

# The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension

Seonghyun Jeong<sup>\*1,2</sup> and Veronika Ročková<sup>†3</sup>

<sup>1</sup>Department of Statistics and Data Science, Yonsei University

<sup>2</sup>Department of Applied Statistics, Yonsei University

<sup>3</sup>Booth School of Business, University of Chicago

February 22, 2022

## Abstract

Considerable effort has been directed to developing asymptotically minimax procedures in problems of recovering functions and densities. These methods often rely on somewhat arbitrary and restrictive assumptions such as isotropy or spatial homogeneity. This work enhances theoretical understanding of Bayesian forests (including BART) under substantially relaxed smoothness assumptions. In particular, 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. Next, we find that the default 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 which would render other default machine learning tools, such as Gaussian processes, suboptimal. Our numerical study shows that Bayesian forests 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 show that Bayesian forests can be successfully applied to many other problems including density estimation and binary classification.

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

---

<sup>\*</sup>Seonghyun Jeong is Assistant Professor at the Department of Statistics and Data Science and the Department of Applied Statistics of Yonsei University. ([sjeong@yonsei.ac.kr](mailto:sjeong@yonsei.ac.kr)).

<sup>†</sup>Veronika Ročková is Associate Professor in Econometrics and Statistics and James S. Kemper Faculty Scholar at the Booth School of Business of the University of Chicago ([veronika.rockova@chicagobooth.edu](mailto:veronika.rockova@chicagobooth.edu)). The authors gratefully acknowledge the support from the James S. Kemper Faculty Research Fund at the Booth School of Business and the National Science Foundation (Grant number DMS 1944740).

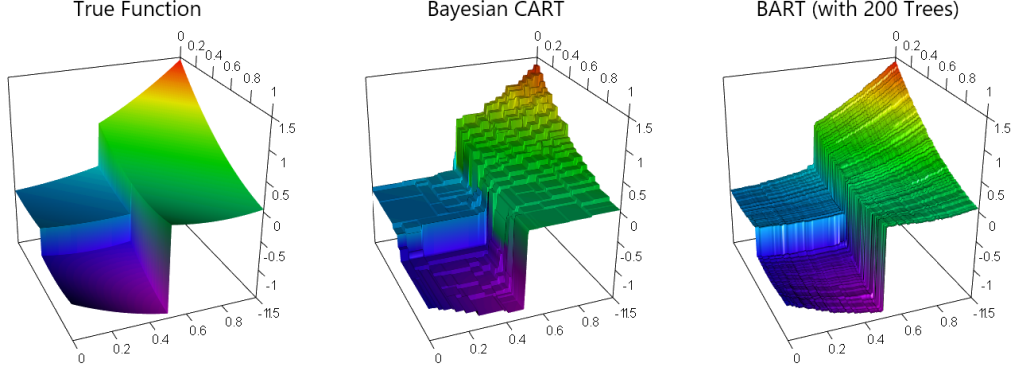
# 1 Introduction

## 1.1 Motivation

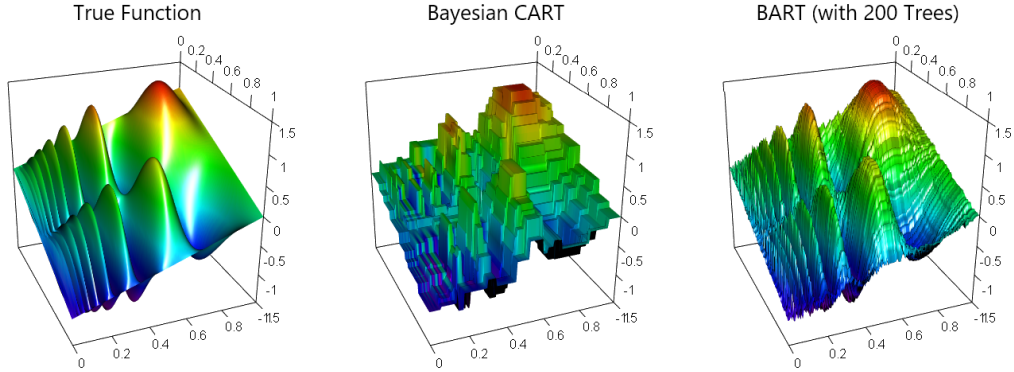
Many of the existing asymptotic minimaxity results for estimating regression functions are predicated on the assumption that certain smoothness conditions hold, which can be rarely satisfied/verified when confronted with real data. This creates a disconnect between theory and practice, limiting the scope of many of the theoretical results. For example, in non-parametric 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 work is motivated by the desire to evaluate theoretical performance of Bayesian forests, one of the workhorses of Bayesian machine learning, in such broad scenarios.

Bayesian trees and their ensembles have achieved a 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) and time series (Taddy et al., 2011), to list a few. Despite the remarkable success in empirical studies, theoretical properties of Bayesian forests were unavailable for a long time with first studies emerging only very recently (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 very well, they are limited to isotropic regression function surfaces, which exhibit the same level of smoothness in every direction. Isotropy is one of the archetypal assumptions in theoretical studies, but can be restrictive in real-world applications. This assumption is particularly unattractive in higher dimensions where the function can behave very poorly in certain directions.

However, from the plentiful empirical evidence, 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 for at least three reasons: (i) tree methods are based on top-down recursive partitioning, where 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 split is data-driven, making the domain divided 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*.



(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.

## 1.2 Our contribution

The main goal of this paper is to study optimality and posterior contraction of Bayesian forests under relaxed smoothness assumptions. More specifically, we introduce a class of functions whose domain has been cleaved into hyper-rectangles where each rectangular piece has its 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 establish 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 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, it is unclear whether they can adapt to piecewise heterogeneous anisotropic spaces without substantial modification. Bayesian forests, on the other hand, 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. The dyadic CART (Donoho, 1997), on the other hand, 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 which are sufficiently evenly distributed. By allowing the split-points 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 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 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 func-

tion 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 that our obtained contraction rates for Bayesian forests are indeed minimax-optimal, up to a logarithmic factor.

We summarize the contribution of this paper as follows.

- **Approximation theory:** The true function should be approximable by tree-based learners in order 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 due to the limitation of piecewise constant learners).
- **Posterior contraction:** For functions belonging to piecewise heterogeneous anisotropic spaces, 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 proving technique 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 preparatory phase, here we provide a preview of our main results. Let us focus on a fixed design regression setup,

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

with a response  $Y_i \in \mathbb{R}$  and a covariate  $x_i \in [0, 1]^p$ , where  $f_0 : [0, 1]^p \mapsto \mathbb{R}$  and  $\sigma_0^2 < \infty$ . Assume that  $f_0$  depends only on  $d$  variables among the  $p$  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 BART priors on  $f_0$ , the posterior contraction rate is obtained as  $\sqrt{(d \log p)/n} + (\log n)^{c - \bar{\alpha}/(2\bar{\alpha} + d)}$  for some  $c > 0$  ([Theorem 2](#)). It is also shown that this rate is minimax-optimal up to a log factor ([Theorem 3](#)). The same contraction rates are also achieved in other statistical setups ([Theorems 4–7](#)).

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. Section 4 sheds light on 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 with discussion. All technical proofs are collected in Appendix.

## 2 Modus operandi

### 2.1 Notation and terminology

Although the main focus of this paper is BART for regression in (1), we work with a general statistical experiment  $P_f$  indexed by a measurable function  $f : [0, 1]^p \mapsto \mathbb{R}$  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  $\mathbb{E}_0$  and  $\mathbb{P}_0$  denote the expectation and probability operators 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  $a_n \leq Cb_n$  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  $a_n/b_n \rightarrow 0$  as  $n \rightarrow \infty$ . For a subspace  $E$  of the Euclidean space,  $\mathcal{C}(E)$  denotes a class of continuous functions  $f : E \mapsto \mathbb{R}$ . For a given measure  $\mu$  and a measurable function  $f$ , we denote by  $\|f\|_{v,\mu} = (\int |f|^v d\mu)^{1/v}$  the  $L_v(\mu)$ -norm,  $1 \leq v < \infty$ . We particularly denote by  $\mathcal{L}_2(\mu)$  the linear space of real valued functions equipped with inner product  $\langle f, g \rangle_\mu = \int fgd\mu$  and norm  $\|f\|_{2,\mu} = \langle f, f \rangle_\mu^{1/2}$ . The support of a measure  $\mu$  is denoted by  $\text{supp}(\mu)$ . The supremum norm of a function  $f$  is denoted by  $\|f\|_\infty$ . For convenience, we write  $\|f\|_v$  for the  $L_v$ -norm with the Lebesgue measure on a unit hypercube. For a given vector  $u$ , the notations  $\|u\|_v$  and  $\|u\|_\infty$  represent the  $\ell_v$ -norms,  $1 \leq v < \infty$ , and the maximum-norm, respectively. For a semimetric space  $(\mathcal{F}, \rho)$  endowed with a semimetric  $\rho$ , the expressions  $D(\epsilon, \mathcal{F}, \rho)$  and  $N(\epsilon, \mathcal{F}, \rho)$  are  $\epsilon$ -packing and  $\epsilon$ -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  $p$ -dimensional hyper-rectangle  $\Psi \subseteq [0, 1]^p$  with any  $p > 0$  is simply called a *box*. Boxes can be closed (the form of  $\prod_{j=1}^p [a_j, b_j]$ ) or half-closed (the form of  $\prod_{j=1}^p (a_j, b_j]$ ) depending on the context. A partition  $\mathfrak{V} = (\Psi_1, \dots, \Psi_J)$  of  $[0, 1]^p$ , consisting of  $J$  disjoint boxes  $\Psi_r \subseteq [0, 1]^p$ ,  $r = 1, \dots, J$ , is called a *box partition*. For a  $p$ -dimensional Cartesian product  $E \subseteq \mathbb{R}^p$  with any  $p > 0$ , we denote the  $j$ th projection mapping of  $E$  by  $[E]_j = \{x_j \in \mathbb{R} : (x_1, \dots, x_p)^\top \in E\}$ . The length and interior of an interval  $I \in \mathbb{R}$  is denoted by  $\text{len}(I)$  and  $\text{int}(I)$ , 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  $f_0 : [0, 1]^p \mapsto \mathbb{R}$  is  $d$ -sparse, i.e., it depends on a small subset of  $d$  variables. This means that there exist a function  $h_0 : [0, 1]^d \mapsto \mathbb{R}$  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. Moreover, 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 a smoothness parameter  $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in (0, 1]^d$ , a box  $\Psi \subseteq [0, 1]^d$ , and a Hölder coefficient  $\lambda < \infty$ , we denote by  $\mathcal{H}_\lambda^{\alpha, d}(\Psi)$  an anisotropic  $\alpha$ -Hölder space on  $\Psi$ , i.e.,

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

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 since 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 several disjoint boxes of the domain with the same harmonic mean (an important assumption for obtaining the minimax rate). 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 anisotropic smoothness of  $h_0$  is specified on an unknown underlying box partition  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  of  $[0, 1]^d$  with  $R \geq 1$  boxes. If  $R = 1$ , we write  $\mathfrak{X} = ([0, 1]^d)$  with  $\Xi_1 = [0, 1]^d$ . The function space is formed by agglomerating anisotropic Hölder spaces for all 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 box partition  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  of  $[0, 1]^d$  with boxes  $\Xi_r \subseteq [0, 1]^d$  and a smoothness parameter  $A_{\bar{\alpha}} =$

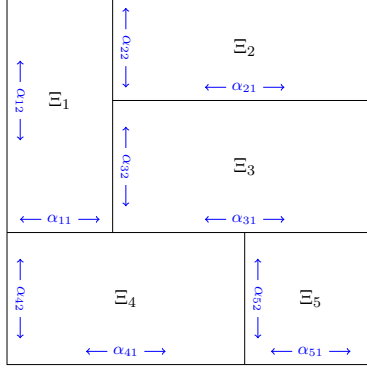


Figure 2: A graphical illustration of a piecewise heterogeneous anisotropic Hölder space with five boxes. Each piece has its own smoothness parameter, but the harmonic mean is assumed to be the same.

$(\alpha_r)_{r=1}^R \in \mathcal{A}_{\bar{\alpha}}^{R,d}$  for some  $\bar{\alpha} \in (0, 1]$ . We define a piecewise heterogeneous anisotropic Hölder space as

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

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  $h \in \mathcal{H}_{\lambda}^{A_{\bar{\alpha}},d}(\mathfrak{X})$  is anisotropic on each  $\Xi_r$  with a smoothness parameter  $\alpha_r \in (0, 1]^d$  and the same harmonic mean  $\bar{\alpha}$  for all  $\Xi_r$ . We again emphasize that discontinuities are allowed at the boundaries of boxes  $\Xi_r$ ,  $r = 1, \dots, R$ .

Definition 2 does not impose a specific structure on the partition  $\mathfrak{X}$  other than a box partition. However, we will later see that depending on the approximation metric, our approximation theory will require it to be a tree-based recursive structure defined in the next section (see Figure 4 below). Nonetheless, since every box partition can be extended to the required form by adding more splits, this discrepancy is not practically an issue. We refer the reader to Section 2.3 and 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 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. On the other hand, our definition deals with different anisotropic smoothness parameters for the boxes in a box partition, and hence seems takes some flexibility. Our jump surfaces, however, are restricted to hyper-planes parallel to the coordinates.

Note that Definition 2 is 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  $f_0 : [0, 1]^p \mapsto \mathbb{R}$ . For any  $S \subseteq \{1, \dots, p\}$ , we denote with  $W_S^p : \mathcal{C}(\mathbb{R}^{|S|}) \mapsto \mathcal{C}(\mathbb{R}^p)$  the map that transmits  $h \in \mathcal{C}(\mathbb{R}^{|S|})$  onto  $W_S^p h : x \mapsto h(x_S)$ . Similarly to Yang and Tokdar (2015) for the isotropic cases, we now formalize  $d$ -sparse function spaces as follows.



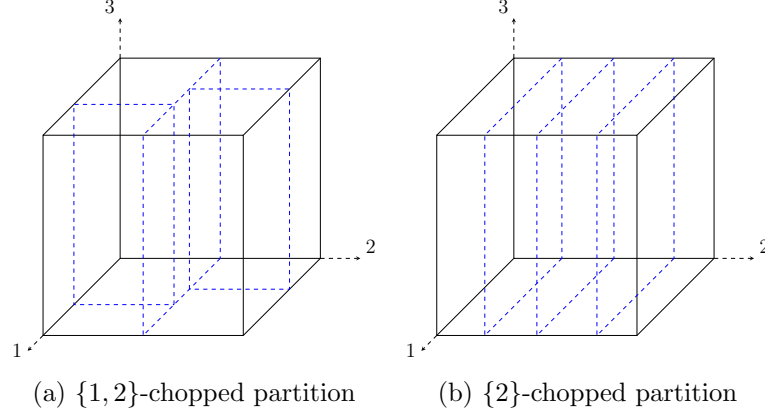


Figure 3: Examples of sparse partitions in three dimensions.

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

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

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  $\lambda$  (which are possibly increasing with  $n$ ), the true function  $f_0$  is assumed to belong to the class  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  which allows for discontinuities, or to its continuous variant  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ .<sup>1</sup> This means that there exists a function  $h_0 : [0, 1]^d \mapsto \mathbb{R}$  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  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  achieves approximability under more relaxed assumptions (see Theorem 1 below). The two spaces are identical if  $R = 1$ .

Note that  $\mathfrak{X}$  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  $\mathfrak{X}$  to the corresponding box partition of the  $p$ -dimensional cube  $[0, 1]^p$ . To this end, we extend each  $\Xi_r$  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 sparsity index  $S_0$ ; that is,  $\Xi_r$  is the projection of  $\Xi_r^*$  onto the coordinates in  $S_0$ . The boxes  $\Xi_r^*$  then constitute the box partition  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_R^*)$  of  $[0, 1]^p$ . It should be noted that  $\mathfrak{X}^*$  depends on the unknown sparsity index  $S_0$  of the true function  $f_0$ . Observe also that our definition gives rise to  $\mathfrak{X}^* = ([0, 1]^p)$  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  $\Psi_r \subseteq [0, 1]^p$ ,  $r = 1, \dots, J$ .<sup>2</sup> For a subset  $S \subseteq \{1, \dots, p\}$ , the partition  $\mathfrak{Y}$  is called  $S$ -chopped if  $\max_{j \in S} \text{len}([\Psi_r]_j) < 1$  and  $\min_{j \notin S} \text{len}([\Psi_r]_j) = 1$  for every  $r = 1, \dots, J$ .

<sup>1</sup>Although  $\bar{\alpha}$ ,  $R$ ,  $d$ ,  $p$ , and  $\lambda$  can be sequences of  $n$  rather than fixed constants, we suppress subscript  $n$  for the sake of notational simplicity.

<sup>2</sup>We often write  $\mathfrak{Y} = (\Psi_r)_r$  to denote an arbitrary box partition of  $[0, 1]^p$  with boxes  $\Psi_r \subseteq [0, 1]^p$ ,  $r = 1, 2, \dots$ , and write  $\Psi \subseteq [0, 1]^p$  to denote an arbitrary  $p$ -dimensional box. The notations  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  and  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_R^*)$  are used only for the box partitions associated with the piecewise heterogeneous anisotropic spaces in Definitions 2–3.

