to the L_2-risk) up to a logarithmic factor.

To deploy the conventional minimax theory, we consider the model with random design given by

$$Y_i = f_0(X_i) + \varepsilon_i, \quad X_i \sim Q, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_0^2), \quad i = 1, \dots, n, \tag{8}$$

where $X_i = (X_{i1}, \dots, X_{ip})$, $i = 1, \dots, n$, are p-dimensional random covariates and Q is a probability measure such that $\mathrm{supp}(Q) \subseteq [0,1]^p$. We assume (without loss of generality) that σ_0^2 is fixed to 1. To obtain a lower bound of the minimax rate, we use the Le Cam equation (Birgé and Massart, 1993; Wong and Shen, 1995; Barron et al., 1999). Now the density q of Q is assumed to satisfy the following assumption under which the $L_2(Q)$-norm is replaced by the L_2-norm.

(M) There exist constants $0 < \underline{q} \le \overline{q} \le \infty$ such that the density q satisfies $\underline{q} \le \inf_x q(x) \le \sup_x q(x) \le \overline{q}$.

We define the L_2-minimax risk for any function space $\mathcal{F} \in \mathcal{L}_2$ as

$$r_n^2(\mathcal{F}) = \inf_{\hat{f} \in \mathcal{B}_n} \sup_{f_0 \in \mathcal{F}} \mathbb{E}_{f_0,Q} \|\hat{f} - f_0\|_2^2, \tag{9}$$

where $\mathcal{B}_n$ is the space of all $\mathcal{L}_2$-measurable function estimators and $\mathbb{E}_{f,Q}$ is the expectation operator under the model with f and Q. The Le Cam equation requires suitable upper and lower bounds of the metric entropy of the target function space. We thus define the bounded function space $\overline{\Gamma}_{\lambda,M}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) = \{f \in \Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) : \|f\|_\infty \le M\lambda\}$ for any $M > 0$. As our contraction rate is the same for both $\Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$ and $\Gamma_{\lambda}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$, we aim to construct a lower bound of $r_n\big(\overline{\Gamma}_{\lambda,M}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)\big)$ close enough to ϵ_n.

Theorem 3 (Minimax lower bound). *Consider model* (8) *for* $\sigma_0^2 = 1$ *with Assumption* (M). *For* $d > 0$, $\lambda > 0$, $R > 0$, *a partition* $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$ *of* $[0,1]^d$, *and a smoothness parameter* $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ *for* $\bar{\alpha} \in (0,1]$ *such that* $\log \mathsf{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$, $1 \le r \le R$, $1 \le j \le d$, *there exists* $M_d > 0$ *depending only on* d *such that*

$$r_n\big(\overline{\Gamma}_{\lambda,M}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)\big) \gtrsim \sqrt{\frac{1}{n} \log \binom{p}{d}} + M_d \left(\frac{\lambda^{d/\bar{\alpha}}}{n}\right)^{\bar{\alpha}/(2\bar{\alpha}+d)}.$$

Proof. See Section A.4 in Appendix. ■

As M_d can be dependent on d, the correct interpretation of the result is with a bounded d. Also, our contraction rate ϵ_n is derived under the condition $\|f_0\|_\infty \lesssim \sqrt{\log n}$, and hence we assume that $\lambda \lesssim \sqrt{\log n}$ to match the two spaces. One can easily verify that the condition $\log \mathsf{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$, $1 \le r \le R$, $1 \le j \le d$, leads to the restriction $\log R \lesssim d/\bar{\alpha}$, which removes the term R from our rate ϵ_n in (7). Putting the bounds together, ϵ_n matches the lower bound up to a logarithmic factor.

5.3 Numerical Study

In this section, we conduct a numerical study that shows the successful performance of BART with a variety of multivariate functions. For competitors we consider Gaussian process (GP) prior regression, gradient boosting (GB), random forest (RF), and neural network (NN) models with the rectified linear unit (ReLU) activation function. GP prior regression is widely exploited for multiple nonparametric regression and ensures theoretical optimality for smooth functions (van der Vaart and van Zanten, 2008). GB is expected to work similarly to BART. RF is expected to satisfactorily detect discontinuous boundaries along the coordinates, as it is based on the additive tree ensembles. We know that NN models adapt well to complicated function classes with the guaranteed optimal properties (e.g., Petersen and Voigtlaender, 2018; Imaizumi and Fukumizu, 2019; Schmidt-Hieber, 2020; Hayakawa and Suzuki, 2020). Our numerical study shows that BART outperforms these competitors in adapting to complicated smoothness structures.

Our synthetic datasets are generated from model (6) with a few different functions $f_0 : [0,1]^p \to \mathbb{R}$. To specify the simulation setups, we first introduce the following functions that maps $[0,1]^p$ to $\mathbb{R}$:

$$\begin{aligned}
\mathsf{base}_p : (x_1,\dots,x_p) &\mapsto \sin\left(\frac{10}{\sqrt{p}}\left\{\sum_{j=1}^p (x_j-0.5)^2 - \frac{p}{12}\right\}\right),\\
\mathsf{discont1} : (x_1,\dots,x_p) &\mapsto \mathbb{1}(x_1 \le 0.5, x_2 > 0.5) + \mathbb{1}(x_1 > 0.5, x_2 \le 0.5),\\
\mathsf{discont2}_p : (x_1,\dots,x_p) &\mapsto \mathbb{1}\Bigg(\sum_{j=1}^p (x_j - 0.5) \le 0, \sum_{j=1}^p (-1)^j (x_j-0.5) > 0\Bigg)\\
&\qquad + \mathbb{1}\Bigg(\sum_{j=1}^p (x_j - 0.5) > 0, \sum_{j=1}^p (-1)^j (x_j-0.5) \le 0\Bigg).
\end{aligned}$$

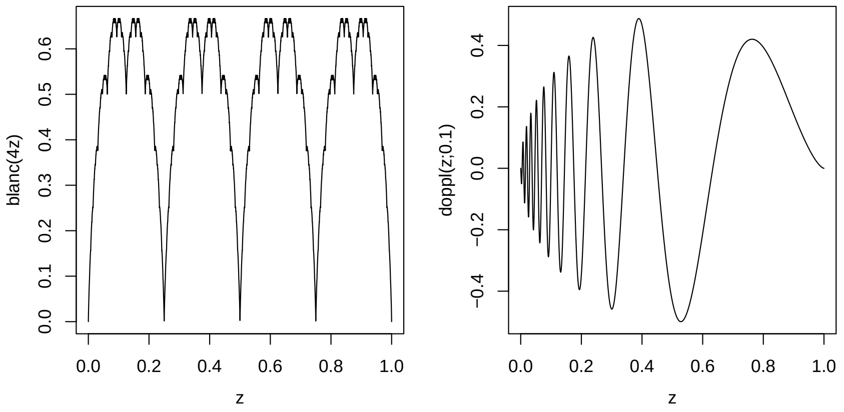
The function base_p is viewed as having an isotropic smoothness and is used as the base component for f_0.[8] The functions $\mathsf{discont1}$ and $\mathsf{discont2}_p$ render discontinuous jumps along hyperplanes in different directions. To account for non-Lipschitz continuity and spatially varying smoothness, we also define the blancmange function and the Doppler function as,

$$\begin{aligned}
\mathsf{blanc}(z) &= \sum_{k=0}^{\infty} \frac{|2^k z - \lfloor 2^k z + 0.5\rfloor|}{2^k}, \quad z \in [0,1],\\
\mathsf{doppl}(z;a) &= \sqrt{z(1-z)}\sin\left(\frac{2\pi(1+a)}{z+a}\right), \quad z \in [0,1],
\end{aligned}$$

which are illustrated in Figure 9.

Using the above functions, we describe six simulation scenarios. Specifically, Scenario k is defined by model (6) with $f_0 = f_0^{(k)}$, $k = 1,\dots,6$, where the true functions $f_0^{(k)} : [0,1]^p \to$

[8] The argument of the sine function is chosen so that it is centered at zero and has a reasonable scale for every p, allowing the period of the sine function to be roughly maintained with p. In particular, if X_j has a uniform distribution on $[0,1]$ independently, one can easily see that $(10/\sqrt{p})\{\sum_{j=1}^p (X_j - 0.5)^2 - p/12\}$ weakly converges to $\mathrm{N}(0,5/9)$ as $p \to \infty$.



(a) Blancmange function $\mathsf{blanc}(4z)$ (b) Doppler function $\mathsf{doppl}(z; 0.1)$

Figure 9: Functions that exhibit non-Lipschitz continuity and spatially varying smoothness.

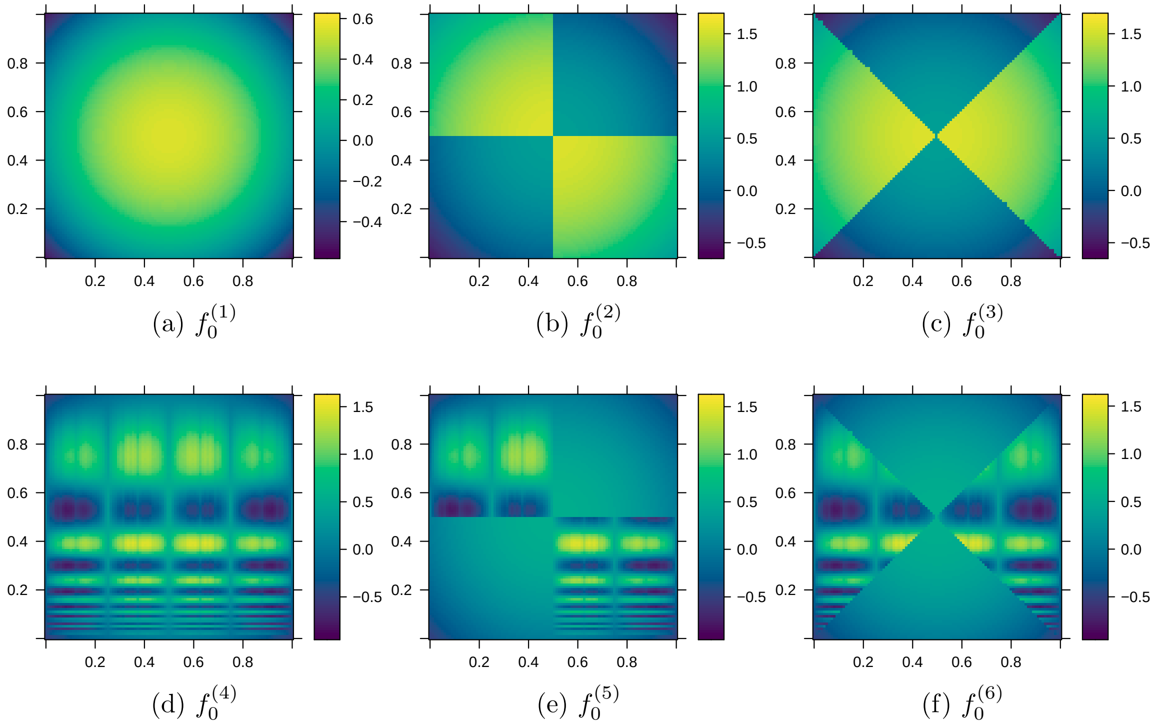
$\mathbb{R}$ are defined as

$$
\begin{aligned}
&f_0^{(1)} : (x_1, \ldots, x_p) \mapsto \mathsf{base}_p(x_1, \ldots, x_p),\\
&f_0^{(2)} : (x_1, \ldots, x_p) \mapsto \mathsf{base}_p(x_1, \ldots, x_p) + \mathsf{discont1}(x_1, \ldots, x_p),\\
&f_0^{(3)} : (x_1, \ldots, x_p) \mapsto \mathsf{base}_p(x_1, \ldots, x_p) + \mathsf{discont2}_p(x_1, \ldots, x_p),\\
&f_0^{(4)} : (x_1, \ldots, x_p) \mapsto \mathsf{base}_p(x_1, \ldots, x_p) + 3\mathsf{blanc}(4x_1)\mathsf{doppl}(x_2; 0.1),\\
&f_0^{(5)} : (x_1, \ldots, x_p) \mapsto \mathsf{base}_p(x_1, \ldots, x_p) + 3\mathsf{blanc}(4x_1)\mathsf{doppl}(x_2; 0.1)\mathsf{discont1}(x_1, \ldots, x_p),\\
&f_0^{(6)} : (x_1, \ldots, x_p) \mapsto \mathsf{base}_p(x_1, \ldots, x_p) + 3\mathsf{blanc}(4x_1)\mathsf{doppl}(x_2; 0.1)\mathsf{discont2}_p(x_1, \ldots, x_p).
\end{aligned}
$$

The functions $f_0^{(1)}$ and $f_0^{(4)}$ represent globally isotropic and anisotropic functions, respectively. The other functions produce discontinuous jumps that are either parallel or oblique to the coordinate system. Specifically, $f_0^{(2)}$ and $f_0^{(5)}$ are regarded as piecewise isotropic and anisotropic functions, respectively, as defined in Definition 2. The remaining functions $f_0^{(3)}$ and $f_0^{(6)}$ are similarly piecewise isotropic and anisotropic, but they differ from Definition 2 in that the jumps are not parallel to the coordinates. The two-dimensional case of each $f_0^{(k)}$ is visualized in Figure 10.

We generate the synthetic datasets under Scenarios 1–6. For each scenario, we consider two sample sizes $n \in \{1000, 5000\}$ and five dimension values $p \in \{2, 5, 10, 20, 50\}$, while fixing $\sigma_0^2 = 0.5^2$ for reasonable signal to noise ratios. Therefore, each scenario has 10 synthetic datasets generated with all possible combinations of n and p. For given predictor variables X_i generated uniformly on $[0, 1]^p$, the response variable Y_i is generated from model (6), $i = 1, \ldots, n$.

All datasets are fitted by BART and the other competitors. For a fair comparison to the other methods, we do not use the Dirichlet sparse prior in (3) for BART. Instead, we assign a uniform prior that corresponds to the Dirichlet prior with concentration parameter 1, with a priori assumption that all predictor variables contribute equally to the observations.



(a) $f_0^{(1)}$ (b) $f_0^{(2)}$ (c) $f_0^{(3)}$

(d) $f_0^{(4)}$ (e) $f_0^{(5)}$ (f) $f_0^{(6)}$

Figure 10: Level plots of the true functions $f_0^{(k)}$, $k = 1, \dots, 6$, for $p = 2$.

We fit BART with 200 trees using the prior that splits a node at depth ℓ with probability $\alpha(\ell+1)^{-\beta}$ for $\alpha \in (0,1)$ and $\beta \in [0,\infty]$, the original construction by Chipman et al. (2010), which is implemented in the R package BART. However, as our theory resorts to the exponentially decaying prior for splits as mentioned in Section 3.1, we also consider BART with the prior that splits a node at depth ℓ with probability $\nu^{\ell+1}$ for $\nu \in (0,1/2)$. We choose $\alpha = 0.3$, $\beta = 2$, and $\nu = 0.3$ to make the two priors roughly similar for small ℓ. We will see that the two priors exhibit similar empirical behavior. For GP prior regression, the squared exponential covariance kernel $k(x,x') = \tau^2 \exp(\|x - x'\|^2/l^2)$ is employed with half normal priors $\tau \sim \mathrm{N}_+(0,1)$ and $l \sim \mathrm{N}_+(0,1)$. Optimizing other parameters in the posterior distribution, the posterior mode of f_0 is obtained in a closed-form expression (we also tried other informative priors for τ and l and observed no significant difference). GB is trained by the gbm package with trees of five splits and the number of trees determined via cross validation (CV). RF is fitted by the randomForest package with 200 trees and the maximal node size 5 or 50 for each tree. The NN models are trained by TensorFlow with the Keras interface. We consider two NN models with two and four hidden layers with $(64, 32)$ and $(256, 128, 64, 32)$ hidden units. All hidden units take the ReLU activation function with the dropout of rate 0.3 for regularization. The description of the methods is summarized in Table 1.

Figures 11 and 12 show the root mean squared prediction error (RMSPE) obtained by the methods described in Table 1. The RMSPEs are estimated by randomly drawn out-of-samples. For Scenario 1 with the global isotropic function $f_0^{(1)}$, BART, GP regression, and

Table 1: The description of the methods for simulation.

Method	Description
BART1	BART with 200 trees Node at depth ℓ is split with prior probability $\alpha(\ell+1)^{-\beta}$, $\alpha = 0.3$, $\beta = 2$
BART2	BART with 200 trees Node at depth ℓ is split with prior probability $\nu^{\ell+1}$, $\nu = 0.3$
GP	GP prior regression with the squared exponential covariance kernel
GB	GB with trees of five splits and the number of trees determined via CV
RF1	RF of 200 trees with maximal node size 5 for each tree
RF2	RF of 200 trees with maximal node size 50 for each tree
NN1	NN model with two hidden layers and $(64, 32)$ hidden units
NN2	NN model with four hidden layers and $(256, 128, 64, 32)$ hidden units

GB perform similarly well in relatively lower dimensions ($p = 2, 5, 10$), but the performance of GP degrades as p increases. For Scenario 2 with the piecewise isotropic function $f_0^{(2)}$, BART clearly outperforms the other methods as expected. Interestingly, GB performs substantially worse than BART in this situation, implying that BART detects discontinuous jumps along the coordinates better. RF falls behind BART and GB although it is also based on binary tree ensembles. For Scenario 3 with $f_0^{(3)}$, GP and NN perform better than BART and GB in lower dimensions; this makes sense given that BART cannot detect such discontinuous jumps efficiently using the coordinate parallel splitting rule. However, the performance of GP and NN deteriorates as p increases, and BART and GB beat the competition in higher dimensions ($p = 50$). The interpretation of the results is similar for the remaining scenarios. The major difference is that BART produces the best prediction error in almost all cases of Scenarios 4–6. Given that BART is designed to capture local anisotropy very effectively, this finding appears to be a natural consequence. Overall, GB performs slightly worse than BART. As well as the setups used in our simulation, we also tested many other tuning parameter setups and network structures for GB, RF, and NN, but found no clear improvement.

Based on Figures 11 and 12, we can also compare the performance of the two BART priors. BART with the polynomially decaying prior (the original BART prior by Chipman et al. (2010)) works slightly better in lower dimensions ($p = 2, 5$), whereas the exponentially decaying prior is marginally preferred in higher dimensions ($p = 10, 20, 50$). However, because the difference is not significant, we conclude that there are no substantial differences in empirical behavior between the two BART priors.

