POSTERIOR CONCENTRATION FOR BAYESIAN REGRESSION TREES AND THEIR ENSEMBLES

By Veronika Ročková and[*,‡], Stéphanie van der Pas[†,§]

29 August 2017

University of Chicago [‡]
Leiden University [§]

Since their inception in the 1980's, regression trees have been one of the more widely used non-parametric prediction methods. Tree-structured methods yield a histogram reconstruction of the regression surface, where the bins correspond to terminal nodes of recursive partitioning. Trees are powerful, yet susceptible to over-fitting. Strategies against overfitting have traditionally relied on pruning greedily grown trees. The Bayesian framework offers an alternative remedy against overfitting through priors. Roughly speaking, a good prior charges smaller trees where overfitting does not occur. While the consistency of random histograms, trees and their ensembles has been studied quite extensively, the theoretical understanding of the Bayesian counterparts has been missing. In this paper, we take a step towards understanding why/when do Bayesian trees and their ensembles not overfit. To address this question, we study the speed at which the posterior concentrates around the true smooth regression function. We propose a *spike-and-tree* variant of the popular Bayesian CART prior and establish new theoretical results showing that regression trees (and their ensembles) (a) are capable of recovering smooth regression surfaces, achieving optimal rates up to a log factor, (b) can adapt to the unknown level of smoothness and (c) can perform effective dimension reduction when $p > n$. These results provide a piece of missing theoretical evidence explaining why Bayesian trees (and additive variants thereof) have worked so well in practice.

1. Non-parametric Regression Setup. The remarkable empirical success of Bayesian tree-based regression [14, 16, 13] has raised considerable interest in understanding why and when these methods produce good results. Despite their extensive use in practice, theoretical justifications have, thus far, been unavailable. To narrow this yawning gap, we consider the funda-

[*] veronika.rockova@chicagobooth.edu;
This work was supported by the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.
[†] svdpas@math.leidenuniv.nl
Keywords and phrases: Additive Regression, Asymptotic Minimaxity, BART, Bayesian CART, Posterior Concentration, Recursive Partitioning, Regression Trees

mental problem of making inference about an unknown regression function.

Our setup consists of the nonparametric regression model

$$Y_i = f_0(\boldsymbol{x}_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \tag{1.1}$$

where output variables $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)'$ are related in a stochastic fashion to a set of p potential covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})', 1 \leq i \leq n$. We assume that the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ are fixed and have been rescaled so that $\boldsymbol{x}_i \in [0,1]^p$. The true unknown regression surface $f_0(\boldsymbol{x})$ will be assumed to be smooth, possibly involving only a small fraction q_0 of the p potential covariates.

In the absence of a parametric model, a natural strategy to estimate the unknown regression function is by partitioning the covariate space into cells and then estimating the function locally within each cell from available observations. Such strategies yield histogram reconstructions of the regression surface and have been analyzed theoretically by multiple authors [30, 17, 36]. Regression trees [9] are among the most popular data-dependent histogram methods, where the partitioning scheme is obtained through nested parallel axis splitting. Trees are an integral constituent of ensemble methods that aggregate single tree learners into forests to boost prediction [8]. Tree-based regression, either single or ensemble, is arguably one of the most popular machine learning tools today. In particular, Bayesian variants of these methods [13, 16] have earned a prominent role as one of the top machine learners. While consistency results for classical trees and random forests have been available [22, 23, 19, 9, 4, 5, 31, 40], theory on the also very widely used Bayesian counterparts is non-existent. Our goal in this paper is to provide the first frequentist optimality results for *Bayesian* trees, and their ensembles.

Most of the work on Bayesian nonparametric regression has revolved around Gaussian processes [38, 3, 41]. While there are multiple results on recursive partitioning priors for Bayesian density estimation [10, 32, 29], the literature on Bayesian regression histograms is far more deserted. One fundamental contribution is due to Coram and Lalley [15], who showed consistency of Bayesian binary regression with uniform mixture priors on step functions and with one predictor. More recently, van der Pas and Ročková [37] considered a similar setup for estimating step mean functions in Gaussian regression, again with a single predictor. This paper goes far beyond that framework, addressing (a) the full-fledged high-dimensional setup with a diverging number of potential covariates, (b) tree ensembles for additive regression.

The purpose of this paper is to study the rate of convergence of posterior

distributions induced by step function priors on the regression surface when $p > n$. The speed of convergence is measured by the size of the smallest shrinking ball around f_0 that contains most of the posterior probability. In pioneering works, Ghosal, Ghosh and Van der Vaart [20] and Shen and Wasserman [33] obtained rates of convergence for infinite-dimensional parametric models with iid observations in terms of the size of the model (measured by the metric entropy) and concentration rate of the prior around f_0. These results were later extended to infinite-dimensional models that are *not iid* by Ghosal and Van der Vaart [21]. Their general conceptual framework serves as an umbrella for our development.

1.1. *Our Contributions.* We initially assume that f_0 is Hölder continuous and may depend only on a small fraction of q_0 predictors. The optimal rate of estimation of a q_0-variable function, which is known to be α-smooth, is $n^{-\alpha/(2\alpha+q_0)}$ [34]. Our first result shows that, with suitable regularization priors, single Bayesian regression trees achieve this minimax rate (up to a log factor). In other words, the posterior behaves nearly as well as if we knew α and the number of active covariates q_0, concentrating at a rate that only depends on the number of active predictors. This is the first optimality result for Bayesian regression trees, demonstrating their adaptability and reluctance to overfit in high-dimensional scenarios with $p > n$. The regularization is achieved through our proposed *spike-and-tree* prior, a new variant of the Bayesian CART prior for dimension reduction and model-free variable selection. Going further, we show that Bayesian *additive regression trees* also achieve (near) minimax-rate optimal performance when approximating a single smooth function. Finally, the tree ensembles are *also* shown to be certifiably optimal when the true function is an actual sum of smooth functions, again concentrating at a near minimax rate.

1.2. *Notation.* The notation $\lesssim$ will be used to denote inequality up to a constant. The ε-covering number of a set Ω for a semimetric d, denoted by $N(\varepsilon, \Omega, d)$, is the minimal number of d-balls of radius ε needed to cover set Ω. We denote by $\phi(\cdot; \sigma^2)$ the normal density with zero mean and variance σ^2 and by $P_f^n = \bigotimes P_{f,i}$ the n-fold product measure of the n independent observations under (1.1) with a regression function f. By $\mathbb{P}_n^{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$ we denote the empirical distribution of the observed covariates and let $||\cdot||_n$ denote the empirical norm on $L_2(\mathbb{P}_n^{\boldsymbol{x}})$. With $||\cdot||_2$ we denote the standard Euclidean norm. For $\boldsymbol{x} \in \mathbb{R}^p$ we denote with $\boldsymbol{x}_{\mathcal{S}}$ the subvector of $\boldsymbol{x}$ indexed by $\mathcal{S} \subseteq \{1, \ldots, p\}$.

1.3. *Outline.* We outline our goals and strategy in Section 2. We then review several useful concepts for analyzing recursive partitioning schemes in Section 3. In Section 4, we state our first main result on the posterior concentration for Bayesian CART. In Section 5, we develop tools for analyzing Bayesian additive regression trees and show their optimal posterior concentration in non-additive regression. Section 6 presents the final development of our theory concerning the recovery of an additive regression function with additive trees. We conclude with a discussion in Section 7. The proofs of our main theorems are presented in Sections 8, 9 and 10.

2. Background. In this section we lay down rudiments of our modus operandi. Our setup comprises a sequence of statistical experiments with observations $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)'$ and models P_f^n defined in (1.1). Each model P_f^n is parametrized by a regression function $f : [0,1]^p \to \mathbb{R}$ that lives in an infinite-dimensional space $\mathcal{F}$ endowed with a prior distribution. With adequate priors, the posterior can exhibit nice frequentist properties, which get passed onto its location/scale summary measures. One such property is the ability to pile up in shrinking neighborhoods around the true regression function f_0. The speed at which the shrinking occurs is the posterior concentration rate and assesses the quality of the posterior beyond just the mere fact that it is consistent. In our setup, we investigate such concentration properties in terms of $||\cdot||_n$ neighborhoods of f_0, where

$$||f - f_0||_n^2 = \frac{1}{n}\sum_{i=1}^{n}[f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)]^2.$$

The key to our approach will be drawing upon the foundational posterior concentration theory for non-iid observations, laid down in the seminal paper by Ghosal and Van der Vaart [21]. Our results are obtained under a unifying hat of a general result which requires three conditions to hold. Namely, suppose that for a sequence $\varepsilon_n^2 \to 0$ such that $n\,\varepsilon_n^2$ is bounded away from zero and sets $\mathcal{F}_n \subset \mathcal{F}$ we have

$$\sup_{\varepsilon > \varepsilon_n} \log N\left(\tfrac{\varepsilon}{36}, \{f \in \mathcal{F}_n : \|f - f_0\|_n < \varepsilon\}, \|.\|_n\right) \leq n\varepsilon_n^2 \tag{2.1}$$

$$\Pi(f \in \mathcal{F} : \|f - f_0\|_n \leq \varepsilon_n) \geq \mathrm{e}^{-d\,n\,\varepsilon_n^2} \tag{2.2}$$

$$\Pi(\mathcal{F} \backslash \mathcal{F}_n) = o(\mathrm{e}^{-(d+2)\,n\varepsilon_n^2}) \tag{2.3}$$

for some $d > 2$. Then it follows from Theorem 4 in [21] that the posterior distribution concentrates at the rate ε_n^2, i.e.

$$\Pi\left(f \in \mathcal{F} : ||f_0 - f||_n > M_n\,\varepsilon_n \,|\, \boldsymbol{Y}^{(n)}\right) \to 0 \tag{2.4}$$

in $P^n_{f_0}$-probability, as $n \to \infty$, for any $M_n \to \infty$.

A few remarks are in place. The condition (2.3) ensures that the prior zooms in on smaller, and thus more manageable, sets of models $\mathcal{F}_n$ by assigning only a small probability outside these sets. The condition (2.1) is known as "the entropy condition" and controls the combinatorial richness of the approximating sets $\mathcal{F}_n$. Finally, condition (2.2) requires that the prior charges an ε_n neighborhood of the true function. We would like to remind the reader that the statement (2.4) is fundamentally a frequentist statement as it relates to a typical behavior of the posterior distribution under the generative model $P^n_{f_0}$.

The conditions (2.1), (2.2) and (2.3) provide a very general recipe for showing posterior concentration in infinite-dimensional models. Our goal in this paper is to obtain tailored statements for Bayesian regression trees and their ensembles. The major challenge will be (a) designing a sequence of approximating spaces (a sieve) $\mathcal{F}_n \subset \mathcal{F}$ and (b) endowing $\mathcal{F}$ with a prior distribution such that the three conditions hold simultaneously for ε_n as small as possible. To this end, we will build on, and develop, tools for an analysis of recursive partitioning schemes.

Throughout this work, we assume that the true regression function f_0 is Hölder continuous (or an additive composition thereof), in the sense made precise below.

DEFINITION 2.1. *With $\mathcal{H}^\alpha_p$ we denote the space of uniformly α-Hölder continuous functions, i.e.*

$$\mathcal{H}^\alpha_p = \left\{ f : [0,1]^p \to \mathbb{R};\ ||f||_{\mathcal{H}^\alpha} \equiv \sup_{\boldsymbol{x},\boldsymbol{y} \in [0,1]^p} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{||\boldsymbol{x} - \boldsymbol{y}||_2^\alpha} < \infty \right\},$$

where $\alpha \in (0,1]$ and where $||f||_{\mathcal{H}^\alpha}$ is the Hölder coefficient.

In some applications, it is reasonable to expect that the regression function f_0 depends only on a small fraction of input covariates. For a set of indices $\mathcal{S} \subseteq \{1,\dots,p\}$, we define

$$\mathcal{C}(\mathcal{S}) = \{ f : [0,1]^p \to \mathbb{R}; f \text{ is constant in directions } \{1,\dots,p\} \backslash \mathcal{S} \}.$$

We will consider two estimation regimes:

(R1) Regime 1: f_0 is α-Hölder continuous and depends on an unknown subset $\mathcal{S}_0$ of $|S_0| = q_0$ covariates, i.e. $f_0 \in \mathcal{H}^\alpha_p \cap \mathcal{C}(\mathcal{S}_0)$.

(R2) Regime 2: f_0 is an aggregate of T_0 α^t-Hölder continuous functions f^t_0, $1 \le t \le T_0$, each depending on an unknown subset $\mathcal{S}^t_0$ of $|S^t_0| = q^t_0$ covariates, i.e. $f_0(\boldsymbol{x}) = \sum_{t=1}^{T_0} f^t_0(\boldsymbol{x})$, where $f^t_0 \in \mathcal{H}^{\alpha^t}_p \cap \mathcal{C}(\mathcal{S}^t_0)$.

For an estimation procedure to be successful in Regime 1, it needs to be doubly adaptive (with respect to α and q_0). We will show that both single trees and tree ensembles adapt accordingly, performing at a near-minimax rate. Regime 2 makes the performance discrepancies more apparent, where the additive structure of f_0 is appreciated by tree ensembles, which can achieve a faster convergence rate than single trees. A variant of Regime 2 was studied by [41] who derived minimax rates for estimating additive smooth functions and showed optimal concentration of additive Gaussian processes. Here, we approximate f_0 with step functions and their aggregates. We limit considerations to step functions that are underpinned by recursive partitioning schemes.

3. On Recursive Partitions. Tree-based regression consists of first finding an underlying partitioning scheme that hierarchically subdivides a dataset into more homogeneous subsets, and then learning a piece-wise constant function on that partition. This section serves to review several useful concepts for analyzing such nested partitioning rules that will be instrumental in our analysis.

3.1. *General Partitions.* Given $K \in \mathbb{N}$, we define a *K-partition* of $[0,1]^p$ as a sequence $\mathbf{\Omega} = \{\Omega_k\}_{k=1}^K$ of K contiguous *rectangles* $\Omega_k \subset [0,1]^p$, where $\cup_{k=1}^K \Omega_k = [0,1]^p$. Sufficient conditions for consistency of regression histograms have traditionally revolved around two requirements on the partitioning cells Ω_k (Section 6.3 in [18]). The first one pertains to cell counts, where Ω_k's should be large enough to capture a sufficient number of points to render local estimation meaningful. The second one pertains to cell sizes, where Ω_k's should be small enough to detect local changes in the regression surface. We borrow some of these concepts for our theoretical analysis.

For the first requirement, let us formalize the notion of the cell size in terms of the empirical measure induced by observations $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. For each cell Ω_k, we define by

$$\mu(\Omega_k) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\boldsymbol{x}_i \in \Omega_k) \tag{3.1}$$

the proportion of observations falling inside Ω_k. Throughout this work, we focus on partitions whose boxes can adaptively stretch and shrink, allowing the splits to arrange themselves in a data-dependent way [35]. The simplest data-adaptive partition is based on statistically equivalent blocks [1, 18], where all cells have exactly the same number of points, i.e. $\mu(\Omega_k) = 1/K$. We deviate from such a strict rule by allowing for imbalance and define the so called *valid* partitions.

DEFINITION 3.1. *(Valid Partitions) Denote by* $\mathbf{\Omega} = \{\Omega_k\}_{k=1}^K$ *a partition of* $[0,1]^p$*. We say that* $\mathbf{\Omega}$ *is valid if*

$$\mu(\Omega_k) \geq \bar{C}^2/n \qquad \text{for all} \quad k = 1, \ldots, K \tag{3.2}$$

for some constant $\bar{C}^2 \in \mathbb{N}\backslash\{0\}$.

Valid partitions have non-empty cells, where the allocation *does not* need to be balanced. In *balanced partitions* (introduced in van der Pas and Ročková [37]), the cells satisfy $\frac{C_{min}^2}{K} \leq \mu(\Omega_k) \leq \frac{C_{max}^2}{K}$ for some $C_{min} < 1 < C_{max}$. One prominent example of such balanced partitions is the median tree (or a k-d tree) [2], which will be discussed in the next section and will be used as a benchmark tree approximation towards establishing Condition 2.2.

For the second requirement, let us formalize the notion of the cell size in terms of the local spread of the data. To this end, we introduce the partition *diameter* [39, 25].