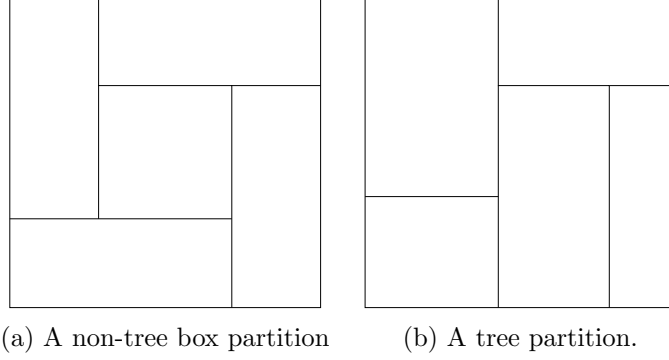


Figure 4: Examples of non-tree box partitions and tree partitions

A graphical illustration of sparse partitions is provided in Figure 3. According to Definition 4, the extended box partition  $\mathfrak{X}^*$  is  $S$ -chopped for some  $S \subseteq S_0$ . Observe that  $\mathfrak{X}^*$  is not always  $S_0$ -chopped, since  $\mathfrak{X}$  may not have been cleaved in some coordinates. For example, if  $f_0(x_1, x_2, x_3) = h_0(x_1, x_3) = \sin(x_1) \cos(x_3) \mathbb{1}(0 \leq x_1 \leq 0.5) \mathbb{1}(0 \leq x_2 \leq 1)$ , then  $S_0 = \{1, 3\}$  but  $\mathfrak{X}^* = ([0, 0.5] \times [0, 1]^2, (0.5, 1] \times [0, 1]^2)$  is  $\{1\}$ -chopped. In particular,  $\mathfrak{X}^*$  is  $\emptyset$ -chopped if  $R = 1$  no matter what  $S_0$  is. From this it is clear that sparsity of  $\mathfrak{X}^*$  is not the same as sparsity of  $f_0$ . In what follows, we use the notation  $S(\mathfrak{X}^*) \subseteq S_0$  to denote sparsity of  $\mathfrak{X}^*$ ; that is,  $\mathfrak{X}^*$  is  $S(\mathfrak{X}^*)$ -chopped.

### 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*  $\tau_j \in \text{int}([\Psi]_j)$ . The pair  $(j, \tau_j)$  then dichotomizes  $\Psi$  along the  $j$ th coordinate into two boxes:  $\{x \in \Psi : x_j \leq \tau_j\}$  and  $\{x \in \Psi : x_j > \tau_j\}$ , where  $x_j$  is  $j$ th 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 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 pre-determined 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.

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

For a given split-net  $\mathcal{Z}$ , we call each point  $z_i = (z_{i1}, \dots, z_{ip})^\top$  a *split-candidate*. For a given splitting coordinate  $j$  and a split-net  $\mathcal{Z}$ , a split-point will be chosen from  $[\mathcal{Z}]_j \cap \text{int}([\Psi]_j)$  to dichotomize a box  $\Psi$ . Note that  $[\mathcal{Z}]_j = \{z_{ij} \in [0, 1], i = 1, \dots, b_n\}$  may have fewer elements than  $\mathcal{Z}$  due 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  $(z_{1j}, \dots, z_{b_n j})$ . Then we have  $\max_{1 \leq j \leq p} b_j(\mathcal{Z}) \leq b_n$  by definition. For example, consider a regular (equidistant) grid system illustrated in Figure 5a, where  $b_j(\mathcal{Z}) = b_n^{1/p} < b_n, j = 1, \dots, p$ . This simplest split-net will be further discussed in Section 4.3.1. It is also possible to construct a split-net such that  $b_j(\mathcal{Z}) = b_n, j = 1, \dots, p$ , 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  $\tau_j$  is chosen such that  $\tau_j \in [\mathcal{Z}]_j \cap \text{int}([\Psi]_j)$ . Since 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 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  $\mathcal{T} = (\Omega_1, \dots, \Omega_K)$  of  $[0, 1]^p$  with boxes  $\Omega_k \subseteq [0, 1]^p, k = 1, \dots, K$ , is called a  $\mathcal{Z}$ -tree partition if every split occurs at points  $z_i \in \mathcal{Z}$ .<sup>3</sup>

In summary, we have 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  $S \subseteq \{1, \dots, p\}$  no matter what  $\mathcal{Z}$  is employed. Since we aim to do sparse estimation in high-dimensional setups, we are mostly interested in  $S$ -chopped  $\mathcal{Z}$ -tree partitions for some low-dimensional  $S$ . In what follows, we denote by  $\mathcal{T}_{S,K,\mathcal{Z}}$  the set of all  $S$ -chopped  $\mathcal{Z}$ -tree partitions with  $K$  boxes.

**Remark 2.** 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).

<sup>3</sup>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.

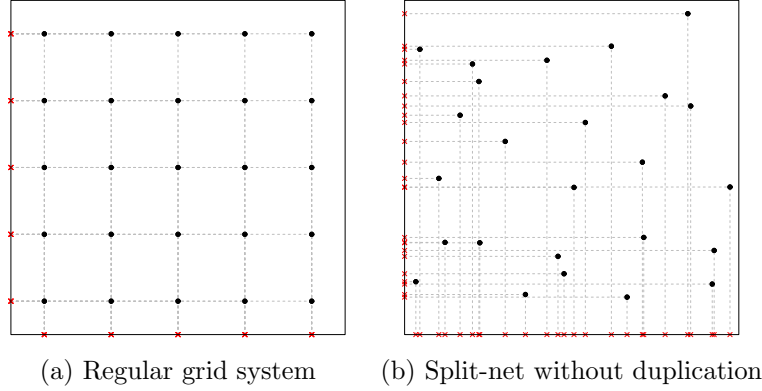


Figure 5: Examples of the split-net with  $b_n = 25$  in two dimensions. For the regular grid in (a), one can easily see that  $b_j(\mathcal{Z}) = 5$ ,  $j = 1, 2$ ; hence, initial splits eliminate the possibility of other splits. The split-candidates of the split-net in (b) are unique in every coordinate,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

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 way 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). This paper only considers putting 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  $\mathcal{T}^t = (\Omega_1^t, \dots, \Omega_{K^t}^t)$  a  $\mathcal{Z}$ -tree partition of size  $K^t$  and with  $\beta^t = (\beta_1^t, \dots, \beta_{K^t}^t)^\top \in \mathbb{R}^{K^t}$  the heights of the step function, called the *step-heights*. An additive tree-based learner is then fully described by a tree ensemble  $\mathcal{E} = (\mathcal{T}^1, \dots, \mathcal{T}^T)$  and terminal node parameters  $B = (\beta^{1^\top}, \dots, \beta^{T^\top})^\top \in \mathbb{R}^{\sum_{t=1}^T K^t}$  through

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

That is,  $f_{\mathcal{E}, B}$  is constant on the boxes constructed by overlapping  $\mathcal{Z}$ -tree partitions  $\mathcal{T}^1, \dots, \mathcal{T}^T$ . Chipman et al. (2010) recommended 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  $\mathcal{F}_{\mathcal{E}}$ .

Our objective is to characterize the posterior asymptotic properties of the tree learners in (2) in estimating the true function  $f_0$  belonging to  $\Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  or  $\Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ . 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. The following two sections aim to 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 due 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. In this paper, we deploy the sparse Dirichlet prior developed by Linero (2018) for ease of computation of the posterior.

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  $\mathbb{S}^p = \{(x_1, \dots, x_p)^\top \in \mathbb{R}^p : \sum_{j=1}^p x_j = 1, x_j \geq 0, j = 1, \dots, p\}$ . A proportion vector  $\eta$  has a Dirichlet prior with  $\zeta > 0$  and  $\xi > 1$ ,

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

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  $\eta$ , 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 of depth  $\ell$  is split with prior probability  $\nu^{\ell+1}$  for  $\nu \in (0, 1/2)$ . If a node corresponding to a box  $\Omega$  is split, a splitting coordinate  $j$  is drawn from the proportion vector  $\eta$  and a split-point  $\tau_j$  will be chosen randomly from  $[\mathcal{Z}]_j \cap \text{int}([\Omega]_j)$  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. Since 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  $K^1, \dots, K^T$  induced by  $\mathcal{E}$ , Chipman et al. (2010) suggested using a Gaussian prior on  $B$ :

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

where  $\phi(\cdot; \mu, \sigma^2)$  is the Gaussian density with mean  $\mu$  and variance  $\sigma^2$ . The variance  $1/T$  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 put 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 due 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), we mostly 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 which 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, it is well-known that balanced  $k$ -d trees (Bentley, 1979) 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  $f_0$ . In this section, we 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 two properties: (i) it should be dense enough so that the boundaries of the box partition  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_R^*)$ , extended from  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$ , 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  $\Xi_r^*$ . We elucidate these two properties.

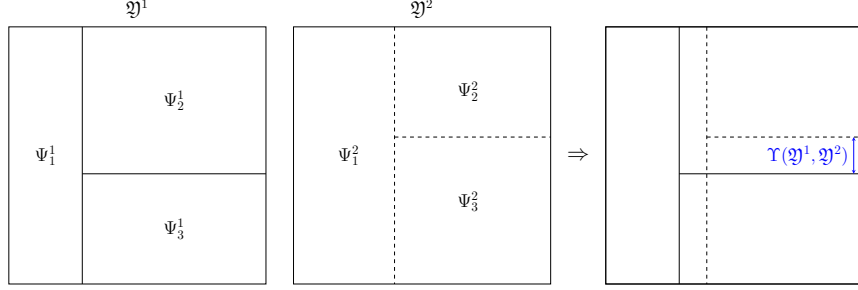


Figure 6: A two-dimensional example of the Hausdorff-type divergence in Definition 7. The divergence is the maximum dependency of the boxes in the partitions.

#### 4.1.1 Dense split-nets: Global approximability

Recall that the underlying partition  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_R^*)$  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  $\mathfrak{X}^*$  is a flexible tree partition. If the prior rewards partitions that are sufficiently close to  $\mathfrak{X}^*$ , Bayesian CART (BART) is expected to adapt to unknown  $\mathfrak{X}^*$  without much loss of efficiency. We examine when this adaptivity can be achieved in more detail below.

The ability to detect  $\mathfrak{X}^*$  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  $\mathfrak{X}^*$ . Therefore, we need a gadget to measure the closeness between two partitions. To this end, below 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

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

where  $P_\pi[J]$  denotes the set of all permutations  $(\pi(1) \dots \pi(J))$  of  $\{1, \dots, J\}$  and  $\text{Haus}(\cdot, \cdot)$  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  $\Upsilon(\mathfrak{X}^*, \mathcal{T})$  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  $(\mathfrak{X}^*, c_n)$ -dense for some suitable  $c_n \geq 0$ . Note that the ideal case  $c_n = 0$  can be achieved only when  $\mathfrak{X}^*$  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 have  $\Upsilon(\mathfrak{Y}, \mathcal{T}) = 0$  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  $\mathfrak{X}^*$  should be at least a flexible tree partition, but not necessarily a  $\mathcal{Z}$ -tree partition. If  $\mathfrak{X}^*$  is a box partition but not a flexible tree partition, we can redefine  $\mathfrak{X}^*$  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. In Section 4.3, we present some examples of dense split-nets.

Dense split-nets have nested properties. That is to say 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 more *global* approximability of underlying partitions which requires split-nets to be suitably dense. Now, we focus on *local* approximability.

Assume that  $\mathfrak{X}^*$  can be approximated well (as discussed in the previous section) by an  $S(\mathfrak{X}^*)$ -chopped  $\mathcal{Z}$ -tree partition  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_R^*)$ .<sup>4</sup> More specifically, for a  $(\mathfrak{X}^*, c_n)$ -dense split-net  $\mathcal{Z}$  with a given  $c_n > 0$ ,  $\mathcal{T}^*$  is defined as  $\operatorname{argmin}_{\mathcal{T} \in \mathcal{T}_{S(\mathfrak{X}^*), R, \mathcal{Z}}} \Upsilon(\mathfrak{X}^*, \mathcal{T})$ . We now focus on local approximability inside each box  $\Omega_r^*$ . Ideally, one would like 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  $\Omega_r^*$ . Recall that the true function  $f_0$  has anisotropic smoothness on each of  $\Xi_r^*$ . Intuitively, denser subdivisions are required for less smooth coordinates to capture the local features. Allowing splits to occur more often in certain directions, below we define the *anisotropic  $k$ -d tree*, which achieves the desired approximation error for anisotropic smoothness. The definition requires the notion of *midpoint-splits* defined as follows. For a given box  $\Psi$  and a splitting coordinate  $j$ , a midpoint-split picks up the  $\lfloor \tilde{b}_j(\mathcal{Z}, \Psi)/2 \rfloor$ th split-candidate in  $[\mathcal{Z}]_j \cap \operatorname{int}([\Psi]_j)$  as a split-point  $\tau_j$ , where  $\tilde{b}_j(\mathcal{Z}, \Psi)$  is the cardinality of  $[\mathcal{Z}]_j \cap \operatorname{int}([\Psi]_j)$ .

**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*  $AKD(\Psi; \mathcal{Z}, \alpha, L, S)$  as the iterative splitting procedure that partitions  $\Psi$  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  $j$ s, choose the smallest  $j$  among such  $j$ 's.

<sup>4</sup>The notation  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_R^*)$  with an asterisk is only used to denote an  $S(\mathfrak{X}^*)$ -chopped  $\mathcal{Z}$ -tree partition approximating  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_R^*)$ .

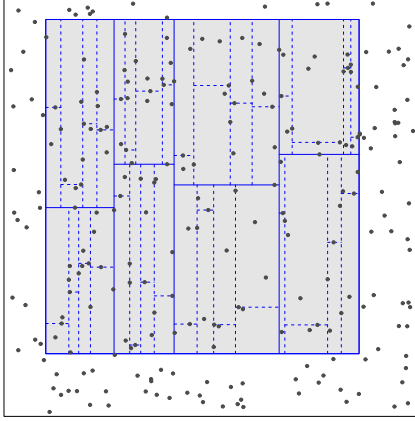


Figure 7: A realization of the anisotropic  $k$ -d tree with smoothness parameters  $\alpha_1 = 0.25$  (for the horizontal axis) and  $\alpha_2 = 0.5$  (for the vertical axis), and some box  $\Psi$  (the shaded box) that is a subspace of  $[0, 1]^2$  (the outer square). Since  $2\alpha_1 = \alpha_2$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

3. For all boxes  $\Omega_k^\circ$ ,  $k = 1, \dots, 2^{\sum_{j=1}^d l_j}$ , 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  $\Omega_k^\circ$ ,  $k = 1, \dots, 2^{1+\sum_{j=1}^d l_j}$ , 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  $AKD(\Psi; \mathcal{Z}, \alpha, L, S)$  returns a tree partition  $\mathcal{T}^\circ = (\Omega_1^\circ, \dots, \Omega_{2^{L^\circ}}^\circ)$  of  $\Psi$  and a vector  $(l_1, \dots, l_d)^\top$  such that  $L^\circ = \sum_{j=1}^d l_j \leq L$ .<sup>5</sup> Although these returned items clearly depend on the inputs of the anisotropic  $k$ -d tree procedure (i.e.,  $\Psi$ ,  $\mathcal{Z}$ ,  $\alpha$ ,  $L$ , and  $S$ ), we suppress them throughout the paper. Each  $l_j$  is a counter of how many times the  $j$ th coordinate has been used. The procedure is designed so that every  $l_j$  is approximately proportional to  $\alpha_j^{-1}$  after enough iterations. The total number of splits for the  $j$ th coordinate is thus close to  $2^{C/\alpha_j}$  for every  $j$  with some  $C > 0$ . In the proof of Theorem 1 below, one can see that this matching is indeed optimal and minimizes the induced bias.

To play a role as a ‘sieve’ for approximation,  $\Psi$  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  $L^\circ < L$  due 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

<sup>5</sup>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.

$j \in S$ , all  $\text{len}([\Omega_k^\circ]_j)$ ,  $k = 1, \dots, 2^L$ , 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 sufficiently 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  $(\Psi, \alpha, L, S)$ -regular if  $\mathcal{T}^\circ = (\Omega_1^\circ, \dots, \Omega_{2^{L^\circ}}^\circ)$  and  $(l_1, \dots, l_d)^\top$  returned by  $AKD(\Psi; \mathcal{Z}, \alpha, L, S)$  satisfy  $L^\circ = L$  and  $\max_k \text{len}([\Omega_k^\circ]_{s_j}) \lesssim \text{len}([\Psi]_{s_j}) 2^{-l_j}$  for every  $j = 1, \dots, d$ .

The condition  $\max_k \text{len}([\Omega_k^\circ]_{s_j}) \lesssim \text{len}([\Psi]_{s_j}) 2^{-l_j}$  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  $\max_k \text{len}([\Omega_k^\circ]_{s_j}) = \text{len}([\Psi]_{s_j}) 2^{-l_j}$ . 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  $(\Psi, \alpha, L, S)$ -regular for some  $\Psi$ ,  $\alpha$ ,  $L$ , and  $S$ , then it is also  $(\Psi, \alpha, \tilde{L}, S)$ -regular for any  $\tilde{L} \leq L$ . This can be easily shown by noting that the latter is determined only by a sub-tree of a blown tree for the former. We are particularly interested in the largest possible  $L$ .

