On Semi-parametric Inference for BART

Veronika Ročková [1]

Abstract

There has been a growing realization of the potential of Bayesian machine learning as a platform that can provide both flexible modeling, accurate predictions as well as coherent uncertainty statements. In particular, Bayesian Additive Regression Trees (BART) have emerged as one of today's most effective general approaches to predictive modeling under minimal assumptions. Statistical theoretical developments for machine learning have been mostly concerned with approximability or rates of estimation when recovering infinite dimensional objects (curves or densities). Despite the impressive array of available theoretical results, the literature has been largely silent about uncertainty quantification. In this work, we continue the theoretical investigation of BART initiated recently by (Ročková and van der Pas, 2017). We focus on statistical inference questions. In particular, we study the Bernstein-von Mises (BvM) phenomenon (i.e. asymptotic normality) for smooth linear functionals of the regression surface within the framework of non-parametric regression with fixed covariates. Our semi-parametric BvM results show that, beyond rate-optimal estimation, BART can be also used for valid statistical inference.

1. Introduction

With visible successes on a wide range of predictive tasks, the role of machine learning has become increasingly recognized across a wide array of application domains ranging from economics to electronic commerce. Bayesian approaches have been particularly appealing as they provide a structured approach to uncertainty assessment via hierarchical modeling. Uncertainty quantification for inference (hypothesis testing and confidence statements) is a fundamental goal of statistics that goes beyond mere prediction. This note studies the inferential potential of Bayesian Additive Regression Trees (BART) of (Chipman et al., 2010), one of the workhorses of Bayesian machine learning (Hahn et al., 2017; Hill, 2011; He et al., 2019; Bleich, 2014; Linero, 2016; Linero and Yang, 2017).

In particular, we study the Gaussian approximability of certain aspects of posterior distributions in non-parametric regression with trees/forest priors. Results of this type, initially due to Laplace (1810) but most commonly known as Bernstein-von Mises (BvM) theorems, imply that posterior-based inference asymptotically coincides with the one based on traditional frequentist $1/\sqrt{n}$-consistent estimators. In this vein, BvM theorems provide a rigorous frequentist justification of Bayesian inference. The main thrust of this work is to understand the extent to which this phenomenon holds for various incarnations of BART.

In simple words, the BvM phenomenon occurs when, as the number of observations increases, the posterior distribution has approximately the shape of a Gaussian distribution centered at an efficient estimator of the parameter of interest. Moreover, the posterior credible sets, i.e. regions with prescribed posterior probability, are then *also* confidence regions with the same asymptotic coverage. This property has important practical implications in the sense that constructing confidence regions via variation in MCMC draws is relatively straightforward compared to direct constructions. Under fairly mild assumptions, BvM statements can be expected to hold in regular finite-dimensional models (van der Vaart, 2000).

Unfortunately, the frequentist theory on asymptotic normality does not generalize fully to semi- and non-parametric estimation problems (Bickel and Kleijn, 2012). Freedman initiated the discussion on the consequences of unwisely chosen priors in the 1960's, providing a negative BvM result in a basic ℓ_2-sequence Gaussian conjugate setting. The warnings against seemingly innocuous priors that may lead to posterior inconsistency were then reiterated many times in the literature, including (Cox, 1993) and (Diaconis and Freedman, 1998; 1986). Other counterexamples and anomalies of the BvM behavior in infinite-dimensional

[1]Booth School of Business, University of Chicago. Correspondence to: Veronika Ročková <Veronika.Rockova@ChicagoBooth.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

problems can be found in (Johnstone, 2010) and (Leahu, 2011). While, as pointed out by (Bickel and Kleijn, 1999), analogues of the BvM property for infinite-dimensional parameters are not immediately obvious, rigorous notions of non-parametric BvM's have been introduced in several pioneering works (Leahu, 2011; Ghosal, 2000; Castillo and Nickl, 2014; 2013).