DEFINITION 3.2. *(Diameter) Denote by* $\mathbf{\Omega} = \{\Omega_k\}_{k=1}^K$ *a partition of* $[0,1]^p$ *and by* $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *a collection of data points in* $[0,1]^p$*. For an index set* $\mathcal{S} \subseteq \{1, \ldots, p\}$*, we define a diameter of* Ω_k *w.r.t.* $\mathcal{S}$ *as*

$$\operatorname{diam}(\Omega_k; \mathcal{S}) = \max_{\boldsymbol{x}, \boldsymbol{y} \in \Omega_k \cap \mathcal{X}} \|\boldsymbol{x}_{\mathcal{S}} - \boldsymbol{y}_{\mathcal{S}}\|_2$$

and a diameter of the entire partition $\mathbf{\Omega}$ *w.r.t.* $\mathcal{S}$ *as*

$$\operatorname{diam}(\mathbf{\Omega}; \mathcal{S}) = \sqrt{\sum_{k=1}^K \mu(\Omega_k) \operatorname{diam}^2(\Omega_k; \mathcal{S})}.$$

The diameter of Ω_k corresponds to the largest $||\cdot||_2$ distance between $\mathcal{S}$-coordinate projections of two points that fall inside Ω_k. This is one of the more flexible notions of a cell diameter, which takes into account the data itself, not just the physical cell size. Traditionally, bias and convergence rate analysis of tree-based estimators have been characterized in terms of the cell diameters. The rate depends on how fast the diameters shrink as we move down the tree: the more rapidly, the better. As will be seen later in Section 3.3, controlling the diameter will be essential for obtaining tight bounds on the approximation error.

3.2. *Tree Partitions.* We are ultimately interested in partitions that can be obtained with *nested* axis-parallel splits. Such partitions relate to a tree growing process in the following way. Starting from a parent node $[0,1]^p$, a binary tree is grown by successively applying a splitting rule on a chosen internal node, say Ω_k. The newborn cells $\{\boldsymbol{x} \in \Omega_k : x_j \leq \tau\}$ and $\{\boldsymbol{x} \in \Omega_k : x_j > \tau\}$ are polytopes obtained by perpendicular bisection of Ω_k along one of the p coordinates, say j. Terminal cells of the splitting process after $K-1$ splits then yield a box-shaped partition $\{\Omega_k\}_{k=1}^K$.

Unlike *dyadic* trees, where the threshold τ is preset at a midpoint of the rectangle Ω_k along the j^{th} direction, we allow for splits at available observations $\mathcal{X}$. With more split opportunities, many more tree topologies can be generated. However, since the tree partitions are arranged in a nested fashion, their combinatorial complexity is not too large (as shown in Lemma 3.1 below).

We will denote each tree-structured K-partition by $\mathcal{T}^K = \{\Omega_k\}_{k=1}^K$. With $\mathcal{V}_{\mathcal{S}}^K$ we denote a *family of valid tree partitions* of $[0,1]^p$, obtained by splitting $K-1$ times at observed values in $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ along each coordinate direction inside $\mathcal{S} \subseteq \{1, \ldots, p\}$ at least once. In particular, each tree $\mathcal{T} \in \mathcal{V}_{\mathcal{S}}^K$ is valid according to Definition 3.1 and uses up all covariates in $\mathcal{S}$ for splits, where $\mathcal{S}$ can be regarded as an index set of active predictors. We will refer to the partitioning number $\Delta(\mathcal{V}_{\mathcal{S}}^K)$ (in a similar vein as in [30]) as the overall number of distinct partitions of $\mathcal{X}$ that can be induced by members of the partitioning family $\mathcal{V}_{\mathcal{S}}^K$.

LEMMA 3.1. *Denote by $\mathcal{S} \subseteq \{1, \ldots, p\}$ an index set of active covariates. Let $\mathcal{V}_{\mathcal{S}}^K$ denote the set of valid tree partitions. Then*

$$\Delta(\mathcal{V}_{\mathcal{S}}^K) < (K-1)!|\mathcal{S}|^K n^K.$$

PROOF. Let $\mathcal{N}_K$ denote the number of trees with K leaves. We have the following recursive formula: $\mathcal{N}_K < \mathcal{N}_{K-1}(K-1)|\mathcal{S}|\, n$, where at the $(K-1)^{th}$ split, there are $K-1$ possible cells to choose from and up to n potential splits along one of the $|\mathcal{S}|$ coordinates. □

Lemma 3.1 will be fundamental for understanding the combinatorial richness of trees and for obtaining bounds on the covering numbers towards establishing Condition (2.1).

REMARK 3.1. *(The k-d tree) We now pause to revisit one of the most popular space partitioning structures, the k-d tree [2]. Such a tree $\widehat{\mathcal{T}}$ is constructed by cycling over coordinate directions in $\mathcal{S}$, where all nodes at the*

same level are split along the same axis. For a given direction $j \in \mathcal{S}$, each internal node, say Ω_k, will be split at a median of the point set (along the j^{th} axis). This split will pass $\lfloor \mu(\Omega_k)n/2 \rfloor$ and $\lceil \mu(\Omega_k)n/2 \rceil$ observations onto its two children, thereby roughly halving the number of points. After s rounds of splits on each variable, all K terminal nodes have at least $\lfloor n/K \rfloor$ observations, where $K = 2^{s\,|\mathcal{S}|}$. The k-d trees are thus balanced in light of Definition 2.4 of van der Pas and Ročková [37]. In addition, the diameters of k-d trees can shrink fast, as long as the number of directions is not too large. In particular, $\operatorname{diam}(\widehat{\mathcal{T}}; \mathcal{S}) \leq |\mathcal{S}|/K^{1/(2\,|\mathcal{S}|)}$ *for* $K = 2^{s\,|\mathcal{S}|}$ *for some* $s \geq 1$ *according to Proposition 6 of [39]. The k-d tree construction will be instrumental in establishing Condition 2.2.*

3.3. *On Tree-structured Step Functions.* The second critical ingredient in building a tree regressor is learning the piecewise function on a given partition. In this section, we describe some facts about the approximating properties of tree-structured step functions. The understanding of how well we can approximate a smooth regression surface will be vital for establishing Condition (2.2).

For a family of *valid* tree partitions $\mathcal{V}_{\mathcal{S}}^K$, we denote by

(3.3)
$$\mathcal{F}(\mathcal{V}_{\mathcal{S}}^K) = \left\{ f_{\mathcal{T},\boldsymbol{\beta}} : [0,1]^p \to \mathbb{R}; f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \beta_k \, \mathbb{I}_{\Omega_k}(\boldsymbol{x}); \mathcal{T} \in \mathcal{V}_{\mathcal{S}}^K, \boldsymbol{\beta} \in \mathbb{R}^K \right\}$$

the set of all step functions supported by members of the tree partitioning family $\mathcal{V}_{\mathcal{S}}^K$. Each function is underpinned by a valid tree partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}_{\mathcal{S}}^K$ and a vector of step heights $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)' \in \mathbb{R}^K$.

The partition diameter $\operatorname{diam}(\mathcal{T}; \mathcal{S})$ oversees how closely we can approximate $f \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S})$ with $f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$. Ideally, the diameter decay should be no slower than $q^\gamma/K^{1/q}$ for some $\gamma > 0$, where $q = |\mathcal{S}|$. To get an intuitive insight into this condition, consider the following perfectly regular partition: a q-dimensional "chess-board" that splits $[0,1]^q$ into $K = s^q$ cubes of length $1/s = 1/K^{1/q}$. The maximal interpoint distance inside each cube will be at most $\sqrt{q}/K^{1/q}$. This partition is, however, not adaptive and thereby less practical. It turns out that, under a mild requirement on the spread of the data points $\mathcal{X}$, the fast diameter decay is *also* guaranteed by the data-adaptive k-d trees mentioned in Remark 3.1. The "mild requirement" is formalized in our definition below.

DEFINITION 3.3. *Denote by* $\widehat{\mathcal{T}} = \{\widehat{\Omega}_k\}_{k=1}^K \in \mathcal{V}_{\mathcal{S}}^K$ *the* k*-*d *tree where* $\mathcal{S} \subseteq \{1, \dots, p\}$ *and* $K = 2^{s\,|\mathcal{S}|}$. *We say that a dataset* $\mathcal{X}$ *is* $(M, \mathcal{S})$*-regular if*

$$\max_{1 \le k \le K} \operatorname{diam}(\widehat{\Omega}_k; \mathcal{S}) < M \sum_{k=1}^{K} \mu(\widehat{\Omega}_k) \operatorname{diam}(\widehat{\Omega}_k; \mathcal{S})$$

for some large enough constant $M > 0$ *and all* $s \in \mathbb{N} \backslash \{0\}$.

The definition states that in a regular dataset, the maximal diameter in the k-d tree partition should not be much larger than a "typical" diameter. This condition assures a local control over individual cell diameters. The requirement can be roughly translated as follows. As more and more data points are collected, the data conform to a structure that does not permit outliers in active directions $\mathcal{S}$. Moreover, to render learning about the active set $\mathcal{S}_0$ in high dimensions meaningful, the observations $\mathcal{X}$ must not have any lower dimensional structure (such as an alignment at coordinate axes). As will be seen in the following lemma, for regular datasets, k-d trees have a faster diameter decay and thereby attain better approximation error. The following lemma is an important element of our proof skeleton, showing that there *exists* a tree (a k-d tree) that approximates well.

LEMMA 3.2. *Assume* $f \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S})$, *where* $|\mathcal{S}| = q$. *Then there exists a tree-structured step function* $f_{\widehat{\mathcal{T}}, \widehat{\boldsymbol{\beta}}} \in \mathcal{V}_{\mathcal{S}}^K$ *for some* $K = 2^{s\,q}$ *with* $s \in \mathbb{N} \backslash \{0\}$ *such that*

$$\|f - f_{\widehat{\mathcal{T}}, \widehat{\boldsymbol{\beta}}}\|_n \le ||f||_{\mathcal{H}^\alpha} q / K^{1/(2q)}. \tag{3.4}$$

Moreover, if $\mathcal{X}$ *is* $(M, \mathcal{S})$*-regular, then*

$$\|f - f_{\widehat{\mathcal{T}}, \widehat{\boldsymbol{\beta}}}\|_n \le ||f||_{\mathcal{H}^\alpha} C\, M^\alpha\, q / K^{\alpha/q} \tag{3.5}$$

for some $C > 0$.

PROOF. Section 8.3. □

Controlling the approximation error is only one of the critical aspects in our theoretical study. The second will be monitoring the complexity of our approximating function classes. Finding the right balance between the two will be instrumental for obtaining optimal performance.

Now that we have developed the necessary tools, we can dive into the main results of the paper.

4. Adaptive Dimension Reduction with Trees. In Regime 1, we assume that the target regression surface $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$, although initially conceived as a function of $\boldsymbol{x} = (x_1, \ldots, x_p)'$, in fact depends only on a small fraction of q_0 features $\boldsymbol{x}_{\mathcal{S}_0}$, where $\mathcal{S}_0 \subset \{1, \ldots, p\}$. If an oracle were able to isolate $\mathcal{S}_0$, the minimax rate would improve from $n^{-\alpha/(2\alpha+p)}$ to $n^{-\alpha/(2\alpha+q_0)}$ and it would be the fastest rate possible. When no such oracle information is available, [41] characterized the minimax risk as follows: $\varepsilon_n^2 \sim \lambda^2 \left(\sqrt{n}\lambda\right)^{-4\alpha/(2\alpha+q_0)} + \frac{q_0}{n} \log\left(\frac{p}{q_0}\right)$, where λ is a bound parameter on the isotropic Hölder norm. The first term is the classical minimax risk for estimating a q_0-dimensional smooth function. The second term is the penalty incurred by variable selection uncertainty. While the number of features p is not forbidden from growing to infinity much faster than n, we keep in mind that consistent estimation will only be possible when q_0 is quite small relative to p and n (in which case the complexity penalty will be dominated by the first term in the minimax rate).

4.1. *Spike-and-Tree Priors.* Bayesian regression tree implementations that do not embrace sparsity (when in fact present) are bound to be less appropriate for the high-dimensional setup. In particular, the traditional Bayesian CART prior [14] grows trees by splitting each node, indexed by the depth level d, with a probability $g(d) = \gamma/(1+d)^\alpha$, where $\alpha > 1, \gamma \in (0, 1)$ are tuning parameters. The splitting variable is picked from $\{x_1, \ldots, x_p\}$ *uniformly* (i.e. with a fixed probability $\theta_i = 1/p$). Recently, [27] proposed an adaptive variant of this prior by placing a sparsity-inducing Dirichlet prior on the splitting proportions θ_i. This prior uses fewer variables in the tree construction and thereby is more reluctant to overfit. In another popular Bayesian CART implementation, [16] suggest directly placing a prior on K and a conditionally uniform prior on tree topologies with K bottom leaves. Again, in its original form, this prior will likely suffer from the curse of dimensionality, failing to harvest the intrinsic lower-dimensional structure. Here, we propose a fix to this problem. To make the Bayesian CART prior of [16] appropriate for high-dimensional setups, we propose a *spike-and-tree* variant by injecting one more layer: a complexity prior over the active set of predictors.

Bayesian models for feature selection have traditionally involved a hierarchy of priors over subset sizes $q = |\mathcal{S}|$ and subsets $\mathcal{S} \subseteq \{1, \ldots, p\}$ [12, 11]. Rather than modeling the mean outcome as a linear functional of active predictors $\boldsymbol{x}_{\mathcal{S}}$, here we grow a tree from $\boldsymbol{x}_{\mathcal{S}}$. We begin by treating q_0 as unknown with an exponentially decaying prior [12]

$$\text{(T1)} \qquad \pi(q) \propto c^{-q}\, p^{-a\,q},\ q = 0, 1, \ldots, p, \qquad \text{for some} \quad a, c > 0.$$

Next, given the dimensionality q, we assume that all $\binom{p}{q}$ subsets $\mathcal{S}$ of $q = |\mathcal{S}|$ covariates are a-priori equally likely, i.e.

$$\pi(\mathcal{S} \mid q) = 1/\binom{p}{q}. \tag{T2}$$

Given q and the feature set $\mathcal{S}$, a tree is grown by splitting at least once on every coordinate inside $\mathcal{S}$, all the way down to K terminal nodes. If we knew α and q_0, the optimal choice of K would be $K \sim n^{q_0/(2\alpha+q_0)}$, for which the actual minimax rate could be achieved. For the more practical case when α and q_0 are both unknown, we shall assume that K arrived from a prior $\pi(K)$. As noted by [15], the tail behavior of $\pi(K)$ is critical for controlling the vulnerability/resilience to overfitting. The anticipation is that with less smooth f_0, more rapid posterior concentration takes place when $\pi(K)$ has a heavier tail. However, too heavy tails make it easier to overfit when the true function is less smooth. To achieve an equilibrium, we consider the Poisson distribution (constrained to $\mathbb{N}\backslash\{0\}$), as suggested by [16] in their Bayesian CART implementation. Namely, we have

$$\pi(K) = \frac{\lambda^K}{(\mathrm{e}^\lambda - 1)K!},\ K = 1, 2, \ldots, \quad \text{for some} \quad \lambda \in \mathbb{R}. \tag{T3}$$

For its practical implementation, one would truncate its support to the maximum number of splits that can be made with n observations. When λ is small, (T3) is concentrated on models with smaller complexity where overfitting does not occur. Increasing λ leads to smearing the prior mass over partitions with more jumps. Similar complexity priors with an exponential decay $\mathrm{e}^{-C\,k\log k}$ have been deployed previously in nonparametric problems [26, 15, 28, 16, 37].

Given $q, \mathcal{S}$ and K, we assign a uniform prior over *valid* tree topologies $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}_{\mathcal{S}}^K$, i.e.

$$\pi(\mathcal{T} \mid \mathcal{S}, K) \propto \frac{1}{\Delta(\mathcal{V}_{\mathcal{S}}^K)} \mathbb{I}\left(\mathcal{T} \in \mathcal{V}_{\mathcal{S}}^K\right). \tag{T4}$$

Similar constraints on trees where each terminal node is assigned a minimal number of data points have been implemented in stochastic search algorithms [16]. At the very minimum, we can choose $\bar{C} = 1$ in (3.1), merely requiring that the cells be non-empty. Finally, given the partition of size K, we assign an iid Gaussian prior on the step heights (similarly as in [14])

$$\pi(\boldsymbol{\beta} \mid K) = \prod_{k=1}^{K} \phi(\beta_k; 1). \tag{T5}$$

REMARK 4.1. *The name spike-and-tree prior deserves a bit of explanation. It follows from the fact that (T1) and (T2) will be satisfied if each covariate has a prior probability* $\theta \sim \mathcal{B}(1, p^u)$ *of contributing to the mean regression surface for some* $u > 1$ *[12]. Endowing each covariate* x_j *with a Bernoulli indicator* γ_j*, where* $\Pi(\gamma_j = 1 \,|\, \theta) = \theta$*, and building a tree on* $\mathcal{S} = \{j : \gamma_j = 1\}$*, one obtains a mixture prior on* $f(\boldsymbol{x})$ *that pertains to spike-and-slab variable selection. Here, the slab is a tree prior built on active covariates rather than an independent product prior on active regression coefficients. This hierarchical construction has distinct advantages for variable selection. In linear regression, it is customary to select variables by thresholding marginal posterior inclusion probabilities* $\Pi(\gamma_j = 1 \,|\, \boldsymbol{Y}^{(n)})$*. These will be available also under our spike-and-tree construction. Inspecting the posterior probabilities* $\Pi(\gamma_j = 1 \,|\, \boldsymbol{Y}^{(n)})$ *obtained with our hierarchical tree prior will be a new avenue for conducting variable selection in Bayesian CART and BART, an alternative to [7]. Thus, our prior has important practical implications for performing principled model-free variable selection.*

4.2. *Posterior Concentration for Bayesian CART.* The difficulty in properly analyzing Bayesian CART stems from the nonparametric combinatorial nature of the prior that makes it less tractable analytically. By building on our developments from previous sections, we are now fully equipped to present the first theoretical result concerning this method.