**Remark 3.** Since 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  $L \leq \lfloor \log_2(\tilde{b}_j(\mathcal{Z}; \Psi) + 1) \rfloor$  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.

## 4.2 Approximation theory

Our goal is to establish asymptotic properties of the posterior distribution. This requires that tree learners be able to approximate functions in the spaces  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  and  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  appropriately. Here, we establish 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  $(\mathfrak{X}^*, c_n)$ -dense for some suitable  $c_n$ . The boundaries of  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_R^*)$  should thus be detected well by the binary tree partitioning rule. Since  $\mathfrak{X}^*$  is approximated by a  $\mathcal{Z}$ -tree partition with a given  $\mathcal{Z}$ , the underlying partition  $\mathfrak{X}^*$  should be at least a flexible tree partition, but a stronger result is obtained if  $\mathfrak{X}^*$  is a  $\mathcal{Z}$ -tree partition (see [Theorem 1](#) below). Denoting by  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_R^*)$  the  $S(\mathfrak{X}^*)$ -chopped  $\mathcal{Z}$ -tree partition approximating  $\mathfrak{X}^*$ , each box  $\Omega_r^*$  should be appropriately subdivided to capture the local/global nature of the true function on  $\Xi_r^*$ . (If  $R = 1$ , we write  $\mathcal{T}^* = \mathfrak{X}^* = ([0, 1]^p)$  with  $\Omega_1^* = [0, 1]^p$ .) Hence, for a smoothness parameter  $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R, d}$  and  $L_0$  specified below,  $\mathcal{Z}$  should also be  $(\Omega_r^*, \alpha_r, L_0, S_0)$ -regular,  $r = 1, \dots, R$ . The integer  $L_0$  is chosen below such that the approximation error is balanced with our target rate. Let  $\mathcal{T}_r^\circ = (\Omega_{r1}^\circ, \dots, \Omega_{r2^{L_0}}^\circ)$  be the tree partition of  $\Omega_r^*$  returned by  $AKD(\Omega_r^*; \mathcal{Z}, \alpha_r, L_0, S_0)$ ,  $r = 1, \dots, R$ . Then, the approximating partition  $\hat{\mathcal{T}}$  is

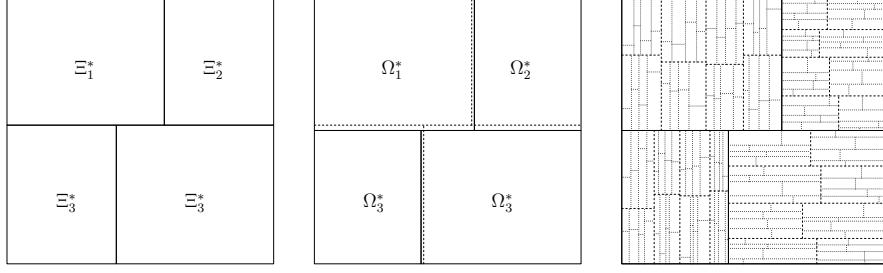


Figure 8: An example of constructing  $\hat{\mathcal{T}}$ . First,  $\mathfrak{X}^* = (\Xi_1^*, \dots, \Xi_4^*)$  is approximated by  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_4^*)$ . Then, each  $\Omega_r^*$  is subdivided by the anisotropic  $k$ -d tree, producing  $\mathcal{T}_r^\circ$  a constituent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

formed by agglomerating all sub-tree partitions  $\mathcal{T}_r^\circ$ , leading to an  $S_0$ -chopped  $\mathcal{Z}$ -tree partition  $\hat{\mathcal{T}} = (\Omega_{11}^\circ, \dots, \Omega_{12^{L_0}}^\circ, \dots, \Omega_{R1}^\circ, \dots, \Omega_{R2^{L_0}}^\circ)$ . A graphical illustration of constructing  $\hat{\mathcal{T}}$  is given in Figure 8.

The strongest approximation results relative to the supremum norm for  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  are of particular interest. Due to the possible discontinuity or heterogeneity at the unknown boundaries of  $\mathfrak{X}^*$ , however, such results are not practically obtained except for the case  $R = 1$ . As the following lemma 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  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  require even milder conditions.

**Theorem 1** (Approximation theory). *For a split-net  $\mathcal{Z}$  that is  $(\mathfrak{X}^*, c_n)$ -dense for  $c_n > 0$  specified below, let  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_R^*) = \operatorname{argmin}_{\mathcal{T} \in \mathcal{T}_{S(\mathfrak{X}^*), R, \mathcal{Z}}} \Upsilon(\mathfrak{X}^*, \mathcal{T})$ . For every  $r = 1, \dots, R$ , assume that  $\mathcal{Z}$  is  $(\Omega_r^*, \alpha_r, L_0, S_0)$ -regular for a smoothness parameter  $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R, d}$  and an integer sequence  $L_0 > 0$  such that  $2^{L_0} \asymp (\lambda^2 d^2 n / (R \log n))^{d/(2\bar{\alpha}+d)}$ . Let  $\bar{\epsilon}_n = (\lambda d)^{d/(2\bar{\alpha}+d)} ((R \log n)/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}$  and construct the  $S_0$ -chopped  $\mathcal{Z}$ -tree partition  $\hat{\mathcal{T}}$  as above. Then, the following assertions hold.*

- (i) *If  $c_n^{\min_{r,j} \alpha_{rj}} \lesssim \bar{\epsilon}_n / (\lambda |S(\mathfrak{X}^*)|)$ , then for every  $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ , there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  such that  $\|f_0 - \hat{f}_0\|_\infty \lesssim \bar{\epsilon}_n$ .*
- (ii) *Fix  $v \geq 1$ . If*

$$c_n \lesssim \left( \frac{\bar{\epsilon}_n}{\|f_0\|_\infty} \right)^v \frac{\min_{r,j} \operatorname{len}([\Xi_r^*]_j)}{|S(\mathfrak{X}^*)|}, \quad (4)$$

*then for every  $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$ , there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  such that  $\|f_0 - \hat{f}_0\|_v \lesssim \bar{\epsilon}_n$ . If (4) is not satisfied but*

$$c_n^{1+v \min_{r,j} \alpha_{rj}} \lesssim \left( \frac{\bar{\epsilon}_n}{\lambda} \right)^v \frac{\min_{r,j} \operatorname{len}([\Xi_r^*]_j)}{|S(\mathfrak{X}^*)|^{v+1}}, \quad (5)$$

*then for every  $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ , there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  such that  $\|f_0 - \hat{f}_0\|_v \lesssim \bar{\epsilon}_n$ .*

- (iii) *Fix  $v \geq 1$ . For every  $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  and  $c_n = 1$ , there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  such that  $\|f_0 - \hat{f}_0\|_{v, P_{\mathcal{Z}}} \lesssim \bar{\epsilon}_n$ , where  $P_{\mathcal{Z}}(\cdot) = b_n^{-1} \sum_{i=1}^{b_n} \delta_{z_i}(\cdot)$ .*

*Proof.* See Section A.1 in Appendix.  $\square$

The assertion in (i) is given for the supremum norm which is the most comprehensive and universally used in many statistical estimation problems. This, however, imposes the continuity restriction. The assertion in (ii) is with respect to the  $L_v$ -norm,  $v \geq 1$ , which is useful in many statistical setups. The result allows for the discontinuity of  $f_0$  if (4) is satisfied. Notice that (5) is not always milder than (4) despite the continuity restriction. It should be understood that the expression of (ii) is adopted for the sake of simplicity. One may easily see that (5) is milder than (4) only if  $\lambda|S(\mathfrak{X}^*)|c_n^{\min_{r,j} \alpha_{rj}} \lesssim \|f_0\|_\infty$ . This is often satisfied since we are usually interested in a decreasing polynomial in  $n$  for  $c_n$ . The assertion in (iii) is particularly useful in regression setups with  $\mathcal{Z}$  chosen by fixed covariates (see Section 4.3.2). Note that (iii) only requires the regularity of a split-net  $\mathcal{Z}$ , as all split-nets are  $(\mathfrak{X}^*, 1)$ -dense.

Observe that  $|S(\mathfrak{X}^*)| = 0$  if  $R = 1$ . In this particular case, the conditions for  $c_n$  in (i) and (ii) are automatically satisfied since all split-nets are  $(([0, 1]^p), 0)$ -dense. The assertions then only require a split-net to be suitably regular. This fact certainly makes sense because there are no boundaries to be detected if  $R = 1$ . Note also that  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) = \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  if  $R = 1$ , and hence (i) is the strongest in this case.

If  $R > 1$ , the conditions in (i) and (ii) are not trivially satisfied. The conditions are oracle-type conditions since they depend on unknown model components, e.g.,  $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R, d}$ ,  $\lambda$ , and  $|S(\mathfrak{X}^*)|$ . More practical conditions can be obtained by plugging in reasonable bounds of the unknown components. For example, we cannot hope for better than  $\bar{\epsilon}_n \gtrsim (\lambda d R (\log n)/n)^{1/3}$  due to the fundamental limitation of piecewise constant learners. For the procedure to be consistent, we need the necessary conditions  $d/\bar{\alpha} \ll \log n$  and  $\lambda^{\bar{\alpha}/d} R \ll n$  (see the rate in (7) below). We can also assume that  $\min_{r,j} \text{len}([\Xi_r^*]_j)$  is bounded away from zero or decreases sufficiently slowly. Putting everything together, the conditions in (i) and (ii) can be easily satisfied if  $c_n$  is a decreasing polynomial in  $n$  with a suitable exponent (a suitable decreasing order of  $\min_{r,j} \alpha_{rj}$  is required for (i)). Note again that (iii) does not require a decreasing  $c_n$ .

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  $\max_{1 \leq j \leq p} \log b_j(\mathcal{Z}) \lesssim \log n$ . Hence, we are ultimately interested in split-nets with  $b_j(\mathcal{Z})$ ,  $j = 1, \dots, p$ , balanced very well.

### 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, it remains unclear how to obtain such a good split-net in practice. We will show that the two split-nets in Figure 5 are suitably dense and regular, and hence fulfill the requirements of Theorem 1. Throughout this section,  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_R^*)$  is defined as  $\mathcal{T}^* = \arg\min_{\mathcal{T} \in \mathcal{T}_{S(\mathfrak{X}^*), R, \mathcal{Z}}} \Upsilon(\mathfrak{X}^*, \mathcal{T})$  for a given  $\mathcal{Z}$ .

#### 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. 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  $b_n = n^{c_p}$  for a constant  $c \geq 1$ . If  $\min_{r,j} \text{len}([\Xi_r^*]_j) \gg n^{-c}$  and  $\lambda d \mathbb{1}(R > 1) / \min_{r,j} \text{len}([\Xi_r^*]_j)^{\bar{\alpha}/d+1/2} \lesssim n^{c\bar{\alpha}/d+(c-1)/2} \sqrt{R \log n}$ , then  $\mathcal{Z}$  is  $(\mathfrak{X}^*, c_n)$ -dense and  $(\Omega_r^*, \alpha_r, L_0, S_0)$ -regular for  $r = 1, \dots, R$ , where  $c_n = n^{-c} \mathbb{1}(R > 1)$ .*

*Proof.* See Section A.2 in Appendix.  $\square$

The second condition is simplified as  $\lambda d \mathbb{1}(R > 1) / \sqrt{\min_{r,j} \text{len}([\Xi_r^*]_j)} \lesssim n^{(c-1)/2} \sqrt{R \log n}$  by taking  $\bar{\alpha} \rightarrow 0$ . We pay particular attention to the case  $R = 1$ , i.e.,  $\mathfrak{X}^* = ([0, 1]^p)$ , where the conditions are trivially satisfied. In this case, we obtain the strongest result in (i) of Theorem 1 (recall that  $\Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) = \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  if  $R = 1$ ). For the case  $R > 1$ , the conditions are very mild if  $c$  is large enough with the necessary conditions  $d \ll \log n$  and  $\lambda^{\bar{\alpha}/d} R \ll n$  for consistent estimation. The choice  $c = 1$  can even be sufficient with stronger boundedness conditions. Since  $c_n$  is a decreasing polynomial in  $n$  with our choice of  $b_n$ , the assertions in (i) and (ii) of Theorem 1 hold with a large enough  $c$ . Note that (iii) of Theorem 1 also holds trivially with this  $\mathcal{Z}$ .

Since  $\max_{1 \leq j \leq p} \log b_j(\mathcal{Z}) = p^{-1} \log b_n \lesssim \log n$ , a regular grid satisfies the condition for the optimal posterior contraction specified in Section 5. This nice property makes a regular grid very appealing for practical use given its simplicity, and there is no much 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 this 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).

### 4.3.2 Fixed design points

Now we focus on a fixed design regression setup, where observed covariate values are readily available. 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 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 unknown  $S_0$ . Thus we make the following assumption on  $\mathcal{Z}$ .

- (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  $(\Psi, \alpha, L, S_0)$ -regular with  $L = \lfloor \log_2(cnP_{\mathcal{Z}}(\Psi)) \rfloor$  for some constant  $c > 0$ .

The condition  $nP_{\mathcal{Z}}(\Psi) \gg 1$  implies that the number of split-candidates contained in  $\Psi$  increases. 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  $(\Psi, \alpha, L, S_0)$ -regular for  $L = \lfloor \log_2(nP_{\mathcal{Z}}(\Psi) + 1) \rfloor$  (see Remark 3). Our requirement is thus milder.

**Lemma 2** (Fixed design points). *Consider fixed design points  $\mathcal{Z} = \{z_i, i = 1, \dots, n\}$  with Assumption in (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  $(\Omega_r^*, \alpha_r, L_0, S_0)$ -regular for  $r = 1, \dots, R$ .*

*Proof.* See Section A.3 in Appendix.  $\square$

Since  $nP_{\mathcal{Z}}(\Xi_r^*)$  is the number of split-candidates in  $\Xi_r^*$ , 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 two conditions are comparable if we take  $\bar{\alpha} \rightarrow 0$  from the practical perspective. We see that (iii) of Theorem 1 directly follows from this lemma. Since 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 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  $\max_{1 \leq j \leq p} \log b_j(\mathcal{Z}) \lesssim \log n$  for the optimal posterior contraction is trivially satisfied in this case.

## 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  $\sigma^2$  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

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

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 (Salomond, 2018). We also refer to Ning et al. (2020) and Jeong and Ghosal (2021) for the construction of relevant test functions with respect to the Rényi divergence. 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  $\sigma_0^2$ .



- (A1) For  $d > 0$ ,  $\lambda > 0$ ,  $R > 0$ ,  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$ , and  $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$  with  $\bar{\alpha} \in (0, 1]$ , the true function satisfies  $f_0 \in \Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  or  $f_0 \in \Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ .
- (A2) It is assumed that  $d$ ,  $p$ ,  $\lambda$ ,  $R$ , and  $\bar{\alpha}$  satisfy  $\epsilon_n \ll 1$ , where

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

- (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 some 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  $\epsilon_n$  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  $\sigma^2$ , 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  $\mathcal{Z}$ . Below the notation  $\text{dep}$  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  $\max_{1 \leq j \leq p} \log b_j(\mathcal{Z}) \lesssim \log n$ .
- (A6) The split-net  $\mathcal{Z}$  is suitably dense and regular to construct a  $\mathcal{Z}$ -tree partition  $\hat{\mathcal{T}}$  such that there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  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  $\mathfrak{X}^*$  satisfies  $\max_r \text{dep}(\Omega_r^*) \lesssim \log n$ .

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. Due to 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, a 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  $\sigma^2$  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  $\epsilon_n$  in (7),*

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

*Proof.* See Section A.4 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  $\lambda$  and  $R$  can be a polynomial in  $n$  with a suitably small power to satisfy  $\epsilon_n \rightarrow 0$ , 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 or not 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 4.** 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. Since this restriction is not required, we keep the rate in the form of (7).

## 5.2 Minimax lower bounds

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 the 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

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

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  $\text{supp}(Q) \subseteq [0, 1]^p$ . We assume (without loss of generality) that  $\sigma_0^2$  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  $L_2$ -norm.

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

We define the minimax risk for any function space  $\mathcal{F} \in L_2(Q)$  as

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

where  $\mathcal{B}_n$  is the space of all  $\mathcal{L}_2(Q)$ -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  $\bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) = \{f \in \Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) : \|f\|_{\infty} \leq M\lambda\}$  for any  $M > 0$ . Since our contraction rate is the same for both  $\Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  and  $\Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ , we aim to construct a lower bound of  $r_n(\bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), Q)$  close enough to  $\epsilon_n$ .

**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  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  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  $\log \text{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$ ,  $1 \leq r \leq R$ ,  $1 \leq j \leq d$ , there exists  $M_d > 0$  depending only on  $d$  such that*

$$r_n(\bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), Q) \geq M_d \gamma_n,$$

where  $\gamma_n = \sqrt{n^{-1} \log \binom{p}{d}} + (\lambda^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}$ , if  $\gamma_n \ll 1/d$ .

*Proof.* See Section A.5 in Appendix. □

Since  $M_d$  can be dependent on  $d$ , the correct interpretation of the result is with a bounded  $d$ . Also, our contraction rate  $\epsilon_n$  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  $\log \text{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$ ,  $1 \leq r \leq R$ ,  $1 \leq j \leq d$ , leads to the restriction  $\log R \lesssim d/\bar{\alpha}$  which removes the term  $R$  is removed from our rate  $\epsilon_n$  in (7). Putting the bounds together,  $\epsilon_n$  matches  $\gamma_n$  up to a logarithmic factor. Even if  $d$  is increasing, recall that we must have  $d/\bar{\alpha} \ll \log n$  to guarantee consistent estimation. Therefore, our rate  $\epsilon_n$  matches the lower bound  $\gamma_n$  up to a logarithmic factor even for increasing  $d$ , implying that the rate  $\epsilon_n$  is near-minimax optimal.