Unwisely chosen priors leave room for unintended consequences also in semi-parametric contexts (Rivoirard and Rousseau, 2012). Castillo (2012a) provided an interesting counterexample where the posterior does not display the BvM behavior due to a non-vanishing bias in the centering of the posterior distribution. Various researchers have nevertheless documented instances of the BvM limit in semi-parametric models (a) when the parameter can be separated into a finite-dimensional parameter of interest and an infinite-dimensional nuisance parameter (Castillo, 2012b; Shen, 2002; Bickel and Kleijn, 2012; Cheng and Kosorok, 2008; Johnstone, 2010; Jonge and van Zanten, 2013), and (b) when the parameter of interest is a functional of the infinite-dimensional parameter (Rivoirard and Rousseau, 2012; Castillo and Rousseau, 2015). In this work, we focus on the latter class of semi-parametric BvM's and study the asymptotic behavior of smooth linear functionals of the regression function.

We consider the standard non-parametric regression setup, where a vector of responses $\boldsymbol{Y}^{(n)} = (Y_1, \dots, Y_n)'$ is linked to fixed (rescaled) predictors $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})' \in [0,1]^p$ for each $1 \leq i \leq n$ through

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

where f_0 is an unknown α-Hölder continuous function on a unit cube $[0,1]^p$. The true generative model giving rise to (1) will be denoted with $\mathbb{P}_0^n$.

Each model is parametrized by $f \in \mathcal{F}$, where $\mathcal{F}$ is an infinite-dimensional set of possibilities of f_0. Let $\Psi : \mathcal{F} \to \mathbb{R}$ be a measurable functional of interest and let Π be a probability distribution on $\mathcal{F}$. Given observations $\boldsymbol{Y}^{(n)}$ from (1), we study the asymptotic behavior of the posterior distribution of $\Psi(f)$, denoted with $\Pi[\Psi(f) \mid \boldsymbol{Y}^{(n)}]$. Let $\mathcal{N}(0, V)$ denote the centered normal law with a covariance matrix V. In simple words, we want to show that under the Bayesian CART or BART priors on $\mathcal{F}$, the *posterior distribution satisfies the BvM-type property* in the sense that

$$\Pi[\sqrt{n}\left(\Psi(f) - \widehat{\Psi}\right) \mid \boldsymbol{Y}^{(n)}] \rightsquigarrow N(0, V) \tag{2}$$

as $n \to \infty$ in $\mathbb{P}_0^n$-probability, where $\widehat{\Psi}$ is a random centering point estimator and where $\rightsquigarrow$ stands for weak convergence. We make this statement more precise in Section 2

Castillo and Rousseau (2015) provide general conditions on the model and on the functional Ψ to guarantee that (2) holds. These conditions describe how the choice of the prior influences the bias $\widehat{\Psi}$ and variance V. Building on this contribution, we provide (a) tailored statements for various incarnations of tree/forest priors that have occurred in the literature, (b) extend the considerations to *adaptive* priors under self-similarity.

1.1. Notation

The notation $\lesssim$ will be used to denote inequality up to a constant, $a \asymp b$ denotes $a \lesssim b$ and $b \lesssim a$ and $a \vee b$ denotes $\max\{a, b\}$. 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 With $\|\cdot\|_2$ we denote the standard Euclidean norm and with $\lambda_{min}(\Sigma)$ and $\lambda_{max}(\Sigma)$ the extremal eigenvalues of a matrix Σ. The set of α-Hölder continuous functions (i.e. Hölder smooth with $0 < \alpha \leq 1$) on $[0,1]^p$ will be denoted with $\mathcal{H}_p^\alpha$.

2. Rudiments

Before delving into our development, we first make precise the definition of asymptotic normality.

Definition 2.1. *We say that the posterior distribution for the functional $\Psi(f)$ is asymptotically normal with centering Ψ_n and variance V if, for β the bounded Lipschitz metric for weak convergence, and the real-valued mapping $\tau_n : f \to \sqrt{n}[\Psi(f) - \Psi_n]$, it holds that*

$$\beta(\Pi[\cdot \mid \boldsymbol{Y}^{(n)}] \circ \tau_n^{-1}, \mathcal{N}(0, V)) \to 0 \tag{3}$$

in $\mathbb{P}_0^n$ probability as $n \to \infty$, which we denote as $\Pi[\cdot \mid \boldsymbol{Y}^{(n)}] \circ \tau_n^{-1} \rightsquigarrow \mathcal{N}(0, V)$.