The following theorem solidifies the optimality properties of Bayesian CART by showing that under the hierarchical prior (T1)-(T5), the posterior adapts to both smoothness and sparsity, concentrating at the (near) minimax rate that depends only on the number of strong covariates regardless of how many noise variables are present. The near-minimaxity refers to an additional log factor, an inevitable penalty for not knowing q_0 and α. The result holds under the high-dimensional regime, where $p > n$ and p can grow much faster than n. We will denote by

$$\mathcal{F}_{\mathcal{T}} = \bigcup_{q=0}^{\infty} \bigcup_{K=1}^{\infty} \bigcup_{\mathcal{S}:|\mathcal{S}|=q} \mathcal{F}(\mathcal{V}_{\mathcal{S}}^{K})$$

the collection of all tree-structured step functions (with various tree sizes and split subsets) that can be obtained by partitioning $\mathcal{X}$.

THEOREM 4.1. *Assume* $f_0 \in \mathcal{H}_p^{\alpha} \cap \mathcal{C}(\mathcal{S}_0)$ *with* $0 < q_0 = |\mathcal{S}_0|$ *such that* $q_0 \lesssim \log^{\beta} n$ *and* $\|f_0\|_{\infty} \lesssim \log^{\beta} n$ *for some* $\beta \geq 1/2$*. Moreover, we assume that* $p > n$ *and* $\log p \lesssim n^{q_0/(2\alpha+q_0)}$ *and that* $\mathcal{X}$ *is* $(M, \mathcal{S}_0)$*-regular. We endow*

$\mathcal{F}_{\mathcal{T}}$ with priors (T1)-(T5). Then with $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)} \log^{\beta} n$ we have

$$\Pi\left(f \in \mathcal{F}_{\mathcal{T}} : \|f_0 - f\|_n > M_n\, \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

for any $M_n \to \infty$ in P_0^n-probability, as $n, p \to \infty$.

PROOF. Section 8. □

The assumption of a regular design is an inevitable consequence of treating $\boldsymbol{x}$'s as fixed. As noted by [6] in their study of random forests, point-wise consistency results have been complicated by the difficulty in controlling local (cell) diameters. The regularity assumption guarantees this control and is apt to be satisfied for most realizations of $\boldsymbol{x}$ from reasonable distributions on $[0,1]^p$. However, we should also point out that this assumption can be completely avoided if the function f_0 is α-smooth with $\alpha \in (0, 1/2)$.

COROLLARY 4.1. *Under the assumptions of Theorem 4.1 with $\beta = 1/2$ we have*

$$\Pi\left(K > C_k n^{q_0/(2\alpha+q_0)} \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

for a suitable constant $C_k > 0$.

PROOF. This statement follows from Lemma 1 of [21] and holds upon the satisfaction of the condition $\Pi(K > C_k n^{q_0/(2\alpha+q_0)}) = o(\mathrm{e}^{-(d+2)\, n\varepsilon_n^2})$ for some suitably large $d > 2$. This is shown in Section 8.4. □

REMARK 4.2. *(Bayesian regression trees do not overfit!) Corollary 4.1 highlights the remarkable feature of Bayesian CART that, under a suitable complexity prior on the number of terminal nodes, the trees are reluctant to overfit. This seen from the behavior of the posterior, which concentrates on values K that are only a constant multiple larger than the optimal oracle value $n^{q_0/(2\alpha+q_0)}$.*

REMARK 4.3. *(Bayesian CART à la Chipman, George and McCulloch [14]) A closer look at the proof reveals that Theorem 4.1 also holds for priors $\pi(k \mid q)$ such that $\mathrm{e}^{-c\, k \log k} \lesssim \pi(k \mid q) \lesssim e^{-c\, k \log(n)^{\eta}}$ for some $0 < \eta < 2\beta$. At the lower end are the complexity priors deployed in similar contexts [32, 37]. The Bayesian CART prior of [14] (discussed in Section 4.1) with $\alpha = 0$ is at the upper end. To see this, note that their tree prior then corresponds to a homogeneous Galton-Watson process, where the number of terminal nodes K satisfies $\Pi(K > k) \leq \mathrm{e}^{-(k-0.5)\log[(1-\gamma)/\gamma]}$. With $\gamma = c/n$ for some $c < n$ we have $\Pi(K = k) \propto \mathrm{e}^{-c\, k \log n}$. While our result concerns the Bayesian CART of [16], this connection shows that the prior of [14] would be amenable to a similar analysis.*

5. Tree Ensembles. Combining multiple trees through additive aggregation has proved to be remarkably effective for enhancing prediction [8, 13]. This section offers new theoretical insights into the mechanics behind the Bayesian variants of such tree ensemble methods. Our approach rests upon a detailed analysis of the *collective behavior* of partitioning cells generated by individual trees. We will see that the overall performance is affected not only by the quality of single trees but also how well they can collaborate [9].

5.1. *Bayesian Additive Regression Trees.* Additive regression trees grow an ensemble predictor by binding together T tree-shaped regressors. For subsets $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^1, \ldots, \mathcal{S}^T\}$ and tree sizes $\boldsymbol{K} = (K^1, \ldots, K^T)'$, we define a sum-of-trees model (forest) as

$$f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) = \sum_{t=1}^{T} f_{\mathcal{T}^t,\boldsymbol{\beta}^t}(\boldsymbol{x}) = \sum_{t=1}^{T}\sum_{k=1}^{K^t} \beta_k^t \, \mathbb{I}(\boldsymbol{x} \in \Omega_k^t), \tag{5.1}$$

where $f_{\mathcal{T}^t,\boldsymbol{\beta}^t} \in \mathcal{F}(\mathcal{V}_{\mathcal{S}^t}^{K^t})$ is a tree learner associated with step sizes $\boldsymbol{\beta}^t \in \mathbb{R}^{K^t}$. Each learner is allowed to use different splitting variables $\mathcal{S}^t$ and different number of bottom nodes K^t. With $\mathcal{E} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ we denote an ensemble of tree partitions and with $\mathcal{B} = (\boldsymbol{\beta}^{1\prime}, \ldots, \boldsymbol{\beta}^{T\prime})' \in \mathbb{R}^{T\bar{K}}$ the terminal node parameters, where $\bar{K} = \frac{1}{T}\sum_{t=1}^{T} K^t$.

Sum-of-trees models offer an improved representation flexibility by chopping up the predictor space into more refined segmentations. These segmentations are obtained by superimposing multiple tree partitions, yielding what we define below as a *global* partition.

DEFINITION 5.1. *For a partition ensemble $\mathcal{E}$, we define a global partition*

$$\widetilde{\mathcal{T}}(\mathcal{E}) = \{\widetilde{\Omega}_k\}_{k=1}^{K(\mathcal{E})}$$

as the partition obtained by merging all cuts in $\mathcal{T}^1, \ldots, \mathcal{T}^T$. We refer to $\widetilde{\Omega}_k$'s as global cells in the ensemble.

The concept of the global partition can be better understood from Figure 1, where splits from $T = 2$ trees (each having 3 leaves) are merged to obtain a global partition with $K(\mathcal{E}) = 7$ global cells. Generally, the global partition $\mathcal{E}$ itself is *not* necessarily a tree and can have as many as $K(\mathcal{E}) \leq \prod_{t=1}^{T} K^t$ cells. This upper bound can be attained if each trees splits $K^t - 1$ times on a single variable, where each tree uses a different one. Without loss of generality, the global partition will be assumed to have non-empty cells.

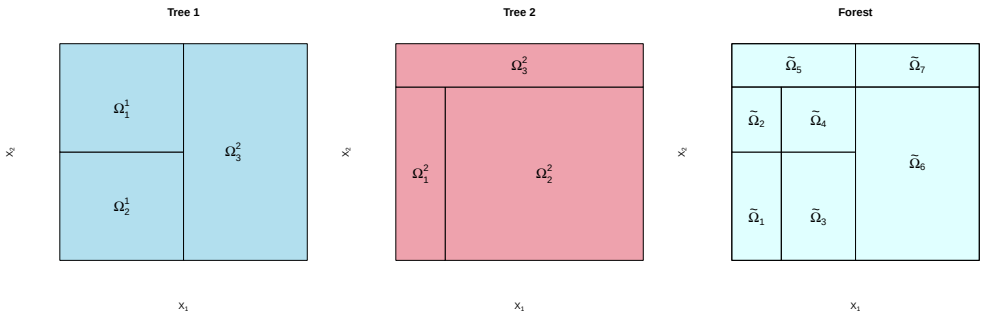

FIG 1. *Ensemble of tree partitions and their global partition.*

This requirement can be met by merging vacuous cells with their nonempty neighbors.

Bayesian additive regression trees were conceived as a collection of weak learners that capture *different* aspects of the predictor space [13]. To characterize the amount of diversity/correlation between trees in the ensemble, we introduce the so-called stretching matrix.

DEFINITION 5.2. *For a partition ensemble* $\mathcal{E} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$*, we define the stretching matrix* $\boldsymbol{A}(\mathcal{E}) = (a_{ij})_{i,j}$ *as follows: for each* $1 \leq i \leq K(\mathcal{E})$ *and* $1 \leq j \leq T\bar{K}$ *we have*

$$a_{ij} = \mathbb{I}\{\widetilde{\Omega}_i \cap \Omega_m^t \neq \emptyset\}, \tag{5.2}$$

where $1 \leq t \leq T$ *and* $1 \leq m \leq K^t$ *are such that* $j = \sum_{l=1}^{t-1} K^l + m$ *and where* Ω_m^t *is the* m^{th} *(local) cell in the* t^{th} *tree* $\mathcal{T}^t$*.*

Each row of the stretching matrix $\boldsymbol{A}(\mathcal{E})$ corresponds to one global cell and each column to one local cell. The row entries sum to T, indicating which local cells overlap with that global cell (as shown in Figure 2 for partitions from Figure 1). To further characterize the pattern of overlap between trees, we introduce the $(T\bar{K}) \times (T\bar{K})$ Gram matrix

$$\widetilde{\boldsymbol{A}}(\mathcal{E}) \equiv \boldsymbol{A}(\mathcal{E})'\boldsymbol{A}(\mathcal{E}) = (\widetilde{a}_{ij})_{i,j}\,. \tag{5.3}$$

The off-diagonal elements measure the "similarity" between local cells, say Ω_j^t and Ω_k^u, in terms of the number of global cells that they share. More formally, let $r[\widetilde{\mathcal{T}}(\mathcal{E}), V] = |\{\widetilde{\Omega}_k : \widetilde{\Omega}_k \cap V \neq \emptyset\}|$ be the restricted cell count [30], measuring the number of global cells that intersect with a compact set V. For $i = \sum_{s=1}^{t-1} K^s + j$ and $l = \sum_{s=1}^{u-1} K^s + k$ we can write $\widetilde{a}_{il} = r[\widetilde{\mathcal{T}}(\mathcal{E}), \Omega_j^t \cap \Omega_k^u]$.

$$\boldsymbol{A}(\mathcal{E}) = \begin{array}{c} \\ \widetilde{\Omega}_1 2 \\ \widetilde{\Omega}_2 \\ \widetilde{\Omega}_3 \\ \widetilde{\Omega}_4 \\ \widetilde{\Omega}_5 \\ \widetilde{\Omega}_6 \\ \widetilde{\Omega}_7 \end{array} \begin{array}{c} \begin{array}{cccccc} \Omega_1^1 & \Omega_2^1 & \Omega_3^1 & \Omega_1^2 & \Omega_2^2 & \Omega_3^2 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}, \quad \boldsymbol{A}(\mathcal{E})'\boldsymbol{A}(\mathcal{E}) = \left(\begin{array}{ccc|ccc} 3 & 0 & 0 & 1 & 1 & 1 \\ 0 & 2 & 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 \\ \hline 1 & 1 & 0 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 3 & 0 \\ 1 & 0 & 1 & 0 & 0 & 2 \end{array}\right)$$

FIG 2. *Stretching matrix* $\boldsymbol{A}(\mathcal{E})$ *and the "Gram matrix"* $\boldsymbol{A}(\mathcal{E})'\boldsymbol{A}(\mathcal{E})$ *from Figure 1.*

Small off-diagonal entries indicate less overlap, where the individual trees capture more diverse aspects of the predictor space. The diagonal elements, on the other hand, quantify the "persistence" of each local cell, say Ω_k^u, counting the number of global cells it stretches over. More formally, for $i = \sum_{s=1}^{u-1} K^s + k$ we have $\widetilde{a}_{ii} = r[\widetilde{\mathcal{T}}(\mathcal{E}), \Omega_k^u]$. The amount of diversity (or tree dis-similarity) in the ensemble can be quantified with eigenvalues of $\widetilde{\boldsymbol{A}}(\mathcal{E})$. We denote by $\lambda_{min}(\mathcal{E})$ (resp. $\lambda_{max}(\mathcal{E})$) the minimal (resp. maximal) singular values of $\boldsymbol{A}(\mathcal{E})$ (i.e. square roots of extremal nonzero eigenvalues of $\widetilde{\boldsymbol{A}}(\mathcal{E})$). If some trees in the ensemble are redundant, the conditioning number $\kappa(\mathcal{E}) \equiv \lambda_{max}(\mathcal{E})/\lambda_{min}(\mathcal{E})$ will be large. The idea of diversifying trees was originally introduced by Breiman [8] via subsampling. To encourage the trees to collaborate and diversify, we can impose a restriction on $\lambda_{min}(\mathcal{E})$ in our prior. For our theoretical study, we focus on the so called valid ensembles that have two properties.

DEFINITION 5.3. *An ensemble* $\mathcal{E} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ *is* δ*-valid if*

1. *Each* $\mathcal{T}^t$ *is valid according to Definition 3.1 and*
2. $\lambda_{min}^2(\mathcal{E}) \gtrsim n^{-\delta}$ *for some* $\delta \geq 0$ *or* $K(\mathcal{E}) < D \sum_{t=1}^T K^t$ *for some* $D > 0$.

For tree sizes $\boldsymbol{K} = (K^1, \ldots, K^T)'$ *and subsets* $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^1, \ldots, \mathcal{S}^T\}$*, we denote the set of all valid ensembles by* $\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$.

The second condition deserves a bit of attention. This technical condition is required for obtaining a bound on the entropy and we expect that it will be inconsequential for the practical deployment of the prior. The minimal singular value will be large for most tall rectangular stretching matrices. For wide stretching matrices with $K(\mathcal{E}) < D \sum_{t=1}^T K^t$, we can bound the entropy without any condition on the singular value.

The representation flexibility of additive trees also pertains to jump sizes. The global step size coefficients under additive trees are intertwined due

to the tree overlap. This can be seen from the following, more compact, representation of (5.1):

$$f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) = \sum_{k=1}^{K(\mathcal{E})} \bar{\beta}_k \, \mathbb{I}(\boldsymbol{x} \in \widetilde{\Omega}_k) \quad \text{with} \quad \bar{\beta}_k \equiv \sum_{t=1}^{T} \sum_{l=1}^{K^t} \beta_l^t \mathbb{I}(\Omega_l^t \cap \widetilde{\Omega}_k \neq \emptyset),$$

where $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_{K(\mathcal{E})})' \in \mathbb{R}^{K(\mathcal{E})}$ are aggregated step sizes. A closer look reveals the following key connection between $\bar{\boldsymbol{\beta}}$ and $\mathcal{B}$

$$\bar{\boldsymbol{\beta}} = \boldsymbol{A}(\mathcal{E})\mathcal{B}, \tag{5.4}$$

where $\boldsymbol{A}(\mathcal{E})$ is the stretching matrix defined in (5.2). This link unfolds the theoretical analysis of tree ensembles using tools that we have already developed for single trees. Note that the condition number $\kappa(\mathcal{E})$ determines how much the relative change in $\mathcal{B}$ influences the relative change in $\bar{\boldsymbol{\beta}}$.