### 5.3 Numerical study

In this section, we provide a numerical study that shows the successful performance of BART for model (6) with a piecewise smooth function. For competitors we consider the random forest and deep neural network (DNN) models with the rectified linear unit (ReLU) activation functions, which are believed to work well for discontinuous or complicated smooth functions. In particular, the recent literature has reported that DNN models are instrumental in adapting 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). The random forest is expected to perform similarly to BART, as it is also based on the additive tree structure. Our numerical study shows that BART outperforms these competitors in estimating piecewise smooth classes.

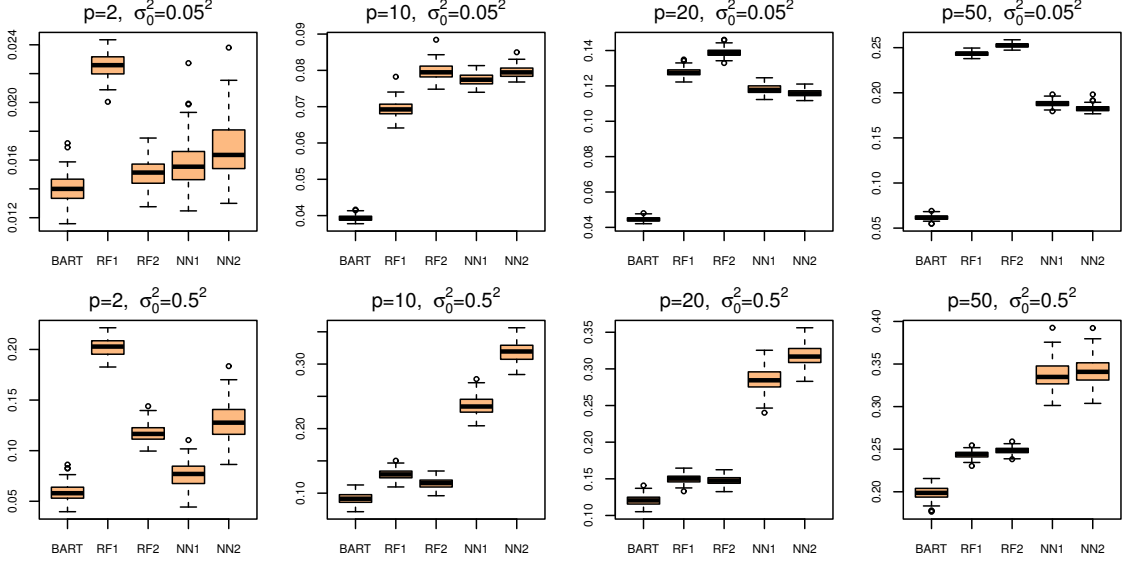


Figure 9: RMSPEs obtained from 100 replicated datasets for each simulation setting. RF1 stands for the random forest of 200 trees with maximal node size 5. RF2 stands for the random forest of 200 trees with maximal node size 50 for each tree. NN1 stands for the DNN model with two hidden layers and (64, 32) hidden units. NN2 stands for the DNN model with four hidden layers and (256, 128, 64, 32) hidden units.

The test datasets are generated from model (6) with the true function  $f_0 : [0, 1]^p \mapsto \mathbb{R}$ ,

$$f_0(x_1, \dots, x_p) = 1 + \frac{1}{p} \left\{ \sum_{j=1}^p (-1)^j \mathbb{1}(x_j \geq 1/2) \right\} \left\{ \sum_{j=1}^p (x_j - 1/2)^2 \right\}, \quad (9)$$

with given  $p$  and  $\sigma_0^2$ . The function  $f_0$  is discontinuous on  $2^p$  pieces of  $[0, 1]^p$ . For a fair comparison to the other methods, we do not use the Dirichlet sparse prior in (3). 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. Hence,  $p$  must be much smaller than  $n$ . For given predictor variables  $X_i$  generated uniformly on  $[0, 1]^p$ , the response variable  $Y_i$  is generated from model (6),  $i = 1, \dots, n$ . With the sample size  $n = 10^3$ , we consider the simulation settings  $p \in \{2, 10, 20, 50\}$  and  $\sigma_0^2 \in \{0.05^2, 0.5^2\}$ . We fit BART with 200 trees using the R package BART. The random forest is fitted by the `randomForest` package with 200 trees and the maximal node size 5 or 50 for each tree. The DNN models are trained by TensorFlow with the Keras interface. We consider two DNN models with 2 and 4 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.

Figure 9 shows the root mean squared prediction error (RMSPE) obtained by the five methods. The RMSPEs are calculated with randomly drawn out-of-samples in 100 replicated datasets for each simulation setting. The result shows that BART clearly outperforms the other learning algorithms for the piecewise smooth function in (9). In particular, the DNN models often fall behind in the competition, while requiring long computation time. We surmise that this is because the training procedure of DNN models often gets stuck at a local

mode due to the susceptibility; it is very challenging to fully optimize complicated DNN models. We also tested many other tuning parameter setups and network structures for the random forest and the DNN, but found no improvement.

## 6 Further applications

Section 5 establishes the posterior contraction rate of BART for the nonparametric regression model and justifies its near-minimax optimality. Since our approximation theory only requires very general 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 nonparametric binary classification. Moreover, since 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.

### 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). The random design assumption is often necessary, for example, in measurement error models (Tuo and Wu, 2015) or causal inference models (Hahn et al., 2020; Ray and van der Vaart, 2020). Here we establish the posterior contraction rate of BART for the random design model in (8). It is important to note that fixed design points in Section 4.3.2 can not be used for a split-net, as the procedure is not truly Bayesian if the prior is dependent on the data. Instead, a regular grid in Section 4.3.1 can be useful.

We consider model (8) for  $Q$  a probability measure that satisfies  $\text{supp}(Q) \subseteq [0, 1]^p$  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  $\rho_H(\cdot, \cdot)$  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. Below we summarize the required modifications of (A3), (A6), (P2), and (P3).

- (A3\*) The true function  $f_0$  satisfies  $\|f_0\|_\infty \leq C_0^*$  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  $\hat{\mathcal{T}}$  such that there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  satisfying  $\|f_0 - \hat{f}_0\|_2 \lesssim \bar{\epsilon}_n$  by Theorem 1.
- (P2\*) A prior on the compact support  $[-\bar{C}_1, \bar{C}_1]$  is assigned to the step-heights  $B$  for some  $\bar{C}_1 > C_0^*$ .
- (P3\*) A prior on the compact support  $[\bar{C}_2^{-1}, \bar{C}_2]$  is assigned to  $\sigma^2$  for some  $\bar{C}_2 > C_0$ .

Assumption (A6\*) requires good approximability with respect to the  $L_2$ -norm. Due to Theorem 1, a regular grid in Section 4.3.1 can be useful to meet this requirement. 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  $\epsilon_n$  in (7),*

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

*Proof.* See Section A.6 in Appendix. □

## 6.2 Density estimation

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

$$X_i \sim P, \quad i = 1, \dots, n. \quad (10)$$

Assume that  $P$  is absolutely continuous with respect to the Lebesgue measure with the true density  $p_0$ . We assign a prior on  $p_f$  indexed by  $f$  such that  $p_f = \exp(f) / \int_{[0,1]^p} \exp(f(x)) dx$  with  $f$  assigned the forest priors in Section 3. We write  $f_0 = \log p_0$  while assuming (A1)–(A3). We leverage the existence of an exponentially powerful test for the Hellinger metric  $\rho_H(\cdot, \cdot)$ . Due to the relationship between Hellinger balls and supremum-norm balls in density estimation with the exponential link, we need an approximation result with respect to the supremum-norm. This is obtained by (i) of Theorem 1 with an appropriate split-net, but requires the continuity restriction on the true function. We make the following assumptions to satisfy this requirement.

- (A1<sup>‡</sup>) For  $d > 0$ ,  $\lambda > 0$ ,  $R > 0$ ,  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$ , and  $A_{\bar{\alpha}} \in \mathcal{A}_{\bar{\alpha}}^{R,d}$  with  $\bar{\alpha} \in (0, 1]$ , the true function satisfies  $f_0 \in \Gamma_{\lambda}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ .
- (A6<sup>‡</sup>) The split-net  $\mathcal{Z}$  is suitably dense and regular to construct a  $\mathcal{Z}$ -tree partition  $\hat{\mathcal{T}}$  such that there exists  $\hat{f}_0 \in \mathcal{F}_{\hat{\mathcal{T}}}$  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<sup>‡</sup>), (A2)–(A3), (A5), (A6<sup>‡</sup>), and (A7), and the prior assigned through (P1)–(P2). Then, there exists a constant  $M > 0$  such that for  $\epsilon_n$  in (7),*

$$\mathbb{E}_0 \Pi \left\{ f : \rho_H(p_f, p_0) > M\epsilon_n \mid X_1, \dots, X_n \right\} \rightarrow 0.$$

*Proof.* See Section A.6 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  $\|f\|_\infty$ . Since 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.

### 6.3 Nonparametric binary classification

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:

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

for some  $\varphi_0 : [0, 1]^p \mapsto [0, 1]$  and some probability measure  $Q$  such that  $\text{supp}(Q) \subseteq [0, 1]^p$  with a bounded density. We thus consider a binary classification problem with random design.

We parameterize the probability function using the logistic link function  $H : \mathbb{R} \mapsto [0, 1]$  such that  $\varphi_f = H(f)$  for  $f$  on which the forest priors in Section 3 are assigned. For true function  $\varphi_0$ , we write  $f_0 = H^{-1}(\varphi_0)$  while assuming (A1)–(A3) as in the density estimation problem. In the proof, it will be shown 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 4.1.2 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  $\epsilon_n$  in (7),*

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

*Proof.* See Section A.6 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. Since 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, though 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<sup>§</sup>) For  $d_t > 0$ ,  $\lambda_t > 0$ ,  $R_t > 0$ ,  $\mathfrak{X}_t = (\Xi_{t1}, \dots, \Xi_{tR})$ , and  $A_{t,\bar{\alpha}_t} \in \mathcal{A}_{\bar{\alpha}_t}^{R_t, d_t}$  with  $\bar{\alpha}_t \in (0, 1]$ ,  $1 \leq t \leq T_0$ , the true function satisfies  $f_0 = \sum_{t=1}^{T_0} f_{0t}$  for  $f_{0t} \in \Gamma_{\lambda_t}^{A_{t,\bar{\alpha}_t}, d_t, p}(\mathfrak{X}_t)$  or  $f_{0t} \in \Gamma_{\lambda_t}^{A_{t,\bar{\alpha}_t}, d_t, p}(\mathfrak{X}_t) \cap \mathcal{C}([0, 1]^p)$ .
- (A2<sup>§</sup>) It is assumed that  $d_t$ ,  $p_t$ ,  $\lambda_t$ ,  $R_t$ , and  $\bar{\alpha}_t$  satisfy  $\epsilon_{t,n} \ll 1$ , where  $\epsilon_{t,n} = \sqrt{(d_t \log p)/n} + (\lambda_t d_t)^{d_t/(2\bar{\alpha}_t + d_t)} ((R_t \log n)/n)^{\bar{\alpha}_t/(2\bar{\alpha}_t + d_t)}$ .
- (A6<sup>§</sup>) The split-net  $\mathcal{Z}$  is suitably dense and regular to construct a  $\mathcal{Z}$ -tree partition  $\hat{\mathcal{T}}^t$  such that for  $\bar{\epsilon}_{t,n} = (\lambda_t d_t)^{d_t/(2\bar{\alpha}_t + d_t)} ((R_t \log n)/n)^{\bar{\alpha}_t/(2\bar{\alpha}_t + d_t)}$ , there exists  $\hat{f}_{0t} \in \mathcal{F}_{\hat{\mathcal{T}}^t}$  satisfying  $\|f_{0t} - \hat{f}_{0t}\|_n \lesssim \bar{\epsilon}_{t,n}$  by Theorem 1,  $1 \leq t \leq T_0$ .
- (A7<sup>§</sup>) The  $\mathcal{Z}$ -tree partition  $\mathcal{T}_t^* = (\Omega_{t1}^*, \dots, \Omega_{tR}^*)$  approximating  $\mathfrak{X}_t^*$  satisfies  $\max_r \text{dep}(\Omega_{tr}^*) \lesssim \log n$ ,  $1 \leq t \leq T_0$ .

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 fact is due 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<sup>§</sup>)–(A2<sup>§</sup>), (A3)–(A5), and (A6<sup>§</sup>)–(A7<sup>§</sup>) and the prior assigned through (P1)–(P3). If  $T_0 \leq T$ , there exists a constant  $M > 0$  such that for  $\epsilon_n^* = \sqrt{\sum_{t=1}^{T_0} \epsilon_{t,n}^2}$ ,*

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

*Proof.* See Section A.6 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  $\epsilon_n^*$  matches the minimax rate for high-dimensional additive regression (Yang and Tokdar, 2015). We strongly believe that  $\epsilon_n^*$  is indeed near-minimax optimal and this 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, however, we do not pursue this direction in this study.

## 7 Discussion

In this paper, we have 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 have derived the minimax rate of estimation of these functions in high-dimensional regression setups, extending existing results obtained

earlier *only* for isotropic functions. Next, we have formalized that Bayesian forests attain the near-optimal posterior concentration rate for these general function classes without any need for prior modification. Our results apply to a general class of estimation problems including nonparametric regression with a fixed and random design, binary classification, and density estimation.

## Appendix

### A Proofs of the main results

#### A.1 Proof of Theorem 1

*Proof of Theorem 1.* 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 we have

$$f_0(x) - \hat{f}_0(x) = \sum_{r=1}^R \sum_{k=1}^{2^{L_0}} \mathbb{1}_{\Omega_{rk}^\circ}(x) (f_0(x) - f_0(y_{rk})).$$

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

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

$$x \mapsto x^* : x^* = \operatorname{argmin}_{z \in \operatorname{clo}(\Omega_{rk}^\circ \cap \Xi_r^*)} \|x - z\|_1, \quad (12)$$

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

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

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

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

Let  $(l_{r1}, \dots, l_{rd})^\top$  be a vector returned by  $AKD(\Omega_r^*; \mathcal{Z}, \alpha, L_0, S_0)$  such that  $L_0 = \sum_{j=1}^d l_{rj}$ ,  $r = 1, \dots, R$ . Since  $x^*, y_{rk} \in \operatorname{clo}(\Omega_{rk}^\circ \cap \Xi_r^*) \subseteq \operatorname{clo}(\Xi_r^*)$  and  $f_0$  is continuous,

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

Let  $\tilde{l}_{rj} = L_0 \bar{\alpha} / (d \alpha_{rj})$  for  $r = 1, \dots, R$ ,  $j = 1, \dots, d$ , such that  $\alpha_{r1} \tilde{l}_{r1} = \dots = \alpha_{rd} \tilde{l}_{rd}$  and  $L_0 = \sum_{j=1}^d \tilde{l}_{rj}$  for every  $r$  (note that  $\tilde{l}_{rj}$  may not be integers). Then it can be seen that  $l_{rj} > \tilde{l}_{rj} - 1$  for every  $r, j$ , and hence

$$\lambda \sum_{j=1}^d 2^{-\alpha_{rj} l_{rj}} \leq 2\lambda \sum_{j=1}^d 2^{-\alpha_{rj} \tilde{l}_{rj}} \leq 2\lambda d 2^{-\bar{\alpha} L_0 / d}. \quad (16)$$

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

$$\|f_0 - \hat{f}_0\|_\infty \lesssim \lambda |S(\mathfrak{X}^*)| c_n^{\min_{r,j} \alpha_{rj}} + \lambda d 2^{-\bar{\alpha} L_0 / d}.$$

This verifies (i).