6. Further Applications

Section 5 establishes the posterior contraction rate of BART for the nonparametric regression model and justifies its near-minimax optimality. As our approximation theory only requires conditions on a split-net, the results can be extended to statistical models beyond nonparametric regression with fixed design. In this section, we consider other applications such as nonparametric regression with random design, density estimation, and nonparamet-

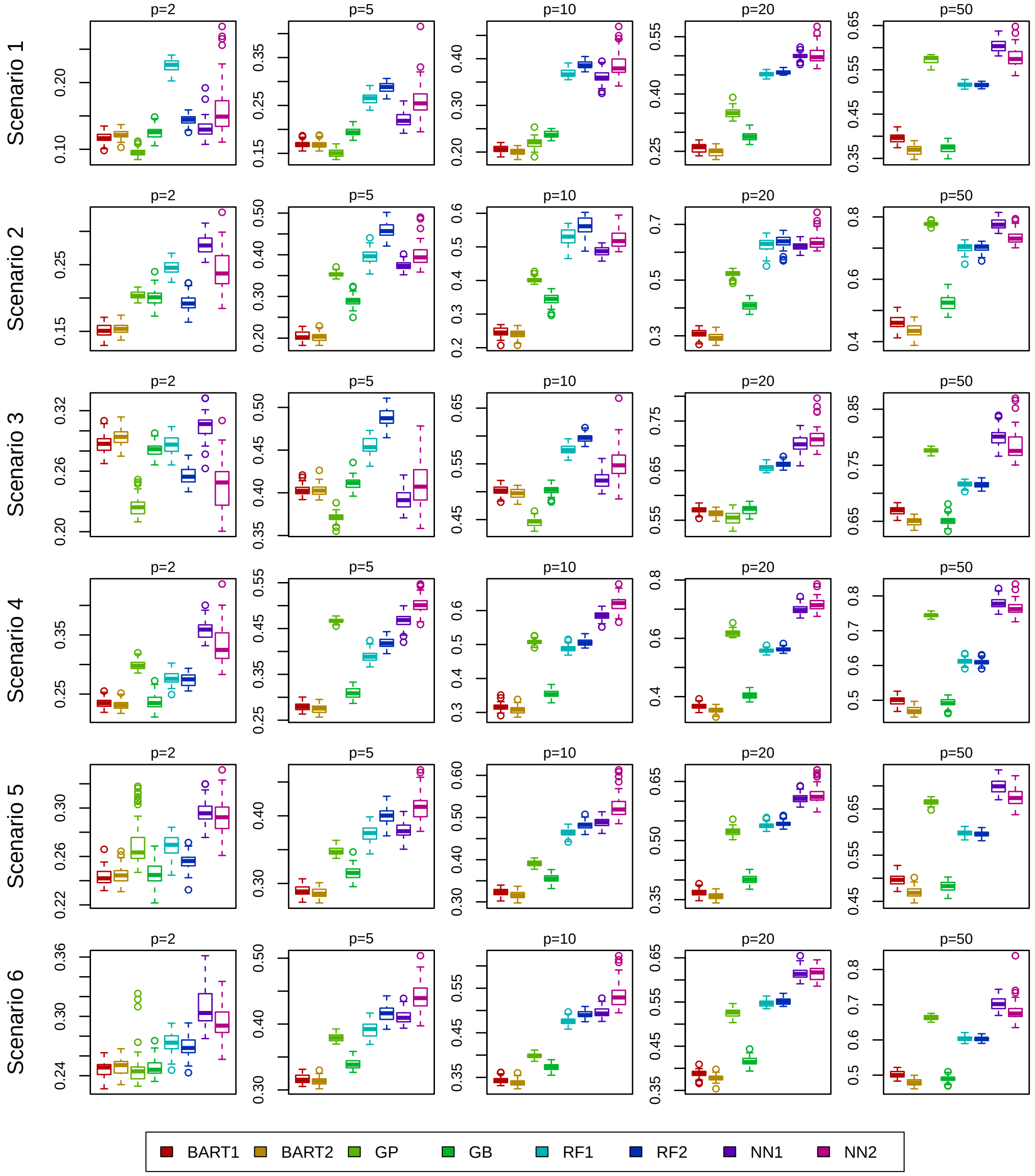


Figure 11: RMSPEs obtained from 50 replicated datasets of size $n = 1000$.

ric binary classification. Moreover, as the technical results in Section 5 hold even with the single tree model ($T = 1$), one can find no theoretical advantages of BART over Bayesian CART. A theoretical advantage of BART can be recognized if the true function has an additive structure (Linero and Yang, 2018; Ročková and van der Pas, 2020). Such an extension is also considered in this section.

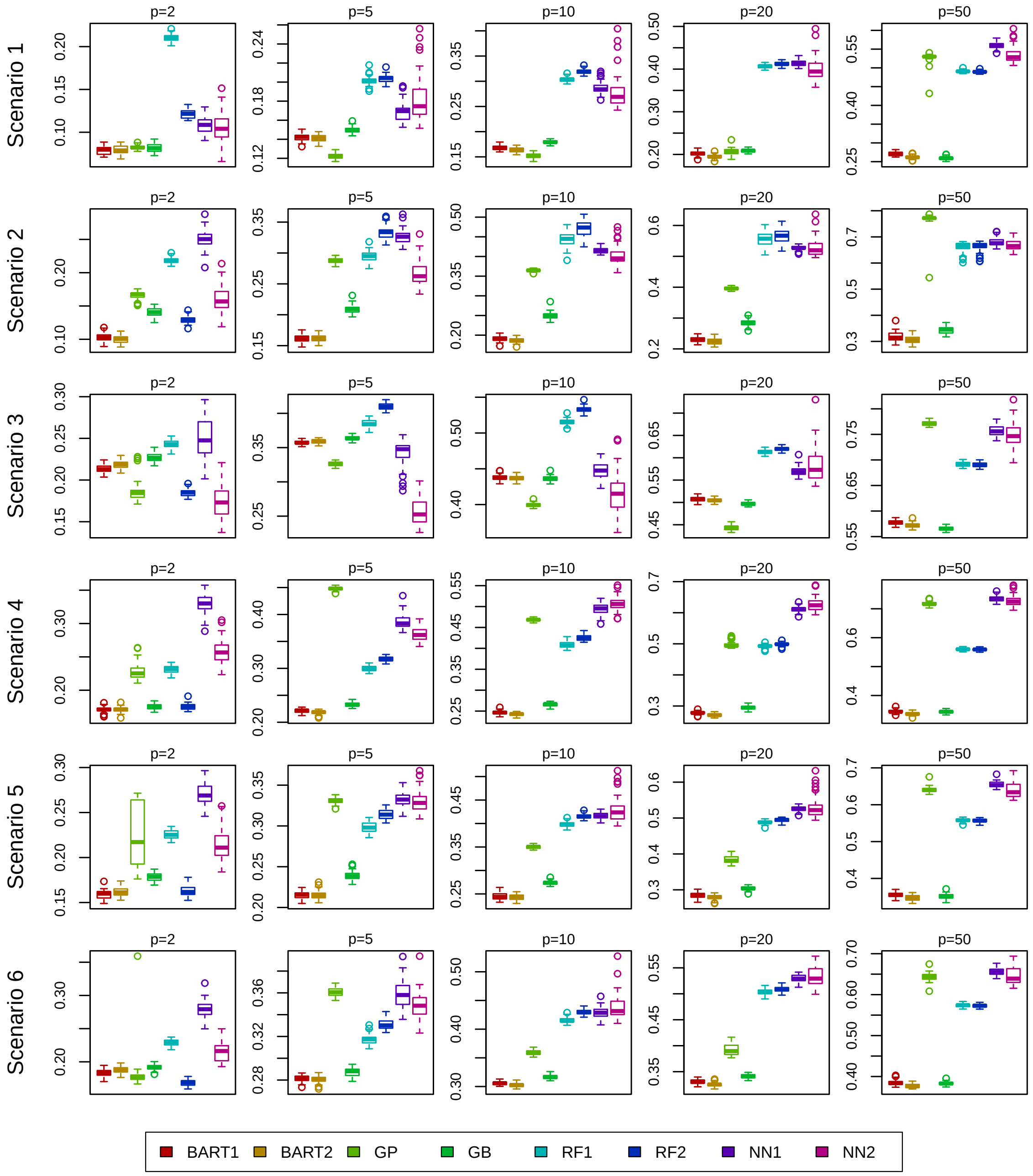


Figure 12: RMSPEs obtained from 50 replicated datasets of size $n = 5000$.

6.1 Nonparametric Regression with Random Design

Theorem 2 quantifies the posterior contraction rate of nonparametric regression with fixed design where the predictor variables are not random variables. Now we consider a random design regression in (8) in which the model is treated as independent and identically distributed. We establish the posterior contraction rate of BART for the random design model in (8). The main advantage of considering random design is that it provides the $L_{2,Q}$-contraction rate without empirical process theory, where Q is a probability measure for X_i, whereas fixed design essentially provides the contraction rate with respect to the empirical L_2-norm as in Section 5. The random design assumption is also often necessary in certain statistical models, for example, in measurement error models (Tuo and Wu, 2015) or causal inference models (Hahn et al., 2020; Ray and van der Vaart, 2020). Note that fixed design points in Section 4.3.2 cannot be used for a split-net, as the procedure is not truly Bayesian if the prior is dependent on the data (X_i is now considered a part of the observation.). Instead, a regular grid in Section 4.3.1 can be useful for this framework.

We consider model (8) for Q a probability measure that satisfies $\operatorname{supp}(Q) \subseteq [0,1]^p$ with a bounded density. Unlike model (6), model (8) is independent and identically distributed. The well-known fact that exponentially powerful tests exist with respect to the Hellinger metric $\rho_{\mathrm{H}}(\cdot,\cdot)$ allows one to establish the contraction rate for the corresponding metric (Ghosal et al., 2000). However, in normal models, the Hellinger distance is matched to the L_2-type metric only when $\|f\|_\infty$ and $|\log\sigma^2|$ are bounded in the entire parameter space, not only for the true values (e.g., Xie and Xu, 2018). Unlike in Theorem 2, this restriction requires that f_0 be uniformly bounded and a prior be appropriately truncated. Note also that we need a good approximation error with respect to the integrated L_2-norm. We summarize the required modifications of (A3), (A6), (P2), and (P3).

(A3*) The true function f_0 satisfies $\|f_0\|_\infty \le C_0^*$ for some sufficiently large $C_0^* > 0$.

(A6*) The split-net $\mathcal{Z}$ is suitably dense and regular to construct a $\mathcal{Z}$-tree partition $\widehat{\mathcal{T}}$ such that there exists $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ satisfying $\|f_0 - \hat{f}_0\|_2 \lesssim \bar{\epsilon}_n$ by Theorem 1.

(P2*) A prior on the compact support $[-\overline{C}_1, \overline{C}_1]$ is assigned to the step-heights B for some $\overline{C}_1 > C_0^*$.

(P3*) A prior on the compact support $[\overline{C}_2^{-1}, \overline{C}_2]$ is assigned to σ^2 for some $\overline{C}_2 > C_0$.

Assumption (A6*) requires good approximability with respect to the L_2-norm. Owing to (ii) of Corollary 1 and Lemma 1, a regular grid in Section 4.3.1 can be useful to meet this requirement (see Remark 9). We wrap up this section with a theorem that formalizes the posterior contraction of BART for model (8).

Theorem 4 (Nonparametric regression, random design)**.** *Consider model* (8) *with Assumptions* (A1)*,* (A2)*,* (A3*)*,* (A4)*,* (A5)*,* (A6*)*, and* (A7)*, and the prior assigned through* (P1)*,* (P2*)*, and* (P3*)*. Then, there exists a constant $M > 0$ such that for ϵ_n in* (7)*,*

$$\mathbb{E}_0\Pi\Big\{(f,\sigma^2) : \|f - f_0\|_{2,Q} + |\sigma^2 - \sigma_0^2| > M\epsilon_n \,\Big|\, (X_1,Y_1),\ldots,(X_n,Y_n)\Big\} \to 0.$$

Proof. See Section A.5 in Appendix. ■

6.2 Density Estimation

In addition to classical nonparametric regression, density estimation is an interesting branch of nonparametric inference. There exist a few studies employing the Bayesian tree ensembles for density regression (Orlandi et al., 2021; Li et al., 2022). Here we consider a more traditional density estimation problem. With the Bayesian tree ensembles, we only provide a theoretical flavor for density estimation rather than practical implementation. It may be difficult to develop an efficient algorithm for the setup considered here.

For some probability measure P that satisfies $\mathrm{supp}(P) \subseteq [0,1]^p$, suppose n independent observations X_i, $i = 1, \dots, n$, are drawn from P, i.e.,

$$X_i \sim P, \quad i = 1, \dots, n. \tag{10}$$

Assume that P is absolutely continuous with respect to the Lebesgue measure μ with the true density p_0. We assign a prior on p_f indexed by f such that $p_f = e^f / \int_{[0,1]^p} e^f d\mu$ with f assigned the forest priors in Section 3. We write $f_0 = \log p_0$ while assuming (A1)–(A3). That is, our d-sparsity for density estimation implies that the remaining $p - d$ variables are independent and uniformly distributed on $[0,1]^{p-d}$. This sparsity setup is useful in high dimensions because the density cannot be estimated effectively without a stronger assumption for $p > n$, such as isotropy. A similar sparsity structure was also imposed in Liu et al. (2007) for high-dimensional density estimation. We leverage the existence of an exponentially powerful test for the Hellinger metric $\rho_{\mathrm{H}}(\cdot,\cdot)$. Owing to the relationship between Hellinger balls and L_∞ balls in density estimation with the exponential link, we need an approximation result with respect to the L_∞-norm. This is obtained by (iv) of Theorem 1 with the continuity restriction on the true function. As (i) of Corollary 1 and Lemma 1 show, a regular grid in Section 4.3.1 is useful to obtain the L_∞-approximation (see Remark 9). We make the following assumptions to satisfy this requirement.

- (A1‡) For $d > 0$, $\lambda > 0$, $R > 0$, $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$, and $A_{\bar\alpha} \in \mathcal{A}_{\bar\alpha}^{R,d}$ with $\bar\alpha \in (0,1]$, the true function satisfies $f_0 \in \Gamma_\lambda^{A_{\bar\alpha},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$.
- (A6‡) The split-net $\mathcal{Z}$ is suitably dense and regular to construct a $\mathcal{Z}$-tree partition $\widehat{\mathcal{T}}$ such that there exists $\hat f_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ satisfying $\|f_0 - \hat f_0\|_\infty \lesssim \bar\epsilon_n$ by Theorem 1.

We assign the tree prior with Dirichlet sparsity and a normal prior on the step-heights. Under suitable assumptions, the following theorem provides the posterior contraction rate for p_f with respect to the Hellinger distance.

Theorem 5 (Density estimation)**.** *Consider model* (10) *with Assumptions* (A1‡), (A2)–(A3)*,* (A5)*,* (A6‡)*, and* (A7)*, and the prior assigned through* (P1)–(P2)*. Then, there exists a constant* $M > 0$ *such that for* ϵ_n *in* (7)*,*

$$\mathbb{E}_0 \Pi\Big\{ f : \rho_{\mathrm{H}}(p_f, p_0) > M\epsilon_n \,\Big|\, X_1, \dots, X_n \Big\} \to 0.$$

Proof. See Section A.5 in Appendix. ■

As mentioned in Section 5.1, the normal prior in (P2) is not necessary and a heavy-tailed prior can relax the assumption on $\|f\|_\infty$. As normal priors are not conjugate to the model likelihood in the density estimation example, there is no clear benefit of adopting (P2) anymore. This is also the case in the example of binary classification given in the next subsection. Nevertheless, we employ (P2) for the sake of simplicity.

Remark 11. As previously stated, the practical implementation of the density estimation problem here is not as straightforward as the Gaussian regression case. We do not believe that there is a highly efficient algorithm for density estimation with the type of Bayesian forest considered here. One possible option is employing the idea of reversible jump moves (Green, 1995), as in Linero (2022) for generalized BART for exponential family models.

6.3 Nonparametric Binary Classification

Nonparametric classification is useful for modeling categorical response variables. In the original work by Chipman et al. (2010), BART for Gaussian regression was readily adapted to probit regression using the latent variable expression (Albert and Chib, 1993). Later, Kindo et al. (2016) devised a BART algorithm for multi-category response variables using multinomial probit models. Although the probit models are particularly simple to implement, we consider nonparametric binary classification with the logistic link function to make use of the classical theory (van der Vaart and van Zanten, 2008). The computation is still straightforward owing to the latent variable expression with a Pólya-gamma distribution (Polson et al., 2013).

For a binary response $Y_i \in \{0,1\}$ and a random covariate $X_i \in \mathbb{R}^p$, assume that we have n independent observations $(X_1,Y_1),\dots(X_n,Y_n)$ from the binary classification model,

$$\mathbb{E}_0[\mathbb{1}(Y_i=1)|X_i=x]=\varphi_0(x), \quad X_i \sim Q, \quad i=1,\dots,n, \tag{11}$$

for some $\varphi_0 : [0,1]^p \to [0,1]$ and some probability measure Q such that $\operatorname{supp}(Q) \subseteq [0,1]^p$ with a bounded density. We thus consider a binary classification problem with random design. We parameterize the probability function using the logistic link function $H : \mathbb{R} \to [0,1]$ such that $\varphi_f = H(f)$ for f on which the forest priors in Section 3 are assigned. For true function φ_0, we write $f_0 = H^{-1}(\varphi_0)$ while assuming (A1)–(A3) as in the density estimation problem. The proof shows that the Hellinger metric is bounded by the $L_2(Q)$-distance in this example, and hence (A6*) is assumed. Similar to Section 6.1, fixed design points are not available for a split-net, but a regular grid in Section 4.3.1 can be useful. The following theorem formalizes the posterior contraction rate with respect to the $L_2(Q)$-distance.

Theorem 6 (Binary classification)**.** *Consider model* (11) *with Assumptions* (A1)–(A3), (A5)*,* (A6*)*, and* (A7)*, and the prior assigned through* (P1)–(P2)*. Then, there exists a constant $M > 0$ such that for ϵ_n in* (7)*,*

$$\mathbb{E}_0\Pi\Big\{f : \|H(f)-H(f_0)\|_{2,Q} > M\epsilon_n \,\Big|\, (X_1,Y_1),\dots,(X_n,Y_n)\Big\} \to 0.$$

Proof. See Section A.5 in Appendix. ■

6.4 Additive Nonparametric Regression

Thus far we have considered statistical models with the true function f_0 that belongs to the piecewise heterogeneous anisotropic Hölder space with sparsity. As Theorems 2-6 hold even with the single tree model ($T = 1$), the empirical success of BART is not well explained by the previous examples, although the empirical performance of BART should be attributed to its fast mixing to some extent. However, Linero and Yang (2018) and Ročková and van der Pas (2020) observed that BART optimally adapts to a larger class of additive functions which single tree models do not adapt to. In this section, we consider additive nonparametric regression to show theoretical advantages of BART over Bayesian CART.

We consider the nonparametric regression model with fixed design in (6), but the true function f_0 is assumed to have an additive structure with T_0 components, $f_0 = \sum_{t=1}^{T_0} f_{0t}$, where each f_{0t} belongs to the piecewise heterogeneous anisotropic Hölder space with sparsity. We also need suitable conditions on a split-net $\mathcal{Z}$ such that the approximation theory works for every additive component. We thus make the following modifications of the conditions used in Section 5.1. In what follows, the subscript or superscript t stands for additive component-specific extensions of the model elements used in Section 5.1.