The mapping (5.4) can be in principle many-to-one in the sense that many tree-structured step functions can sum towards the same target (5.1). Such over-parametrization occurs, for instance, when $K(\mathcal{E}) < T\bar{K}$ or, more generally, when $\widetilde{\boldsymbol{A}}(\mathcal{E})$ has zero eigenvalues. This redundancy is not entirely unwanted and, in fact, it endows sum-of-trees models with a lot of modeling freedom. Excessive over-parametrization will be controlled with our diversification prior which limits consideration to δ-valid ensembles.

We now formally define the space of approximating additive trees. For variable sets $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^1, \ldots, \mathcal{S}^T\}$ and a vector of tree sizes $\boldsymbol{K} = (K^1, \ldots, K^T)'$, we denote by

$$\mathcal{F}(\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}) = \left\{ f_{\mathcal{E},\mathcal{B}} : [0,1]^p \to \mathbb{R} : f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) = \sum_{t=1}^{T} f_{\mathcal{T}^t,\boldsymbol{\beta}^t}(\boldsymbol{x}); \mathcal{E} \in \mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}, \boldsymbol{\beta}^t \in \mathbb{R}^{K^t} \right\} \tag{5.5}$$

the set of all additive tree step functions supported on *valid* ensembles $\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$. The union of these over the number of trees T, all possible sets $\boldsymbol{\mathcal{S}}$ of sizes $\boldsymbol{q} = (q^1, \ldots, q^T)'$ and tree sizes $\boldsymbol{K}$ gives rise to

$$\mathcal{F}_{\mathcal{E}} = \bigcup_{T=1}^{\infty} \bigcup_{\boldsymbol{q}} \bigcup_{\boldsymbol{\mathcal{S}}:|\mathcal{S}^t|=q^t} \bigcup_{\boldsymbol{K}} \mathcal{F}(\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}), \tag{5.6}$$

our approximating space of additive regression tree functions.

5.2. *Additive Regression Trees are Adaptive.* This section provides an interesting initial perspective on the behavior of Bayesian additive regression

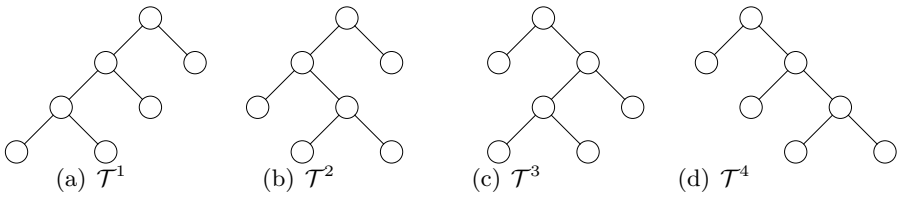

FIG 3. *Trees in the approximating ensemble* $\widehat{\mathcal{E}}$.

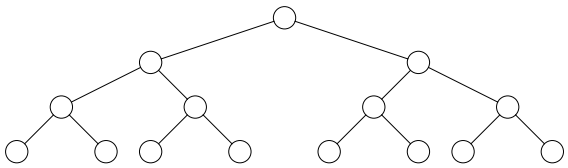

FIG 4. *Approximating k-d tree partition* $\widehat{\mathcal{T}} = \widetilde{\mathcal{T}}(\widehat{\mathcal{E}})$

trees in Regime 1. We will continue with the more general Regime 2 in the next section. We will focus on a variant of the popular Bayesian Additive Regression Trees (BART) model [13], modified in three ways. First, the tree prior will be according to [16] rather than [14]. The second modification is that the trees are built on the *same set of variables*, endowed with a subset selection prior construction. In the next section, we allow for the fully general case where each tree builds on a potentially different set of variables. Third, rather than fixing the number of trees, we endow T with a prior distribution.

We will see that having a good control of the regression function variation inside each *global* cell together with a good choice of the prior on the total number of leaves $\sum K^t$ will be sufficient to ensure optimal behavior. The approximation ability of tree ensembles hinges on the diameter of the *global* partition. Each tree partition does not need to have a small diameter (i.e. can be a weak learner), as long as the global one does. An important building block in our proof will be the construction of a single tree ensemble that can approximate well. As will be shown in Lemma 10.1, we can construct such ensemble by first finding a single *k-d* tree from Lemma 3.2 (a strong learner) and then redistributing the cuts among small trees (weak learners) in a way that the global partition is exactly equal to the *k-d* tree. An example of this deconstruction is depicted in Figure 3, where a full symmetric tree from Figure 4, say $\widehat{\mathcal{T}}$, has been trimmed into many smaller imbalanced trees which add up towards $\widehat{\mathcal{T}}$. More details on this decomposition are in the proof of Lemma 10.1.

The following theorem is an ensemble variant of Theorem 4.1 which will

serve as a useful stepping stone towards the full-fledged result presented in the next section. Instead of approximating $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$ with one large tree in Regime 1, we build a forest made up of many smaller trees (weak learners). The first two layers of the prior (T1) and (T2) are the same. Next, we assign a prior on the number of trees

$$\text{(T3*)} \qquad \pi(T) \propto \mathrm{e}^{-C_T\, T} \quad \text{for} \quad T \in \mathbb{N}\backslash\{0\} \quad \text{with} \quad C_T > \log 2,$$

which is sufficiently diffuse so as to promote ensembles with many trees. Next, conditionally on T, we assign a prior on $\boldsymbol{K} = (K^1, \ldots, K^T)'$ as an independent product of Poisson distributions (T3). An important distinction between the Poisson prior deployed for single trees in Section 4 and the one deployed here for ensembles is that the hyper-parameter depends on T. Namely, our prior on leaves satisfies

$$\text{(T4*)} \qquad \pi(\boldsymbol{K} \,|\, T) \;=\; \prod_{t=1}^{T} \frac{(\lambda/T)^{K^t}}{(\mathrm{e}^{\lambda/T} - 1) K^t!} \quad \text{for} \quad K^t \in \mathbb{N}\backslash\{0\}.$$

As a prior on $\mathcal{E}$, given $(\mathcal{S}, \boldsymbol{K})$, we use the uniform prior over δ-valid ensembles

$$\text{(T5*)} \qquad \pi(\mathcal{E} \,|\, \mathcal{S}, \boldsymbol{K}) \propto \frac{1}{\Delta(\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}})} \mathbb{I}\left(\mathcal{E} \in \mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}}\right),$$

where $\Delta(\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}})$ is the overall number of ensembles consisting of T valid trees that can be obtained by splitting on data points $\mathcal{X}$. Recall that throughout this section, all trees are constrained to use split variables in the same set $\mathcal{S}$. To mark this difference, we have denoted the partition ensembles with $\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}}$ instead of $\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$.

Finally, given $(T, \boldsymbol{K})$, we assign an iid Gaussian prior on the step heights $\mathcal{B} \in \mathbb{R}^{\sum_{t=1}^{T} K^t}$ with a variance $1/T$ (as suggested by [13])

$$\text{(T6*)} \qquad \pi(\mathcal{B} \,|\, T, \boldsymbol{K}) = \prod_{t=1}^{T} \prod_{k=1}^{K^t} \phi(\beta_k^t; 1/T).$$

The following theorem shows that additive regression trees, in combination with a subset selection prior, can nicely adapt to the ambient dimensionality and smoothness, *also* achieving the optimal concentration rate in Regime 1.

THEOREM 5.1. *Assume $f_0 \in \mathcal{H}_p^\alpha \cap \mathcal{C}(\mathcal{S}_0)$ with $0 < q_0 = |\mathcal{S}_0|$ such that $q_0 \lesssim \log^\beta n$ and $\|f_0\|_\infty \lesssim \log^\beta n$ for some $\beta \geq 1$. Moreover, we assume that*

$p > n$ and $\log p \lesssim n^{q_0/(2\alpha+q_0)}$ and that $\mathcal{X}$ is $(M, \mathcal{S}_0)$-regular. We endow $\mathcal{F}_\mathcal{E}$ with priors (T1),(T2),(T3)-(T6*). With $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)} \log^\beta n$ we have*

$$\Pi\left(f \in \mathcal{F}_\mathcal{E} : \|f_0 - f\|_n > M_n\, \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

for any $M_n \to \infty$ in P_0^n-probability, as $n, p \to \infty$.

PROOF. Section 10. □

Similarly as for single trees, we obtain the following corollary which states that the posterior concentrates on ensembles whose *overall number of leaves* is not much larger than the optimal value $n^{q_0/(2\alpha+q_0)}$.

COROLLARY 5.1. *Under the assumptions of Theorem 5.1 with $\beta = 1$ we have*

$$\Pi\left((T, \boldsymbol{K}) : \sum_{t=1}^{T} K^t > C_k n^{q_0/(2\alpha+q_0)} \log n \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

for a suitable constant $C_k > 0$.

PROOF. This statement follows from Lemma 1 of [21] and holds upon the satisfaction of $\Pi\left((T, \boldsymbol{K}) : \sum_{t=1}^{T} K^t > C_k n^{q_0/(2\alpha+q_0)} \log n\right) = o(\mathrm{e}^{-(d+2)\, n\varepsilon_n^2})$ for some suitably large $d > 2$. This condition is shown in Section 10.3. □

Corollary 5.1 shows that the posterior distribution charges additive regression trees consisting of either many weak learners or a few strong ones. The compromise between the two is regulated by the prior (T3*), where stronger shrinkage (i.e. larger C_T) will result in fewer trees. This corollary provides an important theoretical justification for why Bayesian additive regression trees have been so resilient to overfitting in practice.

6. Tree Ensembles in Additive Regression. In Section 4.2 we have shown that the posterior distribution under the Bayesian CART prior has optimal properties. However, it is now well known that the practical deployments of Bayesian CART suffer from poor MCMC mixing. Additive aggregations of single small trees [13] have proven to have far more superior mixing properties. One may wonder whether the benefits of additive trees are purely computational or whether there are some aspects that make them more attractive also theoretically. We will address this fundamental question.

For estimating a single smooth function, we were not able to tell apart single trees from tree ensemble in terms of their convergence rate (besides perhaps a small difference in the log factor). They are both optimal in Regime 1. Tree ensembles are inherently additive and, as such, are well-equipped for approximating additive f_0 (Regime 2). Throughout this section, we assume

$$f_0(\boldsymbol{x}) = \sum_{t=1}^{T_0} f_0^t(\boldsymbol{x}), \tag{6.1}$$

where $f_0^t(\boldsymbol{x}) \in \mathcal{H}_p^{\alpha^t} \cap \mathcal{C}(\mathcal{S}_0^t)$. Note that each component f_0^t depends only on a potentially very small subset $\mathcal{S}_0^t$ of covariates, where $|\mathcal{S}_0^t| = q_0^t$. However, the additive structure allows f_0 to depend on a larger number of variables, say q_0, where $\max_{1\le t\le T_0} q_0^t \le q_0 \le \sum_{t=1}^{T_0} q_0^t$. The minimax rate for estimating f_0 in Regime 2 is known to be $\varepsilon_n^2 = \sum_{t=1}^{T_0} \varepsilon_n^{t2}$, where $\varepsilon_n^{t2} = \lambda^t 2\left(\sqrt{n}\lambda^t\right)^{-4\alpha^t/(2\alpha^t+q_0^t)} + \frac{q_0^t}{n}\log\left(\frac{p}{q_0^t}\right)$ [41]. The sparsity constraint in Regime 2 is less strict than in Regime 1, where q_0 can be potentially larger than $\log n$ while still allowing for consistent estimation. For the isotropic case ($\alpha^t = \alpha$ and $q_0^t = \widetilde{q}_0$), single trees can achieve the slower rate $n^{-2\alpha/(2\alpha+q_0)}$ where $\widetilde{q}_0 \le q_0 \le T_0\widetilde{q}_0$. As will be shown below, tree ensembles can achieve the actual minimax rate, up to a log factor.

We will approximate f_0 with tree ensembles $f_{\mathcal{E},\mathcal{B}} \in \mathcal{F}_{\mathcal{E}}$. The ensembles here differ from the ones considered in Section 5.2. The crucial difference is that now we allow each of the trees $\mathcal{T}^t$ to depend on a *different set of variables* $\mathcal{S}^t$.

Now we have a vector of subset sizes $\boldsymbol{q} = (q^1, \ldots, q^T)'$ and a set of subsets $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^1, \ldots, \mathcal{S}^T\}$, one for each tree. We consider the following independent product variant of the complexity prior (T1)

$$\pi(\boldsymbol{q} \,|\, T) \propto \prod_{t=1}^{T} c^{-q^t} p^{-aq^t},\ q^t = 0, 1, \ldots, p, \quad \text{for} \quad a > 2 \tag{T1*}$$

and a product prior variant of (T2), given $(T, \boldsymbol{q})$,

$$\pi(\boldsymbol{\mathcal{S}} \,|\, T, \boldsymbol{q}) \propto \prod_{t=1}^{T} 1/\binom{p}{q^t}. \tag{T2*}$$

The prior on the number of trees, the number of leaves, ensembles and step sizes is the same as in (T3*), (T4*), (T5*) (where the ensembles are δ_n-valid with some $\delta_n > 0$) and (T6*). The value δ_n is defined as the smallest constant $\widetilde{\delta}$ so that we can construct a $\widetilde{\delta}$-valid ensemble $\widehat{\mathcal{E}}$ consisting of k-d

trees that approximate well. The construction of $\widehat{\mathcal{E}}$ is detailed in Section 9.2. We are now ready to present our final result showing that the posterior concentration for Bayesian additive regression trees is near-minimax rate optimal when f_0 has an additive structure and when $\delta_n < \delta$ for some suitably large $\delta > 0$.

THEOREM 6.1. *Assume that f_0 is as in (6.1) where $f_0^t \in \mathcal{H}_p^{\alpha^t} \cap \mathcal{C}(\mathcal{S}_0^t)$ with $q_0^t = |\mathcal{S}_0^t|$ such that $0 < q_0^t \lesssim \log^{\beta^t} n$ and $\|f_0^t\|_\infty \lesssim \log^{\beta^t} n$. Moreover, assume $p > n$ and $\log p \lesssim \min_{1\le t\le T_0} n^{q_0^t/(2\alpha^t+q_0^t)}$ and that $\mathcal{X}$ is $(M, \mathcal{S}_0^t)$-regular for $1 \le t \le T_0$. We endow $\mathcal{F}_\mathcal{E}$ with priors (T1*)-(T6*), where the prior focuses on δ_n-valid ensembles. With $\varepsilon_n^2 = \sum_{t=1}^{T_0} n^{-2\alpha^t/(2\alpha^t+q_0^t)} \log^{2\beta^t} n$, we have for any $\beta^t \ge 1/2$*

$$\Pi\left(f \in \mathcal{F}_\mathcal{E} : \|f_0 - f\|_n > M_n\, \varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

for any $M_n \to \infty$ in P_0^n-probability, as $n, p \to \infty$, when $\delta_n < \delta$ for some $\delta > 0$.

Failing to recognize the additive structure in f_0, single regression trees achieve the slower rate $n^{-2\alpha/(2\alpha+q_0)} \log^\beta n$ according to Theorem 4.1. Theorem 6.1 thus provides an additional theoretical justification for Bayesian additive tree models suggesting their performance superiority over single trees when f_0 is additive.

7. Discussion. In this work, we laid down foundations for the theoretical study of Bayesian regression trees and their additive variants. We have shown an optimal behavior of Bayesian CART, the first theoretical result on this method. Our approach rests upon a careful analysis of recursive partitioning schemes and associated sieves of approximating step functions. We have developed several useful tools for analyzing additive regression trees (variants of the BART method), showing their optimal performance in both additive and non-additive regression. The smoothness order of studied functions is restricted to values not exceeding one, a main limitation of our approach due to the fact that our approximations are piece-wise constants. This limitation could be overcome by extending our approach to piece-wise polynomials, an elaboration that we leave for future investigation.