*Verification of (ii):* Observe that (ii) consists of two parts. To show the first part of (ii), we first show that when  $f_0 \in \Gamma_\lambda^{A\bar{\alpha}, d, p}(\mathfrak{X})$ , for any finite measure  $\mu$  and any fixed  $v \geq 1$ ,

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

Observe that

$$\int |f_0(x) - \hat{f}_0(x)|^v d\mu(x) = \sum_{r=1}^R \sum_{k=1}^{2^{L_0}} \int_{\Omega_{rk}^\circ} |f_0(x) - f_0(y_{rk})|^v d\mu(x). \quad (18)$$

The integral term in each summand is bounded by

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

Let  $(l_{r1}, \dots, l_{rd})^\top$  be the counter vector returned by  $AKD(\Omega_r^*; \mathcal{Z}, \alpha, L_0, S_0)$  such that  $L_0 = \sum_{j=1}^d l_{rj}$ ,  $r = 1, \dots, R$ . Using (15) and (16), observe that for every  $x, y_{rk} \in \Omega_{rk}^\circ \cap \Xi_r^* \subseteq \Xi_r^*$ ,

$$|f_0(x) - f_0(y_{rk})| \lesssim \lambda d 2^{-\bar{\alpha} L_0 / d}.$$

It follows that the first term of (19) is bounded by a constant multiple of  $\mu(\Omega_{rk}^\circ \cap \Xi_r^*) (\lambda d 2^{-\bar{\alpha} L_0 / d})^v$ . Note also that  $\sum_k \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) = \mu(\Omega_r^* \cap \Xi_r^{*c})$ . Therefore,

$$\begin{aligned} \|f_0 - \hat{f}_0\|_{v, \mu}^v &\lesssim \sum_{r=1}^R \sum_{k=1}^{2^{L_0}} \left\{ \mu(\Omega_{rk}^\circ \cap \Xi_r^*) \left( \lambda d 2^{-\bar{\alpha} L_0 / d} \right)^v + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \|f_0\|_\infty^v \right\} \\ &\leq \mu([0, 1]^p) \left( \lambda d 2^{-\bar{\alpha} L_0 / d} \right)^v + \|f_0\|_\infty^v \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}). \end{aligned} \quad (20)$$

This leads to the assertion in (17). Now, to verify the first part of (ii), it suffices to show that  $\sum_{r=1}^R \text{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \lesssim (\bar{\epsilon}_n / \|f_0\|_\infty)^v$  for  $\text{Leb}_p$  the Lebesgue measure on a  $p$ -dimensional space. For each  $r$ , we only need to consider the case  $\Xi_r^* \subsetneq \Omega_r^*$ , as  $\text{Leb}_p(\Omega_r^* \cap \Xi_r^{*c})$  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

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

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. Since  $\text{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) = 0$  for  $j \notin S(\mathfrak{X}^*) \subseteq S_0$  and  $\max_j \text{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) \leq 2c_n$ , the last expression is bounded by

$$|S(\mathfrak{X}^*)| \max_j \left\{ \text{Leb}_1([\Omega_r^* \cap \Xi_r^{*c}]_j) \prod_{k \neq j} \text{len}([\Omega_r^*]_k) \right\} \leq \frac{2c_n |S(\mathfrak{X}^*)| \text{vol}(\Omega_r^*)}{\min_j \text{len}([\Omega_r^*]_j)},$$

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

$$\sum_{r=1}^R \text{Leb}_p(\Omega_r^* \cap \Xi_r^{*c}) \leq \frac{2c_n |S(\mathfrak{X}^*)|}{\min_{r,j} \text{len}([\Xi_r^*]_j) - 2c_n} \leq \frac{3c_n |S(\mathfrak{X}^*)|}{\min_{r,j} \text{len}([\Xi_r^*]_j)}, \quad (21)$$

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

Now we verify the second part of (ii). Similar to (17), we first show that when  $f_0 \in \Gamma_\lambda^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$ , for any finite measure  $\mu$  and any  $v \geq 1$ ,

$$\|f_0 - \hat{f}_0\|_{v, \mu} \lesssim \bar{\epsilon}_n, \quad \text{if} \quad c_n^{v \min_{r,j} \alpha_{rj}} \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}) \lesssim (\bar{\epsilon}_n / (\lambda |S(\mathfrak{X}^*)|))^v. \quad (22)$$

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

$$\mu(\Omega_{rk}^\circ \cap \Xi_r^*) \left( \lambda \sum_{j=1}^d \text{len}([\Omega_{rk}^\circ]_{s_{0,j}})^{\alpha_{rj}} \right)^v + \int_{\Omega_{rk}^\circ \cap \Xi_r^{*c}} |f_0(x) - f_0(y_{rk})|^v d\mu(x). \quad (23)$$

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

$$\begin{aligned} & 2^{v-1} \int_{\Omega_{rk}^\circ \cap \Xi_r^{*c}} (|f_0(x) - f_0(x^*)|^v + |f_0(x^*) - f_0(y_{rk})|^v) d\mu(x) \\ & \leq 2^{v-1} \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \left\{ \left( \lambda |S(\mathfrak{X}^*)| c_n^{\min_{r,j} \alpha_{rj}} \right)^v + \left( \lambda \sum_{j=1}^d \text{len}([\Omega_{rk}^\circ]_{s_{0,j}})^{\alpha_{rj}} \right)^v \right\}, \end{aligned}$$

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

$$\mu(\Omega_{rk}^\circ) \left( \lambda d 2^{-\bar{\alpha} L_0 / d} \right)^v + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \left( \lambda |S(\mathfrak{X}^*)| c_n^{\min_{r,j} \alpha_{rj}} \right)^v,$$

and we have that

$$\begin{aligned} \|f_0 - \hat{f}_0\|_{v, \mu}^v & \lesssim \sum_{r=1}^R \sum_{k=1}^{2L_0} \left\{ \mu(\Omega_{rk}^\circ) \left( \lambda d 2^{-\bar{\alpha} L_0 / d} \right)^v + \mu(\Omega_{rk}^\circ \cap \Xi_r^{*c}) \left( \lambda |S(\mathfrak{X}^*)| c_n^{\min_{r,j} \alpha_{rj}} \right)^v \right\} \\ & \leq \mu([0, 1]^p) \left( \lambda d 2^{-\bar{\alpha} L_0 / d} \right)^v + \left( \lambda |S(\mathfrak{X}^*)| c_n^{\min_{r,j} \alpha_{rj}} \right)^v \sum_{r=1}^R \mu(\Omega_r^* \cap \Xi_r^{*c}). \end{aligned} \quad (24)$$

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

*Verification of (iii):* We again use the result in (17). Now, take  $P_{\mathcal{Z}}$  for  $\mu$ . 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. Since we have  $\sum_{r=1}^R P_{\mathcal{Z}}(\Omega_r^* \cap \Xi_r^{*c}) = 0$  in this case, (iii) easily follows.  $\square$

## A.2 Proof of Lemma 1

To prove Lemma 1, we first give 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  $J \geq 2$ ,  $\mathcal{Z}$  is  $(\mathfrak{Y}, 1/b_n^{1/p})$ -dense if  $\min_{r,j} \text{len}([\Psi_r]_j) \geq b_n^{-1/p}$ .*
- (ii) *For any  $S \subseteq \{1, \dots, p\}$ ,  $\alpha \in (0, 1]^d$ ,  $\Psi \subseteq [0, 1]^p$ , and  $L = \lfloor \log_2(b_n^{1/p} \min_j \text{len}([\Psi]_j) - 1) \rfloor$  with any constant  $M > 0$ ,  $\mathcal{Z}$  is  $(\Psi, \alpha, L, S)$ -regular if  $\min_{r,j} \text{len}([\Psi_r]_j) \geq 3b_n^{-1/p}$ .*

*Proof. Verification of (i):* Consider the  $p$ -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. Since 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  $\Upsilon(\mathfrak{Y}, \mathcal{T}) \leq 1/b_n^{1/p}$  if  $\min_{r,j} \text{len}([\Psi_r]_j) \geq 1/b_n^{1/p}$ . The assertion easily follows.

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

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

Thus, in every coordinate, midpoint-splits can occur  $\lfloor b_n^{1/p} \min_j \text{len}([\Psi]_j) \rfloor - 1$  times without choosing the leftmost and rightmost split-points (these two points may produce too small cells). This allows us to choose  $L = \lfloor \log_2(b_n^{1/p} \min_j \text{len}([\Psi]_j) - 1) \rfloor$  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 a splitting coordinate  $j \in S$ , a mid-point split choose  $\lceil \tilde{b}_j(\mathcal{Z}, \Psi)/2 \rceil$ th split-candidate in  $[\mathcal{Z}]_j \cap \text{int}([\Psi]_j)$  as a split-point  $\tau_j$ . The resulting two cells have at most  $\lfloor \tilde{b}_j(\mathcal{Z}, \Psi)/2 \rfloor$  split-points in coordinate  $j$ . Therefore, using (25),

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

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

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

This leads to the assertion.  $\square$

Now, we prove Lemma 1 below.

*Proof of Lemma 1.* If  $R = 1$ , it is obvious that  $\mathcal{Z}$  is  $(\mathfrak{X}^*, 0)$ -dense. If  $R > 1$ , by (i) of Lemma 3,  $\mathcal{Z}$  is  $(\mathfrak{X}^*, 1/b_n^{1/p})$ -dense since  $b_n^{1/p} \min_{r,j} \text{len}([\Xi_r^*]_j) \gg 1$ . Then, there exists an  $S_0$ -chopped  $\mathcal{Z}$ -tree partition approximating  $\mathfrak{X}^*$ , denoted by  $\mathcal{T}^* = (\Omega_1^*, \dots, \Omega_R^*)$ , such that  $\Upsilon(\mathfrak{X}^*, \mathcal{T}^*) \leq 1/b_n^{1/p}$ .

Observe that  $\min_{r,j} \text{len}([\Omega_r^*]_j) \geq \min_{r,j} \text{len}([\Xi_r^*]_j) - 2/b_n^{1/p}$ . This inequality gives us  $b_n^{1/p} \min_{r,j} \text{len}([\Omega_r^*]_j) \gg 1$  since  $b_n^{1/p} \min_{r,j} \text{len}([\Xi_r^*]_j) \gg 1$ . Therefore, by (ii) of Lemma 3,  $\mathcal{Z}$  is  $(\Omega_r^*, \alpha_r, L_r, S_0)$ -regular for  $L_r = \lfloor \log_2(b_n^{1/p} \min_j \text{len}([\Omega_r^*]_j) - 1) \rfloor$ ,  $r = 1, \dots, R$ . 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  $L_0 \leq \min_r L_r$ . Since  $2^{L_0} \asymp (n(\lambda d)^2 / (R \log n))^{d/(2\bar{\alpha}+d)}$ ,  $L_0$  can be chosen to be  $2^{L_0} \leq C_1(n(\lambda d)^2 / (R \log n))^{d/(2\bar{\alpha}+d)}$  for small enough  $C_1 > 0$  as desired. Thus, a sufficient condition for  $L_0 \leq \min_r L_r$  is  $(n(\lambda d)^2 / (R \log n))^{d/(2\bar{\alpha}+d)} \lesssim b_n^{1/p} \min_j \text{len}([\Omega_r^*]_j)$ . Since  $\min_{r,j} \text{len}([\Omega_r^*]_j) \geq \min_{r,j} \text{len}([\Xi_r^*]_j) - 2/b_n^{1/p}$ , it suffices to show that  $(n(\lambda d)^2 / (R \log n))^{d/(2\bar{\alpha}+d)} \lesssim b_n^{1/p} \min_{r,j} \text{len}([\Xi_r^*]_j)$  as soon as  $b_n^{1/p} \min_{r,j} \text{len}([\Xi_r^*]_j) \gg 1$ .  $\square$

### A.3 Proof of Lemma 2

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

### A.4 Proof of Theorem 2

We deploy the standard theory on posterior contraction (Ghosal et al., 2000; Ghosal and van der Vaart, 2007). The general theory requires enough prior mass around the true parameters. To be in line with this, we first show that the approximating tree partition is supported with sufficient prior mass.

**Lemma 4** (Prior concentration). *Let  $\hat{\mathcal{T}}$  be the  $\mathcal{Z}$ -tree partition defined in Section 4.2. Under Assumptions (A5) and (A7),  $\log \Pi(\hat{\mathcal{T}}) \gtrsim -\hat{K} \log n - d \log p$ .*

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

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

Fix  $\eta^* = (\eta_1^*, \dots, \eta_p^*)^\top \in \mathbb{S}^p$  such that  $\eta_j^* = 1/d$ ,  $j \in S_0$ , and  $\eta_j^* = 0$ ,  $j \notin S_0$ . Then it can be easily shown that  $V_1 \supseteq \{\eta \in \mathbb{S}^p : \|\eta - \eta^*\|_1 \leq 1/(2d)\}$ . By (55) of Lemma 8, 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  $\mathfrak{X}^*$ , and the remaining splits generate  $\mathcal{T}_r^\circ$ , the tree partition of  $\Omega_r^*$  constructed by an anisotropic  $k$ -d tree,  $r = 1, \dots, R$ . Hence, we can write

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

We first focus on the prior probability  $\Pi(\mathcal{T}^* \text{ is a subtree of } \widehat{\mathcal{T}}|V_1)$ . To generate  $\mathcal{T}^*$ , the root node is subdivided  $R - 1$  times in a top-down manner. Since each node splits with probability  $\nu^{\ell+1}$  for depth  $\ell$ , this happens with probability at least  $\nu^{(R-1) \max_r \text{dep}(\Omega_r^*)}$  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  $\eta_j$ ,  $j \in S_0$ , 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:

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

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

Now, we obtain a lower bound of  $\Pi(\mathcal{T}_r^\circ | \Omega_r^*, V_1)$ . In splitting each  $\Omega_r^*$ , observe that  $2^k$  cells split at depth  $k = 0, \dots, L_0 - 1$ , and each cell splits with probability  $\nu^{\text{dep}(\Omega_r^*) + k + 1}$  at depth  $k$ . Note also 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 above,

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

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 gives that  $\sum_{r=1}^R \log \Pi(\mathcal{T}_r^\circ | \Omega_r^*, V_1) \gtrsim -R 2^{L_0} \log(2d \max_{1 \leq j \leq p} b_j(\mathcal{Z})) - R 2^{L_0} \max_r \text{dep}(\Omega_r^*) - R L_0 2^{L_0} \gtrsim -R 2^{L_0} \log n$  since  $L_0 \lesssim \log n$ .

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

Now we prove Theorem 2 below.

*Proof of Theorem 2.* We write  $\rho_n^2((f_1, \sigma_1^2), (f_2, \sigma_2^2)) = \|f_1 - f_2\|_n^2 + |\sigma_1^2 - \sigma_2^2|^2$  for any  $f_1, f_2 : \mathbb{R}^p \mapsto \mathbb{R}$  and any  $\sigma_1^2, \sigma_2^2 \in (0, \infty)$ . By Lemma 1 in the supplementary material of Salomond (2018), if the  $\epsilon_n$ -covering number of  $\Theta_{j,n} = \{(f, \sigma^2) : j\epsilon_n \leq \rho_n((f, \sigma^2), (f_0, \sigma_0^2)) \leq (j+1)\epsilon_n\}$



is bounded by  $e^{Cn\epsilon_n^2 j^2}$  for some  $C > 0$ , there exists a test function  $\varphi_n$  such that for some  $K_1 > 0$  and every large  $j$ ,

$$\mathbb{E}_0 \varphi_n \rightarrow 0, \quad \sup_{(f, \sigma^2) \in \Theta_{j,n}} \mathbb{E}_{f, \sigma^2} (1 - \varphi_n) \leq e^{-K_1 n \epsilon_n^2 j^2}.$$

We write  $\mathcal{F}_* = \cup_{\mathcal{E}} \mathcal{F}_{\mathcal{E}}$ , where the union is taken over all  $\mathcal{E}$  generated by a given  $\mathcal{Z}$ . By Theorem 8.12 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  $\bar{c} > 0$ ,

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

$$\log N(\epsilon_n, \Theta_n, \rho_n) \lesssim n \epsilon_n^2, \quad (27)$$

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

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

We first verify (26). By direct calculations,

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

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

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

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

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

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

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

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

with  $\beta^1 \in \mathbb{R}^{\hat{K}}$  and  $\beta^t \in \mathbb{R}$ ,  $t = 2, \dots, T$ . Hence, letting  $\hat{B} = (\hat{\beta}^\top, 0, \dots, 0)^\top$ , we can write  $f_{0, \hat{\tau}, \hat{\beta}} = f_{0, \hat{\mathcal{E}}, \hat{B}}$  for  $f_{0, \hat{\mathcal{E}}, \hat{B}}$  defined with the ensemble components  $(\hat{\mathcal{E}}, \hat{B})$ . Putting the bounds together, for some  $C_2 > 0$ ,

$$\Pi(f \in \mathcal{F}_{\hat{\mathcal{E}}} : \|f - f_0\|_\infty \leq C_1 \epsilon_n) \geq \Pi(f \in \mathcal{F}_{\hat{\mathcal{E}}} : \|f - f_{0, \hat{\mathcal{E}}, \hat{B}}\|_\infty \leq C_2 \epsilon_n). \quad (30)$$

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

$$\begin{aligned} \|f_{\hat{\mathcal{E}}, B_1} - f_{\hat{\mathcal{E}}, B_2}\|_\infty &= \left\| \beta_1^1 + \sum_{t=2}^T \beta_1^t - \beta_2^1 - \sum_{t=2}^T \beta_2^t \right\|_\infty \\ &\leq \|\beta_1^1 - \beta_2^1\|_1 + \sum_{t=2}^T |\beta_1^t - \beta_2^t| \\ &\leq \|B_1 - B_2\|_2 \sqrt{\hat{K}_*}, \end{aligned}$$

where  $\hat{K}_* = \hat{K} + T - 1$ . It follows that (30) is bounded below by

$$\Pi\left(B \in \mathbb{R}^{\hat{K}_*} : \|B - \hat{B}\|_2 \leq C_2 \epsilon_n / \sqrt{\hat{K}_*}\right).$$

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  $D \in \mathbb{R}^{\hat{K}_* \times \hat{K}_*}$  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

$$\begin{aligned} \Pi\left(B \in \mathbb{R}^{\hat{K}_*} : \|D(B - \hat{B})\|_2 \leq C_2 \epsilon_n \sigma_{\max}^{-1}(D^{-1}) / \sqrt{\hat{K}_*}\right) \\ \geq 2^{-\hat{K}_*/2} e^{-\|D\hat{B}\|_2^2} \Pi\left(B \in \mathbb{R}^{\hat{K}_*} : \|DB\|_2 \leq C_2 \epsilon_n \sigma_{\max}^{-1}(D^{-1}) / \sqrt{2\hat{K}_*}\right), \end{aligned} \quad (31)$$

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

$$\Pi(B \in \mathbb{R}^{\hat{K}_*} : \|DB\|_2 \leq C_2 \tilde{\epsilon}_n / \sqrt{2}) \geq \frac{2/\hat{K}_*}{2^{\hat{K}_*} \Gamma(\hat{K}_*/2)} (C_2 \tilde{\epsilon}_n)^{\hat{K}_*} e^{-C_2^2 \tilde{\epsilon}_n^2 / 4}.$$

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

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

as soon as  $\|f_0\|_\infty \lesssim \sqrt{\log n}$ . Putting everything together, we conclude that there exists a constant  $\bar{c}$  such that  $\Pi(B_n) \geq e^{-\bar{c}n\epsilon_n^2}$ .