(A1§) For $d_t > 0$, $\lambda_t > 0$, $R_t > 0$, $\mathfrak{X}_{0t} = \{\Xi_{t1}, \dots, \Xi_{tR}\}$, and $A_{t,\bar{\alpha}_t} \in \mathcal{A}_{\bar{\alpha}_t}^{R_t,d_t}$ with $\bar{\alpha}_t \in (0,1]$, $t = 1, \dots, T_0$, the true function satisfies $f_0 = \sum_{t=1}^{T_0} f_{0t}$ for $f_{0t} \in \Gamma_{\lambda_t}^{A_{t,\bar{\alpha}_t},d_t,p}(\mathfrak{X}_{0t})$ or $f_{0t} \in \Gamma_{\lambda_t}^{A_{t,\bar{\alpha}_t},d_t,p}(\mathfrak{X}_{0t}) \cap \mathcal{C}([0,1]^p)$.

(A2§) It is assumed that d_t, p_t, λ_t, R_t, and $\bar{\alpha}_t$ satisfy $\epsilon_{t,n} \ll 1$, where $\epsilon_{t,n} = \sqrt{(d_t \log p)/n} + (\lambda_t d_t)^{d_t/(2\bar{\alpha}_t + d_t)} \left((R_t \log n)/n\right)^{\bar{\alpha}_t/(2\bar{\alpha}_t+d_t)}$.

(A6§) The split-net $\mathcal{Z}$ is suitably dense and regular to construct a $\mathcal{Z}$-tree partition $\widehat{\mathcal{T}}^t$ such that for $\bar{\epsilon}_{t,n} = (\lambda_t d_t)^{d_t/(2\bar{\alpha}_t + d_t)} \left((R_t \log n)/n\right)^{\bar{\alpha}_t/(2\bar{\alpha}_t+d_t)}$, there exists $\hat{f}_{0t} \in \mathcal{F}_{\widehat{\mathcal{T}}^t}$ satisfying $\|f_{0t} - \hat{f}_{0t}\|_n \lesssim \bar{\epsilon}_{t,n}$ by Theorem 1, $t = 1, \dots, T_0$.

(A7§) The $\mathcal{Z}$-tree partition $\mathcal{T}_t^* = \{\Omega_{t1}^*, \dots, \Omega_{tR}^*\}$ approximating $\mathfrak{X}_{0t}^*$ satisfies $\max_r \mathsf{dep}(\Omega_{tr}^*) \lesssim \log n$, $t = 1, \dots, T_0$.

These simply mean that the assumptions in Section 5.1 hold for every additive component f_{0t}. It is worth noting that we do not need to modify the prior distribution for additive regression, which makes BART very appealing in that the procedure truly adapts to the unknown true function. This is owing to the use of the Dirichlet prior in (3); the spike-and-slab prior does not yield such a nice property (Ročková and van der Pas, 2020). The next theorem provides the posterior contraction rate for the additive regression model.

Theorem 7 (Additive nonparametric regression). *Consider model* (6) *with Assumptions* (A1§)–(A2§), (A3)–(A5)*, and* (A6§)–(A7§) *and the prior assigned through* (P1)–(P3)*. If $T_0 \le T$, there exists a constant $M > 0$ such that for $\epsilon_n^* = \sqrt{\sum_{t=1}^{T_0} \epsilon_{t,n}^2}$,*

$$\mathbb{E}_0 \Pi\Big\{(f,\sigma^2) : \|f - f_0\|_n + |\sigma^2 - \sigma_0^2| > M\epsilon_n^* \,\Big|\, Y_1, \dots, Y_n\Big\} \to 0.$$

Proof. See Section A.5 in Appendix. ■

Theorem 7 shows that the posterior contraction rate for additive regression is the sum of the rates for the additive components. If the function space is reduced to a high-dimensional isotropic class, then our rate ϵ_n^* matches the minimax rate for high-dimensional additive regression (Yang and Tokdar, 2015). We believe that ϵ_n^* is indeed near-minimax optimal, which can be formally justified by combining the proof technique of our Theorem 3 and the tools for additive scenarios developed in Yang and Tokdar (2015). Considering the length of the paper, we do not pursue this direction in this study.

7. Discussion

In this study, we enlarged the scope of theoretical understanding of Bayesian forests in the context of function estimation by considering relaxed smoothness assumptions. We introduced a new class of piecewise anisotropic sparse functions, which form a blend of anisotropy and spatial inhomogeneity. We derived a minimax lower bound for estimation of these functions in high-dimensional regression setups, extending existing results obtained earlier *only* for isotropic functions. We formalized that Bayesian forests attain the near-optimal posterior concentration rate for these general function classes without any need for prior modification.

Our results are extended to a general class of estimation problems including nonparametric regression with a fixed and random design, binary classification, and density estimation. Although we do not consider further nonparametric statistical models with BART priors in view of the length of the work, there are many other possible directions, such as mean-variance function estimation (Pratola et al., 2020) and causal inference (Hahn et al., 2020). Refer to Linero (2017), Tan and Roy (2019), and Hill et al. (2020) for extensive surveys of the application of BART to various nonparametric models. Because our Lemmas 4–7 enjoy a model-free framework, they will also be useful in investigating the posterior contraction rates for other statistical models.

Acknowledgments

We are grateful to two referees and the associate editor for valuable comments and constructive suggestions. We also thank Qurie Moon for sharing the code for BART with the exponentially decaying prior distribution. Seonghyun Jeong was supported by the Yonsei University Research Fund of 2021-22-0032 and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2022R1C1C1006735). Veronika Ročková gratefully acknowledges support from the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business and the National Science Foundation (DMS:1944740).

Appendix

Appendix A. Technical Proofs

A.1 Proof of Theorem 1 and Corollary 1

We can prove Theorem 1 and Corollary 1 using a suitably chosen approximator $\hat{f}_0$ of f_0. The following proof shows that an approximator can be constructed with the step-heights evaluated at any $y_{rk} \in \Omega_{rk}^{\circ} \cap \Xi_r^*$, $r = 1, \ldots, R$, $k = 1, \ldots, 2^L$.

Proof of Theorem 1. As $\mathcal{T}^*$ is chosen as in (4) and every Ξ_r^* is regular, note that $\Omega_{rk}^\circ \cap \Xi_r^*$ is not empty for every r and k. We fix any $y_{rk} \in \Omega_{rk}^\circ \cap \Xi_r^*$ and let $\hat{f}_0(x) = \sum_{r,k} \mathbb{1}_{\Omega_{rk}^\circ}(x)\beta_{rk,0}$ for $\beta_{rk,0} = f_0(y_{rk})$, so that

$$f_0(x) - \hat{f}_0(x) = \sum_{r=1}^{R} \sum_{k=1}^{2^L} \mathbb{1}_{\Omega_{rk}^\circ}(x)(f_0(x) - f_0(y_{rk})).$$

In what follows, we write $S_0 = \{s_{0,1}, \ldots, s_{0,d}\}$. We verify the assertion for each of the given metrics.

Verification of (i) *and* (iv): We first prove (iv). Fix r and k. For any $x \in \Omega_{rk}^\circ$, define

$$x \mapsto x^* : x^* = \operatorname*{arg\,min}_{z \in \mathsf{cl}(\Omega_{rk}^\circ \cap \Xi_r^*)} \|x - z\|_1, \tag{12}$$

where $\mathsf{cl}(\cdot)$ denotes the closure of a set. If $x \in \mathsf{cl}(\Omega_{rk}^\circ \cap \Xi_r^*)$, it is trivial that $x^* = x$, which gives $|f_0(x) - f_0(x^*)| = 0$. If $x \notin \mathsf{cl}(\Omega_{rk}^\circ \cap \Xi_r^*)$, there exists $r' \neq r$ such that $x \in \Xi_{r'}^*$ for $\Xi_{r'}^*$ that is contiguous to Ξ_r^*, and hence, $x^* \in \mathsf{cl}(\Xi_r^*) \cap \mathsf{cl}(\Xi_{r'}^*)$. In this case, we have $x \neq x^*$ but $x_j = x_j^*$ for $j \notin S_0^* \subseteq S_0$, where x_j and x_j^* are the jth entries of x and x^*, respectively. As f_0 is continuous and $x, x^* \in \mathsf{cl}(\Xi_{r'}^*)$, we obtain $|f_0(x) - f_0(x^*)| = |h_0(x_{S_0}) - h_0(x_{S_0}^*)| \leq \lambda \sum_{j=1}^d |x_{s_{0,j}} - x_{s_{0,j}}^*|^{\alpha_{r'j}}$. It follows that, for any $x \in \Omega_{rk}^\circ$ with given r and k,

$$|f_0(x) - f_0(x^*)| \leq \lambda \sum_{j=1}^{d} |x_{s_{0,j}} - x_{s_{0,j}}^*|^{\min_{r,j} \alpha_{rj}} \leq \lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}}, \tag{13}$$

since $\|x - x^*\|_\infty \leq c_n$ and $x_j = x_j^*$ for $j \notin S_0^*$. Hence, by the triangle inequality, for any $x \in \Omega_{rk}^\circ$,

$$|f_0(x) - f_0(y_{rk})| \leq \lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}} + |f_0(x^*) - f_0(y_{rk})|. \tag{14}$$

Let $\{\tilde{\Omega}_{r1}^\circ, \ldots, \tilde{\Omega}_{r2^L}^\circ\}$ be the tree partition and $(l_{r1}, \ldots l_{rd})^\top$ be the counter vector returned by $\mathsf{Akd}(\Xi_r^*; \mathcal{Z}, \alpha, L, S_0)$ such that $L = \sum_{j=1}^d l_{rj}$, $r = 1, \ldots, R$. As $x^*, y_{rk} \in \mathsf{cl}(\Omega_{rk}^\circ \cap \Xi_r^*) \subseteq \mathsf{cl}(\Xi_r^*)$ and f_0 is continuous,

$$|f_0(x^*) - f_0(y_{rk})| \leq \lambda \sum_{j=1}^{d} \mathsf{len}([\tilde{\Omega}_{rk}^\circ]_{s_{0,j}})^{\alpha_{rj}} \lesssim \lambda \sum_{j=1}^{d} 2^{-\alpha_{rj} l_{rj}}. \tag{15}$$

Let $\tilde{l}_{rj} = L\bar{\alpha}/(d\alpha_{rj})$ for $r = 1, \ldots, R$, $j = 1, \ldots, d$, such that $\alpha_{r1}\tilde{l}_{r1} = \cdots = \alpha_{rd}\tilde{l}_{rd}$ and $L = \sum_{j=1}^d \tilde{l}_{rj}$ for every r (note that $\tilde{l}_{rj}$ may not be integers). Then, it can be easily seen that $l_{rj} > \tilde{l}_{rj} - 1$ for every r, j, and hence

$$\lambda \sum_{j=1}^{d} 2^{-\alpha_{rj} l_{rj}} \leq 2\lambda \sum_{j=1}^{d} 2^{-\alpha_{rj} \tilde{l}_{rj}} \leq 2\lambda d 2^{-\bar{\alpha} L/d}. \tag{16}$$

Putting the bounds together for every r and k, we obtain

$$\|f_0 - \hat{f}_0\|_\infty \lesssim \lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}} + \lambda d 2^{-\bar{\alpha} L/d}.$$

This verifies (iv).

Now, to prove (i), note that $c_n = 0$ implies $\mathsf{cl}(\Omega_r^*) = \mathsf{cl}(\Xi_r^*)$ although possibly $\Omega_r^* \neq \Xi_r^*$. That is, $(\Omega_r^* \cup \Xi_r^*) \cap (\Omega_r^* \cap \Xi_r^*)^c$ is a null set with measure zero for every r. Therefore, in evaluating the L_∞-norm $\|f_0 - \hat{f}_0\|_\infty$ with the essential supremum, we can ignore such a null set and focus on $\Omega_r^* \cap \Xi_r^*$. If $x \in \Omega_{rk}^\circ \cap \Xi_r^*$, then similar to (15) and (16), we obtain that $|f_0(x) - f_0(y_{rk})| \lesssim \lambda \sum_{j=1}^d 2^{-\alpha_{rj} l_{rj}} \lesssim \lambda d 2^{-\bar{\alpha} L/d}$ since $x, y \in \Xi_r^*$. Putting the bounds together for every r and k, we conclude the assertion.

Verification of (ii) *and* (v): To verify (ii), we first show that when $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0)$, for any finite measure μ and any fixed $v \geq 1$,

$$\|f_0 - \hat{f}_0\|_{v,\mu} \lesssim \tilde{\epsilon}_n, \quad \text{if} \quad \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}) \lesssim (\tilde{\epsilon}_n / \|f_0\|_\infty)^v. \tag{17}$$

Observe that

$$\int |f_0(x) - \hat{f}_0(x)|^v d\mu(x) = \sum_{r=1}^R \sum_{k=1}^{2^L} \int_{\Omega_{rk}^\circ} |f_0(x) - f_0(y_{rk})|^v d\mu(x). \tag{18}$$

The integral term in each summand is bounded by

$$\int_{\Omega_{rk}^\circ \cap \Xi_r^*} |f_0(x) - f_0(y_{rk})|^v d\mu(x) + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c})(2\|f_0\|_\infty)^v. \tag{19}$$

Using (15) and (16), observe that, for every $x, y_{rk} \in \Omega_{rk}^\circ \cap \Xi_r^*$,

$$|f_0(x) - f_0(y_{rk})| \lesssim \lambda d 2^{-\bar{\alpha} L/d}.$$

The first term of (19) is thus bounded by a constant multiple of $\mu(\Omega_{rk}^\circ \cap \Xi_r^*)(\lambda d 2^{-\bar{\alpha} L/d})^v$. Note also that $\sum_k \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) = \mu(\Omega_r^* \cap \Xi_r^{*c})$. Therefore,

$$\begin{aligned} \|f_0 - \hat{f}_0\|_{v,\mu}^v &\lesssim \sum_{r=1}^R \sum_{k=1}^{2^L} \left\{ \mu(\Omega_{rk}^\circ \cap \Xi_r^*) \left(\lambda d 2^{-\bar{\alpha} L/d}\right)^v + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \|f_0\|_\infty^v \right\} \\ &\leq \mu([0,1]^p) \left(\lambda d 2^{-\bar{\alpha} L/d}\right)^v + \|f_0\|_\infty^v \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}). \end{aligned} \tag{20}$$

This leads to the assertion in (17). Now, to verify the first part of (ii), it suffices to show that $\sum_{r=1}^R \mathrm{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \lesssim (\tilde{\epsilon}_n / \|f_0\|_\infty)^v$ for Leb_p, the Lebesgue measure on a p-dimensional space. For each r, we only need to consider the case $\Xi_r^* \subsetneq \Omega_r^*$, as $\mathrm{Leb}_p(\Omega_r^* \cap \Xi_r^{*c})$ is maximized in this case. Then, $\Omega_r^* \cap \Xi_r^{*c}$ is not a box but a p-dimensional orthogonal polyhedron (for example, with a rectangular hole). One can easily see that

$$\mathrm{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \leq \sum_{j=1}^p \mathrm{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) \prod_{k \neq j} \mathsf{len}([\Omega_r^*]_k).$$

It should be noticed that $[\Omega_r^* \cap \Xi_r^{*c}]_j$ may not be an interval but can be an empty set or a union of two isolated intervals. As $\mathrm{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) = 0$ for $j \notin S_0^* \subseteq S_0$ and $\max_j \mathrm{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) \le 2c_n$, the last expression is bounded by

$$|S_0^*| \max_j \left\{ \mathrm{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) \prod_{k \neq j} \mathsf{len}([\Omega_r^*]_k) \right\} \le \frac{2c_n |S_0^*| \mathsf{vol}(\Omega_r^*)}{\min_j \mathsf{len}([\Omega_r^*]_j)},$$

where we use the notation $\mathsf{vol}(\cdot)$ to denote the volume of a box. As $\mathsf{len}([\Omega_r^*]_j) \ge \mathsf{len}([\Xi_r^*]_j) - 2c_n$ for every j,

$$\sum_{r=1}^R \mathrm{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \le \frac{2c_n |S_0^*|}{\min_{r,j} \mathsf{len}([\Xi_r^*]_j) - 2c_n} \le \frac{3c_n |S_0^*|}{\min_{r,j} \mathsf{len}([\Xi_r^*]_j)}, \tag{21}$$

for every small $c_n > 0$. It follows from this that $\sum_{r=1}^R \mathrm{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \lesssim (\tilde\epsilon_n / \|f_0\|_\infty)^v$ if $c_n \lesssim (\tilde\epsilon_n/\|f_0\|_\infty)^v \min_{r,j} \mathsf{len}([\Xi_r^*]_j)/|S_0^*|$. The first part of (ii) is verified.

We now verify (v). Similar to (17), we first show that when $f_0 \in \Gamma_\lambda^{A_{\bar\alpha},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$, for any finite measure μ and any $v \ge 1$,

$$\|f_0 - \hat f_0\|_{v,\mu} \lesssim \tilde\epsilon_n, \quad \text{if} \quad c_n^{v \min_{r,j} \alpha_{rj}} \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}) \lesssim (\tilde\epsilon_n / (\lambda |S_0^*|))^v. \tag{22}$$

We start from the identity in (18). Similar to the above, one can observe that the integral term in (18) is bounded by

$$\mu(\Omega_{rk}^\circ \cap \Xi_r^*) \left(\lambda d 2^{-\bar\alpha L/d} \right)^v + \int_{\Omega_{rk}^\circ \cap \Xi_r^{*c}} |f_0(x) - f_0(y_{rk})|^v d\mu(x). \tag{23}$$

Using $x^* \in \mathsf{cl}(\Omega_{rk}^\circ \cap \Xi_r^*)$ in (12), the second term of (23) is bounded by

$$\begin{aligned} &2^{v-1} \int_{\Omega_{rk}^\circ \cap \Xi_r^{*c}} (|f_0(x) - f_0(x^*)|^v + |f_0(x^*) - f_0(y_{rk})|^v) d\mu(x) \\ &\quad \le 2^{v-1} \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \left\{ \left(\lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}} \right)^v + \left(\lambda d 2^{-\bar\alpha L/d} \right)^v \right\}, \end{aligned}$$

where the inequality holds by (13) combined with the fact that $x, x^* \in \mathsf{cl}(\Xi_{r'}^*)$ and $x^*, y_{rk} \in \mathsf{cl}(\Xi_r^*)$ for some $r' \neq r$. Hence, (23) is further bounded by a constant multiple of

$$\mu(\Omega_{rk}^\circ) \left(\lambda d 2^{-\bar\alpha L/d} \right)^v + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \left(\lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}} \right)^v,$$

and we obtain that

$$\begin{aligned} \|f_0 - \hat f_0\|_{v,\mu}^v &\lesssim \sum_{r=1}^R \sum_{k=1}^{2^L} \left\{ \mu(\Omega_{rk}^\circ) \left(\lambda d 2^{-\bar\alpha L/d} \right)^v + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \left(\lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}} \right)^v \right\} \\ &\le \mu([0,1]^p) \left(\lambda d 2^{-\bar\alpha L/d} \right)^v + \left(\lambda |S_0^*| c_n^{\min_{r,j} \alpha_{rj}} \right)^v \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}). \end{aligned} \tag{24}$$

This leads to (22). Now, to verify the second part of (ii), we take the Lebesgue measure for μ. Then using the bound in (21), we have that $(\lambda|S_0^*|c_n^{\min_{r,j}\alpha_{rj}})^v \sum_{r=1}^R \mathrm{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \lesssim \tilde{\epsilon}_n^v$ if $c_n^{1+v\min_{r,j}\alpha_{rj}} \lesssim (\tilde{\epsilon}_n/\lambda)^v \min_{r,j} \mathsf{len}([\Xi_r^*]_j)/|S_0^*|^{v+1}$. This proves the assertion.