When efficiency theory at the rate $\sqrt{n}$ is available, we say that the posterior distribution satisfies the BvM theorem if (3) holds with $\Psi_n = \widehat{\Psi} + o_P(1/\sqrt{n})$ for $\widehat{\Psi}$ a linear efficient estimator of $\Psi(f_0)$. The proof of such a statement typically requires of a few steps (a) a semi-local step where one proves that the posterior distribution concentrates on certain sets and (b) a strictly local study on these sets, which can be carried out under the LAN (local asymptotic normality) conditions. Denoting the log-likelihood with

$$\ell_n(f) = \frac{n}{2}\log 2\pi - \sum_{i=1}^{n} \frac{[Y_i - f(\boldsymbol{x}_i)]^2}{2},$$

we define the log-likelihood ratio $\Delta_n(f) = \ell_n(f) - \ell_n(f_0)$ and express it as a sum of a quadratic term and a stochastic term via the LAN expansion as follows

$$\Delta_n(f) = -\frac{n}{2}\|f - f_0\|_L^2 + \sqrt{n}W_n(f - f_0), \tag{4}$$

where

$$\|f - f_0\|_L^2 = \frac{1}{n}\sum_{i=1}^{n}[f_0(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)]^2,$$
$$W_n(f - f_0) = \langle f - f_0, \sqrt{n}\,\boldsymbol{\varepsilon}\rangle_L = \frac{1}{n}\sum_{i=1}^{n}\sqrt{n}\,\varepsilon_i\,[f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)].$$

Recall that the phrase "semi-parametric" here refers to the problem of estimating functionals in an infinite-dimensional model rather than Euclidean parameters in the presence of infinite-dimensional nuisance parameters. In this paper, we consider the smooth linear functional

$$\Psi(f) = \frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{x}_i) f(\boldsymbol{x}_i) \tag{5}$$

for some smooth uniformly bounded γ-Hölder continuous function $a(\cdot)$, i.e. $\|a\|_\infty < C_a$ and $a \in \mathcal{H}_p^\gamma$ for some $0 < \gamma \leq 1$. Were $\gamma > 1$, Hölder continuity would imply that $a(\cdot)$ is a constant function and (5) boils down to a constant multiple of the average regression surface evaluated at fixed design points. This quantity is actually of independent interest and has been studied in a different setup by (Ray and van der Vaart, 2018) who focus on the posterior distribution of the "half average treatment effect estimator" (the mean regression surface) in the presence of missing data and random covariates. Our results can be extended to this scenario.

3. Tree and Forest Priors

Regression trees provide a piecewise constant reconstruction of the regression surface f_0, where the pieces correspond to terminal nodes of recursive partitioning (Donoho, 1997). Before introducing the tree function classes, we need to define the notion of tree-shaped partitions.

Starting from a parent node $[0,1]^p$, a binary tree partition is obtained by successively applying a splitting rule on a chosen internal node. Each such internal node is divided into two daughter cells with a split along one of the p coordinates at a chosen observed value. These daughter cells define two new rectangular subregions of $[0,1]^p$, which can be split further (to become internal nodes) or end up being terminal nodes. The terminal cells after $K-1$ splits then yield a tree-shaped partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ consisting of boxes $\Omega_k \subset [0,1]^p$. We denote with $\mathcal{V}^K$ the set of all tree-shaped partitions that can be obtained by recursively splitting $K-1$ times with each split made at one of the observed values $\mathcal{X} \equiv \{\boldsymbol{x}_i\}_{i=1}^p$.