Next we verify the entropy condition (27). We denote by  $\mathcal{E}_{S,K^1,\dots,K^T}$  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  $\mathcal{F}_{\mathcal{E},M}^{(1)} = \{f_{\mathcal{E},B} \in \mathcal{F}_{\mathcal{E}} : \|B\|_\infty \leq M\}$  and  $\mathcal{F}_{\mathcal{E},M}^{(2)} = \{f_{\mathcal{E},B} \in \mathcal{F}_{\mathcal{E}} : \|B\|_\infty > M\}$  such that  $\mathcal{F}_{\mathcal{E},M}^{(1)} \cup \mathcal{F}_{\mathcal{E},M}^{(2)} = \mathcal{F}_{\mathcal{E}}$ . We also define

$$\mathcal{F}_{\bar{s}_n, \bar{K}_n, M}^{(\ell)} := \bigcup_{\mathcal{E} \in \mathcal{E}_{S,K^1,\dots,K^T} : |S| \leq \bar{s}_n, K^t \leq \bar{K}_n, t=1,\dots,T} \mathcal{F}_{\mathcal{E},M}^{(\ell)}, \quad \ell = 1, 2, \quad (32)$$

for  $\bar{K}_n \asymp n\epsilon_n^2 / \log n$  and  $\bar{s}_n \asymp n\epsilon_n^2 / \log p$ . That is,  $\mathcal{F}_{\bar{s}_n, \bar{K}_n, M}^{(\ell)}$  is the collection of all  $\mathcal{F}_{\mathcal{E},M}^{(\ell)}$  such that  $K^t \leq \bar{K}_n$  and  $|S| \leq \bar{s}_n$ . We take  $\Theta_n = \mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(1)} \times (n^{-M_2}, e^{M_2 n \epsilon_n^2})$  for large  $M_1, M_2 > 0$ . First it is easy to see that  $\log N(\epsilon_n, (n^{-M_2}, e^{M_2 n \epsilon_n^2}), \|\cdot\|) \lesssim n\epsilon_n^2$ . We also obtain that

$$\begin{aligned} & N\left(\epsilon_n, \mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(1)}, \|\cdot\|_n\right) \\ & \leq \sum_{S: |S| \leq \bar{s}_n} \sum_{(K^1, \dots, K^T): K^t \leq \bar{K}_n, t=1, \dots, T} \sum_{\mathcal{E} \in \mathcal{E}_{S,K^1,\dots,K^T}} N\left(\epsilon_n, \mathcal{F}_{\mathcal{E}, n^{M_1}}^{(1)}, \|\cdot\|_\infty\right). \end{aligned} \quad (33)$$

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

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

Observe that the cardinality of the set  $\mathcal{E}_{S, \bar{K}^1, \dots, \bar{K}^T}$  is equal to  $\prod_{t=1}^T |\mathcal{T}_{S, K^t, \mathcal{Z}}| \leq |\mathcal{T}_{S, \bar{K}_n, \mathcal{Z}}|^T$ . Hence, (33) is further bounded by

$$(\bar{K}_n)^T \times N\left(\frac{\epsilon_n}{T \bar{K}_n}, \left\{B \in \mathbb{R}^{T \bar{K}_n} : \|B\|_\infty \leq n^{M_1}\right\}, \|\cdot\|_\infty\right) \sum_{S: |S| \leq \bar{s}_n} |\mathcal{T}_{S, \bar{K}_n, \mathcal{Z}}|^T. \quad (34)$$

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

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

Therefore, (34) is further bounded by  $(\bar{K}_n)^T s_n p^{\bar{s}_n} (\bar{s}_n \max_{1 \leq j \leq p} b_j)^{T \bar{K}_n} (3T \bar{K}_n n^{M_1} / \epsilon_n)^{T \bar{K}_n}$ . This shows that the entropy, the logarithm of the covering number, 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  $\max_{1 \leq j \leq p} b_j \lesssim \log n$ . This verifies (27).

Next, we verify (28). First, it is easy to see that  $\Pi(\sigma^2 \notin (n^{-M_2}, e^{M_2 n \epsilon_n^2})) e^{\bar{c} n \epsilon_n^2} \rightarrow 0$  if  $M_2$  is large enough, using the tail probabilities of inverse gamma distributions. Now we

show that  $\Pi(\mathcal{F}_* \setminus \mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(1)})e^{\bar{c}n\epsilon_n^2} \rightarrow 0$ . Note first that using the tail probability of normal distributions,

$$\begin{aligned} \Pi(\mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(2)}) &\leq \sum_{S: |S| \leq \bar{s}_n} \sum_{(K^1, \dots, K^T): K^t \leq \bar{K}_n, t=1, \dots, T} \sum_{\mathcal{E} \in \mathcal{E}_{S, K^1, \dots, K^T}} \Pi(\mathcal{F}_{\mathcal{E}, M}^{(2)}) \\ &\leq (\bar{K}_n)^T \bar{s}_n p^{\bar{s}_n} \left( \bar{s}_n \max_{1 \leq j \leq p} b_j \right)^{T\bar{K}_n} 2T\bar{K}_n e^{-n^2 M_1/2}. \end{aligned}$$

The right most side is  $o(e^{-\bar{c}n\epsilon_n^2})$  as soon as  $\max_{1 \leq j \leq p} b_j \lesssim \log n$  and  $M_1 > 1/2$ . Now observe that

$$\begin{aligned} &\Pi\left(\mathcal{F}_* \setminus (\mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(1)} \cup \mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(2)})\right) \\ &\leq \sum_{t=1}^T \Pi(K^t > \bar{K}_n) + \Pi(S : s > \bar{s}_n | K^t \leq \bar{K}_n, t = 1, \dots, T). \end{aligned} \tag{35}$$

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  $\epsilon_n \gtrsim n^{-1/2}$ , 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 (35), we define the set

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

for  $\kappa_n$  specified below. By (56) of Lemma 8, it can be shown that the prior satisfies  $\Pi(V_2) \leq e^{-C(\xi-1)\bar{s}_n \log p - \log \kappa_n}$ . Hence,

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

The term  $\Pi(S : s > \bar{s}_n | K^t \leq \bar{K}_n, t = 1, \dots, T, V_2^c)$  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  $\eta$  is available, this probability is

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

Conditional on  $V_2^c$ , the last expression is further bounded by  $1 - (1 - \kappa_n)^{T\bar{K}_n} \leq \kappa_n T\bar{K}_n$ . Choose  $\kappa_n = e^{-(\bar{c}+1)n\epsilon_n^2}$ ,  $\bar{K}_n = \lfloor M_3 n\epsilon_n^2 / \log n \rfloor$ , and  $\bar{s}_n = \lfloor M_3 n\epsilon_n^2 / \log p \rfloor$  for a sufficiently large  $M_3 > 0$ . Then, a union bound for (35) gives  $\Pi(\mathcal{F}_* \setminus \mathcal{F}_{\bar{s}_n, \bar{K}_n, n^{M_1}}^{(1)}) \ll e^{-\bar{c}n\epsilon_n^2}$ .  $\square$

## A.5 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](#)). The 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). That is, the minimax lower bound  $\hat{\gamma}_n$  is the solution to  $\log D(\hat{\gamma}_n, \mathcal{F}, \|\cdot\|_{2,Q}) = n\hat{\gamma}_n^2$  for the target function space  $\mathcal{F}$ . We will show that the solution satisfies  $\hat{\gamma}_n = M_d \gamma_n$  for some  $M_d > 0$  depending on  $d$ . This requires suitable upper and lower bounds of the metric entropy. Below we write  $\overline{\mathcal{H}}_{\lambda,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}) = \{h \in \mathcal{H}_{\lambda}^{A_{\bar{\alpha}},d}(\mathfrak{X}) : \|h\|_{\infty} \leq M\lambda\}$  for  $M > 0$ . (We only directly obtain the lower bound of the minimax rate, and hence  $\lambda$  needs not be bounded; see Theorem 6 of Yang and Barron (1999).)

In the following two lemmas, we first 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. We hence first formalize this in the following lemma. We obtain the metric entropy bound for the space  $\overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X})$ , which is not worse than that for  $\overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}) \cap \mathcal{C}([0,1]^d)$ . This result implies that the Le Cam equation gives the same lower bound for the two spaces.

**Lemma 5** (Covering number, upper bound). *For  $d > 0$ ,  $R > 0$ , a partition  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  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  $\log \text{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$ ,  $1 \leq r \leq R$ ,  $1 \leq j \leq d$ , there exist constants  $\epsilon_0 > 0$  and  $M_0 > 1$  such that for any  $\epsilon < \epsilon_0$ ,*

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

*Proof.* Let  $C_1 > 0$  be a constant such that  $\log \text{len}([\Xi_r]_j) \geq -C_1/\alpha_{rj}$  for every  $r$  and  $j$  by assumption. For a sufficiently small  $C_2 > 0$ , choose  $\delta \in (0, \min\{e^{-C_1}, C_2/d\})$  such that  $\min_{r,j} \text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} > 1$ . On each  $\text{clo}(\Xi_r)$ , consider a grid that is a Cartesian product  $\prod_{j=1}^d \{I_{rj}^L, I_{rj}^L + u_{rj}, I_{rj}^L + 2u_{rj}, \dots, I_{rj}^L + \text{len}([\Xi_r]_j)\}$ , where  $u_{rj} = \text{len}([\Xi_r]_j) / \lceil \text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} \rceil$  is the mesh-size and  $I_{rj}^L$  is the left-boundary of  $\Xi_r$  in coordinate  $j$ . Then for each  $r$ , we can index the grid as  $\tilde{\mathcal{G}}_r = \{x_r^\ell \in \text{clo}(\Xi_r), \ell = 1, \dots, \tilde{m}_r\}$  with  $\tilde{m}_r = \prod_{j=1}^d (1 + \lceil \text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} \rceil)$ . One can easily see that  $\tilde{m}_r \leq \prod_{j=1}^d (2 + \text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}}) \leq \text{vol}(\Xi_r) 3^d \delta^{-d/\bar{\alpha}}$ .

For every  $h \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X})$ , we define the vector

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

Note that for every  $x = (x_1, \dots, x_d)^\top \in \Xi_r$  with given  $r$ , there exists  $x_r^\ell = (x_{r1}^\ell, \dots, x_{rd}^\ell)^\top \in \tilde{\mathcal{G}}_r$  such that  $\sum_{j=1}^d |x_j - x_{rj}^\ell|^{\alpha_{rj}} \leq d\delta$  since  $u_{rj} \leq \delta^{1/\alpha_{rj}}$ . Hence for such  $x$  and  $x_r^\ell$ , all functions  $h_1, h_2 \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X})$  such that  $Gh_1 = Gh_2$  satisfy

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

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

ranges over  $\overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X})$ . Without loss of generality, we assume that  $(x_r^\ell)_{\ell=1}^{\tilde{m}_r}$  is appropriately sorted such that for every  $\ell > 1$ , there exists  $\ell' < \ell$  such that  $\sum_{j=1}^d |x_{rj}^{\ell'} - x_{rj}^\ell|^{\alpha_{rj}} \leq \delta$ . For the enumeration, we begin with the first element of  $Gh$ , which is defined with  $x_1^1 \in \tilde{\mathcal{G}}_1$ . Since  $\|h\|_\infty \leq M$ , the number of possible values of  $\lfloor h(x_1^1)/\delta \rfloor$  does not exceed  $2M/\delta + 1$ . For the remainder defined with  $x_1^\ell \in \tilde{\mathcal{G}}_1$ ,  $2 \leq \ell \leq \tilde{m}_1$ , there exists  $\ell' < \ell$  such that

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

It follows that for a given  $\lfloor h(x_1^{\ell'})/\delta \rfloor$ , the number of possible values of  $\lfloor h(x_1^\ell)/\delta \rfloor$  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  $(2M/\delta + 1)7^{\tilde{m}_1 - 1}$ . Next, since  $h \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X})$  can be discontinuous at the boundaries of the pieces of  $\mathfrak{X}$ , the  $(\tilde{m}_1 + 1)$ th element  $Gh$ , defined with  $x_2^1 \in \tilde{\mathcal{G}}_2$ , has no restriction. Hence the number of possible values of  $\lfloor h(x_2^1)/\delta \rfloor$  is at most  $2M/\delta + 1$  similar to the case with  $r = 1$  above. The number of possible values of  $\lfloor h(x_2^\ell)/\delta \rfloor$  is at most 7 for every  $2 \leq \ell \leq \tilde{m}_2$ . This concludes that the number of possible values of the next  $\tilde{m}_2$  elements of  $Gh$  is bounded by  $(2M/\delta + 1)7^{\tilde{m}_2 - 1}$ . Concatenating this for all  $r$ , one can easily see that the number of possible vectors  $Gh$  is at most  $\prod_{r=1}^R (2M/\delta + 1)7^{\tilde{m}_r - 1} = (2M/\delta + 1)^R 7^{\tilde{m} - R}$ . Taking  $\epsilon = 3d\delta$ ,

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

Since  $\log(6Md/\epsilon + 1) \lesssim (6Md/\epsilon)^{d/\bar{\alpha}}$  and  $\log R \lesssim d/\bar{\alpha}$  (by the condition  $\log \text{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$ ,  $1 \leq r \leq R$ ,  $1 \leq j \leq d$ ), the last expression is bounded by  $(M_0 d/\epsilon)^{d/\bar{\alpha}}$  for some  $M_0 > 0$ . In order to complete the proof, it only remains to show that there exists a small constant  $\epsilon_0 > 0$  such that  $\epsilon = 3d\delta < \epsilon_0$ . This is easily satisfied with a sufficiently small  $C_2 > 0$  since  $\delta < C_2/d$ .  $\square$

**Lemma 6** (Packing number, lower bound). *For  $d > 0$ ,  $R > 0$ , a partition  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  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  $\log \text{len}([\Xi_r]_j) \gtrsim -1/\alpha_{rj}$ ,  $1 \leq r \leq R$ ,  $1 \leq j \leq d$ , there exist constants  $\epsilon_1 > 0$  and  $M_1 > 1$  such that for any  $\epsilon < \epsilon_1/d$ , there are  $N \geq \exp\{1/(M_1 d \epsilon)^{d/\bar{\alpha}}\}$  functions  $h_i \in \overline{\mathcal{H}}_{1,M}^{A_{\bar{\alpha}},d}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^d)$ ,  $i = 1, \dots, N$ , and  $h_0 = 0$  satisfying*

$$\int_{[0,1]^d} h_i(x) dx_j = 0, \quad 0 \leq i \leq N, \quad 1 \leq j \leq d, \quad (37)$$

$$\|h_i - h_k\|_2 \geq \epsilon, \quad 0 \leq i < k \leq N. \quad (38)$$

*Proof.* Let  $C_1 \geq \log 8$  be a constant such that  $\log \text{len}([\Xi_r]_j) \geq -C_1/\alpha_{rj}$  for every  $r$  and  $j$  by assumption. Choose a constant  $\delta \in (0, \min\{e^{-C_1}, M\}]$  such that  $\text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} \geq 1$ ,  $1 \leq r \leq R$ ,  $1 \leq j \leq d$ . On each  $\Xi_r$ , consider a grid that is a Cartesian product  $\prod_{j=1}^d \{I_{rj}^L + u_{rj}/2, I_{rj}^L + 3u_{rj}/2, I_{rj}^L + 5u_{rj}/2, \dots, I_{rj}^L + \text{len}([\Xi_r]_j) - u_{rj}/2\}$ , where  $u_{rj} = \text{len}([\Xi_r]_j)/\lceil \text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} \rceil$

is the mesh-size and  $I_{rj}^L$  is the left-boundary of  $\Xi_r$  in coordinate  $j$ . (Compare with the grid used in the proof of Lemma 5.) Then we can index each grid as  $\mathcal{G}_r = \{x_r^\ell \in \Xi_r, \ell = 1, \dots, m_r\}$  with  $m_r = \prod_{j=1}^d \lceil \text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} \rceil$ . Combining all  $x_{(r)}^\ell$  for all  $r$ , we can then index the full grid by  $\mathcal{G} = \{x^\ell, \ell = 1, \dots, m\}$  with  $m = \sum_{r=1}^R m_r$ . Observe that  $m_r \geq \text{vol}(\Xi_r) \delta^{-d/\bar{\alpha}}$ , and hence  $m \geq \delta^{-d/\bar{\alpha}}$ .