Verification of (iii): We again use the result in (17). Take $P_{\mathcal{Z}}$ for μ. Then, it can be seen that split-points can be picked up such that there are no z_i on $\cup_r(\Omega_r^* \cap \Xi_r^{*c})$ by choosing the points closest to the boundaries in every split. As we have $\sum_{r=1}^R P_{\mathcal{Z}}(\Omega_r^* \cap \Xi_r^{*c}) = 0$ in this case, (iii) easily follows. ■

Proof of Corollary 1. *Verification of* (i): As $\bar{\epsilon}_n \gtrsim (\lambda dR(\log n)/n)^{1/3}$ and $|S_0^*| \le d$, we obtain $\bar{\epsilon}_n/(\lambda|S_0^*|) \gtrsim (n^{-1}R\log n)^{1/3}(\lambda d)^{-2/3} \ge n^{-(1+2a_2)/3}\log^{-1/3} n$. The assertion in (i) follows by combining (iv) of Theorem 1 and the bound $\min_{r,j}\alpha_{rj} \ge a_1$.

Verification of (ii): Similar to above, we obtain

$$\begin{aligned}(\bar{\epsilon}_n/\|f_0\|_\infty)^v \min_{r,j} \mathsf{len}([\Xi_r^*]_j)/|S_0^*| &\gtrsim (n^{-1}\lambda dR\log n)^{v/3}(\log n)^{-v/2} n^{-a_3}/d \\ &\gtrsim \lambda^{v/3} n^{-(v/3+a_3)}(\log n)^{-(\max\{0,1-v/3\}+v/6)}.\end{aligned}$$

As $\lambda \gtrsim 1$, we can verify the assertion in (ii) using (ii) of Theorem 1. ■

A.2 Proof of Lemmas 1–2

To prove Lemma 1, we first provide the following lemma, which shows that a regular grid is dense and regular for arbitrary inputs under mild conditions.

Lemma 3 (Regular grid, general case)**.** *For a regular grid $\mathcal{Z}$, we have the following assertions.*

(i) *For any $S \subseteq \{1,\dots,p\}$ and any S-chopped flexible tree partition $\mathfrak{Y} = \{\Psi_1,\dots,\Psi_J\}$ with $J \ge 2$, $\mathcal{Z}$ is $(\mathfrak{Y}, 1/b_n^{1/p})$-dense if* $\min_{r,j}\mathsf{len}([\Psi_r]_j) \ge b_n^{-1/p}$.

(ii) *For any $S \subseteq \{1,\dots,p\}$, $\alpha \in (0,1]^d$, $\Psi \subseteq [0,1]^p$, and $L = \lfloor \log_2(b_n^{1/p}\min_j \mathsf{len}([\Psi]_j) - 1)\rfloor$, $\mathcal{Z}$ is (Ψ,α,L,S)-regular if* $\min_{r,j}\mathsf{len}([\Psi_r]_j) \ge 3b_n^{-1/p}$.

Proof. *Verification of* (i): Consider a p-dimensional checkerboard $\prod_{j=1}^p[(i_j-1)/b_n^{1/p}, i_j/b_n^{1/p}]$, $i_j = 1,\dots,b_n$. Note that each point z_i in $\mathcal{Z}$ is located at the center of each box of this checkerboard. As the mesh-size of the checkerboard is $1/b_n^{1/p}$, there exists an S-chopped $\mathcal{Z}$-tree partition $\mathcal{T}$ such that $\Upsilon(\mathfrak{Y},\mathcal{T}) \le 1/b_n^{1/p}$ if $\min_{r,j}\mathsf{len}([\Psi_r]_j) \ge 1/b_n^{1/p}$. The assertion easily follows.

Verification of (ii): The condition $\min_{r,j}\mathsf{len}([\Psi_r]_j) \ge 3b_n^{-1/p}$ is made to ensure that there is at least one split-point that is sufficiently far away from the boundaries of Ψ in every coordinate. Observe that for any box $\Psi \subseteq [0,1]^p$, we obtain

$$\tilde{b}_j(\mathcal{Z},\Psi) \le b_n^{1/p}\mathsf{len}([\Psi]_j) \le \tilde{b}_j(\mathcal{Z},\Psi) + 1, \quad j = 1,\dots,p. \tag{25}$$

Thus, in every coordinate, midpoint-splits can occur $\lfloor b_n^{1/p}\min_j \mathsf{len}([\Psi]_j)\rfloor - 1$ times without choosing the leftmost and rightmost split-points (these two points may produce too small cells). This allows us to choose $L = \lfloor \log_2(b_n^{1/p} \min_j \mathsf{len}([\Psi]_j) - 1)\rfloor$ for an anisotropic k-d tree (note that $\lfloor \log\lfloor x\rfloor\rfloor = \lfloor \log x\rfloor$, $x > 0$).

For any $\Psi \subseteq [0,1]^p$ and $j \in S$, a mid-point split chooses $\lceil \tilde{b}_j(\mathcal{Z},\Psi)/2\rceil$th split-candidate in $[\mathcal{Z}]_j \cap \mathsf{int}([\Psi]_j)$ as a split-point τ_j. The resulting two cells have at most $\lfloor \tilde{b}_j(\mathcal{Z},\Psi)/2\rfloor$ split-points in coordinate j. Therefore, using (25),

$$\begin{aligned}\max_k \mathsf{len}([\Omega_k^\circ]_{s_j}) &\le \frac{\tilde{b}_j(\mathcal{Z},\Psi)2^{-l_j}+1}{b_n^{1/p}} \\ &\le \mathsf{len}([\Psi]_{s_j})2^{-l_j} + 1/b_n^{1/p} \\ &\le \mathsf{len}([\Psi]_{s_j})\left(2^{-l_j} + \frac{1}{b_n^{1/p}\min_{r,j}\mathsf{len}([\Psi_r]_j)}\right).\end{aligned}$$

As $L \le \log_2(b_n^{1/p}\min_j \mathsf{len}([\Psi]_j) - 1) \le \log_2(b_n^{1/p}\min_j \mathsf{len}([\Psi]_j)) - 1$ and $l_j \le L$ for every $j = 1,\dots,d$, the last expression is bounded by

$$\mathsf{len}([\Psi]_{s_j})(2^{-l_j} + 2^{1-L}) \le 3\mathsf{len}([\Psi]_{s_j})2^{-l_j}.$$

This leads to the assertion. ■

Proof of Lemma 1. If $R = 1$, it is obvious that $\mathcal{Z}$ is $(\mathfrak{X}_0^*, 0)$-dense. If $R > 1$, by (i) of Lemma 3, $\mathcal{Z}$ is $(\mathfrak{X}_0^*, 1/b_n^{1/p})$-dense since $b_n^{1/p}\min_{r,j}\mathsf{len}([\Xi_r^*]_j) \gg 1$. Also, (ii) of Lemma 3 shows that $\mathcal{Z}$ is $(\Xi_r^*, \alpha_r, L_r, S_0)$-regular for $L_r = \lfloor \log_2(b_n^{1/p}\min_j \mathsf{len}([\Xi_r^*]_j) - 1)\rfloor$, $r = 1,\dots,R$. To conclude that $\mathcal{Z}$ is $(\Omega_r^*, \alpha_r, L_0, S_0)$-regular for $r = 1,\dots,R$, we only need to show that $L_0 \le \min_r L_r$. As $2^{L_0} \asymp (n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)}$, L_0 can be chosen to be $2^{L_0} \le C_1(n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)}$ for small enough $C_1 > 0$ as desired. Therefore, a sufficient condition for $L_0 \le \min_r L_r$ is $(n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)} \lesssim b_n^{1/p}\min_j \mathsf{len}([\Xi_r^*]_j)$. Plugging in $b_n = n^{cp}$, the conditions in the lemma are obtained. ■

Proof of Lemma 2. Assumption (F) implies that $\mathcal{Z}$ is $(\Xi_r^*, \alpha_r, L_r, S_0)$-regular for $L_r = \lfloor \log_2(C_1 n P_{\mathcal{Z}}(\Xi_r^*))\rfloor$, for some $C_1 > 0$, $r = 1,\dots,R$. It remains to show that $L_0 \le \min_r L_r$. Recall that $2^{L_0} \asymp (n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)}$. As L_0 can be chosen to be $2^{L_0} \le C_2(n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)}$ for small enough $C_2 > 0$ as desired, a sufficient condition for $L_0 \le \min_r L_r$ is given by $(n(\lambda d)^2/(R\log n))^{d/(2\bar{\alpha}+d)} \lesssim nP_{\mathcal{Z}}(\Xi_r^*)$ no matter what C_1 is. Using that $P_{\mathcal{Z}}(\Xi_r^*) \gtrsim R^{-1}$, the inequality is translated into $\lambda d \lesssim (n/R)^{\bar{\alpha}/d}\sqrt{\log n}$. ■

A.3 Proof of Theorem 2

We deploy the standard theory on posterior contraction (Ghosal et al., 2000; Ghosal and van der Vaart, 2007). The required conditions for the general theory are deferred to Lemmas 4–7.

Proof of Theorem 2. As σ_0^2 is bounded below and above, $|\sigma^2 - \sigma_0^2|$ and $|\sigma - \sigma_0|$ have the same rate. We will work with the latter for convenience. We write $\rho_n^2((f_1,\sigma_1),(f_2,\sigma_2)) = \|f_1 - f_2\|_n^2 + |\sigma_1 - \sigma_2|^2$ for any $f_1, f_2 : \mathbb{R}^p \to \mathbb{R}$ and any $\sigma_1, \sigma_2 \in (0,\infty)$. (Observe that

$\sqrt{\|\cdot\|_n^2 + |\cdot|^2}$ and $\|\cdot\|_n + |\cdot|$ have the same order.) By Lemma 1 of Lim and Jeong (2023), for every $\epsilon > 0$ and (f_1, σ_1) with $\|f_1 - f_0\|_n^2 + |\sigma_1 - \sigma_0|^2 \geq \epsilon^2$, there exists a test ϕ_n such that, for a universal constant $K > 0$,

$$\mathbb{E}_0 \phi_n \leq e^{-Kn\epsilon^2}, \qquad \sup_{(f,\sigma^2): \|f - f_1\|_n^2 + |\sigma - \sigma_1|^2 \leq \epsilon^2/36} \mathbb{E}_{f,\sigma^2}(1 - \phi_n) \leq e^{-Kn\epsilon^2}.$$

We write $\mathcal{F}_* = \cup_{\mathcal{E}} \mathcal{F}_{\mathcal{E}}$, where the union is taken over all $\mathcal{E}$ generated by a given $\mathcal{Z}$. For the Kullback-Leibler (KL) divergence $K(p_1, p_2) = \int \log(p_1/p_2) p_1$ and its second order variation $V(p_1, p_2) = \int |\log(p_1/p_2) - K(p_1, p_2)|^2 p_1$, define

$$B_n = \left\{ (f, \sigma) : \sum_{i=1}^n K(p_{0,i}, p_{f,\sigma,i}) \leq n\epsilon_n^2, \sum_{i=1}^n V(p_{0,i}, p_{f,\sigma,i}) \leq n\epsilon_n^2 \right\}.$$

By Theorem 8.19 of Ghosal and van der Vaart (2017), we only need to verify that there exists a sieve $\Theta_n \subseteq \mathcal{F} \times (0, \infty)$ such that for some $\bar{c} > 0$ and a sufficiently large $\bar{c}' > 0$,

$$\Pi(B_n) \geq e^{-\bar{c} n \epsilon_n^2}, \tag{26}$$

$$\log N(\epsilon_n, \Theta_n, \rho_n) \lesssim n\epsilon_n^2, \tag{27}$$

$$\Pi((f, \sigma) \notin \Theta_n) \ll e^{-\bar{c}' n \epsilon_n^2}, \tag{28}$$

We first verify (26). By direct calculations,

$$\frac{1}{n} \sum_{i=1}^n K(p_{0,i}, p_{f,\sigma,i}) = \frac{1}{2} \log\left(\frac{\sigma^2}{\sigma_0^2}\right) - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{\|f - f_0\|_n^2}{2\sigma^2},$$

$$\frac{1}{n} \sum_{i=1}^n V(p_{0,i}, p_{f,\sigma,i}) = \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right)^2 + \frac{\sigma_0^2 \|f - f_0\|_n^2}{\sigma^2}.$$

Using the Taylor expansion, it is easy to see that, for any $\epsilon_n \to 0$, there exists a constant $C_1 > 0$ such that

$$B_n \supseteq \{(f, \sigma) : \|f - f_0\|_n \leq C_1 \epsilon_n, |\sigma - \sigma_0| \leq C_1 \epsilon_n\}.$$

First, note that $\log \Pi(\sigma^2 : |\sigma - \sigma_0| \leq C_1 \epsilon_n) \gtrsim -\log n$ if σ_0 lies on a compact subset of $(0, \infty)$. We will construct a good approximating ensemble denoted by $\widehat{\mathcal{E}} = (\widehat{\mathcal{T}}^1, \ldots, \widehat{\mathcal{T}}^T)$. By restricting the function space to the one constructed by $\widehat{\mathcal{E}}$, we obtain

$$\Pi(f \in \mathcal{F}_* : \|f - f_0\|_n \leq C_1 \epsilon_n) \geq \Pi(\widehat{\mathcal{E}}) \Pi(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_0\|_n \leq C_1 \epsilon_n). \tag{29}$$

Assumption (A6) states that, for a given split-net $\mathcal{Z}$ there exists a $\mathcal{Z}$-tree partition $\widehat{\mathcal{T}}$ producing $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$ satisfying $\|f_0 - \hat{f}_0\|_n \lesssim \bar{\epsilon}_n$. An approximating ensemble $\widehat{\mathcal{E}}$ can be constructed by setting $\widehat{\mathcal{T}}^1$ to be $\widehat{\mathcal{T}}$ and $\widehat{\mathcal{T}}^t$, $t = 2, \ldots, T$, to be root nodes with no splits, i.e., $\widehat{\mathcal{T}}^t = \{[0,1]^p\}$, $t = 2, \ldots, T$. Then,

$$\log \Pi(\widehat{\mathcal{E}}) = \sum_{t=1}^T \log \Pi(\widehat{\mathcal{T}}^t) = \log \Pi(\widehat{\mathcal{T}}^1) + (T - 1) \log(1 - \nu) \gtrsim -n\epsilon_n^2,$$

by Lemma 4. It remains to bound the second term of (29). By (A6), we have $\|f - f_0\|_n \lesssim \|f - \hat{f}_0\|_\infty + \epsilon_n$ for some $\hat{f}_0 \in \mathcal{F}_{\widehat{\mathcal{T}}}$. We can construct $\hat{f}_0$ as in the proof of Theorem 1. We denote this $\hat{f}_0$ by $f_{0,\widehat{\mathcal{T}},\widehat{\beta}}$, where $\widehat{\beta}$ is the corresponding step-heights, to emphasize the dependence on $\widehat{\mathcal{T}}$ and $\widehat{\beta}$. We shall now express $f_{0,\widehat{\mathcal{T}},\widehat{\beta}}$ using the approximating ensemble $\widehat{\mathcal{E}}$ with corresponding step-heights $\widehat{B}$. As all trees in $\widehat{\mathcal{E}}$ are the root nodes except for the first one $\widehat{\mathcal{T}}^1$, every step-heights vector B for $\widehat{\mathcal{E}}$ has the form $B = (\beta^{1\top}, \beta^2, \dots, \beta^T)^\top \in \mathbb{R}^{\widehat{K}+T-1}$ with $\beta^1 \in \mathbb{R}^{\widehat{K}}$ and $\beta^t \in \mathbb{R}$, $t = 2, \dots, T$, where $\widehat{K}$ is the size of $\widehat{\mathcal{T}}$. Hence, letting $\widehat{B} = (\widehat{\beta}^\top, 0, \dots, 0)^\top$, we can write $f_{0,\widehat{\mathcal{T}},\widehat{\beta}} = f_{0,\widehat{\mathcal{E}},\widehat{B}}$ for $f_{0,\widehat{\mathcal{E}},\widehat{B}}$ defined with the ensemble components $(\widehat{\mathcal{E}}, \widehat{B})$. Putting the bounds together, for some $C_2 > 0$,

$$\Pi\left(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_0\|_n \le C_1 \epsilon_n\right) \ge \Pi\left(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_{0,\widehat{\mathcal{E}},\widehat{B}}\|_\infty \le C_2 \epsilon_n\right).$$

By Lemma 5, the right-hand side is bounded below as desired. Putting everything together, we conclude that there exists a constant $\bar{c}$ such that $\Pi(B_n) \ge e^{-\bar{c} n \epsilon_n^2}$.

Next, we verify the entropy condition (27). We denote by $\mathscr{E}_{S,K^1,\dots,K^T}$ the collection of $\mathcal{E} = \{\mathcal{T}^1, \dots, \mathcal{T}^T\}$ with given $S, K^1, \dots, K^T$; that is, each $\mathcal{T}^t$ is an S-chopped $\mathcal{Z}$-tree partition of size K^t. With given $\mathcal{E}$ and $M > 0$, we first define the function spaces $\mathcal{F}^{(1)}_{\mathcal{E},M} = \{f_{\mathcal{E},B} \in \mathcal{F}_{\mathcal{E}} : \|B\|_\infty \le M\}$ and $\mathcal{F}^{(2)}_{\mathcal{E},M} = \{f_{\mathcal{E},B} \in \mathcal{F}_{\mathcal{E}} : \|B\|_\infty > M\}$ such that $\mathcal{F}^{(1)}_{\mathcal{E},M} \cup \mathcal{F}^{(2)}_{\mathcal{E},M} = \mathcal{F}_{\mathcal{E}}$. We also define

$$\mathcal{F}^{(\ell)}_{\bar{s}_n,\bar{K}_n,M} := \bigcup_{\mathcal{E} \in \mathscr{E}_{S,K^1,\dots,K^T} : |S| \le \bar{s}_n, K^t \le \bar{K}_n, t=1,\dots,T} \mathcal{F}^{(\ell)}_{\mathcal{E},M}, \quad \ell = 1, 2, \tag{30}$$

for $\bar{K}_n \asymp n\epsilon_n^2/\log n$ and $\bar{s}_n \asymp n\epsilon_n^2/\log p$. That is, $\mathcal{F}^{(\ell)}_{\bar{s}_n,\bar{K}_n,M}$ is the collection of all $\mathcal{F}^{(\ell)}_{\mathcal{E},M}$ such that $K^t \le \bar{K}_n$ and $|S| \le \bar{s}_n$. We take $\Theta_n = \mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^{M_1}} \times (n^{-M_2}, e^{M_2 n\epsilon_n^2})$ for large $M_1, M_2 > 0$. It is easy to see that $\log N(\epsilon_n, (n^{-M_2}, e^{M_2 n\epsilon_n^2}), |\cdot|) \lesssim n\epsilon_n^2$. Combining this with Lemma 6, we conclude that (27) is verified.