The tree functions can be then written as

$$f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{x} \in \Omega_k)\beta_k, \tag{6}$$

where $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}^K$ and where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)' \in \mathbb{R}^K$ is a vector of jump sizes. Solitary trees are not as good performers as ensembles of trees/forests (Breiman, 2001; Chipman et al., 2010). The forest mapping underpinning the BART method of (Chipman et al., 2010) is the following sum-of-trees model indexed by a collection of T tree-shaped partitions $\mathcal{E} = [\mathcal{T}^1, \ldots, \mathcal{T}^T]$ and heights $\mathcal{B} = [\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^T]$:

$$f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) = \sum_{t=1}^{T} f_{\mathcal{T}^t,\boldsymbol{\beta}^t}(\boldsymbol{x}) \quad \text{for } \mathcal{T}^t \in \mathcal{V}^{K^t} \text{ and } \boldsymbol{\beta}^t \in \mathbb{R}^{K^t}. \tag{7}$$

The prior distribution is assigned over the class of forests

$$\mathcal{F} = \{f_{\mathcal{E},\mathcal{B}}(\boldsymbol{x}) \text{ as in (7) for some } \mathcal{E} \text{ and } \mathcal{B} \text{ and } T \in \mathbb{N}\}$$

in a hierarchical manner. One first assigns a prior distribution over T (or sets it equal to some value) and then a prior over the tree-shaped partitions $\mathcal{T}^t$ as well as heights $\boldsymbol{\beta}^t$ for $1 \leq t \leq T$.

3.1. Tree Partition Priors $\pi(\mathcal{T})$

In 1998, there were two Bayesian CART developments that surfaced independently of each other: Chipman et al. (1997) and Denison et al. (1998). Albeit related, the two methods differ in terms of the proposed tree partition prior $\pi(\mathcal{T})$.

The prior of Denison et al. (1998) is equalitarian in the sense that trees with the same number of leaves are a-priori equally likely, regardless of their shape. To prioritize smaller trees (that do not overfit), one assigns a complexity prior $\pi(K)$ on the tree size (i.e. the number of bottom nodes) K. They considered the Poisson distribution

$$\pi(K) = \frac{\lambda^K}{(\mathrm{e}^\lambda - 1)K!},\ K = 1, 2, \ldots, \quad \text{for some} \quad \lambda > 0. \tag{8}$$

Given the tree size K, one assigns a uniform prior over tree topologies

$$\pi(\mathcal{T} \mid K) = \frac{\mathbb{I}\left(\mathcal{T} \in \mathcal{V}^K\right)}{\mid \mathcal{V}^K \mid}, \tag{9}$$

where $\mid \mathcal{V}^K \mid$ is the cardinality of $\mathcal{V}^K$. This prior can be straightforwardly implemented using Metropolis-Hastings strategies (Denison et al., 1998) and was studied theoretically by Ročková and van der Pas (2017).

The Bayesian CART prior of Chipman et al. (1997), on the other hand, specifies the prior implicitly as a tree-generating Galton-Watson (GW) process. This process provides a mathematical representation of an evolving population of individuals who reproduce and die subject to laws of chance. The tree growing process is characterized as follows. Starting with a root node $\Omega_{00} = [0,1]^p$, one decides to split each node Ω_{lk} into two children by flipping a coin. We are tacitly using the two-index numbering of nodes (l,k), where

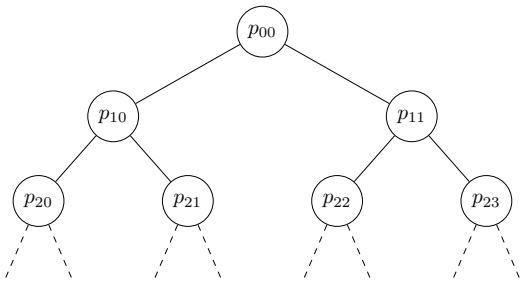

Figure 1: A binary tree of prior cut probabilities in Bayesian CART by Chipman et al. (1998).

l corresponds to the tree layer and k denotes the $(k+1)^{st}$ left-most node at the l^{th} layer. In order to prevent the trees from growing indefinitely, the success probability decays with l, making bushy trees far less likely. Let us denote with $\gamma_{lk} \in \{0, 1\}$ the binary splitting indicator for whether or not a node Ω_{lk} was divided. Chipman et al. (1997) assume

$$\mathbb{P}(\gamma_{lk} = 1) = p_{lk} = \frac{\alpha}{(1+l)^{\delta}} \tag{10}$$

for some $\alpha \in (0, 1)$ and $\delta \geq 0$. A plot of the hierarchically organized split probabilities is in Figure 1. Ročková and Saha (2019) propose a modification $p_{lk} = \alpha^l$ for some $1/n < \alpha < 1$, which yields optimal posterior concentration in the ℓ_2 sense. If the node Ω_{lk} splits (i.e. $\gamma_{lk} = 1$), one chooses a splitting rule by first picking a variable j uniformly from available directions $\{1, \ldots, p\}$ and then picking a split point c uniformly from available data values $x_{1j}, \ldots, x_{nj}$.