We define the kernel  $\mathcal{K} : \mathbb{R}^d \mapsto [-1/2, 1/2]$  as

$$\mathcal{K}(x) = \sum_{a=(a_1, \dots, a_d)^\top \in \{0,1\}^d} K_a(x), \quad (39)$$

for

$$K_a(x) = (-1)^{\sum_{j=1}^d a_j} \left\{ \frac{1}{2} - \max_{1 \leq j \leq d} \left| (-1)^{a_j} x_j - \frac{1}{2} \right| \right\} \prod_{j=1}^d \mathbb{1}(0 \leq (-1)^{a_j} x_j \leq 1). \quad (40)$$

Observe that each  $K_a$  is in the shape of a hyper-pyramid which is disjointly supported on a quadrant of  $[-1, 1]^d$ , and hence  $\mathcal{K}$  is supported on  $[-1, 1]^d$ . By Lemma 9, we obtain  $\int \mathcal{K}(x) dx = 0$  and  $\|\mathcal{K}\|_2^2 = 2^d / [2(d+1)(d+2)]$ .

For each  $1 \leq \ell \leq m_r$  and  $1 \leq r \leq R$ , we define the shifted and scaled kernel function

$$\phi_r^\ell(x) = \frac{\delta}{4} \mathcal{K} \left( \frac{x_1 - x_{(r)1}^\ell}{u_{r1}/2}, \dots, \frac{x_d - x_{(r)d}^\ell}{u_{rd}/2} \right),$$

which is supported on  $\mathcal{X}_r^\ell = \prod_{j=1}^d [x_{(r)j}^\ell - u_{rj}/2, x_{(r)j}^\ell + u_{rj}/2]$ . Since  $|\mathcal{K}(x) - \mathcal{K}(y)| \leq \sum_{j=1}^d |x_j - y_j|$  for every  $x, y \in [-1, 1]^d$  by Lemma 9, we obtain that for every  $x, y$  on the support  $\mathcal{X}_r^\ell$ ,

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

where we used the inequalities  $x \leq x^a$  for any  $x \in [0, 1]$  and  $a \in [0, 1]$ , and  $u_{rj} \geq 1/(2\delta^{-1/\alpha_{rj}})$  as soon as  $\text{len}([\Xi_r]_j) \delta^{-1/\alpha_{rj}} \geq 1/2$  (note that  $\lfloor x \rfloor \leq 2x$  for  $x \geq 1/2$ ). This implies that  $\phi_r^\ell \in \mathcal{H}_1^{\alpha_r, d}(\mathcal{X}_r^\ell)$  for every  $1 \leq \ell \leq m_r$  and  $1 \leq r \leq R$ . For each 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  $\Xi_r$ . Since  $\int \phi_r^\ell(x) dx_j = 0$  for every  $j$  due to Lemma 9 and each  $\phi_r^\ell$  is a shifted copy of another,  $h_{\tilde{\omega}_r}$  satisfies  $\int h_{\tilde{\omega}_r}(x) dx_j = 0$  for every  $j$  and  $h_{\tilde{\omega}_r} \in \mathcal{H}_1^{\alpha_r, d}(\Xi_r)$ . For  $m \geq \delta^{-d/\bar{\alpha}}$ , 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, since  $\|h_\omega\|_\infty = \max_{r, \ell} \|\phi_r^\ell\|_\infty \leq \delta/8 \leq M$  and each  $h_{\tilde{\omega}_r}$  is zero at all points on the boundary of  $\Xi_r$ , it is easy to see that  $h_\omega \in \overline{\mathcal{H}}_{1, M}^{A_{\bar{\alpha}}, d}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^d)$  and  $\int h_\omega(x) dx_j = 0$ . We also have that for any  $\omega, \omega' \in \{0, 1\}^m$ ,

$$\|h_\omega - h_{\omega'}\|_2^2 \geq \sum_{\ell=1}^m (\omega_\ell - \omega'_\ell)^2 \min_{r, \ell} \int [\phi_r^\ell(x)]^2 dx = \rho(\omega, \omega') \left( \frac{\delta}{4} \right)^2 \frac{\|\mathcal{K}\|_2^2}{2^d} \min_r \prod_{j=1}^d u_{rj}, \quad (41)$$

where  $\rho(\omega, \omega') = \sum_{\ell=1}^m \mathbb{1}(\omega_\ell \neq \omega'_\ell)$  is the Hamming distance between  $\omega$  and  $\omega'$ . Since  $m \geq \delta^{-d/\bar{\alpha}} \geq \delta^{-1} \geq e^{C_1} > 8$ , the Gilbert-Varshamov bound (Lemma 2.9 of [Tsybakov](#)



(2008)) says that there exist  $N \geq 2^{m/8}$  binary strings  $\omega^{(1)}, \dots, \omega^{(N)} \in \{0, 1\}^m$  such that  $\rho(\omega^{(\ell)}, \omega^{(\ell')}) \geq m/8$ ,  $0 \leq \ell < \ell' \leq N$ , with  $\omega^{(0)} = 0$ . Since  $\min_r \prod_{j=1}^d u_{rj} \geq 1/(2\delta^{-d/\bar{\alpha}})$  and  $\|\mathcal{K}\|_2^2 = 2^d/[2(d+1)(d+2)]$  by Lemma 9, the lower bound in (41) gives that for every  $0 \leq \ell < \ell' \leq N$ ,

$$\|h_{\omega^{(\ell)}} - h_{\omega^{(\ell')}}\|_2^2 \geq \frac{m}{8} \left(\frac{\delta}{4}\right)^2 \frac{1}{2(d+1)(d+2)} \times \frac{1}{2\delta^{-d/\bar{\alpha}}} \geq 2^{-10} \left(\frac{\delta}{d}\right)^2.$$

Letting  $\epsilon = 2^{-5}\delta/d$ , the previous lower bound gives  $\|h_{\omega^{(\ell)}} - h_{\omega^{(\ell')}}\|_2 \geq \epsilon$  while  $N \geq 2^{m/8} \geq \exp(\delta^{-d/\bar{\alpha}}(\log 2)/8) \geq \exp(1/(2^9 d\epsilon)^{d/\bar{\alpha}})$ . Since  $\delta$  is a constant, this holds for every  $\epsilon < \epsilon_1/d$  for some  $\epsilon_1 > 0$ .  $\square$

**Lemma 7** (Entropy with sparsity). *For  $d > 0$ ,  $\lambda > 0$ ,  $R > 0$ , a partition  $\mathfrak{X} = (\Xi_1, \dots, \Xi_R)$  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  $\epsilon' < \epsilon_2/d$ ,*

$$\log D(\epsilon, \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}), \|\cdot\|_2) \leq \left(\frac{M_2 d \lambda}{\epsilon}\right)^{d/\bar{\alpha}} + \log \binom{p}{d}, \quad (42)$$

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

*Proof.* We only need to verify the assertion for  $\lambda = 1$  since  $D(\epsilon, \lambda \mathcal{F}, \|\cdot\|_2) = D(\epsilon/\lambda, \mathcal{F}, \|\cdot\|_2)$  for any set  $\mathcal{F}$ . We first verify the upper bound (42). For every  $\epsilon < \epsilon_0$ , Lemma 5 gives  $\log D(\epsilon, \bar{\mathcal{H}}_{1, M}^{A_{\bar{\alpha}}, d}(\mathfrak{X}), \|\cdot\|_2) \leq \log N(\epsilon/2, \bar{\mathcal{H}}_{1, M}^{A_{\bar{\alpha}}, d}(\mathfrak{X}), \|\cdot\|_2) \leq (2M_0 d/\epsilon)^{d/\bar{\alpha}}$ . Since  $\bar{\Gamma}_{1, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X})$  is a union of  $\binom{p}{d}$  many  $\bar{\mathcal{H}}_{1, M}^{A_{\bar{\alpha}}, d}(\mathfrak{X})$ , the assertion easily follows.

Next, we verify (43). By Lemma 6, for every  $\epsilon' < \epsilon_1/d$ , there are functions  $h_0 = 0$ ,  $h_i \in \bar{\mathcal{H}}_{1, M}^{A_{\bar{\alpha}}, d}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^d)$ ,  $1 \leq i \leq N$  satisfying (37) and (38), with  $N \geq \exp\{1/(M_1 d \epsilon')^{d/\bar{\alpha}}\}$ . This means that for any  $S \subseteq \{1, \dots, p\}$  such that  $|S| = d$ , we have that  $W_S^p h_i \in \bar{\Gamma}_{1, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  for every such  $h_i$ ,  $0 \leq i \leq N$ . Therefore,

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

Now, for any  $S \neq S' \subseteq \{1, \dots, p\}$  and  $1 \leq i \leq k \leq N$ , observe that  $\|W_S^p h_i - W_{S'}^p h_k\|_2 = (\|h_i\|_2^2 + \|h_k\|_2^2)^{1/2} \geq \epsilon'$  by (38) since  $\langle W_S^p h_i, W_{S'}^p h_k \rangle = 0$  due to (37), where we used  $h_0 = 0$ . Also for any  $S \subseteq \{1, \dots, p\}$ , it is easy to see that  $\|W_S^p h_i - W_S^p h_k\|_2 = \|h_i - h_k\|_2 \geq \epsilon'$  by (38). These imply that  $\mathcal{W}(\epsilon')$  is  $\epsilon'$ -separated, and hence the packing number  $D(\epsilon', \bar{\Gamma}_{1, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_2)$  is bounded below by the cardinality of  $\mathcal{W}(\epsilon')$ , which is  $\binom{p}{d}N$ . This leads to the assertion.  $\square$

Now we prove Theorem 3 below.

*Proof of Theorem 3.* By Theorem 1 (combined with Theorem 6) of Yang and Barron (1999), the minimax lower bound  $\hat{\gamma}_n$  is the solution to  $\log D(\hat{\gamma}_n, \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_{2, Q}) = n\hat{\gamma}_n^2$ . (We only pursue the lower bound as the upper bound requires the target space  $\bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p)$  to be uniformly bounded; see Theorem 6 of Yang and Barron (1999).) We will

show that the solution  $\hat{\gamma}_n$  satisfies  $\hat{\gamma}_n = M_d \gamma_n$  for some  $M_d > 0$  depending on  $d$ . Assume that  $d$  is not increasing too fast such that  $\gamma_n \ll 1/d$ . We use the relation  $N(\epsilon, \mathcal{F}, d) \leq D(\epsilon, \mathcal{F}, d) \leq N(\epsilon/2, \mathcal{F}, d)$  between packing and covering numbers for a metric space  $(\mathcal{F}, d)$ . Since  $\|\cdot\|_{2,Q} \leq \sqrt{q}\|\cdot\|_2$ , we obtain that

$$\begin{aligned} & \log D \left( \max\{1, 2\sqrt{q}\} M_2 d \gamma_n, \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_{2,Q} \right) \\ & \leq \log N \left( M_2 d \gamma_n, \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_2 \right) \\ & \leq \log D \left( M_2 d (\lambda^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}, \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_2 \right). \end{aligned}$$

Since  $\gamma_n \ll 1/d$ , (42) in Lemma 7 shows that the rightmost side is further bounded by  $n\gamma_n^2 \leq \max\{1, 4\bar{q}\} M_2^2 d^2 n \gamma_n^2$ . Since  $D(\epsilon, \mathcal{F}, d)$  is nondecreasing in  $\epsilon$ , the upper bound implies that  $\hat{\gamma}_n \leq \max\{1, 2\sqrt{q}\} M_2 d \gamma_n$ .

For the other direction, let

$$\kappa_n = 2 \max \left\{ (\lambda^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)}, \sqrt{\frac{1}{n} \log \binom{p}{d}} \right\} \geq \gamma_n.$$

Since  $\|\cdot\|_{2,Q} \geq \sqrt{q}\|\cdot\|_2$ , we obtain that

$$\begin{aligned} & \log D \left( \min\{1, \sqrt{q}/2\} \gamma_n / (2M_2 d), \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_{2,Q} \right) \\ & \geq \log D \left( \kappa_n / (2M_2 d), \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_2 \right). \end{aligned} \tag{44}$$

Since  $\kappa_n \leq 2\gamma_n \ll 1/d$ , we can use (43) in Lemma 7 to find a lower bound of the last expression. If  $(\lambda^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)} \geq \sqrt{n^{-1} \log \binom{p}{d}}$ , the right-hand side of (44) is equal to

$$\log D \left( (\lambda^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)} / (M_2 d), \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_2 \right) \geq n\gamma_n^2$$

If  $(\lambda^{d/\bar{\alpha}}/n)^{\bar{\alpha}/(2\bar{\alpha}+d)} < \sqrt{n^{-1} \log \binom{p}{d}}$ , (44) is equal to

$$\begin{aligned} \log D \left( \sqrt{\frac{1}{n} \log \binom{p}{d}} / (M_2 d), \bar{\Gamma}_{\lambda, M}^{A_{\bar{\alpha}}, d, p}(\mathfrak{X}) \cap \mathcal{C}([0, 1]^p), \|\cdot\|_2 \right) & \geq \left( \frac{\lambda}{\sqrt{n^{-1} \log \binom{p}{d}}} \right)^{d/\bar{\alpha}} + \log \binom{p}{d} \\ & > n\gamma_n^2. \end{aligned}$$

We thus conclude that (44) is bounded below by  $\min\{1, q/4\} n\gamma_n^2 / (4M_2^2 d^2)$ , which implies  $\hat{\gamma}_n \geq \min\{1, \sqrt{q}/2\} \gamma_n / (2M_2 d)$ . Therefore,  $d^{-1} \gamma_n \lesssim \hat{\gamma}_n \lesssim d \gamma_n$ .  $\square$

## A.6 Proof 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  $p_{f,\sigma^2}$  be the density of model (8) with  $f$  and  $\sigma^2$ . By Lemma B.1 of Xie and Xu (2018), the Hellinger distance  $\rho_H$  satisfies

$$\|f_1 - f_2\|_{2,Q}^2 + |\sigma_1^2 - \sigma_2^2|^2 \lesssim \rho_H^2(p_{f_1,\sigma_1^2}, p_{f_2,\sigma_2^2}) \lesssim \|f_1 - f_2\|_{1,Q} + |\sigma_1^2 - \sigma_2^2|^2, \quad (45)$$

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  $\Theta_n \subseteq \mathcal{F} \times [\bar{C}_2^{-1}, \bar{C}_2]$  such that for some  $\bar{c} > 0$ ,

$$\Pi(B_n) \geq e^{-\bar{c}n\epsilon_n^2}, \quad (46)$$

$$\log N(\epsilon_n, \Theta_n, \rho_H) \lesssim n\epsilon_n^2, \quad (47)$$

$$\Pi((f, \sigma^2) \notin \Theta_n) \ll e^{-\bar{c}n\epsilon_n^2}, \quad (48)$$

similar to (26)–(28), where  $B_n = \{f : K(p_0, p_{f,\sigma^2}) \leq \epsilon_n^2, V(p_0, p_{f,\sigma^2}) \leq \epsilon_n^2\}$ . Using (45), the conditions (47) and (48) can be similarly verified as in the proof of Theorem 2; only difference is that we use truncated priors, so (48) is even more easily satisfied. For (46), note that by Lemma B.2 of Xie and Xu (2018),

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

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

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

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  $\log \Pi(\sigma^2 : |\sigma^2 - \sigma_0^2| \leq C_1\epsilon_n^2) \gtrsim -\log n$  since  $|\log \sigma_0^2|$  is bounded. The rest of the proof follows similarly to that of Theorem 2.  $\square$

*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  $\bar{c} > 0$ ,

$$\Pi(B_n) \geq e^{-\bar{c}n\epsilon_n^2}, \quad (49)$$

$$\log N(\epsilon_n/36, \Theta_n, \rho_H) \lesssim n\epsilon_n^2, \quad (50)$$

$$\Pi(f \notin \Theta_n) \ll e^{-\bar{c}n\epsilon_n^2}, \quad (51)$$

where  $B_n = \{f : K(p_0, p_f) \leq \epsilon_n^2, V(p_0, p_f) \leq \epsilon_n^2\}$ . The last condition (51) follows directly from the proof of Theorem 2, so we only need to verify (49) and (50).