Lastly, we verify (28). First, it is easy to see that $\Pi(\sigma^2 \notin (n^{-2M_2}, e^{2M_2 n\epsilon_n^2})) e^{\bar{c}' n\epsilon_n^2} \to 0$ if M_2 is large enough, using the tail probabilities of inverse gamma distributions. Choose $\bar{K}_n = \lfloor M_3 n\epsilon_n^2/\log n \rfloor$ and $\bar{s}_n = \lfloor M_3 n\epsilon_n^2/\log p \rfloor$ for a sufficiently large $M_3 > 0$. As we have $\Pi(\mathcal{F}_* \setminus \mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^{M_1}}) e^{\bar{c}' n\epsilon_n^2} \to 0$ by Lemma 7, the condition is verified. ■

Lemma 4 (Prior concentration of tree sizes)**.** *Let $\widehat{\mathcal{T}}$ be the $\mathcal{Z}$-tree partition defined in* (5)*. Under Assumptions* (A5) *and* (A7)*,* $\log \Pi(\widehat{\mathcal{T}}) \gtrsim -\widehat{K} \log n - d \log p$.

Proof. We will obtain a lower bound of $\Pi(\widehat{\mathcal{T}})$. As this depends on splitting proportions drawn from a Dirichlet prior, we first restrict the proportions to the set

$$V_1 = \left\{ \eta \in \mathbb{S}^p : \eta_j \ge \frac{1}{2d}, j \in S_0, \sum_{j \notin S_0} \eta_j \le \frac{1}{2d} \right\}.$$

Fix $\eta^* = (\eta_1^*, \dots, \eta_p^*)^\top \in \mathbb{S}^p$ such that $\eta_j^* = 1/d$, $j \in S_0$, and $\eta_j^* = 0$, $j \notin S_0$. It can be easily shown that $V_1 \supseteq \{\eta \in \mathbb{S}^p : \|\eta - \eta^*\|_1 \le 1/(2d)\}$. By (54) of Lemma 12, it follows that $\Pi(V_1) \geq e^{-C_1 d \log p}$ for some $C_1 > 0$. Recall that the first $R-1$ splits of $\widehat{\mathcal{T}}$ form $\mathcal{T}^* = \{\Omega_1^*, \dots, \Omega_R^*\}$, the approximating tree partition of $\mathfrak{X}_0^*$, and the remaining splits generate $\mathcal{T}_r^\circ$, the tree partition of Ω_r^* constructed by an anisotropic k-d tree, $r = 1, \dots, R$. Hence, we can write

$$\Pi(\widehat{\mathcal{T}}) \geq e^{-C_1 d \log p} \Pi(\widehat{\mathcal{T}} | V_1) = e^{-C_1 d \log p} \Pi(\mathcal{T}^* \text{ is a pruned tree of } \widehat{\mathcal{T}} | V_1) \prod_{r=1}^{R} \Pi(\mathcal{T}_r^\circ | \Omega_r^*, V_1).$$

We first focus on the prior probability $\Pi(\mathcal{T}^* \text{ is a pruned tree of } \widehat{\mathcal{T}} | V_1)$. To generate $\mathcal{T}^*$, the root node is subdivided $R-1$ times in a top-down manner. As each node splits with probability $\nu^{\ell+1}$ for depth ℓ, this occurs with probability at least $\nu^{(R-1)\max_r \mathsf{dep}(\Omega_r^*)}$ no matter what the partition is. Note also that, for every split, there are at most $\max_{1 \leq j \leq p} b_j(\mathcal{Z})$ splitting points and a splitting coordinate j is chosen by η_j, $j \in S_0$, which is at least $1/(2d)$ on V_1. Hence the prior probability of choosing the correct split is bounded below by $1/(2d \max_{1 \leq j \leq p} b_j(\mathcal{Z}))$ for every split. This gives us a lower bound:

$$\Pi(\mathcal{T}^* \text{ is a pruned tree of } \widehat{\mathcal{T}} | V_1) \geq \frac{\nu^{(R-1)\max_r \mathsf{dep}(\Omega_r^*)}}{(2d \max_{1 \leq j \leq p} b_j(\mathcal{Z}))^{R-1}}.$$

It follows that $\log \Pi(\mathcal{T}^* \text{ is a pruned tree of } \widehat{\mathcal{T}} | V_1) \gtrsim -R \log n$ as $\log(2d \max_{1 \leq j \leq p} b_j(\mathcal{Z})) \lesssim \log n$ by (A5) and $\max_r \mathsf{dep}(\Omega_r^*) \lesssim \log n$ by (A7).

We now obtain a lower bound of $\Pi(\mathcal{T}_r^\circ | \Omega_r^*, V_1)$. In splitting each Ω_r^*, observe that 2^k cells split at depth $k = 0, \dots, L_0 - 1$, and each cell splits with probability $\nu^{\mathsf{dep}(\Omega_r^*)+k+1}$ at depth k. Note that closing each of the terminal nodes is of probability at least $1-\nu$ and there are 2^{L_0} terminal nodes. Hence, similar to the above,

$$\begin{aligned} \Pi(\mathcal{T}_r^\circ | \Omega_r^*, V_1) &\geq (1-\nu)^{2^{L_0}} \prod_{k=0}^{L_0-1} \left(\frac{\nu^{\mathsf{dep}(\Omega_r^*)+k+1}}{2d \max_{1 \leq j \leq p} b_j(\mathcal{Z})} \right)^{2^k} \\ &= (1-\nu)^{2^{L_0}} \frac{\nu^{(\mathsf{dep}(\Omega_r^*)+1)(2^{L_0}-1)+(L_0-2)2^{L_0}+2}}{(2d \max_{1 \leq j \leq p} b_j(\mathcal{Z}))^{2^{L_0}-1}}, \end{aligned}$$

where we used the formulae $\sum_{k=0}^{a-1} 2^k = 2^a - 1$ and $\sum_{k=0}^{a-1} k 2^k = (a-2)2^a + 2$. This yields $\sum_{r=1}^{R} \log \Pi(\mathcal{T}_r^\circ | \Omega_r^*, V_1) \gtrsim -R 2^{L_0} \log(2d \max_{1 \leq j \leq p} b_j(\mathcal{Z})) - R 2^{L_0} \max_r \mathsf{dep}(\Omega_r^*) - R L_0 2^{L_0} \gtrsim -R 2^{L_0} \log n$ since $L_0 \lesssim \log n$.

Putting everything together, we thus obtain $\log \Pi(\widehat{\mathcal{T}}) \gtrsim -R 2^{L_0} \log n - d \log p$. As $\widehat{K} = R 2^{L_0}$, this verifies the assertion. ■

Lemma 5 (Prior concentration of tree learners). *Define $\widehat{\mathcal{E}}$ and $\widehat{B}$ as in the proof of Theorem 2. Under* (A3) *and* (P2), *for any $C > 0$,*

$$-\log \Pi \left(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \| f - f_{0, \widehat{\mathcal{E}}, \widehat{B}} \|_\infty \leq C \epsilon_n \right) \lesssim n \epsilon_n^2.$$

Proof. For any step-heights $B_1 = (\beta_1^{1\top}, \beta_1^2, \dots, \beta_1^T)^\top, B_2 = (\beta_2^{1\top}, \beta_2^2, \dots, \beta_2^T)^\top \in \mathbb{R}^{\widehat{K}+T-1}$ with $\widehat{\mathcal{E}}$, we write $f_{\widehat{\mathcal{E}},B_1}$, $f_{\widehat{\mathcal{E}},B_2} \in \mathcal{F}_{\widehat{\mathcal{E}}}$ to denote two additive tree functions that lie on the same partition ensemble $\widehat{\mathcal{E}}$. Evidently,

$$\|f_{\widehat{\mathcal{E}},B_1} - f_{\widehat{\mathcal{E}},B_2}\|_\infty = \left\| \sum_{t=1}^T \beta_1^t - \sum_{t=1}^T \beta_2^t \right\|_\infty \le \|\beta_1^1 - \beta_2^1\|_1 + \sum_{t=2}^T |\beta_1^t - \beta_2^t| \le \|B_1 - B_2\|_2 \sqrt{\widehat{K}_*},$$

where $\widehat{K}_* = \widehat{K} + T - 1$. It follows that, for some $C_1 > 0$,

$$\Pi\left(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_{0,\widehat{\mathcal{E}},\widehat{B}}\|_\infty \le C\epsilon_n \right) \ge \Pi\left(B \in \mathbb{R}^{\widehat{K}_*} : \|B - \widehat{B}\|_2 \le C_1 \epsilon_n / \sqrt{\widehat{K}_*} \right).$$

Recall that the eigenvalues of the covariance matrix for a normal prior is bounded below and above. This means that there exists an invertible matrix $D \in \mathbb{R}^{\widehat{K}_* \times \widehat{K}_*}$ such that DB has a product of independent standard normal priors. Following the computations in page 216 of Ghosal and van der Vaart (2007), the last display is further bounded below by

$$\begin{aligned} &\Pi\left(B \in \mathbb{R}^{\widehat{K}_*} : \|D(B - \widehat{B})\|_2 \le C_1 \epsilon_n \sigma_{\max}^{-1}(D^{-1}) / \sqrt{\widehat{K}_*} \right) \\ &\qquad \ge 2^{-\widehat{K}_*/2} e^{-\|D\widehat{B}\|_2^2} \Pi\left(B \in \mathbb{R}^{\widehat{K}_*} : \|DB\|_2 \le C_1 \epsilon_n \sigma_{\max}^{-1}(D^{-1}) / \sqrt{2\widehat{K}_*} \right), \end{aligned} \tag{31}$$

where $\sigma_{\max}(D^{-1})$ is the spectral norm of D^{-1}, which is bounded by the assumption. As the induced prior for $\|DB\|_2^2$ is a chi-squared distribution with degree of freedom $\widehat{K}_*$, we obtain that for $\upsilon_n = \epsilon_n \sigma_{\max}^{-1}(D^{-1}) / \sqrt{\widehat{K}_*} \lesssim \epsilon_n$,

$$\Pi(B \in \mathbb{R}^{\widehat{K}_*} : \|DB\|_2 \le C_1 \upsilon_n / \sqrt{2}) \ge \frac{2/\widehat{K}_*}{2^{\widehat{K}_*} \Gamma(\widehat{K}_*/2)} (C_1 \upsilon_n)^{\widehat{K}_*} e^{-C_1^2 \upsilon_n^2 / 4}.$$

The logarithm of the right-hand side is bounded below by a constant multiple of $-(\widehat{K} + T) \log n - \upsilon_n^2 \gtrsim -n\epsilon_n^2$. It only remains to bound $e^{-\|D\widehat{B}\|_2^2}$ in (31). Observe that $\|\widehat{\beta}\|_\infty = \|f_{0,\widehat{\mathcal{T}},\widehat{\beta}}\|_\infty \le \|f_0\|_\infty$, where the inequality follows from our choice of $\hat{f}_0 = f_{0,\widehat{\mathcal{T}},\widehat{\beta}}$ (see the proof of Theorem 1). Therefore,

$$\|D\widehat{B}\|_2^2 \le \sigma_{\max}^2(D) \|\widehat{\beta}\|_2^2 \le \sigma_{\max}^2(D) \widehat{K} \|\widehat{\beta}\|_\infty^2 \lesssim \widehat{K} \log n,$$

as soon as $\|f_0\|_\infty \lesssim \sqrt{\log n}$. ■

Lemma 6 (Metric entropy)**.** *Let $\bar{K}_n \asymp n\epsilon_n^2 / \log n$ and $\bar{s}_n \asymp n\epsilon_n^2 / \log p$. Define $\mathcal{F}_{\bar{s}_n, \bar{K}_n, M}^{(1)}$ for $M > 0$ as in* (30)*. Under* (A5)*, for any $C > 0$,*

$$\log N\left(\epsilon_n, \mathcal{F}_{\bar{s}_n, \bar{K}_n, n^C}^{(1)}, \|\cdot\|_n \right) \lesssim n\epsilon_n^2.$$

Proof. Observe that the exponential of the left-hand side is bounded by

$$\sum_{S:|S|\le \bar{s}_n}\ \sum_{(K^1,\dots,K^T):K^t\le \bar{K}_n,t=1,\dots,T}\ \sum_{\mathcal{E}\in\mathscr{E}_{S,K^1,\dots,K^T}} N\left(\epsilon_n, \mathcal{F}^{(1)}_{\mathcal{E},n^C}, \|\cdot\|_\infty\right). \tag{32}$$

For any given $\mathcal{E}$ and $B_1, B_2 \in \mathbb{R}^{\sum_{t=1}^T K^t}$,

$$\|f_{\mathcal{E},B_1} - f_{\mathcal{E},B_2}\|_\infty = \sup_{x\in[0,1]^p}\left|\sum_{t=1}^{T}\sum_{k=1}^{K^t}(\beta^t_{1k} - \beta^t_{2k})\mathbb{1}(x\in\Omega^t_k)\right| \le \left(\sum_{t=1}^T K^t\right)\|B_1 - B_2\|_\infty.$$

Observe that the cardinality of the set $\mathscr{E}_{S,\bar{K}^1,\dots,\bar{K}^T}$ is equal to $\prod_{t=1}^T |\mathscr{T}_{S,K^t,\mathcal{Z}}| \le |\mathscr{T}_{S,\bar{K}_n,\mathcal{Z}}|^T$. Hence, (32) is further bounded by

$$(\bar{K}_n)^T \times N\left(\frac{\epsilon_n}{T\bar{K}_n}, \left\{B\in\mathbb{R}^{T\bar{K}_n} : \|B\|_\infty \le n^C\right\}, \|\cdot\|_\infty\right)\sum_{S:|S|\le\bar{s}_n}|\mathscr{T}_{S,\bar{K}_n,\mathcal{Z}}|^T. \tag{33}$$

Observe that $|\mathscr{T}_{S,\bar{K}_n,\mathcal{Z}}| \le (|S|\max_{1\le j\le p} b_j)^{\bar{K}_n}$, as all splits are restricted to S and each one has at most $\max_{1\le j\le p} b_j$ split points. It follows that

$$\sum_{S:|S|\le\bar{s}_n}|\mathscr{T}_{S,\bar{K}_n,\mathcal{Z}}|^T \le \sum_{s=1}^{\bar{s}_n}\binom{p}{s}\left(s\max_{1\le j\le p} b_j\right)^{T\bar{K}_n} \le \bar{s}_n p^{\bar{s}_n}\left(\bar{s}_n\max_{1\le j\le p} b_j\right)^{T\bar{K}_n}.$$

Therefore, (33) is further bounded by $(\bar{K}_n)^T s_n p^{\bar{s}_n}(\bar{s}_n\max_{1\le j\le p} b_j)^{T\bar{K}_n}(3T\bar{K}_n n^C/\epsilon_n)^{T\bar{K}_n}$. The logarithm is bounded by a constant multiple of $\bar{s}_n\log p + \bar{K}_n\log n \lesssim n\epsilon_n^2$ as soon as $\max_{1\le j\le p} b_j \lesssim \log n$. ■

Lemma 7 (Prior mass of sieve)**.** *Let $\bar{K}_n = \lfloor M' n\epsilon_n^2/\log n\rfloor$, and $\bar{s}_n = \lfloor M' n\epsilon_n^2/\log p\rfloor$ for a sufficiently large $M' > 0$. Define $\mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,M}$ for $M > 0$ as in* (30)*. Under* (P1) *and* (P2)*, for any $C > 1$ and $C' > 0$,*

$$\Pi(\mathcal{F}_* \setminus \mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^C}) \ll e^{-C'n\epsilon_n^2}.$$

Proof. Note that $\mathcal{F}_* \setminus \mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^C} = \mathcal{F}^{(2)}_{\bar{s}_n,\bar{K}_n,n^C} \cup (\mathcal{F}_* \setminus (\mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^{M_1}} \cup \mathcal{F}^{(2)}_{\bar{s}_n,\bar{K}_n,n^{M_1}}))$. We will give a union bound. First, observe that

$$\Pi(\mathcal{F}^{(2)}_{\mathcal{E},M}) = \Pi(B\in\mathbb{R}^{\sum_{t=1}^T K^t}\colon \|B\|_\infty > M) \le \Pi\left(B\in\mathbb{R}^{\sum_{t=1}^T K^t}\colon \|DB\|_\infty > \frac{M\sigma^{-1}_{\max}(D^{-1})}{\sqrt{\sum_{t=1}^T K^t}}\right),$$

where D is the matrix with bounded singular values that makes the prior for DB the standard normal distribution. Using the tail probability of normal distributions,

$$\begin{aligned}\Pi\big(\mathcal{F}^{(2)}_{\bar{s}_n,\bar{K}_n,n^C}\big) &\le \sum_{S:|S|\le\bar{s}_n}\ \sum_{(K^1,\dots,K^T):K^t\le\bar{K}_n,t=1,\dots,T}\ \sum_{\mathcal{E}\in\mathscr{E}_{S,K^1,\dots,K^T}}\Pi\big(\mathcal{F}^{(2)}_{\mathcal{E},n^C}\big)\\ &\le (\bar{K}_n)^T\bar{s}_n p^{\bar{s}_n}\left(\bar{s}_n\max_{1\le j\le p} b_j\right)^{T\bar{K}_n} 2T\bar{K}_n e^{-\sigma^{-2}_{\max}(D^{-1})n^{2C}/(2T\bar{K}_n)}.\end{aligned}$$

Since $T\bar{K}_n \lesssim n\epsilon_n^2/\log n \ll n$ and $\sigma_{\max}(D^{-1})$ is bounded, if $\max_{1\le j\le p} b_j \lesssim \log n$ and $C > 1$, the right most side of the expression is $o(e^{-C'n\epsilon_n^2})$ for any $C' > 0$. Now observe that

$$\begin{aligned}&\Pi\Big(\mathcal{F}_* \setminus (\mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^{M_1}} \cup \mathcal{F}^{(2)}_{\bar{s}_n,\bar{K}_n,n^{M_1}})\Big)\\ &\quad\le \sum_{t=1}^{T} \Pi(K^t > \bar{K}_n) + \Pi(S : s > \bar{s}_n | K^t \le \bar{K}_n, t = 1,\dots,T).\end{aligned} \tag{34}$$

The prior satisfies $\log \Pi(K^t > \bar{K}_n) \lesssim -\bar{K}_n \log \bar{K}_n$ for every $t = 1,\dots,T$ (see Lemma 5.1 and Corollary 5.2 of Ročková and Saha (2019)). Using that $\bar{K}_n \asymp n\epsilon_n^2/\log n$ and $n\epsilon_n^2 \gtrsim n^{d/(2\bar{\alpha}+d)} \ge n^{1/3}$, we obtain $-\bar{K}_n \log \bar{K}_n \lesssim -\bar{K}_n \log n$. To bound the second term of the right-hand side of (34), we define the set

$$V_2 = \left\{ \eta \in \mathbb{S}^p : \min_{S:|S|=\bar{s}_n} \sum_{j\notin S} \eta_j \ge \kappa_n \right\},$$

for κ_n specified below. By (55) of Lemma 12, we show that the prior satisfies $\Pi(V_2) \le e^{-C_1(\xi-1)\bar{s}_n \log p - \log \kappa_n}$ for some $C_1 > 0$. Hence,

$$\begin{aligned}\Pi(S : s > \bar{s}_n | K^t \le \bar{K}_n, t = 1,\dots,T) &\le e^{-C_1(\xi-1)\bar{s}_n \log p - \log \kappa_n}\\ &\quad + \Pi(S : s > \bar{s}_n | K^t \le \bar{K}_n, t = 1,\dots,T, V_2^c).\end{aligned}$$

The term $\Pi(S : s > \bar{s}_n | K^t \le \bar{K}_n, t = 1,\dots,T, V_2^c)$ is interpreted as the prior probability that splits occur along more than $\bar{s}_n$ coordinates with at most $T\bar{K}_n$ splits given V_2^c. If η is available, this probability is

$$1 - \sum_{S:|S|\le\bar{s}_n} \left(\sum_{j\in S} \eta_j \right)^{T\bar{K}_n} \le 1 - \left(\max_{S:|S|\le\bar{s}_n} \sum_{j\in S} \eta_j \right)^{T\bar{K}_n}.$$

Conditional on V_2^c, the last expression is further bounded by $1 - (1-\kappa_n)^{T\bar{K}_n} \le \kappa_n T\bar{K}_n$. Choosing $\kappa_n = e^{-(C'+1)n\epsilon_n^2}$, the resulting bound for (34) gives $\Pi(\mathcal{F}_* \setminus \mathcal{F}^{(1)}_{\bar{s}_n,\bar{K}_n,n^C}) \ll e^{-C'n\epsilon_n^2}$ as M' is sufficiently large. ■

A.4 Proof of Theorem 3

Our proof is similar to the proof of Theorem 3.1 in Yang and Tokdar (2015), which is based on the Le Cam equation (Birgé and Massart, 1993; Wong and Shen, 1995; Barron et al., 1999). A minimax lower bound of nonparametric regression can be obtained by solving the Le Cam equation with the metric entropy of the target function space (Yang and Barron, 1999). We first formalize this result in the following lemma, which is a corollary induced by Theorem 1 of Yang and Barron (1999).