8. Proof of Theorem 4.1. Our approach consists of establishing conditions (2.1), (2.2) and (2.3) for $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)} \log^\beta n$. The log factor is an inevitable penalty for not knowing q_0 and α. The first step requires constructing the sieve $\mathcal{F}_\mathcal{T}^n \subset \mathcal{F}_\mathcal{T}$. For a given $n \in \mathbb{N}$ and suitably large integers

$q_n < k_n$ (chosen later), we define the sieve $\mathcal{F}_{\mathcal{T}}^n$ as consisting of step functions over small trees that split only on a few variables, i.e.

$$\mathcal{F}_{\mathcal{T}}^n = \bigcup_{q=0}^{q_n} \bigcup_{K=1}^{k_n} \bigcup_{\mathcal{S}:|\mathcal{S}|=q} \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K),$$

where $\mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$ was defined in (3.3). The optimal choice of k_n and q_n will follow from our considerations below.

8.1. *Condition* (2.1). We start with a useful lemma that characterizes a useful upper bound on the covering number of the smaller sets $\mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$.

LEMMA 8.1. *Let $\mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$ be the class of step functions (3.3). Then*

$$N\Big(\tfrac{\varepsilon}{36}, \Big\{f \in \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K) : \|f - f_0\|_n < \varepsilon\Big\}, \|.\|_n\Big) \le \Delta(\mathcal{V}_{\mathcal{S}}^K) \left(\frac{108}{\bar{C}}\sqrt{n}\right)^K, \tag{8.1}$$

where $\Delta(\mathcal{V}_{\mathcal{S}}^K)$ is the partitioning number of $\mathcal{V}_{\mathcal{S}}^K$.

PROOF. Let $f_{\mathcal{T},\boldsymbol{\beta}_1}$ and $f_{\mathcal{T},\boldsymbol{\beta}_2} \in \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$ be two step functions supported on a single valid tree partition $\mathcal{T} \in \mathcal{V}_{\mathcal{S}}^K$ with steps $\boldsymbol{\beta}_1 \in \mathbb{R}^K$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^K$. Then, by the minimum leaf size condition, we have

$$\frac{\bar{C}^2}{n}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \le \|f_{\mathcal{T},\boldsymbol{\beta}_1} - f_{\mathcal{T},\boldsymbol{\beta}_2}\|_n^2 = \sum_{k=1}^K \mu(\Omega_k)(\beta_{1k} - \beta_{2k})^2 \le \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2.$$

Denote by $f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}$ the $\|\cdot\|_n$ projection of f_0 onto $\mathcal{F}(\mathcal{T}) \subset \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$, the set of all step functions that live on a given partition $\mathcal{T}$. Then $\{\boldsymbol{\beta} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_n \le \varepsilon\} \subset \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2 \le \varepsilon\sqrt{n}/\bar{C}\}$ and $\{\boldsymbol{\beta} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_n \le \varepsilon/36\} \supset \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2 \le \varepsilon/36)\}$. This relationship shows that the $\varepsilon/36$ covering number of an $\|\cdot\|_n$ ball $\{\boldsymbol{\beta} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_n \le \varepsilon\}$ can be bounded from above by the $\varepsilon/36$ covering number of an Euclidean ball of a radius $\varepsilon\sqrt{n}/\bar{C}$, which is bounded by $(\frac{108}{\bar{C}}\sqrt{n})^K$. We can repeat this argument by projecting f_0 onto any valid tree topology $\mathcal{T} \in \mathcal{V}_{\mathcal{S}}^K$. The number of such valid trees is no larger than $\Delta(\mathcal{V}_{\mathcal{S}}^K)$, which completes the proof. □

The covering number for the entire sieve $\mathcal{F}_{\mathcal{T}}^n$ is then seen to satisfy

$$N\Big(\tfrac{\varepsilon}{36}, \Big\{f \in \mathcal{F}_{\mathcal{T}}^n : \|f - f_0\|_n < \varepsilon\Big\}, \|.\|_n\Big) < \sum_{q=0}^{q_n} \sum_{K=1}^{k_n} \sum_{\mathcal{S}:|\mathcal{S}|=q} N\Big(\tfrac{\varepsilon}{36}, \Big\{f \in \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K) : \|f - f_0\|_n < \varepsilon\Big\}, \|.\|_n\Big).$$

From Lemma 8.1 and Lemma 3.1, we obtain the following upper bound

$$\sum_{q=0}^{q_n}\sum_{K=1}^{k_n}\binom{p}{q}(K-1)!q^K n^K\left(\frac{108}{\bar{C}}\sqrt{n}\right)^K$$

$$< \sum_{K=0}^{k_n}\left(\frac{108}{\bar{C}}q_n n^{3/2}\,K\right)^K\sum_{q=0}^{q_n}\binom{p}{q}, \tag{8.2}$$

where we used the fact $K! < K^K$. Next, using the regularized incomplete beta function representation of the Binomial *cdf*, we can write

$$\sum_{q=0}^{q_n}\binom{p}{q} = 2^p(p-q_n)\binom{p}{q_n}\int_0^{1/2} x^{p-q_n-1}(1-x)^{q_n}\,\mathrm{d}\,x$$

$$\leq 2^{q_n+1}(p-q_n)\binom{p}{q_n}\frac{1-1/2^{q_n+1}}{q_n+1} \leq 2^{q_n+1}\binom{p}{q_n}\frac{p-q_n}{q_n+1}. \tag{8.3}$$

Using $\binom{p}{q_n} \leq (\mathrm{e}\,p/q_n)^{q_n}$, the quantity in (8.2) can be bounded by

$$\left(\frac{2\,\mathrm{e}\,p}{q_n}\right)^{q_n+1}\sum_{K=0}^{k_n}\left(\frac{108}{\bar{C}}q_n\,n^{3/2}\,k_n\right)^K < \left(\frac{2\,\mathrm{e}\,p}{q_n}\right)^{q_n+1}\frac{(C\,q_n\,n^{3/2}\,k_n)^{k_n+1}-1}{C\,q_n\,n^{3/2}\,k_n-1},$$

where $C = \left(\frac{108}{\bar{C}}\right)$. Finally, the entropy condition requires that the log-covering number, now upper bounded by

$$(q_n+1)\log(2\,\mathrm{e}\,p/q_n) + (k_n+1)\log(C\,q_n\,n^{3/2}\,k_n),$$

is no larger than (a constant multiple of) $n\,\varepsilon_n^2 = n^{q_0/(2\alpha+q_0)}\log^{2\beta} n$. Under our assumption $\log p \lesssim n^{q_0/(2\alpha+q_0)}$, this will be satisfied with $q_n = \lfloor C_q n\,\varepsilon_n^2/\log p\rfloor$ and $k_n = \lfloor C_k n\,\varepsilon_n^2/\log n\rfloor \sim n^{q_0/(2\alpha+q_0)}\log^{2\beta-1} n$. Furthermore, our choice of q_n satisfies $\log^{2\beta} n \lesssim q_n$. The constants C_q and C_k will be determined later.

8.2. *Condition* (2.2). We wish to show that the prior assigns enough mass around the truth in the sense that

$$\Pi(f \in \mathcal{F}_{\mathcal{T}} : \|f - f_0\|_n \leq \varepsilon_n) \geq \mathrm{e}^{-d\,n\varepsilon_n^2} \tag{8.4}$$

for some large enough $d > 0$. We establish this condition by finding a lower bound on the prior probability in (8.4), using all step functions supported on a single good partition. We denote by $\mathcal{S}_0$ the true index set of active

covariates, where $|\mathcal{S}_0| = q_0$. According to Lemma 3.2, there exists a tree-structured step function $f_{\widetilde{\mathcal{T}},\widetilde{\boldsymbol{\beta}}} \in \mathcal{F}(\mathcal{V}^K_{\mathcal{S}_0})$ for some $K = 2^{q_0\, s}$ and $s \in \mathbb{N}\backslash\{0\}$ such that

$$\|f_0 - f_{\widetilde{\mathcal{T}},\widetilde{\boldsymbol{\beta}}}\|_n \leq ||f_0||_{\mathcal{H}^\alpha} C_1\, q_0 / K^{\alpha/q_0} \tag{8.5}$$

for some $C_1 > 0$. Before we delve into the proof of Theorem 4.1, we present a proof of Lemma 3.2.

8.3. *Proof of Lemma 3.2.* We start with an auxiliary lemma showing that when f is α-Hölder continuous, we can grow a step function on any given (tree) partition so that the approximation error will be governed by a partition diameter.

LEMMA 8.2. *Assume $f \in \mathcal{H}^\alpha_p \cap \mathcal{C}(\mathcal{S})$ and a valid tree partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}^K_{\mathcal{S}}$. Then there exists a step function $f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}(\boldsymbol{x}) = \sum_{k=1}^K \widehat{\beta}_k\, \mathbb{I}_{\Omega_k}(\boldsymbol{x})$ such that*

$$\|f - f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}\|_n \leq ||f||_{\mathcal{H}^\alpha}\, \mathrm{diam}(\mathcal{T};\mathcal{S}).$$

PROOF. We give a constructive proof. Given $\mathcal{T}$ and design points $\mathcal{X}$, we take $\widehat{\beta}_k = \frac{1}{n_k}\sum_{i=1}^n f(\boldsymbol{x}_i)\mathbb{I}_{\Omega_k}(\boldsymbol{x}_i)$, where $n_k = \mu(\Omega_k)n$. Then for $\boldsymbol{x}_j \in \Omega_k \cap \mathcal{X}$ we have, from Hölder continuity,

$$\begin{aligned} |f(\boldsymbol{x}_j) - f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}(\boldsymbol{x}_j)| &< \frac{1}{n_k}\sum_{\boldsymbol{x}_i\in\Omega_k} |f(\boldsymbol{x}_j) - f(\boldsymbol{x}_i)| \\ &\leq ||f||_{\mathcal{H}^\alpha}\frac{1}{n_k}\sum_{\boldsymbol{x}_i\in\Omega_k} ||\boldsymbol{x}_{j\mathcal{S}} - \boldsymbol{x}_{i\mathcal{S}}||_2^\alpha \leq ||f||_{\mathcal{H}^\alpha}\, \mathrm{diam}^\alpha(\Omega_k;\mathcal{S}). \end{aligned}$$

Then the approximation error satisfies

$$||f - f_{\mathcal{T},\widehat{\boldsymbol{\beta}}}||_n \leq ||f||_{\mathcal{H}^\alpha}\sqrt{\sum_{k=1}^K \mu(\Omega_k)\mathrm{diam}^{2\alpha}(\Omega_k;\mathcal{S})} < ||f||_{\mathcal{H}^\alpha}\, \mathrm{diam}(\mathcal{T};\mathcal{S}).\ \square$$

Next, we grow a k-d tree partition $\widehat{\mathcal{T}} = \{\widehat{\Omega}_k\}_{k=1}^K$ (as explained in Remark 3.1) and construct an approximating step function $f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}$ according to Lemma 8.2. Then, the statement (3.4) follows directly from Lemma 8.2 and the Proposition 6 in [39], which shows $\mathrm{diam}(\widehat{\mathcal{T}},\mathcal{S}) \leq q/K^{1/(2q)}$ for $K = 2^{s\,q}$ and some $s \in \mathbb{N}\backslash\{0\}$.

The second statement (3.5) is obtained by noting

$$||f - f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}||_n < ||f||_{\mathcal{H}^\alpha} \sqrt{\sum_{k=1}^{K} \mu(\widehat{\Omega}_k) \mathrm{diam}^{2\alpha}(\widehat{\Omega}_k; \mathcal{S})}$$
$$< C_{\max} ||f||_{\mathcal{H}^\alpha} \max_{1 \le k \le K} \mathrm{diam}^{\alpha}(\widehat{\Omega}_k; \mathcal{S}),$$

where we used the fact that $\mu(\widehat{\Omega}_k) \le C_{max}^2/K$ in k-d trees. A minor modification of the proof of Proposition 6 in [39] yields $\sum_{k=1}^{K} \mu(\widehat{\Omega}_k)\mathrm{diam}(\widehat{\Omega}_k; \mathcal{S}) \le \frac{q}{K^{1/q}}$. The rest follows from Definition 3.3. □

Now we continue with the proof of Theorem 4.1. We find the smallest $K = 2^s q_0$ such that the function $f_{\widetilde{\mathcal{T}},\widetilde{\boldsymbol{\beta}}} \in \mathcal{F}(\mathcal{V}_{\mathcal{S}_0}^K)$ in (8.5) safely approximates f_0 with an error that is no larger than $\varepsilon_n/2$, a constant multiple of the target rate. Such a K will be denoted by a_n and is defined as the smallest K such that $\varepsilon_n \ge 2C_0 q_0 K^{-\alpha/q_0}$ for $C_0 = ||f_0||_{\mathcal{H}^\alpha} C_1$. Then we have

$$\left(\frac{2C_0 q_0}{\varepsilon_n}\right)^{\frac{q_0}{\alpha}} \le a_n \le \left(\frac{2C_0 q_0}{\varepsilon_n}\right)^{\frac{q_0}{\alpha}} + 1. \tag{8.6}$$

We denote by $\widehat{\mathcal{T}} = \{\widehat{\Omega}_k\}_{k=1}^{a_n} \in \mathcal{T}_{\mathcal{S}_0}^{a_n}$ the k-d tree partition from Lemma 3.2 obtained with the choice $K = a_n$. The tree $\widehat{\mathcal{T}}$ is not only valid, but also balanced in the sense that $C_{min}^2/a_n < \mu(\widehat{\Omega}_k) \le C_{max}^2/a_n$ for some $C_{min} < 1 < C_{max}$. The associated step sizes of the approximating tree will be denoted by $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{a_n}$. Now, we lower-bound the prior probability of the neighborhood $\{f \in \mathcal{F}_{\mathcal{T}} : \|f - f_0\|_n^2 \le \varepsilon_n^2\}$ by the prior probability of all regression trees supported on $\widehat{\mathcal{T}}$ inside this neighborhood (denoted by $\mathcal{F}(\widehat{\mathcal{T}})$):

$$\Pi(f \in \mathcal{F}_{\mathcal{T}} : \|f - f_0\|_n^2 \le \varepsilon_n^2)$$
$$\ge \pi(q_0)\pi(a_n) \frac{\Pi(f \in \mathcal{F}(\widehat{\mathcal{T}}) : \|f - f_0\|_n^2 \le \varepsilon_n^2)}{\binom{p}{q_0} \Delta(\mathcal{V}_{\mathcal{S}_0}^{a_n})}. \tag{8.7}$$

For any $\boldsymbol{\beta} \in \mathbb{R}^{a_n}$, we have

$$\|f_{\widehat{\mathcal{T}},\boldsymbol{\beta}} - f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_n^2 = \sum_{k=1}^{a_n} \mu(\widehat{\Omega}_k)(\beta_k - \widehat{\beta}_k)^2 \le \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2^2$$

and by the reverse triangle inequality

$$\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2 \ge \|f_{\widehat{\mathcal{T}},\boldsymbol{\beta}} - f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_n \ge \big|\|f_{\widehat{\mathcal{T}},\boldsymbol{\beta}} - f_0\|_n - \|f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}} - f_0\|_n\big|.$$

Then the statement $\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2 < \varepsilon_n/2$ implies $\|f_0 - f_{\widehat{\mathcal{T}},\boldsymbol{\beta}}\|_n < \|f_0 - f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_n + \varepsilon/2 < \varepsilon$, where the last inequality follows from the definition of a_n. Thus, we have

$$\{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2 \le \varepsilon_n/2\} \subset \{f \in \mathcal{F}(\widehat{\mathcal{T}}) : \|f_0 - f\|_n < \varepsilon_n\}.$$

Now we can lower-bound (8.7) with

$$L(q_0, a_n, \mathcal{S}_0, \widehat{\boldsymbol{\beta}}, \varepsilon_n) \equiv \pi(q_0)\pi(a_n)\frac{\Pi(\boldsymbol{\beta}\in\mathbb{R}^{a_n} : \|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2 \le \varepsilon_n/2)}{\binom{p}{q_0}\Delta(\mathcal{V}_{\mathcal{S}_0}^{a_n})}. \tag{8.8}$$

In order to bound $\Pi(\boldsymbol{\beta}\in\mathbb{R}^{a_n} : \|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2 \le \varepsilon_n/2)$, we follow the computations of [21], Theorem 12, to obtain

$$\Pi\Big(\boldsymbol{\beta}\in\mathbb{R}^{a_n} : \|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_2 \le \varepsilon_n/2\Big) \ge \frac{2^{-a_n}\mathrm{e}^{-\|\widehat{\boldsymbol{\beta}}\|_2^2-\varepsilon_n^2/8}}{\Gamma(\frac{a_n}{2})^{\frac{a_n}{2}}}\left(\frac{\varepsilon_n^2}{4}\right)^{\frac{a_n}{2}}. \tag{8.9}$$