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

$$\begin{aligned} K(p_f, p_g) &\lesssim \|f - g\|_\infty^2 e^{\|f - g\|_\infty} (1 + \|f - g\|_\infty), \\ V(p_f, p_g) &\lesssim \|f - g\|_\infty^2 e^{\|f - g\|_\infty} (1 + \|f - g\|_\infty)^2, \\ \rho_H(p_f, p_g) &\leq \|f - g\|_\infty e^{\|f - g\|_\infty/2}. \end{aligned} \quad (52)$$

The first two assertions imply that there exists  $C_1 > 0$  such that  $B_n \supseteq \{f : \|f - f_0\|_\infty \leq C_1 \epsilon_n\}$  if  $\epsilon_n \rightarrow 0$ . Hence we follow the calculation in the proof of Theorem 2 to conclude that there exists a constant  $\bar{c} > 0$  such that  $\Pi(B_n) \geq e^{-\bar{c}n\epsilon_n^2}$ . The last assertion of (52) enables us to work with the supremum norm in the calculation of the Hellinger covering number. The entropy calculation in Theorem 2 also verifies (50), completing the proof.  $\square$

*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 (49)–(51) 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  $\rho_H(p_f, p_g) \lesssim \|f - g\|_{2,Q}$  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.  $\square$

*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

$$\log \Pi(\hat{\mathcal{E}}) + \log \Pi(f \in \mathcal{F}_{\hat{\mathcal{E}}} : \|f - f_0\|_n \leq C_1 \epsilon_n^*) \gtrsim -n(\epsilon_n^*)^2, \quad (53)$$

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

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

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

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

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

$$\Pi(f \in \mathcal{F}_{\hat{\mathcal{E}}} : \|f - f_0\|_\infty \leq C_1 \epsilon_n^*) \geq \Pi(f \in \mathcal{F}_{\hat{\mathcal{E}}} : \|f - f_{0, \hat{\mathcal{E}}, \hat{B}}\|_\infty \leq C_2 \epsilon_n^*). \quad (54)$$

For any  $B_1 = (\beta_1^{1\top}, \dots, \beta_1^{T_0\top}, \beta_1^{T_0+1}, \dots, \beta_1^T)^\top, B_2 = (\beta_2^{1\top}, \dots, \beta_2^{T_0\top}, \beta_2^{T_0+1}, \dots, \beta_2^T)^\top \in \mathbb{R}^{\hat{K}_*}$ , we write  $f_{\hat{\mathcal{E}}, B_1}, f_{\hat{\mathcal{E}}, B_2} \in \mathcal{F}_{\hat{\mathcal{E}}}$  to denote two additive tree functions that lie on the same partition ensemble  $\hat{\mathcal{E}}$ . From (2), it is easy to see that  $\|f_{\hat{\mathcal{E}}, B_1} - f_{\hat{\mathcal{E}}, B_2}\|_\infty \leq \|B_1 - B_2\|_2 \hat{K}_*^{1/2}$ . Since  $\hat{K}_* \log n \lesssim \sum_{t=1}^{T_0} \hat{K}^t \log n \lesssim n(\epsilon_n^*)^2$ , one can follow the proof of Theorem 2 to lower bound the logarithm of (54) by a constant multiple of  $-n(\epsilon_n^*)^2$ . Combined with the lower bound of  $\pi(\hat{\mathcal{E}})$ , this verifies (53). The conditions in (27) and (28) follow directly from the proof of Theorem 2, but with the rate  $\epsilon_n^*$  for the additive regression.  $\square$

## B Auxiliary results

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

**Lemma 8** (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*

$$\Pi(\|\eta - \eta^*\|_1 \leq \epsilon) \geq \exp\{-C\xi s \log(p/\epsilon)\}, \quad (55)$$

$$\Pi\left(\min_{S: |S|=s} \sum_{j \notin S} \eta_j \geq \epsilon\right) \leq \exp\{-C(\xi - 1)s \log p - \log \epsilon\}. \quad (56)$$

*Proof.* We first prove (55). Without loss of generality, we assume that the index set of nonzero entries of  $\eta^*$  is  $\{1, 2, \dots, s-1, p\}$ , i.e.,  $\eta_j^* = 0$ ,  $j = s, s+1, \dots, p-1$ . By the inequality  $|\eta_p - \eta_p^*| = |\sum_{j=1}^{p-1} \eta_j - \sum_{j=1}^{p-1} \eta_j^*| \leq \sum_{j=1}^{p-1} |\eta_j - \eta_j^*|$ , observe that  $\|\eta - \eta^*\|_1 \leq 2 \sum_{j=1}^{p-1} |\eta_j - \eta_j^*| = 2 \sum_{j=1}^{s-1} |\eta_j - \eta_j^*| + 2 \sum_{j=s}^{p-1} \eta_j$ . Hence, for  $b_0 = \epsilon/(4s)$  and  $b_1 = \epsilon/(4p - 4s)$ ,

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

Using this, we obtain that

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

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

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

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

Now, we verify (56). Consider a Dirichlet process  $\text{DP}(\zeta/p^{\xi-1}, Q_0)$  with concentration parameter  $\zeta/p^{\xi-1}$  and uniform measure  $Q_0$  on  $[0, 1]$ . Suppose a random measure  $P \sim \text{DP}(\zeta/p^{\xi-1}, Q_0)$ . Then for the intervals  $\mathcal{I}_j = [j-1)/p, j/p)$ ,  $j = 1, \dots, p$ , we have

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

This allows us to define  $\eta$  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

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

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

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

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

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

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

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

using the Markov inequality. The rightmost side verifies the assertion.  $\square$

The next lemma formalizes the properties of the kernel function used in finding an estimate packing number for the function space in our minimax study.

**Lemma 9** (Packing kernel). *The kernel function  $\mathcal{K}$  defined in (39) and (40) satisfies*

- (i)  $\int \mathcal{K}(x) dx = 0$  and  $\int \mathcal{K}^2(x) dx = 2^d / [2(d+1)(d+2)]$ ;
- (ii)  $|\mathcal{K}(x) - \mathcal{K}(y)| \leq \sum_{j=1}^d |x_j - y_j|$  for every  $x, y \in [-1, 1]^d$ .

*Proof.* We first verify (i). Every  $K_a$  in (40) is a shifted copy of  $\pm K_{(0, \dots, 0)^\top}(x)$ , depending on the sign of  $(-1)^{\sum_{j=1}^d a_j}$ . Hence,  $\int K_a(x) dx = \int K_{(0, \dots, 0)^\top}(x) dx$  if  $\sum_{j=1}^d a_j$  is even, and  $\int K_a(x) dx = -\int K_{(0, \dots, 0)^\top}(x) dx$  if  $\sum_{j=1}^d a_j$  is odd. Note that  $|\{a : \sum_{j=1}^d a_j = k\}| = \binom{d}{k}$ . Therefore,

$$\int \mathcal{K}(x) dx = \left\{ \sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} - \sum_{k=0}^{\lfloor d/2 \rfloor - 1} \binom{d}{2k+1} \right\} \int K_{(0, \dots, 0)^\top}(x) dx = 0,$$

since  $\sum_{k=0}^d \binom{d}{k} = 2^d$  and  $\sum_{k=0}^{\lfloor d/2 \rfloor} \binom{d}{2k} = 2^{d-1}$ .

Also, since each  $K_a$  is disjointly supported on each of the  $2^d$  quadrants of  $[-1, 1]^d$ , we obtain

$$\int \mathcal{K}^2(x) dx = \sum_{a=(a_1, \dots, a_d)^\top \in \{0, 1\}^d} \int K_a^2(x) dx = 2^d \int K_{(0, \dots, 0)^\top}^2(x) dx.$$

A simple direct calculation gives

$$\begin{aligned} \int K_{(0, \dots, 0)^\top}^2(x) dx &= \int_{[0, 1]^d} \left\{ \frac{1}{2} - \max_{1 \leq j \leq d} \left| x_j - \frac{1}{2} \right| \right\}^2 dx \\ &= \frac{1}{4} - 2^d \int_{[1/2, 1]^d} \max_{1 \leq j \leq d} \left| x_j - \frac{1}{2} \right| dx + 2^d \int_{[1/2, 1]^d} \max_{1 \leq j \leq d} \left| x_j - \frac{1}{2} \right|^2 dx. \end{aligned}$$

Using the change of variable followed by the use of a uniform random variable  $U_j$  on  $[0, 1]$ , it is easy to see that

$$\begin{aligned} 2^d \int_{[1/2, 1]^d} \max_{1 \leq j \leq d} \left| x_j - \frac{1}{2} \right| dx &= \frac{1}{2} \int_{[0, 1]^d} \max_{1 \leq j \leq d} x_j dx = \frac{1}{2} \mathbb{E} \max_{1 \leq j \leq d} U_j = \frac{d}{2(d+1)}, \\ 2^d \int_{[1/2, 1]^d} \max_{1 \leq j \leq d} \left| x_j - \frac{1}{2} \right|^2 dx &= \frac{1}{4} \int_{[0, 1]^d} \max_{1 \leq j \leq d} x_j^2 dx = \frac{1}{4} \mathbb{E} \max_{1 \leq j \leq d} U_j^2 = \frac{d}{4(d+2)}. \end{aligned}$$

Putting everything together, we verify that  $\int \mathcal{K}^2(x) dx = 2^d / [2(d+1)(d+2)]$ .

Now we verify (ii). For given  $x = (x_1, \dots, x_d)^\top \in [-1, 1]^d$  and  $y = (y_1, \dots, y_d)^\top \in [-1, 1]^d$ , define  $z^{(0)} = x$ ,  $z^{(d)} = y$  and  $z^{(j)} = (y_1, \dots, y_j, x_{j+1}, \dots, x_d)^\top$ ,  $j = 1, \dots, d$ . Let  $a^{(j)} = (a_1^{(j)}, \dots, a_d^{(j)})^\top \in \{0, 1\}^d$  be the binary vector such that  $z^{(j)} \in \{x : 0 \leq (-1)^{a_k^{(j)}} x_k \leq 1, k = 1, \dots, d\}$ ; hence  $a^{(j)}$  defines the quadrant which  $z^{(j)}$  belongs to. Using the fact that  $\mathcal{K}(z^{(j)}) = K_{a^{(j)}}(z^{(j)})$ , we obtain

$$|\mathcal{K}(x) - \mathcal{K}(y)| \leq \sum_{j=1}^d |\mathcal{K}(z^{(j-1)}) - \mathcal{K}(z^{(j)})| = \sum_{j=1}^d |K_{a^{(j-1)}}(z^{(j-1)}) - K_{a^{(j)}}(z^{(j)})|. \quad (57)$$

Since all elements of  $z^{(j-1)}$  and  $z^{(j)}$  are the same except for the  $j$ th one, either they belong to the same quadrant or two contiguous quadrants. If  $z^{(j-1)}$  and  $z^{(j)}$  are contained in the same quadrant, i.e.,  $a^{(j-1)} = a^{(j)} = (a_1^*, \dots, a_d^*)^\top$ , then

$$\begin{aligned} |K_{a^{(j-1)}}(z^{(j-1)}) - K_{a^{(j)}}(z^{(j)})| &= \left| \max_{1 \leq k \leq d} \left| (-1)^{a_k^*} z_k^{(j-1)} - \frac{1}{2} \right| - \max_{1 \leq k \leq d} \left| (-1)^{a_k^*} z_k^{(j)} - \frac{1}{2} \right| \right| \\ &\leq \max_{1 \leq k \leq d} \left| (-1)^{a_k^*} (z_k^{(j-1)} - z_k^{(j)}) \right| \\ &= |x_j - y_j|, \end{aligned}$$

where we used the inequalities  $|\max f - \max g| \leq \max |f - g|$  and  $||s| - |t|| \leq |s - t|$ . If each of  $z^{(j-1)}$  and  $z^{(j)}$  is contained in each of two contiguous quadrants, then  $a_j^{(j-1)} \neq a_j^{(j)}$  and  $a_k^{(j-1)} = a_k^{(j)}$ ,  $k \neq j$ . Since  $\max_{1 \leq k \leq d} |(-1)^{a_k^{(j-m)}} z_k^{(j-m)} - 1/2| \leq 1/2$  for  $m = 0, 1$ , we obtain



that

$$\begin{aligned}
|K_{a^{(j-1)}}(z^{(j-1)}) - K_{a^{(j)}}(z^{(j)})| &= 1 - \max_{1 \leq k \leq d} \left| (-1)^{a_k^{(j-1)}} z_k^{(j-1)} - \frac{1}{2} \right| - \max_{1 \leq k \leq d} \left| (-1)^{a_k^{(j)}} z_k^{(j)} - \frac{1}{2} \right| \\
&\leq 1 - \left| (-1)^{a_j^{(j-1)}} x_j - \frac{1}{2} \right| - \left| (-1)^{a_j^{(j)}} y_j - \frac{1}{2} \right|.
\end{aligned} \tag{58}$$

Because  $a_j^{(j-1)}, a_j^{(j)} \in \{0, 1\}$  but  $a_j^{(j-1)} \neq a_j^{(j)}$ , the rightmost side of (58) is equal to either  $1 - |x_j - 1/2| - |y_j + 1/2|$  or  $1 - |x_j + 1/2| - |y_j - 1/2|$ . One may easily check that  $1 - |s + 1/2| - |t - 1/2| \leq |s - t|$  for every  $s, t \in \mathbb{R}$  by splitting up the domain  $\mathbb{R}^2$ . This concludes that (58) is bounded by  $|x_j - y_j|$ . Plugging in the bounds into (57), we obtain that  $|\mathcal{K}(x) - \mathcal{K}(y)| \leq \sum_{j=1}^d |x_j - y_j|$ . Since this holds for every  $x, y \in [-1, 1]^d$ , (ii) is verified.  $\square$

## References

- Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013). Efficient gaussian process regression for large datasets. *Biometrika*, 100(1):75–89.
- Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413.
- Bentley, J. L. (1979). Multidimensional binary search trees in database applications. *IEEE Transactions on Software Engineering*, SE-5(4):333–340.
- Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics*, 42(1):352.
- Birgé, L. (1986). On estimating a density using hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71(2):271–291.
- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150.
- Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. (2014). Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, pages 1750–1781.
- Candès, E. J. and Donoho, D. L. (2000). Curvelets: A surprisingly effective nonadaptive representation for objects with edges. In *Curve and surface fitting*, pages 105–120. Vanderbilt University Press.
- Candès, E. J. and Donoho, D. L. (2004). 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.
- Castillo, I. and Ročková, V. (2021). Uncertainty quantification for bayesian cart. *The Annals of Statistics*, 49(6):3482–3509.
- Chandrasekaran, V., Wakin, M. B., Baron, D., and Baraniuk, R. G. (2008). Representation and compression of multidimensional piecewise functions using surflets. *IEEE Transactions on Information Theory*, 55(1):374–400.

- Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.
- Denison, D. G., Mallick, B. K., and Smith, A. F. (1998). A Bayesian CART algorithm. *Biometrika*, 85(2):363–377.
- Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5):1870–1911.
- Du, J. and Linero, A. R. (2019). Interaction detection with Bayesian decision tree ensembles. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 108–117.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.
- Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.
- Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.
- Guo, K. and Labate, D. (2007). Optimally sparse multidimensional representation using shearlets. *SIAM Journal on Mathematical Analysis*, 39(1):298–318.
- Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, to appear.
- Hayakawa, S. and Suzuki, T. (2020). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361.
- He, J., Yalov, S., and Hahn, P. R. (2019). XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.
- Hoffman, M. and Lepski, O. (2002). Random rates in anisotropic regression. *The Annals of Statistics*, 30(2):325–396.
- Ibragimov, I. A. and Hasminskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, volume 16. Springer.
- Imaizumi, M. and Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 869–878.

- Jeong, S. and Ghosal, S. (2021). Unified Bayesian theory of sparse linear regression with nuisance parameters. *Electronic Journal of Statistics*, 15(1):3040–3111.
- Lakshminarayanan, B., Roy, D., and Teh, Y. W. (2013). Top-down particle filtering for bayesian decision trees. In *International Conference on Machine Learning*, pages 280–288.
- Le Pennec, E. and Mallat, S. (2005). Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438.
- Lepski, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *The Annals of Statistics*, 43(3):1178–1242.
- Lepski, O. and Levit, B. Y. (1999). Adaptive non-parametric estimation of smooth multivariate functions. *Mathematical Methods of Statistics*, 8:344–370.
- Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636.
- Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110.
- Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). When gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423.
- Liu, Z. and Guo, W. (2010). Data driven adaptive spline smoothing. *Statistica Sinica*, 20(3):1143–1163.
- Neumann, M. H. and von Sachs, R. a. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *The Annals of Statistics*, 25(1):38–76.
- Ning, B., Jeong, S., and Ghosal, S. (2020). Bayesian linear regression for multivariate responses under group sparsity. *Bernoulli*, 26(3):2353–2382.
- Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in  $L_2$ . *The Annals of Statistics*, 13(3):984–997.
- Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330.
- Pintore, A., Speckman, P., and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika*, 93(1):113–125.
- Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *Journal of Computational and Graphical Statistics*, 29(2):405–417.
- Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994.

- Ray, K. and van der Vaart, A. (2020). Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020.
- Ročková, V. (2020). On semi-parametric Bernstein-von Mises theorems for BART. In *The 37th International Conference on Machine Learning*.
- Ročková, V. and Saha, E. (2019). On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848.
- Ročková, V. and van der Pas, S. (2020). Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131.
- Salomond, J.-B. (2018). Testing un-separated hypotheses by estimating a distance. *Bayesian Analysis*, 13(2):461–484.
- Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897.
- Shen, W. and Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213.
- Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753.
- Taddy, M. A., Gramacy, R. B., and Polson, N. G. (2011). Dynamic trees for learning and design. *Journal of the American Statistical Association*, 106(493):109–123.
- Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.
- Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.
- Tuo, R. and Wu, C. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.
- van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.
- Verma, N., Kpotufe, S., and Dasgupta, S. (2009). Which spatial partition trees are adaptive to intrinsic dimension? In *The 25th Conference on Uncertainty in Artificial Intelligence*, pages 565–574.
- Wang, X., Du, P., and Shen, J. (2013). Smoothing splines with varying smoothing parameter. *Biometrika*, 100(4):955–970.
- Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362.

- Xie, F. and Xu, Y. (2018). Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Analysis*, 15(1):159–186.
- Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.
- Yang, Y. and Dunson, D. B. (2014). Minimax optimal Bayesian aggregation. *arXiv preprint arXiv:1403.1345*.
- Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674.