Lemma 8 (Minimax lower bound in nonparametric regression). *For a function space $\mathcal{F} \subset \mathcal{L}_2(Q)$, suppose there are upper and lower bounds of the metric entropies as*

$$\begin{aligned}\log N(\epsilon, \mathcal{F}, \|\cdot\|_{2,Q}) &\le V^*(\epsilon),\\ \log D(\epsilon, \mathcal{F}, \|\cdot\|_{2,Q}) &\ge V_*(\epsilon).\end{aligned} \tag{35}$$

Suppose that $\bar{\gamma}_n$ is the solution to $V^*(\bar{\gamma}_n) \asymp n\bar{\gamma}_n^2$. Then, for the nonparametric regression model in (8)*, the sequence γ_n such that $V_*(\gamma_n) \asymp n\bar{\gamma}_n^2$ satisfies*

$$r_n(\mathcal{F}, Q) \gtrsim \gamma_n,$$

where r_n is the $L_2(Q)$-minimax risk defined as $r_n(\mathcal{F}, Q) = \inf_{\hat{f}\in\mathcal{B}_n} \sup_{f_0\in\mathcal{F}} \mathbb{E}_{f_0,Q}\|\hat{f} - f_0\|_{2,Q}^2$ with $\mathcal{B}_n$ the space of all $L_2(Q)$-measurable function estimators.

Proof. By Theorem 1 of Yang and Barron (1999), the assertion holds for every statistical model if (35) is replaced by

$$\begin{aligned}\log N(\epsilon, \mathcal{F}, K^{1/2}) &\le V^*(\epsilon),\\ \log D(\epsilon, \mathcal{F}, \|\cdot\|_{2,Q}) &\ge V_*(\epsilon),\end{aligned}$$

for the KL divergence K. Let $p_f(x, y) = (2\pi\sigma_0^2)^{-1/2} \exp\{-(y - f(x))^2/(2\sigma_0^2)\}q(x)$. One can easily observe that $K(p_{f_1}, p_{f_2}) = (2\sigma_0^2)^{-1}\|f_1 - f_2\|_{2,Q}^2$. The assertion in the lemma follows immediately. ■

The key to obtaining a sharp minimax lower bound γ_n is to establish the bounds V^* and V_* as tight as possible. In the Lemmas 9–10 below, we provide entropy estimates for the d-dimensional (non-sparse) piecewise heterogeneous anisotropic Hölder space. While an upper bound of the metric entropy is well known for isotropic classes (e.g., Theorem 2.7.1 of van der Vaart and Wellner (1996)), we believe that there is no available result on more complicated function space in the literature, even for the simple anisotropic classes in Definition 1. Lemma 11 concatenates the results to obtain entropy bounds for the sparse function space.

Below we write $\overline{\mathcal{H}}_{\lambda,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0) = \{h \in \mathcal{H}_\lambda^{A_{\bar{\alpha}},d}(\mathfrak{X}_0) : \|h\|_\infty \le M\lambda\}$ for $M > 0$. For the upper bound of the metric entropy, we consider a bound for the space $\overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0)$, which is not worse than that for $\overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0) \cap \mathcal{C}([0, 1]^d)$. This implies that the Le Cam equation gives the same minimax lower bound for the two spaces.

Lemma 9 (Covering number, upper bound)**.** *For $d > 0$, $R > 0$, a partition $\mathfrak{X}_0 = \{\Xi_1, \ldots, \Xi_R\}$ of $[0, 1]^d$, and a smoothness parameter $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ for $\bar{\alpha} \in (0, 1]$ such that $\log \mathsf{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$, $1 \le r \le R$, $1 \le j \le d$, there exist constants $\epsilon_0 > 0$ and $M_0 > 1$ such that for any $\epsilon < \epsilon_0$,*

$$\log N\big(\epsilon, \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0), \|\cdot\|_\infty\big) \le (M_0 d/\epsilon)^{d/\bar{\alpha}}. \tag{36}$$

Proof. To express the assumption more explicitly, let $C_1 > 0$ be a constant such that $\log \mathsf{len}([\Xi_r]_j) \ge -C_1/\alpha_{rj}$ for every r and j. For a sufficiently small $C_2 > 0$, choose $\delta_d \in (0, \min\{e^{-C_1}, C_2/d\})$ such that $\min_{r,j} \mathsf{len}([\Xi_r]_j)\delta_d^{-1/\alpha_{rj}} > 1$. On each box $\mathsf{cl}(\Xi_r)$, consider a Cartesian product of grid points,

$$\tilde{\mathcal{G}}_r := \prod_{j=1}^d \left\{I_{rj}^L, I_{rj}^L + u_{rj}, I_{rj}^L + 2u_{rj}, \ldots, I_{rj}^L + \mathsf{len}([\Xi_r]_j)\right\},$$

where $u_{rj} = \mathsf{len}([\Xi_r]_j)/\lceil \mathsf{len}([\Xi_r]_j)\delta_d^{-1/\alpha_{rj}}\rceil$ is the mesh-size and I_{rj}^L is the left-boundary of Ξ_r in coordinate j. Observe that

$$\tilde{m}_r := |\tilde{\mathcal{G}}_r| = \prod_{j=1}^d (1 + \lceil \mathsf{len}([\Xi_r]_j)\delta_d^{-1/\alpha_{rj}}\rceil) \leq \prod_{j=1}^d (2 + \mathsf{len}([\Xi_r]_j)\delta_d^{-1/\alpha_{rj}}) \leq \mathsf{vol}(\Xi_r) 3^d \delta_d^{-d/\bar{\alpha}}. \tag{37}$$

We write the elements of $\tilde{\mathcal{G}}_r$ as $x_r^\ell = (x_{r1}^\ell, \dots, x_{rd}^\ell)^\top$, i.e., $x_r^\ell \in \tilde{\mathcal{G}}_r$, $\ell = 1, \dots, \tilde{m}_r$, $r = 1, \dots, R$. For every $h \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0)$, we define the vector

$$Gh = \Big(\lfloor h(x_1^1)/\delta_d \rfloor, \dots, \lfloor h(x_1^{\tilde{m}_1})/\delta_d \rfloor, \dots, \lfloor h(x_R^1)/\delta_d \rfloor, \dots, \lfloor h(x_R^{\tilde{m}_R})/\delta_d \rfloor\Big)^\top.$$

Because mesh-size satisfies $u_{rj} \leq \delta_d^{1/\alpha_{rj}}$, for every $x = (x_1, \dots, x_d)^\top \in \Xi_r$ with given r, there exists a point $x_r^\ell \in \tilde{\mathcal{G}}_r$ such that $\sum_{j=1}^d |x_j - x_{rj}^\ell|^{\alpha_{rj}} \leq d\delta_d$. Hence, for every such x and x_r^ℓ, all functions $h_1, h_2 \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0)$ such that $Gh_1 = Gh_2$ satisfy

$$|h_1(x) - h_2(x)| \leq |h_1(x_r^\ell) - h_2(x_r^\ell)| + 2\sum_{j=1}^d |x_j - x_{rj}^\ell|^{\alpha_{rj}} \leq \delta_d + 2d\delta_d.$$

As this holds for every $1 \leq r \leq R$, it follows that $\|h_1 - h_2\|_\infty \leq 3d\delta_d$ for any h_1, h_2 such that $Gh_1 = Gh_2$. This means that, whenever $3d\delta_d < \epsilon_0$ for some small constant $\epsilon_0 > 0$, the covering number $N(3d\delta_d, \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0), \|\cdot\|_\infty)$ is bounded by the number of possible vectors Gh for h that ranges over $\overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0)$.

Without loss of generality, we now assume that $(x_r^\ell)_{\ell=1}^{\tilde{m}_r}$ in $\tilde{\mathcal{G}}_r$ are appropriately sorted so that every two successive values differ in only one coordinate by mesh-size; that is, for every $\ell > 1$, there exists $\ell' < \ell$ such that $\sum_{j=1}^d |x_{rj}^{\ell'} - x_{rj}^\ell|^{\alpha_{rj}} = u_{rj'}^{\alpha_{rj'}} \leq \delta_d$ for some j'. For the enumeration, we begin with the first element of Gh, which is defined with $x_1^1 \in \tilde{\mathcal{G}}_1$. As $\|h\|_\infty \leq M$, the number of possible values of $\lfloor h(x_1^1)/\delta_d \rfloor$ does not exceed $2M/\delta_d + 1$. For every remainder defined with $x_1^\ell \in \tilde{\mathcal{G}}_1$, $2 \leq \ell \leq \tilde{m}_1$, there exists $\ell' < \ell$ such that

$$\begin{aligned}
&|\lfloor h(x_1^{\ell'})/\delta_d \rfloor - \lfloor h(x_1^\ell)/\delta_d \rfloor| \\
&\quad \leq \delta_d^{-1}|h(x_1^{\ell'}) - h(x_1^\ell)| + |h(x_1^{\ell'})/\delta_d - \lfloor h(x_1^{\ell'})/\delta_d \rfloor| + |h(x_1^\ell)/\delta_d - \lfloor h(x_1^\ell)/\delta_d \rfloor| \\
&\quad \leq \delta_d^{-1}\sum_{j=1}^d |x_j^{\ell'} - x_j^\ell|^{\alpha_{rj}} + 2 \leq 3.
\end{aligned}$$

It follows that, for a given $\lfloor h(x_1^{\ell'})/\delta_d \rfloor$, the number of possible values of $\lfloor h(x_1^\ell)/\delta_d \rfloor$ is at most 7, which is the case for every $\ell > 1$. Putting the bounds together, the number of possible values of the first m_1 elements of Gh is bounded by $(2M/\delta_d + 1)7^{\tilde{m}_1 - 1}$. Next, because $h \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0)$ can be discontinuous at the boundaries of the pieces of $\mathfrak{X}_0$, the $(\tilde{m}_1 + 1)$th element of Gh, defined with $x_2^1 \in \tilde{\mathcal{G}}_2$, has no restriction. Similar to the case with $r = 1$ above, the number of possible values of $\lfloor h(x_2^1)/\delta_d \rfloor$ at most $2M/\delta_d + 1$, and the number of possible values of $\lfloor h(x_2^\ell)/\delta_d \rfloor$ is at most 7 for every $2 \le \ell \le \tilde{m}_2$. This concludes that the number of possible values of the next $\tilde{m}_2$ elements of Gh is bounded by $(2M/\delta_d+1)7^{\tilde{m}_2-1}$. Concatenating this for all r, the number of possible vectors Gh is clearly at most $\prod_{r=1}^R (2M/\delta_d+1)7^{\tilde{m}_r-1} = (2M/\delta_d+1)^R 7^{\tilde{m}-R}$, where $\tilde{m} = \sum_{r=1}^R \tilde{m}_r$. Using (37), it is evident that $\tilde{m} \le 3^d \delta_d^{-d/\bar{\alpha}}$ because $\sum_{r=1}^R \mathsf{vol}(\Xi_r) = 1$. Taking $\epsilon = 3d\delta_d$,

$$\log N(\epsilon, \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^d), \|\cdot\|_\infty) \le R \log(6Md/\epsilon + 1) + 3^d (3d/\epsilon)^{d/\bar{\alpha}} \log 7.$$

As $\log(6Md/\epsilon+1) \lesssim (6Md/\epsilon)^{d/\bar{\alpha}}$ and $\log R \lesssim d/\bar{\alpha}$ (by the condition $\log \mathsf{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$), the last expression is bounded by $(M_0 d/\epsilon)^{d/\bar{\alpha}}$ for some $M_0 > 0$. To complete the proof, we must now show that there exists a small constant $\epsilon_0 > 0$ such that $\epsilon = 3d\delta_d < \epsilon_0$. This is achieved by a sufficiently small $C_2 > 0$ since $\delta_d < C_2/d$. ■

Lemma 10 (Packing number, lower bound). *For $d > 0$, $R > 0$, a partition $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$ of $[0,1]^d$, and a smoothness parameter $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ for $\bar{\alpha} \in (0,1]$ such that $\log \mathsf{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$, $1 \le r \le R$, $1 \le j \le d$, there exist constants $\epsilon_1 > 0$ and $M_1 > 1$ such that for any $\epsilon < \epsilon_1^d$, there are $N \ge \exp\{1/(M_1^d \epsilon)^{d/\bar{\alpha}}\}$ functions $h_i \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^d)$, $i = 1, \dots, N$, and $h_0 = 0$ satisfying*

$$\int_{[0,1]^d} h_i(x) dx_j = 0, \quad 0 \le i \le N, \quad 1 \le j \le d, \tag{38}$$

$$\|h_i - h_k\|_2 \ge \epsilon, \quad 0 \le i \le k \le N. \tag{39}$$

Proof. Similar to above, let $C_1 \ge \log 8$ be a constant such that $\log \mathsf{len}([\Xi_r]_j) \ge -C_1/\alpha_{rj}$ for every r and j and choose a constant $\delta \in (0, \min\{e^{-C_1}, M\}]$ such that $\mathsf{len}([\Xi_r]_j)\delta^{-1/\alpha_{rj}} > 1$, $1 \le r \le R$, $1 \le j \le d$. On each box Ξ_r, consider a Cartesian product of grid points,

$$\mathcal{G}_r := \prod_{j=1}^d \left\{ I_{rj}^L + \frac{u_{rj}}{2}, I_{rj}^L + \frac{3u_{rj}}{2}, I_{rj}^L + \frac{5u_{rj}}{2}, \dots, I_{rj}^L + \mathsf{len}([\Xi_r]_j) - \frac{u_{rj}}{2} \right\},$$

where $u_{rj} = \mathsf{len}([\Xi_r]_j)/\lceil \mathsf{len}([\Xi_r]_j)\delta^{-1/\alpha_{rj}} \rceil$ is the mesh-size and I_{rj}^L is the left-boundary of Ξ_r in coordinate j (cf. the grid $\tilde{\mathcal{G}}_r$ used in the proof of Lemma 9). Note that

$$m_r := |\mathcal{G}_r| = \prod_{j=1}^d \lceil \mathsf{len}([\Xi_r]_j)\delta^{-1/\alpha_{rj}} \rceil \ge \mathsf{vol}(\Xi_r)\delta^{-d/\bar{\alpha}}. \tag{40}$$

We write the elements of $\mathcal{G}_r$ as $x_r^\ell = (x_{r1}^\ell, \dots, x_{rd}^\ell)^\top$, i.e., $x_r^\ell \in \mathcal{G}_r$, $\ell = 1, \dots, m_r$, $r = 1, \dots, R$. We define the univariate kernel $\mathcal{K}(t) = t\mathbb{1}(|t| \le 1/2) + (\mathrm{sgn}(t) - t)\mathbb{1}(1/2 < |t| \le 1)$, $t \in \mathbb{R}$, supported on $[-1,1]$. Clearly, $\mathcal{K}$ is 1-Lipschitz and satisfies $\int \mathcal{K}(t)dt = 0$.