From the triangle inequality (and because $\widehat{\mathcal{T}}$ is balanced) we have

$$\|\widehat{\boldsymbol{\beta}}\|_2 \le \frac{\sqrt{a_n}}{C_{min}}\|f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_n \le \frac{\sqrt{a_n}}{C_{min}}\left(\|f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}} - f_0\|_n + \|f_0\|_\infty\right) \le \frac{\sqrt{a_n}}{C_{min}}\left(\frac{\varepsilon_n}{2} + \|f_0\|_\infty\right).$$

Thereby we can write $\|\widehat{\boldsymbol{\beta}}\|_2^2 \le C_2\|f_0\|_\infty^2 a_n$ for some constant $C_2 > 0$. Now we continue with a lower bound to $L(q_0, a_n, \mathcal{S}_0, \widehat{\boldsymbol{\beta}}, \varepsilon_n)$ defined in (8.8). Using the following facts $\Gamma(x) \le x^x$, $\binom{p}{q_0} \le (ep/q_0)^{q_0}$ and $\Delta(\mathcal{V}_{\mathcal{S}_0}^{a_n}) < (a_n\, q_0\, n)^{a_n}$ (Lemma 3.1) and using (8.9), we arrive at the following lower bound:

$$\frac{\pi(a_n)c^{-q_0}p^{-aq_0}}{\left(\frac{\mathrm{e}\,p}{q_0}\right)^{q_0}(a_n\, q_0\, n)^{a_n}}\mathrm{e}^{-\varepsilon_n^2/8-a_n(C_2\|f_0\|_\infty^2+\log 2)}\left(\frac{\varepsilon_n^2}{4}\right)^{\frac{a_n}{2}}\left(\frac{2}{a_n}\right)^{a_n/2+1}. \tag{8.10}$$

Condition 2.2 will be satisfied if this quantity is at least as large as $\mathrm{e}^{-d\,n\varepsilon_n^2}$ for some large $d > 0$. We denote $c(p, q_0, a_n) = c^{-q_0}p^{-aq_0}(q_0/ep)^{q_0}$. Then we can rewrite (8.10) as

$$\pi(a_n)c(p, q_0, a_n)\left(\sqrt{2}\mathrm{e}^{C_2\|f_0\|_\infty^2+\log 2}a_n^{3/2}n\right)^{-a_n}\frac{2}{a_n}\left(\frac{\varepsilon_n}{q_0}\right)^{a_n}\mathrm{e}^{-\varepsilon_n^2/8}.$$

Taking minus the log of this quantity, Condition (2.2) will be met when

$$-\log c(p, q_0, a_n) - \log\pi(a_n) + a_n\log\left(\sqrt{2}\mathrm{e}^{C_2\|f_0\|_\infty^2+\log 2}a_n^{3/2}n\right) + a_n\log\left(\frac{q_0}{\varepsilon_n}\right) \tag{8.11}$$

is smaller than a constant multiple of $n\varepsilon_n^2$. Above, we omitted the small terms $\varepsilon_n^2/8$ (since $\varepsilon_n \to 0$) and $\log(a_n/2)$. First, we note that

$$-\log c(p, q_0, a_n) \le q_0 \log(c\, p^{a+1} \mathrm{e}/q_0).$$

With $q_0 \lesssim \log^\beta n$ and $\log p \lesssim n^{q_0/(2\alpha+q_0)}$, we obtain $-\log c(p, q_0, a_n) \lesssim n\varepsilon_n^2$. Next, focusing on the last term in (8.11), we obtain (from the left inequality in (8.6)) the following bound

$$q_0/\varepsilon_n \le a_n^{\alpha/q_0}/(2C_0)$$

and hence $q_0/\varepsilon_n \lesssim a_n$ for $\alpha \le 1$ and $q_0 \ge 1$. Moreover, from the right inequality in (8.6) we obtain for $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)} \log^\beta n$ and $2C_0 q_0 \lesssim \log^\beta n$

$$a_n \lesssim n^{q_0/(2\alpha+q_0)}. \tag{8.12}$$

Under our assumption $\|f_0\|_\infty \lesssim \log^\beta n$, (8.12) immediately yields $a_n \|f_0\|_\infty^2 \lesssim n\varepsilon_n^2$. All of these considerations, combined with the fact $-\log \pi(a_n) \lesssim a_n \log a_n$, yield the following leading term behind the last three summands in (8.11): $a_n \log(a_n^{3/2} n)$. Using (8.12), we obtain $a_n \log(a_n^{3/2} n) \lesssim a_n \log n \lesssim n\varepsilon_n^2$ for $\beta \ge 1/2$. Altogether, there exists $d > 0$ such that (8.4) is satisfied.

8.4. *Condition* (2.3). In order to establish Condition 2.3, we begin by noting $\Pi(\mathcal{F}_\mathcal{T} \backslash \mathcal{F}_\mathcal{T}^n) < \Pi(q > q_n) + \Pi(K > k_n)$. Thus, the condition will be met when both $\Pi(q > q_n) = o(\mathrm{e}^{-(d+2)\, n\varepsilon_n^2})$ and $\Pi(K > k_n) = o(\mathrm{e}^{-(d+2)\, n\varepsilon_n^2})$, where d is the constant deployed in Section 8.2. First, we find that

$$\Pi(q > q_n) \lesssim \sum_{k=q_n+1}^{p} (c\, p^a)^{-k} = (c\, p^a)^{-(q_n+1)} \frac{1 - (cp^a)^{-(p-q_n)}}{1 - 1/(cp^a)} < (c\, p^a)^{-q_n}.$$

With our choice $q_n = \lfloor C_q n\varepsilon_n^2 / \log p \rfloor$, it turns out that

$$\Pi(q > q_n) \mathrm{e}^{(d+2)\, n\varepsilon_n^2} < \mathrm{e}^{-q_n[\log c + a \log p] + (d+2)\, n\varepsilon_n^2} \to 0$$

for a large enough constant $C_q > 0$. Next, we apply the Chernoff bound for $\Pi(K > k_n)$. Namely, for any $t > 0$ we can write

$$\Pi(K > k_n) < \mathrm{e}^{-t\,(k_n+1)} \mathbb{E}\, \mathrm{e}^{t\,K} \propto \mathrm{e}^{-t\,(k_n+1)} \sum_{k=1}^{\infty} \frac{\left(\mathrm{e}^t \lambda\right)^k}{k!} \propto \mathrm{e}^{-t\,(k_n+1)} \left(\mathrm{e}^{\mathrm{e}^t \lambda} - 1\right). \tag{8.13}$$

With our choice $k_n = \lfloor C_k n\varepsilon_n^2 / \log n \rfloor \sim C_k n^{q_0/(2\alpha+q_0)} \log^{2\beta-1} n$ (Section 8.1) and with $t = \log k_n$ we obtain

$$\Pi(K > k_n) \mathrm{e}^{(d+2)\, n\varepsilon_n^2} \lesssim \mathrm{e}^{-(k_n+1)\log k_n + \lambda k_n + (d+2)\, n\varepsilon_n^2} \to 0$$

for a large enough constant C_k.

9. Proof of Theorem 6.1. Again, we aim to establish conditions (2.1), (2.2) and (2.3) for $\varepsilon_n^2 = \sum_{t=1}^{T_0} (\varepsilon_n^t)^2$, where $\varepsilon_n^t = n^{-\alpha^t/(2\alpha^t + q_0^t)} \log^{\beta^t} n$. Our sieve consists of valid forests with either (a) many trees that are small (weak learners), or (b) a few large trees (strong learners). We impose a joint requirement on $\sum_{t=1}^T K^t$ so that the overall number of leaves in the ensemble is small. At the same time, we require that $\sum_{t=1}^T q^t$ (the upper bound on the number of active variables in the ensemble) is small as well. The sieve is constructed as follows:

$$\mathcal{F}_{\mathcal{E}}^n = \bigcup_{T=1}^{\infty} \bigcup_{\boldsymbol{q}: \sum_{t=1}^T q^t \le s_n} \bigcup_{\boldsymbol{\mathcal{S}}: |\mathcal{S}^t| = q^t} \bigcup_{\boldsymbol{K}: \sum_{t=1}^T K^t \le z_n} \mathcal{F}(\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}) \tag{9.1}$$

for some integer values s_n and z_n. Throughout this section we denote $\bar{K} = \frac{1}{T}\sum_{t=1}^T K^t$.

9.1. *Condition 2.1.* We first obtain an upper bound on the covering number for $\mathcal{F}(\mathcal{E}) = \left\{ f_{\mathcal{E},\mathcal{B}} : [0,1]^p \to \mathbb{R} : f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) = \sum_{t=1}^T f_{\mathcal{T}^t,\boldsymbol{\beta}^t}(\boldsymbol{x}); \boldsymbol{\beta}^t \in \mathbb{R}^{K^t} \right\}$, the set of all additive step functions supported on a single partition ensemble $\mathcal{E} \in \mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$. We denote by $\widetilde{\mathcal{T}}(\mathcal{E}) = \{\widetilde{\Omega}_k\}_{k=1}^{K(\mathcal{E})}$ the global partition associated with $\mathcal{E}$, consisting of $K(\mathcal{E})$ global cells. For $\mathcal{B}_1, \mathcal{B}_2 \in \mathbb{R}^{T \times \bar{K}}$, we denote by $f_{\mathcal{E},\mathcal{B}_1}, f_{\mathcal{E},\mathcal{B}_2} \in \mathcal{F}(\mathcal{E})$ two additive regression trees that sit on the same partition ensemble $\mathcal{E}$. Let $\bar{\boldsymbol{\beta}}_1 = \boldsymbol{A}(\mathcal{E})\mathcal{B}_1$ and $\bar{\boldsymbol{\beta}}_2 = \boldsymbol{A}(\mathcal{E})\mathcal{B}_2$ be the aggregated step sizes, as defined in (5.4), where $\boldsymbol{A}(\mathcal{E})$ is the stretching matrix. Because $\widetilde{\mathcal{T}}(\mathcal{E})$ has non-empty cells, we can write

$$\frac{1}{n}||\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_2||_2^2 \le ||f_{\mathcal{E},\mathcal{B}_1} - f_{\mathcal{E},\mathcal{B}_2}||_n^2 = \sum_{k=1}^{K(\mathcal{E})} \mu(\widetilde{\Omega}_k)(\bar{\beta}_{1k} - \bar{\beta}_{2k})^2 \le ||\bar{\boldsymbol{\beta}}_1 - \bar{\boldsymbol{\beta}}_2||_2^2.$$

Deploying the singular value decomposition $\boldsymbol{A}(\mathcal{E}) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ we write $\tilde{\mathcal{B}}_1 = \boldsymbol{V}^T\mathcal{B}_1$ and $\tilde{\mathcal{B}}_2 = \boldsymbol{V}^T\mathcal{B}_2$. Using the fact that $\boldsymbol{U}$ is unitary, we have

$$\frac{1}{n}||\boldsymbol{D}(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_2)||_2^2 \le ||f_{\mathcal{E},\mathcal{B}_1} - f_{\mathcal{E},\mathcal{B}_2}||_n^2 \le ||\boldsymbol{D}(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_2)||_2^2.$$

Denote by $\lambda_{max}(\mathcal{E})$ and $\lambda_{min}(\mathcal{E})$ the maximal and minimal singular values of $\boldsymbol{A}(\mathcal{E})$. Because $\boldsymbol{V}$ is unitary, we obtain the following equivalence between norms:

$$\frac{\lambda_{min}^2(\mathcal{E})}{n}||\mathcal{B}_1 - \mathcal{B}_2||_2^2 \le ||f_{\mathcal{E},\mathcal{B}_1} - f_{\mathcal{E},\mathcal{B}_2}||_n^2 \le \lambda_{max}^2(\mathcal{E})||\mathcal{B}_1 - \mathcal{B}_2||_2^2. \tag{9.2}$$

Now we can argue as in Lemma 8.1 and upper-bound $\log N\Big(\frac{\varepsilon}{36}, \Big\{f \in \mathcal{F}(\mathcal{E}) : \|f - f_0\|_n < \varepsilon\Big\}, \|.\|_n\Big)$ with

$$(T \times \bar{K}) \log\left(108\, \kappa(\mathcal{E})\, \sqrt{n}\right), \tag{9.3}$$

where $\kappa(\mathcal{E}) = \lambda_{max}(\mathcal{E})/\lambda_{min}(\mathcal{E})$ is the condition number of $\boldsymbol{A}(\mathcal{E})$. We can control $\kappa(\mathcal{E})$ with an upper bound obtained in the following lemma.

LEMMA 9.1. *Assume an ensemble $\mathcal{E}$ consisting of T trees, each with K^t leaves. When $\lambda^2_{min}(\mathcal{E}) \gtrsim n^{-\delta}$ for some $\delta \geq 0$, we have*

$$\kappa^2(\mathcal{E}) \leq n^{\delta} K(\mathcal{E})(T \times \bar{K}), \tag{9.4}$$

where $\bar{K} = \frac{1}{T}\sum_{t=1}^{T} K^t$.

PROOF. By the Gershgorin circle theorem, all eigenvalues of $\boldsymbol{A}(\mathcal{E})'\boldsymbol{A}(\mathcal{E}) = \widetilde{\boldsymbol{A}}(\mathcal{E}) = (\widetilde{a}_{ij})$ lie inside the union of intervals $[\widetilde{a}_{ii} - \sum_{j\neq i}\widetilde{a}_{ij}, \widetilde{a}_{ii} + \sum_{j\neq i}\widetilde{a}_{ij}]$ for $i = 1, \ldots, T \times \bar{K}$. As explained in Section 5.1, the diagonal and off-diagonal entries of $\widetilde{\boldsymbol{A}}(\mathcal{E})$ quantify the persistence and the overlap in terms of the number of intersecting global partitioning cells. The magnitude $|\widetilde{a}_{ij}|$ is no larger than $K(\mathcal{E})$ for each $1 \leq i, j \leq T \times \bar{K}$. The upper bound on the maximal eigenvalue $\lambda^2_{max}(\mathcal{E})$ is thus $K(\mathcal{E})(T \times \bar{K})$. □

It is worth pointing out that when the stretching matrix $\boldsymbol{A}(\mathcal{E})$ is wide, or more generally when $K(\mathcal{E}) < D\,(T \times \bar{K})$ for some $D > 0$, we can replace (9.3) directly with

$$D(T \times \bar{K}) \log\left(108\, \sqrt{n}\right), \tag{9.5}$$

a bound not involving $\kappa(\mathcal{E})$.

Now we find an upper bound on the number of valid ensembles $\mathcal{E} \in \mathcal{VE}^{\boldsymbol{K}}_{\boldsymbol{\mathcal{S}}}$ inside the sieve $\mathcal{F}^n_{\mathcal{E}}$. To start, we note that given $(T, \boldsymbol{q}, \boldsymbol{K}, \boldsymbol{\mathcal{S}})$, there are at most $\prod_{t=1}^{T}(K^t\, q^t\, n)^{K^t}$ valid ensembles $\mathcal{VE}^{\boldsymbol{K}}_{\boldsymbol{\mathcal{S}}}$. This bound is obtained from Lemma 3.1 by combining all possible T-tuples of trees.[1] Given $(T, \boldsymbol{q})$, there are $\prod_{t=1}^{T}\binom{p}{q^t}$ sets of subsets $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^1, \ldots, \mathcal{S}^T\}$ satisfying the constraint

[1] The order of trees in $\mathcal{E}$ matters.

$|\mathcal{S}^t| = q^t$. This leads to an overall upper bound

$$\begin{aligned}
&\sum_{T=1}^{\min\{s_n,z_n\}} \sum_{\boldsymbol{K}:\sum_{t=1}^T K^t \le z_n} \sum_{\boldsymbol{q}:\sum_{t=1}^T q^t \le s_n} \prod_{t=1}^T \binom{p}{q^t} \left(K^t\, q^t\, n\right)^{K^t} \\
&\qquad < \sum_{T=1}^{\min\{s_n,z_n\}} \sum_{\boldsymbol{K}:\sum_{t=1}^T K^t \le z_n} (z_n\, s_n\, n)^{z_n} \sum_{\boldsymbol{q}:\sum_{t=1}^T q^t \le s_n} \prod_{t=1}^T \left(\frac{p\,\mathrm{e}}{q^t}\right)^{q^t} \\
&\qquad < s_n^{s_n+1} z_n^{z_n+1} (z_n\, s_n\, n)^{z_n} (p\,\mathrm{e})^{s_n}\,.
\end{aligned}$$

Combining this bound with (9.3) and (9.4) (or (9.5)), and because $K(\mathcal{E}) \le n$ (recall that $\widetilde{\mathcal{T}}(\mathcal{E})$ has non-empty cells) we obtain the following bound

$$\begin{aligned}
\log N\Big(\tfrac{\varepsilon}{36}, \Big\{f \in \mathcal{F}^n_{\mathcal{E}} : \|f - f_0\|_n < \varepsilon\Big\}, \|.\|_n\Big) &< (s_n+1)\log s_n + (z_n+1)\log z_n \\
&+ z_n \log(z_n\, s_n\, n) + s_n \log(p\,\mathrm{e}) + D\, z_n \log\left(108\sqrt{z_n} n^{1+\delta/2}\right).
\end{aligned} \tag{9.6}$$

Condition 2.1 will be met when (9.6) is smaller than (a constant multiple of) $n\varepsilon_n^2 = \sum_{t=1}^{T_0} n(\varepsilon_n^t)^2$. With the choice $s_n = \lfloor C_s n\varepsilon_n^2 / \log p \rfloor$ and $z_n = \lfloor C_z n\varepsilon_n^2 / \log n \rfloor$, where C_s and C_z are large enough constants to be determined later, this condition is satisfied.