We define the function

$$\phi_r^\ell(x) = \frac{\delta}{2^{d+1}} \prod_{j=1}^d \mathcal{K}\left(\frac{x_j - x_{rj}^\ell}{u_{rj}/2} \right), \quad 1 \le \ell \le m_r, \quad 1 \le r \le R,$$

which is supported on $\mathcal{X}_r^\ell := \prod_{j=1}^d [x_j^\ell - u_{rj}/2, x_j^\ell + u_{rj}/2]$ with the center x_r^ℓ. As $\|\mathcal{K}\|_\infty = 1/2$, we obtain $\|\phi_r^\ell\|_\infty \le \delta\|\mathcal{K}\|_\infty^d/2^{d+1} \le 1$ for a suitable $C_1 > 0$. Using the Lipschitz continuity of $\mathcal{K}$ and the inequality $|\prod_j a_j - \prod_j b_j| \le \sum_j |a_j - b_j|$ for any $a_j, b_j \in [-1,1]$, we have that for any x, y on the support $\mathcal{X}_r^\ell$,

$$|\phi_r^\ell(x) - \phi_r^\ell(y)| \le \frac{\delta}{2}\sum_{j=1}^d \left|\frac{x_j - y_j}{u_{rj}}\right| \le \frac{\delta}{2}\sum_{j=1}^d \left|\frac{x_j - y_j}{u_{rj}}\right|^{\alpha_{rj}} \le \sum_{j=1}^d |x_j - y_j|^{\alpha_{rj}},$$

where we used the inequalities $x \le x^a$ for any $x \in [0,1]$ and $a \in [0,1]$, and $u_{rj} \ge 1/(2\delta^{-1/\alpha_{rj}})$ as soon as $\mathsf{len}([\Xi_r]_j)\delta^{-1/\alpha_{rj}} \ge 1/2$ (note that $\lceil x \rceil \le 2x$ for $x \ge 1/2$). This shows that $\phi_r^\ell \in \mathcal{H}_1^{\alpha_r,d}(\mathcal{X}_r^\ell)$ for every $1 \le \ell \le m_r$ and $1 \le r \le R$. For a binary vector $\tilde{\omega}_r = (\tilde{\omega}_r^1, \dots, \tilde{\omega}_r^{m_r})^\top \in \{0,1\}^{m_r}$, define the continuous function $h_{\tilde{\omega}_r} = \sum_{\ell=1}^{m_r} \tilde{\omega}_r^\ell \phi_r^\ell$ supported on Ξ_r. As $\int \phi_r^\ell(x) dx_j = 0$ for every j and each ϕ_r^ℓ is a shifted copy of another, we obtain $\int h_{\tilde{\omega}_r}(x) dx_j = 0$ for every j and $h_{\tilde{\omega}_r} \in \mathcal{H}_1^{\alpha_r,d}(\Xi_r)$. Let $m = \sum_{r=1}^R m_r$, which satisfies $m \ge \delta^{-d/\bar{\alpha}}$ by (40). We write $\omega = (\omega_1, \dots, \omega_m)^\top = (\tilde{\omega}_1^\top, \dots, \tilde{\omega}_R^\top)^\top \in \{0,1\}^m$ and define $h_\omega = \sum_{r=1}^R h_{\tilde{\omega}_r}$. Then, as $\|h_\omega\|_\infty = \max_{r,\ell}\|\phi_r^\ell\|_\infty \le \delta \le M$ and each $h_{\tilde{\omega}_r}$ is zero at all points on the boundary of Ξ_r, it is easy to see that $h_\omega \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^d)$ and $\int h_\omega(x) dx_j = 0$. We also have that for any $\omega, \omega' \in \{0,1\}^m$,

$$\|h_\omega - h_{\omega'}\|_2^2 \ge \left[\sum_{b=1}^m (\omega_b - \omega_b')^2\right] \min_{r,\ell} \int [\phi_r^\ell(x)]^2 dx = \rho(\omega, \omega')\left(\frac{\delta^2\|\mathcal{K}\|_2^{2d}}{2^{3d+2}}\right) \min_r \prod_{j=1}^d u_{rj}, \quad (41)$$

where $\rho(\omega, \omega') = \sum_{b=1}^m \mathbb{1}(\omega_b \ne \omega_b')$ is the Hamming distance between ω and ω'. As $m \ge \delta^{-d/\bar{\alpha}} \ge \delta^{-1} \ge e^{C_1} > 8$, the Gilbert-Varshamov bound (Lemma 2.9 of Tsybakov (2008)) says that there exist $N \ge 2^{m/8}$ binary strings $\omega^{(1)}, \dots, \omega^{(N)} \in \{0,1\}^m$ such that $\rho(\omega^{(\ell)}, \omega^{(\ell')}) \ge m/8$, $0 \le \ell < \ell' \le N$, with $\omega^{(0)} = 0$. As $\min_r \prod_{j=1}^d u_{rj} \ge 1/(2^d \delta^{-d/\bar{\alpha}})$ and $\|\mathcal{K}\|_2^2 = 1/6$, the lower bound in (41) gives that for every $0 \le \ell < \ell' \le N$,

$$\|h_{\omega^{(\ell)}} - h_{\omega^{(\ell')}}\|_2^2 \ge \frac{m}{8}\left(\frac{\delta^2}{6^d 2^{3d+2}}\right) \min_r \prod_{j=1}^d u_{rj} \ge \frac{\delta^2}{2^5 96^d}.$$

Letting $\epsilon = \delta/\sqrt{2^5 96^d}$, the previous lower bound gives $\|h_{\omega^{(\ell)}} - h_{\omega^{(\ell')}}\|_2 \ge \epsilon$ while $N \ge 2^{m/8} \ge \exp(\delta^{-d/\bar{\alpha}}(\log 2)/8) \ge \exp(1/(2^8 96^{d/2}\epsilon)^{d/\bar{\alpha}})$. As δ is a constant, this holds for every $\epsilon < \epsilon_1/96^{d/2}$ for some $\epsilon_1 > 0$. ■

Lemma 11 (Entropy with sparsity). *For $d > 0$, $\lambda > 0$, $R > 0$, a partition $\mathfrak{X}_0 = \{\Xi_1, \dots, \Xi_R\}$ of $[0,1]^d$, and a smoothness parameter $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$ for $\bar{\alpha} \in (0,1]$ such that, there exist $\epsilon_2 > 0$ and $M_2 > 1$ such that for any $\epsilon < \epsilon_2$ and $\epsilon' < \epsilon_2^d$,*

$$\log N\big(\epsilon, \overline{\Gamma}_{\lambda,M}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0), \|\cdot\|_2\big) \le \log\binom{p}{d} + \Big(\frac{M_2 \lambda d}{\epsilon}\Big)^{d/\bar{\alpha}}, \quad (42)$$

$$\log D\big(\epsilon', \overline{\Gamma}_{\lambda,M}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p), \|\cdot\|_2\big) \ge \log\binom{p}{d} + \Big(\frac{\lambda}{M_2^d \epsilon'}\Big)^{d/\bar{\alpha}}. \quad (43)$$

Proof. We only need to verify the assertion for $\lambda = 1$ since $D(\epsilon, \lambda\mathcal{F}, \|\cdot\|_2) = D(\epsilon/\lambda, \mathcal{F}, \|\cdot\|_2)$ and $N(\epsilon, \lambda\mathcal{F}, \|\cdot\|_2) = N(\epsilon/\lambda, \mathcal{F}, \|\cdot\|_2)$ for any set $\mathcal{F}$. We first verify the upper bound (42). For every $\epsilon < \epsilon_0$, Lemma 9 gives $\log N(\epsilon, \overline{\mathcal{H}}^{A_{\bar{\alpha}},d}_{1,M}(\mathfrak{X}_0), \|\cdot\|_2) \le (M_0 d/\epsilon)^{d/\bar{\alpha}}$. As $\overline{\Gamma}^{A_{\bar{\alpha}},d,p}_{1,M}(\mathfrak{X}_0)$ is a union of $\binom{p}{d}$ many $\overline{\mathcal{H}}^{A_{\bar{\alpha}},d}_{1,M}(\mathfrak{X}_0)$, the assertion easily follows.

Next, we verify (43). By Lemma 10, for every $\epsilon' < \epsilon_1^d$, there are functions $h_0 = 0$, $h_i \in \overline{\mathcal{H}}^{A_{\bar{\alpha}},d}_{1,M}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^d)$, $1 \le i \le N$ satisfying (38) and (39), with $N \ge \exp\{1/(M_1^d \epsilon')^{d/\bar{\alpha}}\}$. This means that for any $S \subseteq \{1, \dots, p\}$ such that $|S| = d$, we have that $W_S^p h_i \in \overline{\Gamma}^{A_{\bar{\alpha}},d,p}_{1,M}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)$ for every such h_i , $0 \le i \le N$. Therefore,

$$\mathcal{W}(\epsilon') := \bigcup_{S \subseteq \{1,\dots,p\}: |S| = d} \{W_S^p h_i : 1 \le i \le N\} \subseteq \overline{\Gamma}^{A_{\bar{\alpha}},d,p}_{1,M}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p).$$

Now, for any $S \ne S' \subseteq \{1, \dots, p\}$ and $1 \le i \le k \le N$, observe that $\|W_S^p h_i - W_{S'}^p h_k\|_2 = (\|h_i\|_2^2 + \|h_k\|_2^2)^{1/2} \ge \epsilon'$ by (39), as $\langle W_S^p h_i, W_{S'}^p h_k \rangle = 0$ owing to (38), where we used $h_0 = 0$. Also for any $S \subseteq \{1, \dots, p\}$, it is easy to see that $\|W_S^p h_i - W_S^p h_k\|_2 = \|h_i - h_k\|_2 \ge \epsilon'$ by (39). These imply that $\mathcal{W}(\epsilon')$ is ϵ'-separated, and hence the packing number $D(\epsilon', \overline{\Gamma}^{A_{\bar{\alpha}},d,p}_{1,M}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p), \|\cdot\|_2)$ is bounded below by the cardinality of $\mathcal{W}(\epsilon')$, which is $\binom{p}{d} N$. This leads to the assertion. ∎

Proof of Theorem 3. Let the right-hand sides of (42) and (43) be $V^*(\epsilon)$ and $V_*(\epsilon)$, respectively. As $L_2(Q)$-norm can be replaced by L_2-norm under Assumption (M), Lemma 8 implies that a sequence γ_n is a minimax lower bound if $V_*(\gamma_n) = n\bar{\gamma}_n^2$ and $V^*(\bar{\gamma}_n) = n\bar{\gamma}_n^2$ for some $\bar{\gamma}_n$.

Let $\hat{\gamma}_n = \sqrt{n^{-1} \log \binom{p}{d}} + ((\lambda d)^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}$ and $\bar{\gamma}_n$ be the solution to $V^*(\bar{\gamma}_n) = n\bar{\gamma}_n^2$. As $V^*(\epsilon)$ is nondecreasing in ϵ, we obtain

$$V^*(M_2 \hat{\gamma}_n) \le V^*\Big(M_2((\lambda d)^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}\Big) = n\hat{\gamma}_n^2 \le M_2^2 n\hat{\gamma}_n^2.$$

This shows that $\bar{\gamma}_n \le M_2 \hat{\gamma}_n$. Now, define $\kappa_n = \max\Big\{\sqrt{n^{-1}\log\binom{p}{d}}, ((\lambda d)^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}\Big\}$. It follows that $V^*(\hat{\gamma}_n/2) \ge V^*(\kappa_n)$ because $\hat{\gamma}_n/2 \le \kappa_n$. If $\sqrt{n^{-1}\log\binom{p}{d}} \le ((\lambda d)^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}$,

$$V^*(\kappa_n) = V^*\Big(((\lambda d)^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}\Big) \ge M_2^{d/\bar{\alpha}} n\kappa_n^2 \ge n\hat{\gamma}_n^2/4,$$

as $M_2^{d/\bar{\alpha}} \ge 1$. If $\sqrt{n^{-1}\log\binom{p}{d}} > ((\lambda d)^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}$,

$$V^*(\kappa_n) = V^*\left(\sqrt{n^{-1}\log\binom{p}{d}}\right) \ge n\kappa_n^2 \ge n\hat{\gamma}_n^2/4.$$

Putting the bounds together, we obtain $\bar{\gamma}_n \ge \hat{\gamma}_n/2$. This concludes $\bar{\gamma}_n \asymp \hat{\gamma}_n$.

Now, let $\tilde{\gamma}_n = \sqrt{n^{-1}\log\binom{p}{d}} + M_2^{-d}(\lambda^{d/\bar{\alpha}}/(d^2 n))^{\bar{\alpha}/(2\bar{\alpha}+d)}$ and γ_n be the solution to $V_*(\gamma_n) = n\hat{\gamma}_n^2$. Then, it is easy to see that

$$V_*(\tilde{\gamma}_n) \le V_*\Big(M_2^{-d}(\lambda^{d/\bar{\alpha}}/(d^2 n))^{\bar{\alpha}/(2\bar{\alpha}+d)}\Big) = n\hat{\gamma}_n^2,$$

which implies $\gamma_n \le \tilde{\gamma}_n$. Let $\tilde{\kappa}_n = \max\left\{\sqrt{n^{-1}\log\binom{p}{d}}, M_2^{-d}(\lambda^{d/\bar{\alpha}}/(d^2n))^{\bar{\alpha}/(2\bar{\alpha}+d)}\right\}$ and note that $V_*(\tilde{\gamma}_n/2) \ge V_*(\tilde{\kappa}_n)$. Similar to the above, if $\sqrt{n^{-1}\log\binom{p}{d}} \le M_2^{-d}(\lambda^{d/\bar{\alpha}}/(d^2n))^{\bar{\alpha}/(2\bar{\alpha}+d)}$,

$$V_*(\tilde{\kappa}_n) = V_*\Big(M_2^{-d}(\lambda^{d/\bar{\alpha}}/(d^2n))^{\bar{\alpha}/(2\bar{\alpha}+d)}\Big) = n\hat{\gamma}_n^2,$$

and if $\sqrt{n^{-1}\log\binom{p}{d}} > M_2^{-d}(\lambda^{d/\bar{\alpha}}/(d^2n))^{\bar{\alpha}/(2\bar{\alpha}+d)}$,

$$V_*(\tilde{\kappa}_n) = V_*\left(\sqrt{n^{-1}\log\binom{p}{d}}\right) \ge n\tilde{\kappa}_n^2 \ge n\hat{\gamma}_n^2/4.$$

These give $\gamma_n \ge \tilde{\gamma}_n/2$, and hence $\gamma_n \asymp \tilde{\gamma}_n$. Lemma 8 concludes that $r_n\big(\overline{\Gamma}_{\lambda,M}^{A_{\bar{\alpha}},d,p}(\mathfrak{X}_0) \cap \mathcal{C}([0,1]^p)\big) \gtrsim \tilde{\gamma}_n$. As $M_2^{-d}d^{-2\alpha/(2\alpha+d)} \ge M_2^{-d}d^{-2}$, $\tilde{\gamma}_n$ is bounded below by the lower bound in Theorem 3 for some $M_d > 1$ depending only on d. ■

A.5 Proofs of Theorems 4–7

This section provides proofs of Theorems 4–7. The proofs are largely based on the proof of Theorem 2. We often refer to the reader to the proof of Theorem 2 rather than showing all details.

Proof of Theorem 4. Let p_{f,σ^2} be the density of model (8) with f and σ^2. By Lemma B.1 of Xie and Xu (2018), the Hellinger distance ρ_{H} satisfies

$$\|f_1 - f_2\|_{2,Q}^2 + |\sigma_1^2 - \sigma_2^2|^2 \lesssim \rho_{\mathrm{H}}^2(p_{f_1,\sigma_1^2}, p_{f_2,\sigma_2^2}) \lesssim \|f_1 - f_2\|_{1,Q} + |\sigma_1^2 - \sigma_2^2|^2, \tag{44}$$

if $f_1, f_2, \log\sigma_1, \log\sigma_2$ are uniformly bounded (we use variance parameters in place of standard deviations; both are identical up to constants under the boundedness assumption). Hence, it suffices to show the assertion with respect to the Hellinger distance.

By the well-known theory of posterior contraction (e.g., Theorem 2.1 of Ghosal et al. (2000)), we need to verify that there exists $\Theta_n \subseteq \mathcal{F} \times [\overline{C}_2^{-1}, \overline{C}_2]$ such that for some $\bar{c} > 0$ and a sufficiently large $\bar{c}' > 0$,

$$\Pi(B_n) \ge e^{-\bar{c}n\epsilon_n^2}, \tag{45}$$

$$\log N(\epsilon_n, \Theta_n, \rho_{\mathrm{H}}) \lesssim n\epsilon_n^2, \tag{46}$$

$$\Pi((f,\sigma^2) \notin \Theta_n) \ll e^{-\bar{c}'n\epsilon_n^2}, \tag{47}$$

similar to (26)–(28), where $B_n = \{f : K(p_0, p_{f,\sigma^2}) \le \epsilon_n^2, V(p_0, p_{f,\sigma^2}) \le \epsilon_n^2\}$. Using (44), the conditions (46) and (47) can be similarly verified as in the proof of Theorem 2; only difference is that we use truncated priors, so (47) is even more easily satisfied. For (45), note that by Lemma B.2 of Xie and Xu (2018),

$$\max\left\{K(p_0, p_{f,\sigma^2}), V(p_0, p_{f,\sigma^2})\right\} \lesssim \|f - f_0\|_{2,Q}^2 + |\sigma^2 - \sigma_0^2|,$$

as $\|f_0\|_\infty$ and $|\log \sigma_0|$ are bounded and the priors are truncated. Hence, there exists a constant $C_1 > 0$ such that

$$B_n \supseteq \{(f, \sigma^2) : \|f - f_0\|_{2,Q} \le C_1\epsilon_n, |\sigma^2 - \sigma_0^2| \le C_1\epsilon_n^2\}.$$

Note that $\|f - f_0\|_{2,Q} \lesssim \|f - f_0\|_2$ if the density of Q is bounded. It is easy to see that $\log \Pi(\sigma^2 : |\sigma^2 - \sigma_0^2| \le C_1\epsilon_n^2) \gtrsim -\log n$, as $|\log \sigma_0^2|$ is bounded. Uisng Lemmas 4–7, the rest of the proof follows similarly to that of Theorem 2. ∎

Proof of Theorem 5. It is well known that the Hellinger distance possesses an exponentially powerful local test with respect to both the type-I and type-II errors (e.g., Section 7 of Ghosal et al. (2000) or Lemma 2 of Ghosal and van der Vaart (2007)). Therefore by the general posterior contraction theory, it suffices to show that there exists $\Theta_n \subseteq \mathcal{F}$ such that for some $\bar{c} > 0$ and a sufficiently large $\bar{c}' > 0$,

$$\Pi(B_n) \ge e^{-\bar{c}n\epsilon_n^2}, \tag{48}$$

$$\log N(\epsilon_n, \Theta_n, \rho_{\mathrm{H}}) \lesssim n\epsilon_n^2, \tag{49}$$

$$\Pi(f \notin \Theta_n) \ll e^{-\bar{c}' n\epsilon_n^2}, \tag{50}$$

where $B_n = \{f : K(p_0, p_f) \le \epsilon_n^2, V(p_0, p_f) \le \epsilon_n^2\}$. The last condition (50) follows directly from the proof of Theorem 2, so we only need to verify (48) and (49).

By Lemma 3.1 of van der Vaart and van Zanten (2008), for any measurable f, g,

$$\begin{aligned} K(p_f, p_g) &\lesssim \|f - g\|_\infty^2 e^{\|f-g\|_\infty}(1 + \|f - g\|_\infty), \\ V(p_f, p_g) &\lesssim \|f - g\|_\infty^2 e^{\|f-g\|_\infty}(1 + \|f - g\|_\infty)^2, \\ \rho_{\mathrm{H}}(p_f, p_g) &\le \|f - g\|_\infty e^{\|f-g\|_\infty/2}. \end{aligned} \tag{51}$$

(The uniform norm is used in van der Vaart and van Zanten (2008) but can be easily replaced by the L_∞-norm.) The first two assertions imply that there exists $C_1 > 0$ such that $B_n \supseteq \{f : \|f - f_0\|_\infty \le C_1\epsilon_n\}$ if $\epsilon_n \to 0$. Hence we follow the calculation in the proof of Theorem 2 to conclude that there exists a constant $\bar{c} > 0$ such that $\Pi(B_n) \ge e^{-\bar{c}n\epsilon_n^2}$. The last assertion of (51) enables us to work with the supremum norm in the calculation of the Hellinger covering number. The entropy calculation in Theorem 2 also verifies (49), completing the proof. ∎

Proof of Theorem 6. Denote by $p_f(x, y)$ the density of model (11) and by $p_0(x, y)$ the true density. We also write $f_0 = H^{-1}(\varphi_0)$. From the fact that $|p_f(0|x) - p_0(0|x)| = |p_f(1|x) - p_0(1|x)| = |H(f(x)) - H(f_0(x))|$, it follows that $\|p_f - p_0\|_2 = \sqrt{2}\|H(f) - H(f_0)\|_{2,Q}$. The L_2-norm is bounded by a multiple of the Hellinger distance as p_f and p_0 are uniformly bounded, (see, for example, Lemma B.1 of Ghosal and van der Vaart (2017)). Hence, it suffices to show the contraction rate results with respect to the Hellinger distance. This means that the assertion can be verified if there exists $\Theta_n \subseteq \mathcal{F}$ satisfying (48)–(50) for some $\bar{c} > 0$. By Lemma 2.8 of Ghosal and van der Vaart (2017), $K(p_0, p_f) \lesssim \|f - f_0\|_{2,Q}^2$ and $V(p_0, p_f) \lesssim \|f - f_0\|_{2,Q}^2$. We also have that $\rho_{\mathrm{H}}(p_f, p_g) \lesssim \|f - g\|_{2,Q}$ for every measurable f, g by the same lemma. Similar to the proof of Theorem 4, the proof is completed by following that of Theorem 2. ■

Proof of Theorem 7. It suffices to verify (26)–(28) for the given model. Following the proof of Theorem 2, one can easily see that (26) is verified as soon as

$$\log \Pi(\widehat{\mathcal{E}}) + \log \Pi(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_0\|_n \le C_1 \epsilon_n^*) \gtrsim -n(\epsilon_n^*)^2, \tag{52}$$

for an approximating ensemble $\widehat{\mathcal{E}}$. Assumption (A6) says that for each $1 \le t \le T_0$, there exists a $\mathcal{Z}$-tree partition $\widehat{\mathcal{T}}^t$ such that $\|\hat{f}_{0t} - f_{0t}\|_n \lesssim \bar{\epsilon}_{t,n}$ for some $\hat{f}_{0t} \in \mathcal{F}_{\widehat{\mathcal{T}}^t}$. We index $\widehat{\mathcal{E}} = (\widehat{\mathcal{T}}^1, \dots, \widehat{\mathcal{T}}^T)$ with $\widehat{\mathcal{T}}^t = \{[0,1]^p\}$, $t = T_0 + 1, \dots, T$. Then,

$$\log \Pi(\widehat{\mathcal{E}}) = \sum_{t=1}^{T_0} \log \Pi(\widehat{\mathcal{T}}^t) + (T - T_0) \log(1 - \nu) \gtrsim -\sum_{t=1}^{T_0} \widehat{K}^t \log n - \sum_{t=1}^{T_0} d_t \log p \gtrsim -n(\epsilon_n^*)^2,$$

by Lemma 4. Constructing $\hat{f}_{0t}$ as in the proof of Theorem 1, we denote every $\hat{f}_{0t}$ by $f_{0t,\widehat{\mathcal{T}}^t,\widehat{\beta}^t}$, where $\widehat{\beta}^t$ is the corresponding step-heights. Then the approximator of f_0 can be expressed as $f_{0,\widehat{\mathcal{E}},\widehat{B}} = \sum_{t=1}^{T_0} f_{0t,\widehat{\mathcal{T}}^t,\widehat{\beta}^t}$ with the ensemble components $(\widehat{\mathcal{E}}, \widehat{B})$, where $\widehat{B} = (\widehat{\beta}^{1\top}, \dots, \widehat{\beta}^{T_0\top}, 0, \dots, 0)^\top \in \mathbb{R}^{\widehat{K}_*}$ with $\widehat{K}_* = \sum_{t=1}^{T_0} \widehat{K}^t + T - T_0$. This gives us that

$$\|f - f_0\|_\infty \le \|f - f_{0,\widehat{\mathcal{E}},\widehat{B}}\|_\infty + \sum_{t=1}^{T_0} \|f_{0t,\widehat{\mathcal{T}}^t,\widehat{\beta}^t} - f_{0t}\|_\infty \lesssim \|f - f_{0,\widehat{\mathcal{E}},\widehat{B}}\|_\infty + \sum_{t=1}^{T_0} \epsilon_{t,n}.$$

Therefore, using $\sum_{t=1}^{T_0} \epsilon_{t,n} \le \sqrt{T_0}\epsilon_n^*$, we obtain that

$$\Pi(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_0\|_\infty \le C_1 \epsilon_n^*) \ge \Pi\left(f \in \mathcal{F}_{\widehat{\mathcal{E}}} : \|f - f_{0,\widehat{\mathcal{E}},\widehat{B}}\|_\infty \le C_2 \epsilon_n^*\right). \tag{53}$$

For any $B_1 = (\beta_1^{1\top}, \dots, \beta_1^{T_0\top}, \beta_1^{T_0+1}, \dots, \beta_1^{T})^\top, B_2 = (\beta_2^{1\top}, \dots, \beta_2^{T_0\top}, \beta_2^{T_0+1}, \dots, \beta_2^{T})^\top \in \mathbb{R}^{\widehat{K}_*}$, we write $f_{\widehat{\mathcal{E}},B_1}$, $f_{\widehat{\mathcal{E}},B_2} \in \mathcal{F}_{\widehat{\mathcal{E}}}$ to denote two additive tree functions that lie on the same partition ensemble $\widehat{\mathcal{E}}$. From (2), it is easy to see that $\|f_{\widehat{\mathcal{E}},B_1} - f_{\widehat{\mathcal{E}},B_2}\|_\infty \le \|B_1 - B_2\|_2 \widehat{K}_*^{1/2}$. As $\widehat{K}_* \log n \lesssim \sum_{t=1}^{T_0} \widehat{K}^t \log n \lesssim n(\epsilon_n^*)^2$, one can follow the proof of Theorem 2 to lower bound the logarithm of (53) by a constant multiple of $-n(\epsilon_n^*)^2$. Combined with the lower bound of $\Pi(\widehat{\mathcal{E}})$, this verifies (52). The conditions in (27) and (28) follow directly from the proof of Theorem 2, but with the rate ϵ_n^* for the additive regression. ■

Appendix B. Auxiliary Result: Dirichlet Prior Concentration

The following lemma is a slight modification of Theorem 2.1 of Yang and Dunson (2014). We provide the complete proof for a self-contained result. Similar results are also available in the literature (e.g., Lemma G.13 of Ghosal and van der Vaart (2017)).

Lemma 12 (Concentration of Dirichlet priors)**.** *Suppose that $\eta \in \mathbb{S}^p$ has a Dirichlet prior in* (3) *with $\zeta > 0$ and $\xi > 1$. For any $\eta^* \in \mathbb{S}^p$ such that $\sum_{j=1}^p \mathbb{1}(\eta_j^* \neq 0) = s$ and any*

$\epsilon \in (0,1)$*, there exists a constant* $C > 0$ *such that*

$$\Pi(\|\eta - \eta^*\|_1 \le \epsilon) \ge \exp\{-C\xi s \log(p/\epsilon)\}, \tag{54}$$

$$\Pi\left(\min_{S:|S|=s} \sum_{j \notin S} \eta_j \ge \epsilon \right) \le \exp\{-C(\xi - 1)s \log p - \log \epsilon\}. \tag{55}$$

Proof. We first prove (54). Without loss of generality, we assume that the index set of nonzero entries of η^* is $\{1, 2, \dots, s-1, p\}$, i.e., $\eta_j^* = 0$, $j = s, s+1, \dots, p-1$. By the inequality $|\eta_p - \eta_p^*| = |\sum_{j=1}^{p-1} \eta_j - \sum_{j=1}^{p-1} \eta_j^*| \le \sum_{j=1}^{p-1} |\eta_j - \eta_j^*|$, observe that $\|\eta - \eta^*\|_1 \le 2\sum_{j=1}^{p-1} |\eta_j - \eta_j^*| = 2\sum_{j=1}^{s-1} |\eta_j - \eta_j^*| + 2\sum_{j=s}^{p-1} \eta_j$. Hence, for $b_0 = \epsilon/(4s)$ and $b_1 = \epsilon/(4p - 4s)$,

$$\begin{aligned} \mathcal{S} &= \{\eta \in \mathbb{S}^p : |\eta_j - \eta_j^*| \le b_0, j = 1, \dots, s-1, \eta_j \in (0, b_1], j = s, \dots, p-1\} \\ &\subseteq \{\eta \in \mathbb{S}^p : \|\eta - \eta^*\|_1 \le \epsilon\}. \end{aligned}$$

Using this, we obtain

$$\begin{aligned} &\Pi(\|\eta - \eta^*\|_1 \le \epsilon) \ge \Pi(\mathcal{S}) \\ &\qquad = \int_{\mathcal{S}} \frac{\Gamma(\zeta/p^{\xi-1})}{\Gamma^p(\zeta/p^\xi)} \prod_{j=1}^{p-1} \eta_j^{\zeta/p^\xi - 1} \left(1 - \sum_{j=1}^{p-1} \eta_j\right)^{\zeta/p^\xi - 1} d\eta_1 \dots d\eta_{p-1} \\ &\qquad \ge \frac{\Gamma(\zeta/p^{\xi-1})}{\Gamma^p(\zeta/p^\xi)} \left\{ \prod_{j=1}^{s-1} \int_{\max\{0, \eta_j^* - b_0\}}^{\min\{1, \eta_j^* + b_0\}} \eta_j^{\zeta/p^\xi - 1} d\eta_j \right\} \left\{ \prod_{j=s}^{p-1} \int_0^{b_1} \eta_j^{\zeta/p^\xi - 1} d\eta_j \right\}, \end{aligned}$$

where we used the fact that $\eta_p \le 1$ and $\zeta/p^\xi - 1 < 0$ for large enough p. As the Taylor expansion of Γ gives that $x\Gamma(x) = 1 - \gamma_0 x + O(x^2)$ for the Euler-Mascheroni constant γ_0, we obtain $\Gamma(x) \asymp 1/x$ for every small enough x. Therefore, the last display is bounded below by a constant multiple of

$$\begin{aligned} \frac{(\zeta/p^\xi)^p}{\zeta/p^{\xi-1}} (2b_0)^{s-1} \left(\frac{p^\xi}{\zeta} b_1^{\zeta/p^\xi} \right)^{p-s} &= \zeta^{s-1} p^{-\xi(s-1)-1} \left(\frac{\epsilon}{2s}\right)^{s-1} \left(\frac{\epsilon}{4p - 4s}\right)^{\zeta p^{-(\xi-1)}(1-s/p)} \\ &\ge \zeta^{s-1} p^{-\xi(s-1)-1} \left(\frac{\epsilon}{2s}\right)^{s-1} \left(\frac{\epsilon}{4p}\right)^{\zeta}, \end{aligned}$$

where for the inequality we used the fact that $\xi \ge 1$. The logarithm of the rightmost side leads to the desired assertion.

Now, we verify (55). Consider a Dirichlet process $\mathrm{DP}(\zeta/p^{\xi-1}, Q_0)$ with concentration parameter $\zeta/p^{\xi-1}$ and uniform measure Q_0 on $[0,1]$. Suppose a random measure $P \sim \mathrm{DP}(\zeta/p^{\xi-1}, Q_0)$. Then, for the intervals $\mathcal{I}_j = [j-1)/p, j/p)$, $j = 1, \dots, p$, we have

$$(P(\mathcal{I}_1), \dots, P(\mathcal{I}_p)) \sim \mathrm{Dir}(\zeta/p^\xi, \dots, \zeta/p^\xi).$$

This allows us to define η as $\eta = (P(\mathcal{I}_1), \dots, P(\mathcal{I}_p))^\top$ using the Dirichlet process above. The stick-breaking representation of a Dirichlet process gives an expression $P = \sum_{k=1}^\infty w_k \delta_{z_k}$ for

$z_k \sim Q_0$ and

$$w_k = v_k \prod_{j=1}^{k-1}(1-v_j), \quad v_k \sim \mathrm{Beta}(1, \zeta/p^{\xi-1}).$$

For every k, let j_k be the index such that $z_k \in \mathcal{I}_{j_k}$. It follows that

$$\max_{S:|S|\le s} \sum_{j\in S} \eta_j \ge \sum_{k=1}^{s} \eta_{j_k} = \sum_{k=1}^{s} P(\mathcal{I}_{j_k}) = \sum_{1\le \ell<\infty : z_\ell \in \cup_{k=1}^{s} \mathcal{I}_{j_k}} w_\ell \ge \sum_{k=1}^{s} w_k,$$

where the last inequality holds as $z_k \in \mathcal{I}_{j_k}$, $k = 1, \dots, s$. This gives that

$$\min_{S:|S|=s} \sum_{j\notin S} \eta_j \le 1 - \sum_{k=1}^{s} w_k = 1 - \sum_{k=1}^{s} v_k \prod_{j=1}^{k-1}(1-v_j) = \prod_{j=1}^{s}(1-v_j),$$

where the last equality can be verified by induction. Letting $\bar{v}_j = 1 - v_j \sim \mathrm{Beta}(\zeta/p^{\xi-1}, 1)$, $j = 1, \dots, s$, we obtain

$$\Pi\left(\min_{S:|S|=s} \sum_{j\notin S} \eta_j \ge \epsilon \right) \le \Pi\left(\prod_{j=1}^{s} \bar{v}_j \ge \epsilon \right) \le \frac{\zeta^s}{\epsilon(\zeta + p^{\xi-1})^s} \le \epsilon^{-1}\zeta^s p^{-s(\xi-1)},$$

using the Markov inequality. The rightmost side verifies the assertion. ■

References

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Anjishnu Banerjee, David B Dunson, and Surya T Tokdar. Efficient Gaussian process regression for large datasets. *Biometrika*, 100(1):75–89, 2013.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, 1999.

Jon Louis Bentley. Multidimensional binary search trees in database applications. *IEEE Transactions on Software Engineering*, SE-5(4):333–340, 1979.

Anirban Bhattacharya, Debdeep Pati, and David Dunson. Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics*, 42(1):352, 2014.

Lucien Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71(2):271–291, 1986.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

Justin Bleich, Adam Kapelner, Edward I George, and Shane T Jensen. Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics*, 8(3): 1750–1781, 2014.

Emmanuel J Candès and David L Donoho. Curvelets: A surprisingly effective nonadaptive representation for objects with edges. In *Curve and surface fitting*, pages 105–120. Vanderbilt University Press, 2000.

Emmanuel J Candès and David L Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, 2004.

Ismaël Castillo and Veronika Ročková. Uncertainty quantification for Bayesian CART. *The Annals of Statistics*, 49(6):3482–3509, 2021.

Venkat Chandrasekaran, Michael B Wakin, Dror Baron, and Richard G Baraniuk. Representation and compression of multidimensional piecewise functions using surflets. *IEEE Transactions on Information Theory*, 55(1):374–400, 2008.

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.

David GT Denison, Bani K Mallick, and Adrian FM Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

David L Donoho. CART and best-ortho-basis: A connection. *The Annals of Statistics*, 25 (5):1870–1911, 1997.

Junliang Du and Antonio R Linero. Interaction detection with Bayesian decision tree ensembles. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 108–117, 2019.

Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

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.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Kanghui Guo and Demetrio Labate. Optimally sparse multidimensional representation using shearlets. *SIAM Journal on Mathematical Analysis*, 39(1):298–318, 2007.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020.

Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

Jingyu He, Saar Yalov, and P Richard Hahn. XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138, 2019.

Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278, 2020.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

M Hoffman and Oleg Lepski. Random rates in anisotropic regression. *The Annals of Statistics*, 30(2):325–396, 2002.

Il'dar Abdulovich Ibragimov and Rafail Zalmanovich Hasminskii. *Statistical Estimation: Asymptotic Theory*. Springer, 1981.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 869–878, 2019.

Seonghyun Jeong and Subhashis Ghosal. Posterior contraction in sparse generalized linear models. *Biometrika*, 108(2):367–379, 2021a.

Seonghyun Jeong and Subhashis Ghosal. Unified Bayesian theory of sparse linear regression with nuisance parameters. *Electronic Journal of Statistics*, 15(1):3040–3111, 2021b.

Bereket P Kindo, Hao Wang, and Edsel A Peña. Multinomial probit Bayesian additive regression trees. *Stat*, 5(1):119–131, 2016.

Balaji Lakshminarayanan, Daniel Roy, and Yee Whye Teh. Top-down particle filtering for Bayesian decision trees. In *International Conference on Machine Learning*, pages 280–288, 2013.

Erwan Le Pennec and Stéphane Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, 2005.

Oleg Lepski. Adaptive estimation over anisotropic functional classes via oracle approach. *The Annals of Statistics*, 43(3):1178–1242, 2015.

OV Lepski and B Ya Levit. Adaptive non-parametric estimation of smooth multivariate functions. *Mathematical Methods of Statistics*, 8:344–370, 1999.

Yinpu Li, Antonio R Linero, and Jared Murray. Adaptive conditional distribution estimation with Bayesian decision tree ensembles. *Journal of the American Statistical Association*, pages 1–14, 2022.

Sunwoo Lim and Seonghyun Jeong. Synergizing roughness penalization and basis selection in Bayesian spline regression. *arXiv preprint arXiv:2311.13481*, 2023.

Antonio R Linero. A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6), 2017.

Antonio R Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.

Antonio R Linero. Generalized Bayesian additive regression trees models: Beyond conditional conjugacy. *arXiv preprint arXiv:2202.09924*, 2022.

Antonio R Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

Han Liu, John Lafferty, and Larry Wasserman. Sparse nonparametric density estimation in high dimensions using the rodeo. In *The 11th International Conference on Artificial Intelligence and Statistics*, pages 283–290, 2007.

Ziyue Liu and Wensheng Guo. Data driven adaptive spline smoothing. *Statistica Sinica*, 20 (3):1143–1163, 2010.

Jared S Murray. Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769, 2021.

Michael H Neumann and Rainer and von Sachs. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *The Annals of Statistics*, 25(1):38–76, 1997.

Bo Ning, Seonghyun Jeong, and Subhashis Ghosal. Bayesian linear regression for multivariate responses under group sparsity. *Bernoulli*, 26(3):2353–2382, 2020.

Michael Nussbaum. Spline smoothing in regression models and asymptotic efficiency in L_2. *The Annals of Statistics*, 13(3):984–997, 1985.

Vittorio Orlandi, Jared Murray, Antonio Linero, and Alexander Volfovsky. Density regression with Bayesian additive regression trees. *arXiv preprint arXiv:2112.12259*, 2021.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Alexandre Pintore, Paul Speckman, and Chris C Holmes. Spatially adaptive smoothing splines. *Biometrika*, 93(1):113–125, 2006.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108 (504):1339–1349, 2013.

Matthew T Pratola, Hugh A Chipman, Edward I George, and Robert E McCulloch. Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417, 2020.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

Kolyan Ray and Aad van der Vaart. Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.

Veronika Ročková. On semi-parametric Bernstein-von Mises theorems for BART. In *The 37th International Conference on Machine Learning*, 2020.

Veronika Ročková and Enakshi Saha. On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848, 2019.

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

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Weining Shen and Subhashis Ghosal. Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213, 2015.

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *The 7th International Conference on Learning Representations*, 2019.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *The 35th Conference on Neural Information Processing Systems*, pages 3609–3621, 2021.

Matthew A Taddy, Robert B Gramacy, and Nicholas G Polson. Dynamic trees for learning and design. *Journal of the American Statistical Association*, 106(493):109–123, 2011.

Yaoyuan Vincent Tan and Jason Roy. Bayesian additive regression trees and the general BART model. *Statistics in Medicine*, 38(25):5048–5069, 2019.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Rui Tuo and CF Jeff Wu. Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352, 2015.

Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Nakul Verma, Samory Kpotufe, and Sanjoy Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? In *The 25th Conference on Uncertainty in Artificial Intelligence*, pages 565–574, 2009.

Xiao Wang, Pang Du, and Jinglai Shen. Smoothing splines with varying smoothing parameter. *Biometrika*, 100(4):955–970, 2013.

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.

Fangzheng Xie and Yanxun Xu. Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Analysis*, 15 (1):159–186, 2018.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Yun Yang and David B Dunson. Minimax optimal Bayesian aggregation. *arXiv preprint arXiv:1403.1345*, 2014.

Yun Yang and Surya T Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.