9.2. *Condition* (2.2). To establish Condition (2.2) for tree ensembles, we begin by finding a single additive tree that approximates well. We will heavily leverage our findings from Section 8.2, noting that the problem of approximating an additive function f_0 with a sum of trees can be decomposed into smaller problems of approximating each layer f_0^t separately.

Denote by a_n^t the smallest leaf size of a k-d tree (defined in Remark 3.1) needed to approximate f_0^t with an error smaller than $\varepsilon_n^t/2$, where $\varepsilon_n^t = n^{-\alpha^t/(2\alpha^t+q_0^t)} \log^{\beta^t} n$. Such a tree step function approximation exists according to Lemma 3.2 when $\mathcal{X}$ is $(M, \mathcal{S}_0^t)$-regular. We will denote this approximation with $f_{\widehat{\mathcal{T}}^t, \widehat{\boldsymbol{\beta}}^t}(\boldsymbol{x})$. Moreover, with $\boldsymbol{a}_n = (a_n^1, \ldots, a_n^{T_0})'$ we denote the vector of such minimal tree sizes, where each a_n^t satisfies (8.6) with q_0^t, α^t and ε_n^t. Next, we will denote by $\widehat{\mathcal{E}} = \{\widehat{\mathcal{T}}^1, \ldots, \widehat{\mathcal{T}}^{T_0}\}$ the approximating partition ensemble with step heights $\widehat{\mathcal{B}} = (\widehat{\boldsymbol{\beta}}^{1\prime}, \ldots, \widehat{\boldsymbol{\beta}}^{T_0\prime})'$. This ensemble is δ_n-valid for some $0 < \delta_n < \delta$. The individual tree approximations $f_{\widehat{\mathcal{T}}^t, \widehat{\boldsymbol{\beta}}^t}(\boldsymbol{x})$ are woven into an approximating forest $f_{\widehat{\mathcal{E}}, \widehat{\mathcal{B}}}(\boldsymbol{x}) = \sum_{t=1}^{T_0} f_{\widehat{\mathcal{T}}^t, \widehat{\boldsymbol{\beta}}^t}(\boldsymbol{x}) \in \mathcal{F}(\mathcal{V}\mathcal{E}^{\boldsymbol{a}_n}_{\boldsymbol{\mathcal{S}}_0})$, where $\boldsymbol{\mathcal{S}}_0 = \{\mathcal{S}_0^1, \ldots, \mathcal{S}_0^{T_0}\}$.

Arguing as in Section 8.2, the statement $\|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^t\|_2 < \frac{\varepsilon_n^t}{2}$ for all $1 \le t \le T_0$ implies $\|f_0^t - f_{\widehat{\mathcal{T}}^t, \boldsymbol{\beta}^t}\|_n < \varepsilon_n^t$ for all $1 \le t \le T_0$ and $\boldsymbol{\beta}^t \in \mathbb{R}^{a_n^t}$. This further

implies

$$\|f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}\|_n \leq \sum_{t=1}^{T_0} \|f_0^t - f_{\widehat{\mathcal{T}}^t,\boldsymbol{\beta}^t}\|_n \leq \sum_{t=1}^{T_0} \varepsilon_n^t \leq \sqrt{T_0}\varepsilon_n,$$

for any $\mathcal{B} = (\boldsymbol{\beta}^{1\prime}, \ldots, \boldsymbol{\beta}^{T_0\prime})' \in \mathbb{R}^{T_0 \times \bar{\boldsymbol{a}}_n}$, where the final inequality is due to Cauchy-Schwarz and where $\bar{\boldsymbol{a}}_n = \frac{1}{T_0}\sum_{t=1}^{T_0} a_n^t$. Denote by $\mathcal{F}(\widehat{\mathcal{E}})$ the set of all additive trees supported on the partition ensemble $\widehat{\mathcal{E}}$. Then we can write

$$\begin{aligned}
&\Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : \|f_0 - f\|_n \leq \varepsilon_n\right) \\
&\quad \geq \Pi\left(\boldsymbol{\beta} \in \mathbb{R}^{\sum_{t=1}^{T_0} a_n^t} : \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^t\|_2 \leq \frac{\varepsilon_n^t}{2\sqrt{T_0}} \quad \text{for each } t = 1, \ldots, T_0\right) \\
&\quad = \prod_{t=1}^{T_0} \Pi\left(\boldsymbol{\beta}^t \in \mathbb{R}^{a_n^t} : \|\widehat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^t\|_2 \leq \frac{\varepsilon_n^t}{2\sqrt{T_0}}\right).
\end{aligned}$$

Because we assumed $\beta_j^t \sim \mathcal{N}(0, 1/T)$, given T, we can directly use (8.9) to lower-bound the above with

$$\prod_{t=1}^{T_0} \frac{2^{-a_n^t} \mathrm{e}^{-\|\widehat{\boldsymbol{\beta}}^t\|_2^2 - (\varepsilon_n^t)^2/8}}{\Gamma(\frac{a_n^t}{2})^{\frac{a_n^t}{2}}} \left(\frac{(\varepsilon_n^t)^2}{4}\right)^{\frac{a_n^t}{2}}. \tag{9.7}$$

Because each tree $\widehat{\mathcal{T}}^t$ is a k-d tree and is by definition balanced, we have $\|\widehat{\boldsymbol{\beta}}^t\|_n^2 \lesssim a_n^t \|f_0^t\|_\infty^2$. Now we can directly apply all our calculations from Section 8.2. In particular, using (9.7) and noting that $\Delta(\mathcal{V}\mathcal{E}_{\boldsymbol{\mathcal{S}}_0}^{\boldsymbol{a}_n}) < \prod_{t=1}^{T_0} \Delta(\mathcal{V}_{\mathcal{S}_0^t}^{a_n^t})$, we obtain

$$\begin{aligned}
\Pi\left(f \in \mathcal{F}_{\mathcal{E}} : \|f_0 - f\|_n \leq \varepsilon_n\right) &\geq \pi(T_0)\pi(\boldsymbol{q}_0 \,|\, T_0)\pi(\boldsymbol{a}_n \,|\, T_0)\pi(\boldsymbol{\mathcal{S}}_0 \,|\, T, \boldsymbol{q}_0) \times \\
&\quad \times \pi(\widehat{\mathcal{E}} \,|\, \boldsymbol{\mathcal{S}}_0, \boldsymbol{a}_n)\Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : ||f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}||_n \leq \varepsilon_n\right) \\
&> \pi(T_0) \prod_{t=1}^{T_0} L(q_0^t, a_n^t, \mathcal{S}_0^t, \widehat{\boldsymbol{\beta}}^t, \varepsilon_n^t),
\end{aligned}$$

where $L(\cdot)$ was defined in (8.8). It follows from Section (8.2) that

$$-\log L(q_0^t, a_n^t, \mathcal{S}_0^t, \widehat{\boldsymbol{\beta}}^t, \varepsilon_n^t) \lesssim n(\varepsilon_n^t)^2$$

for each $1 \leq t \leq T_0$ when $q_0^t \lesssim \log^{\beta^t} n$, $\log p \lesssim \min_{1 \leq t \leq T_0} n^{q^t/(2\alpha^t + q_0^t)}$ and $T_0 \lesssim n$. The last condition $T_0 \lesssim n$ is needed under our prior $a_n \sim \text{Poisson}(\lambda/T)$ so that $-\log \pi(a_n) \lesssim a_n \log a_n + a_n \log T_0 \lesssim n(\varepsilon_n^t)^2 = n^{q_0^t/(2\alpha^t + q_0^t)} \log^{2\beta^t} n$ for

$\beta^t \geq 1/2$. Putting all the pieces together, we obtain the following lower bound

$$\Pi\left(f \in \mathcal{F}_{\mathcal{E}} : \|f_0 - f\|_n \leq \varepsilon_n\right) \geq \pi(T_0) \mathrm{e}^{-d\, n \sum_{t=1}^{T_0} (\varepsilon_n^t)^2}$$

for some suitably large $d > 0$. The last requirement needed for Condition (2.2) to be satisfied is that $\pi(T_0) \geq \mathrm{e}^{-d\, n\varepsilon_n^2}$. Our prior $\pi(T) \propto \mathrm{e}^{-C_T T}$ safely satisfies this requirement.

9.3. *Condition* (2.3). The condition entails showing that $\Pi(\mathcal{F}_{\mathcal{E}} \backslash \mathcal{F}_{\mathcal{E}}^n) = o(\mathrm{e}^{-(d+2)\, n\varepsilon_n^2})$ for d deployed in the previous section. It suffices to show that

$$\left[\Pi\left((T, \boldsymbol{K}) : \sum_{t=1}^{T} K^t > z_n\right) + \Pi\left((T, \boldsymbol{q}) : \sum_{t=1}^{T} q^t > s_n\right)\right] \mathrm{e}^{(d+2)\, n\varepsilon_n^2} \to 0.$$

Because we assume $K^t \mid T \overset{iid}{\sim} \text{Poisson}(\lambda/T)$ for some $\lambda \in \mathbb{R}$ (according to our definition in (T4*)), we can apply a similar Chernoff bound as in (8.13). Namely, we have for any $\gamma > 0$

$$\begin{aligned}\Pi\left(\sum_{t=1}^{T} K^t > z_n\right) &= \sum_{T=1}^{\infty} \pi(T) \Pi\left(\sum_{t=1}^{T} K^t > z_n \mid T\right) \\ &\lesssim \mathrm{e}^{-\gamma(z_n+1)} \sum_{T=1}^{\infty} \pi(T) \left(\mathrm{e}^{\mathrm{e}^{\gamma}\lambda/T} - 1\right)^T \lesssim \mathrm{e}^{-\gamma(z_n+1)+\mathrm{e}^{\gamma}\lambda}.\end{aligned}$$

With $z_n = \lfloor C_z n\varepsilon_n^2 / \log n \rfloor \sim \sum_{t=1}^{T_0} n^{q_0^t/(2\alpha^t+q_0^t)} \log^{2\beta^t-1} n$ and $\gamma = \log z_n$, we have $\Pi\left(\sum_{t=1}^{T} K^t > z_n\right) \mathrm{e}^{d\, n\varepsilon_n^2} \to 0$ for a large enough constant $C_z > 0$. Next, with the independent product prior (T1*), the Chernoff bound gives

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n \mid T\right) \leq \mathrm{e}^{-\gamma(s_n+1)} \prod_{t=1}^{T} \mathbb{E}\left[\mathrm{e}^{\gamma q^t}\right] \lesssim \mathrm{e}^{-\gamma(s_n+1)} \mathrm{e}^{-T \log[1-\mathrm{e}^{\gamma}/(cp^a)]}$$

for any $\gamma > 0$, where we used the fact

$$\mathbb{E}\left[\mathrm{e}^{\gamma q^t}\right] = \sum_{q=0}^{p} \left[\mathrm{e}^{\gamma}/(cp^a)\right]^q < \frac{1}{1 - \mathrm{e}^{\gamma}/(cp^a)}.$$

With $\gamma = \log p$ and $a > 2$, we can write

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n \mid T\right) \lesssim \mathrm{e}^{-(s_n+1)\, \log p} \mathrm{e}^{-T \log\left[1-1/(cp^{a-1})\right]} < \mathrm{e}^{-(s_n+1) \log p - T \log[1-1/(cp)]}.$$

Next, we have

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n\right) \lesssim \mathrm{e}^{-(s_n+1)\log p} \sum_{T=1}^{\infty} \pi(T)\mathrm{e}^{T\log[1+1/(cp-1)]}.$$

With $\pi(T) \propto \mathrm{e}^{-C_T T}$, where $C_T > \log 2$, we have

$$\Pi\left(\sum_{t=1}^{T} q^t > s_n\right) \lesssim \mathrm{e}^{-(s_n+1)\log p} \sum_{T=1}^{\infty} \mathrm{e}^{-T(C_T-\log 2)} \lesssim \mathrm{e}^{-(s_n+1)\log p}.$$

With $s_n = \lfloor C_q n\varepsilon_n^2/\log p \rfloor$ we have $\Pi\left(\sum_{t=1}^{T} q^t > s_n\right) \mathrm{e}^{(d+2)\, n\varepsilon_n^2} \to 0$ for a large enough constant C_q.

10. Proof of Theorem 5.1. The sieve will be very similar to (9.1). The only difference is that each tree in the ensemble is now constrained to depend on the same set of active variables $\mathcal{S}$. To mark this difference, we have denoted the partition ensembles with $\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}}$ instead of $\mathcal{VE}_{\boldsymbol{\mathcal{S}}}^{\boldsymbol{K}}$. Throughout this section, we use the following sieve:

$$\mathcal{F}_{\mathcal{E}}^{n} = \bigcup_{T=1}^{\infty} \bigcup_{q=0}^{q_n} \bigcup_{\mathcal{S}:|\mathcal{S}|=q} \bigcup_{\boldsymbol{K}:\sum_{t=1}^{T} K^t \le z_n} \mathcal{F}(\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}}). \tag{10.1}$$

10.1. *Condition 2.1.* Our sieve (10.1) is embedded in (9.1), where the number of ensembles $\mathcal{E}$ inside $\mathcal{F}_{\mathcal{E}}^{n}$ is now upper-bounded by

$$\sum_{T=1}^{z_n} \sum_{q=0}^{q_n} \sum_{\boldsymbol{K}:\sum_{t=1}^{T} K^t \le z_n} \binom{p}{q} \prod_{t=1}^{T} \left(K^t\, q\, n\right)^{K^t} < z_n^{z_n+1}(q_n+1)\left(z_n\, q_n\, n\right)^{z_n}\left(p\,\mathrm{e}\right)^{q_n}.$$

Using the same arguments as in Section 9.1, we choose $q_n = \lfloor C_q n\varepsilon_n^2/\log p \rfloor$ and $z_n = \lfloor C_z n\varepsilon_n^2/\log n \rfloor$, for which the condition holds.

10.2. *Condition (2.2).* The key ingredient for establishing Condition (2.2) is the following lemma on the existence of a tree ensemble that approximates f_0 well.

LEMMA 10.1. *Assume $f \in \mathcal{H}_p^{\alpha} \cap \mathcal{C}(\mathcal{S})$, where $|\mathcal{S}| = q$. Then for any $s \in \mathbb{N}\backslash\{0\}$, there exists an additive tree function $f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}} \in \mathcal{F}(\mathcal{VE}_{\mathcal{S}}^{\boldsymbol{K}})$ consisting of $T = 2^{s\,q-1}$ trees, each with $K^t = s\,q+1$ leaves, such that*

$$\|f - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \le ||f||_{\mathcal{H}^{\alpha}} q/K(\widehat{\mathcal{E}})^{1/(2q)},$$

where $K(\widehat{\mathcal{E}}) = 2^{s\,q}$. *Moreover, if* $\mathcal{X}$ *is* $(M, \mathcal{S})$*-regular, then*

$$\|f - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \leq ||f||_{\mathcal{H}^\alpha} C\, M^\alpha\, q / K(\widehat{\mathcal{E}})^{\alpha/q}$$

for some $C > 0$.

Proof. From Lemma 3.2, we know that there exists a single tree function $f_{\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}} \in \mathcal{F}(\mathcal{V}_{\mathcal{S}}^K)$ with $\widehat{K} = 2^{s\,q}$ leaves which approximates well. We regard the full symmetric tree $\widehat{\mathcal{T}}$ as the global partition of the approximating partition ensemble $\widehat{\mathcal{E}}$, i.e. $\widetilde{\mathcal{T}}(\widehat{\mathcal{E}}) = \widehat{\mathcal{T}}$. Moreover, $\widehat{\boldsymbol{\beta}}$ is regarded as the vector of aggregated steps, i.e. $\bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$ (the definition of the aggregated steps is in (5.4)). The actual ensemble $\widehat{\mathcal{E}}$ is obtained from $\widehat{\mathcal{T}}$ by redistributing the cuts among T trees, each with $K^t = K$ leaves, in the following way. We take completely imbalanced trees that keep refining one cell until the resolution reaches the tree depth $\log_2 \widehat{K}$. These trees have $K^t = \log_2 \widehat{K} + 1$ leaves and we need $T = \widehat{K}/2$ of those to sum up towards $\widehat{\mathcal{T}}$. This decomposition is illustrated in Figure 3, where a full symmetric tree $\widehat{\mathcal{T}}$ (Figure 4) with $\widehat{K} = 8$ leaves has been trimmed into $T = \widehat{K}/2 = 4$ smaller trees with $K^t = \log_2 \widehat{K} + 1 = 4$ leaves. The decomposition yields a tree ensemble $\widehat{\mathcal{E}} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ with a stretching matrix $\boldsymbol{A}(\widehat{\mathcal{E}}) = [\mathrm{I}_{\widehat{K}}, \boldsymbol{A}_1]$ (after a suitable permutation of columns), where $\boldsymbol{A}_1$ is some binary matrix. It follows from Lemma 1(g) of Govaerts and Pryce [24] that $\lambda^2_{min}(\widehat{\mathcal{E}}) = 1 + \sigma^2_{min}(\boldsymbol{A}_1) \geq 1$, where $\sigma_{min}(\boldsymbol{A}_1)$ denotes the smallest singular value of $\boldsymbol{A}_1$. Moreover, we have $K(\widehat{\mathcal{E}}) = \widehat{K}$. Finally, we use (5.4) to obtain the individual tree steps via $\widehat{\mathcal{B}} = (\boldsymbol{A}(\widehat{\mathcal{E}})'\boldsymbol{A}(\widehat{\mathcal{E}}))^\dagger \boldsymbol{A}(\widehat{\mathcal{E}})'\bar{\boldsymbol{\beta}}$, where $\boldsymbol{A}^\dagger$ denotes the Moore pseudoinverse of $\boldsymbol{A}$. The rest follows from Lemma 3.2. □

Now we proceed with Condition 2.2. Denote by $\widehat{\mathcal{E}}$ the approximating ensemble from Lemma 10.1. Recall that the global partition $\widetilde{T}(\widehat{\mathcal{E}}) = \{\widetilde{\Omega}_k\}_{k=1}^{K(\widehat{\mathcal{E}})}$ is a k-d tree, which is balanced in the sense that $C^2_{min}/K(\widehat{\mathcal{E}}) \leq \mu(\widetilde{\Omega}_k) \leq C^2_{max}/K(\widehat{\mathcal{E}})$ for some constants $C_{min} < 1 < C_{max}$ and $k = 1, \ldots, K(\widehat{\mathcal{E}})$. Next, we find the smallest $K(\widehat{\mathcal{E}})$ such that $||f||_{\mathcal{H}^\alpha} C\, M^\alpha\, q_0 / K(\widehat{\mathcal{E}})^{\alpha/q_0} < \varepsilon_n/2$. This value will be denoted by a_n and it satisfies (8.6). Next, we denote by $\widehat{T} = a_n/2$ the number of approximating trees and by $\widehat{\boldsymbol{K}} = (\widehat{K}^1, \ldots, \widehat{K}^T)'$ the vector of leaves, where $\widehat{K}^t = \log_2 a_n + 1$ (again we are using the construction from Lemma 10.1). Then, using similar arguments as in Section 9.2 we can lower-bound $\Pi(f \in \mathcal{F}_{\mathcal{E}} : \|f - f_0\|_n \leq \varepsilon_n)$ with

$$\frac{\pi(\widehat{T})\pi(\widehat{\boldsymbol{K}} \mid \widehat{T})\pi(q_0)}{\left(\frac{\mathrm{e}\,p}{q_0}\right)^{q_0} \prod_{t=1}^{\widehat{T}} (\widehat{K}^t\, q_0\, n)^{\widehat{K}^t}} \Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : ||f - f_0||_n \leq \varepsilon_n\right), \tag{10.2}$$

where $\mathcal{F}(\widehat{\mathcal{E}})$ consists of all additive tree functions supported on $\widehat{\mathcal{E}}$. We denote by $\widetilde{a}_n = \sum_{t=1}^{\widehat{T}} \widehat{K}^t = a_n/2(\log_2 a_n + 1)$ and by $\widehat{\mathcal{B}} \in \mathbb{R}^{\widetilde{a}_n}$ the steps of the approximating additive trees from Lemma 10.1. Because $\mu(\widetilde{\Omega}_k) \leq C_{max}^2/K(\widehat{\mathcal{E}})$ we obtain for any arbitrary vector $\mathcal{B} \in \mathbb{R}^{\widetilde{a}_n}$ (similarly as in Section 9.1)

$$\|f_{\widehat{\mathcal{E}},\mathcal{B}} - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \leq C_{max}\lambda_{max}(\widehat{\mathcal{E}})/\sqrt{K(\widehat{\mathcal{E}})}\|\mathcal{B} - \widehat{\mathcal{B}}\|_2$$

and thereby

$$\|\mathcal{B} - \widehat{\mathcal{B}}\|_2 \geq \sqrt{K(\widehat{\mathcal{E}})}/(C_{max}\lambda_{max}(\widehat{\mathcal{E}}))\big|\|f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}\|_n - \|f_0 - f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n\big|.$$

Combined with the fact $\lambda_{\max}^2(\mathcal{E}) \leq K(\mathcal{E})\widetilde{a}_n$ (as shown in the proof of Lemma 9.1), the statement $\|\mathcal{B} - \widehat{\mathcal{B}}\|_2 < \frac{\varepsilon_n}{2}\frac{1}{C_{max}\sqrt{\widetilde{a}_n}}$ implies $\|f_0 - f_{\widehat{\mathcal{E}},\mathcal{B}}\|_n < \varepsilon_n$. Therefore we have

$$\Pi\left(f \in \mathcal{F}(\widehat{\mathcal{E}}) : \|f - f_0\|_n \leq \varepsilon_n\right) > \Pi\left(\mathcal{B} \in \mathbb{R}^{\widetilde{a}_n} : \|\mathcal{B} - \widehat{\mathcal{B}}\|_2 < \frac{\varepsilon_n}{2}\frac{1}{C_{max}\sqrt{\widetilde{a}_n}}\right).$$

Moreover, because $\mu(\widetilde{\Omega}_k) \geq C_{min}^2/K(\widehat{\mathcal{E}})$ for some $C_{min} < 1$, we have

$$\|\widehat{\mathcal{B}}\|_2 \leq \frac{\sqrt{a_n}}{C_{min}\lambda_{min}(\widehat{\mathcal{E}})}\|f_{\widehat{\mathcal{E}},\widehat{\mathcal{B}}}\|_n \leq \frac{\sqrt{a_n}}{C_{min}}\left(\frac{\varepsilon_n}{2} + \|f_0\|_\infty\right),$$

where we used the fact $\lambda_{min}^2(\widehat{\mathcal{E}}) \geq 1$ (proof of Lemma 10.1). Therefore we have $\|\widehat{\mathcal{B}}\|_2^2 \leq C_2 a_n \|f_0\|_\infty^2$ for some $C_2 > 0$. Following the calculations from Section 8.2 (namely (8.10)), we continue to lower-bound (10.2) with

(10.3)
$$\frac{\pi(\widehat{T})\pi(\widehat{\boldsymbol{K}} \mid \widehat{T})\pi(q_0)\mathrm{e}^{-\frac{\varepsilon_n^2}{8C_{max}^2\widetilde{a}_n} - a_n(C_2\|f_0\|_\infty^2 + \log 2)}}{\left(\frac{\mathrm{e}\,p}{q_0}\right)^{q_0}\left[(\log_2 a_n + 1)\,q_0\,n\right]^{\widetilde{a}_n}}\left(\frac{\varepsilon_n^2}{4C_{max}^2\widetilde{a}_n}\right)^{\frac{\widetilde{a}_n}{2}}\left(\frac{2}{\widetilde{a}_n}\right)^{\widetilde{a}_n/2+1}.$$

This quantity should be at least $\mathrm{e}^{-d\,n\varepsilon_n^2}$ for some suitably large $d > 0$. Now, with our prior (T4*) we can write

$$\pi(\widehat{\boldsymbol{K}} \mid \widehat{T}) \gtrsim \prod_{t=1}^{\widehat{T}} \mathrm{e}^{\widehat{K}^t \log(\lambda/\widehat{T}) - \widehat{K}^t \log \widehat{K}^t} \gtrsim \mathrm{e}^{-\widetilde{a}_n \log(\log_2 a_n + 1) + \widetilde{a}_n \log(2\lambda/a_n)}.$$

This quantity can be lower-bounded by $\mathrm{e}^{-D\,\widetilde{a}_n \log a_n}$ for some $D > 0$. Then we can write

$$\frac{\pi(\widehat{T})\pi(\widehat{\boldsymbol{K}} \mid \widehat{T})\pi(q_0)}{\left(\frac{\mathrm{e}\,p}{q_0}\right)^{q_0}\left[(\log_2 a_n + 1)\,q_0\,n\right]^{\widetilde{a}_n}} >$$
$$\mathrm{e}^{-C_T a_n/2}\mathrm{e}^{-D\,\widetilde{a}_n \log a_n - q_0 \log(c\,\mathrm{e}\,p^{a+1}/q_0)}\mathrm{e}^{-\widetilde{a}_n \log(q_0 n(\log_2 a_n + 1))}.$$

By our assumptions $q_0 \lesssim \log^\beta n$ and $p \lesssim n^{q_0/(2\alpha+q_0)}$, the term $\mathrm{e}^{-q_0 \log(c\,\mathrm{e}\,p^{a+1}/q_0)}$ will safely be larger than $\mathrm{e}^{-d_1 n\varepsilon_n^2}$ for some $d_1 > 0$. We take all the remaining important terms in (10.3), aiming to show that (i) $\widetilde{a}_n \log(\widetilde{a}_n/\varepsilon_n^2)$, (ii) $a_n \|f_0\|_\infty^2$, (iii) $\widetilde{a}_n \log(q_0 n \log_2 a_n)$ and (iv) $\widetilde{a}_n \log \widetilde{a}_n$ are bounded by a constant multiple of $n\varepsilon_n^2$. From (8.6), we obtain $a_n \lesssim n^{q_0/(2\alpha+q_0)}$ under our assumption $2C_0 q_0 \lesssim \log^\beta n$. Then we can write

$$\widetilde{a}_n = a_n/2(\log_2 a_n + 1) \lesssim n^{q_0/(2\alpha+q_0)} \log n. \tag{10.4}$$

Using this bound, we verify that (i)-(iv) are bounded by a constant multiple of $n\varepsilon_n^2 = n^{\frac{q_0}{2\alpha+q_0}} \log^{2\beta} n$. First, note that

$$\widetilde{a}_n \log(\widetilde{a}_n/\varepsilon_n^2) \lesssim n^{\frac{q_0}{2\alpha+q_0}} \log n \log\left(n^{\frac{q_0}{2\alpha+q_0}} n^{\frac{2\alpha}{2\alpha+q_0}} \log^{1-2\beta} n\right).$$

This quantity is bounded by a multiple of $n^{\frac{q_0}{2\alpha+q_0}} \log^{2\beta} n$ when $\beta \geq 1$. Next, we can write $a_n \|f_0\|_\infty^2 \lesssim n^{\frac{q_0}{2\alpha+q_0}} \log^{2\beta} n$. Lastly, it follows from (10.4) that $\widetilde{a}_n \log \widetilde{a}_n \lesssim n^{q_0/(2\alpha+q_0)} \log^2 n$ and $\widetilde{a}_n \log(q_0 n \log_2 a_n) \lesssim n^{q_0/(2\alpha+q_0)} \log^2 n$. To sum up, there exists $d > 0$ such that $\Pi(f \in \mathcal{F}_\mathcal{E} : \|f - f_0\|_n \leq \varepsilon_n) > \mathrm{e}^{-d\, n\varepsilon_n^2}$ for $\beta \geq 1$.

10.3. *Condition* (2.3). The condition $\Pi(\mathcal{F}_\mathcal{E} \backslash \mathcal{F}_\mathcal{E}^n)\mathrm{e}^{(d+2)\, n\varepsilon_n^2} \to 0$ is verified similarly as in Section 9.3 and Section 8.4. For $\Pi(q > q_n)$, we use the bound from Section 8.4 with $q_n = \lfloor C_q n\varepsilon_n^2 / \log p \rfloor$ and a large enough constant C_q. For $\Pi\left((T, \boldsymbol{K}) : \sum_{t=1}^T K^t > z_n\right)$ we use the bound from Section 9.3 with $z_n = \lfloor C_z n\varepsilon_n^2 / \log n \rfloor$ and a large enough constant C_z.

References.

[1] T. W. Anderson. Some nonparametric multivariate procedures based on statistically equivalent blocks. *Multivariate Analysis*, 1:5–27, 1966.

[2] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[3] A. Bhattacharya, D. Pati, and D. Dunson. Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of Statistics*, 42:352–381, 2014.

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

[5] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

[6] G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

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

[8] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

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

[10] I. Castillo. Pólya tree posterior distributions on densities. *Annales de l'Institut Henri Poincaré (to appear)*, 2016.
[11] I. Castillo, J. Schmidt-Hieber, and A.W. van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43:1986–2018, 2015.
[12] I. Castillo and A.W. van der Vaart. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40:2069–2101, 2012.
[13] H. Chipman, E. I. George, and R. McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4:266–298, 2010.
[14] H. Chipman, E. I. George, and R. E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93:935–960, 1997.
[15] M. Coram and S. Lalley. Consistency of Bayes estimators of a binary regression function. *Annals of Statistics*, 34:1233–1269, 2006.
[16] D. Denison, B. Mallick, and A. Smith. A Bayesian CART algorithm. *Biometrika*, 85:363–377, 1998.
[17] K. Devroye and L. Györfi. Distribution-free exponential bounds on the l_1 error of partitioning estimates of a regression function. *In proceedings of the fourth Pannonian Symposium of Mathematical Statistics*, pages 67–76, 1985.
[18] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer series: Stochastic Modelling and Applied Probability, 1996.
[19] D. Donoho. CART and best-ortho-basis: a connection. *Annals of Statistics*, 25:1870–1911, 1997.
[20] S. Ghosal, J. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28:500–5311, 2000.
[21] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35:192–223, 2007.
[22] L. Gordon and R. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627, 1980.
[23] L. Gordon and R. Olshen. Almost sure consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 15:147–163, 1984.
[24] W. Govaerts and J. Pryce. A singular value inequality for block matrices. *Linear algebra and its applications*, 125:141–145, 1989.
[25] S. Kpotufe. *The curse of dimension in nonparametric regression.* University of California, San Diego, 2010.
[26] H. Lian. Consistency of Bayesian estimation of a step function. *Statistics & probability letters*, 77(1):19–24, 2007.
[27] A. R. Linero. Bayesian regression trees for high dimensional prediction and variable selection. To appear in the Journal of the American Statistical Association, 2016.
[28] L. Liu and W. H. Wong. Multivariate density estimation via adaptive partitioning (ii): Posterior concentration. arXiv preprint arXiv:1508.04812, 2015.
[29] L. Lu, H. Jiang, and W. Wong. Multivariate density estimation by Bayesian sequential partitioning. *Journal of the American Statistical Association*, 108:1402–1410, 2013.
[30] A. Nobel. Histogram regression estimation using data-dependent partitions. *Annals of Statistics*, 24:1084–1105, 1996.
[31] E. Scornet, G. Biau, and J.P. Vert. Consistency of random forests. *Annals of Statistics*, 43:1716–1741, 2015.
[32] C. Scricciolo. On rates of convergence for Bayesian density estimation. *Scandinavian Journal of Statistics*, 34:626–642, 2007.
[33] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Annals of Statistics*, 29:687–714, 2001.

[34] C. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

[35] C. Stone. An asymptotically optimal histogram selection rule. *In proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, pages 513–530, 1985.

[36] C. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1985.

[37] S. van der Pas and V. Ročková. Bayesian dyadic trees and histograms for regression. *Advances in Neural Information Processing Systems (accepted)*, pages 1–12, 2017.

[38] A. van der Vaart and J. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics*, 36:1435–1463, 2008.

[39] N. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 565–574. AUAI Press, 2009.

[40] S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[41] Y. Yang and S. Tokdar. Minimax-optimal nonparametric regression in high dimensions. *Annals of Statistics*, 43:652–674, 2015.