Unlike in the homogeneous case (where all γ_{lk}'s are iid), (10) defines a *heterogeneous* GW process where the offspring distribution is allowed to vary from generation to generation, i.e. the variables γ_{lk} are independent but *non-identical*.

3.2. Priors on Jump Sizes $\pi(\boldsymbol{\beta} \mid K)$

After partitioning the predictor space into K nested rectangular cells, one needs to assign a prior on the presumed level of the outcome. Throughout this work, we denote with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ the vector of jump sizes associated with K partitioning cells. Both Chipman et al. (1997) and Denison et al. (1998) assumed an independent product of Gaussians $\pi(\boldsymbol{\beta}) \sim \prod_{k=1}^{K} \phi(\beta_k;\ \sigma^2)$ for some $\sigma^2 > 0$. Chipman et al. (2000) argue, however, that the simple independence prior on the bottom leaf coefficients β_k may not provide enough structure. They claim that values β_k that correspond to nearby cells in the predictor space should be more similar so that the prior incorporates local smoothness. They suggest a prior on bottom leaves that aggregates priors on the ancestral internal nodes and, in this way, induces correlation among neighboring cells. Motivated by these considerations, here we allow for general correlation structures by assuming a multivariate Gaussian prior

$$\pi(\boldsymbol{\beta} \mid K) \sim \mathcal{N}_K(\mathbf{0}, \Sigma), \tag{11}$$

with $\lambda_{min}(\Sigma) > c > 0$ and $\lambda_{max}(\Sigma) \lesssim n$ and where Σ is some $K \times K$ positive definite matrix. We also consider an independent product of Laplace priors (which was not yet studied in this context)

$$\pi(\boldsymbol{\beta} \mid K) = \prod_{k=1}^{K} \psi(\beta_k;\ \lambda), \tag{12}$$

where $\psi(\beta;\ \lambda) = \lambda/2\, \mathrm{e}^{-\lambda|\beta|}$ for some $\lambda > 0$.

4. Simple One-dimensional Scenario

To set the stage for our development, we start with a simple toy scenario where (a) $p = 1$, (b) K is regarded as fixed, and (c) when there is *only one* partition $\mathcal{T} = \{\Omega_k\}_{k=1}^{K}$ consisting of K *equivalent blocks*. The equivalent blocks partition $\mathcal{T} = \{\Omega_k\}_{k=1}^{K}$ (Anderson, 1966) comprises K intervals Ω_k with roughly equal number of observations in them, i.e. $\mu(\Omega_k) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\boldsymbol{x}_i \in \Omega_k) \asymp 1/K$. For the sake of simplicity, we will also assume that the data points lie on a regular grid, where $x_i = i/n$ for $1 \leq i \leq n$ in which case the intervals Ω_k have also roughly equal lengths. This setup was studied previously by van der Pas and Ročková (2017) in the study of regression histograms. We relax this assumption in the next section. We denote the class of approximating functions as

$$\mathcal{F}[K] = \left\{ f_{\boldsymbol{\beta}}^{K} : [0,1]^p \to \mathbb{R};\ f_{\boldsymbol{\beta}}^{K}(\boldsymbol{x}) = \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{x} \in \Omega_k)\beta_k \right\}, \tag{13}$$

where we have omitted the subscript $\mathcal{T}$ and highlighted the dependence on K. We denote with $\Pi^K(f)$ the prior distribution over $\mathcal{F}[K]$, obtained by assigning the prior (11) or (12). To further simplify the notation, we will drop the subscript $\boldsymbol{\beta}$ and write f^K when referring to the elements of $\mathcal{F}[K]$.

The aim is to study the posterior distribution of $\Psi(f^K)$ and to derive conditions under which

$$\Pi[\sqrt{n}(\Psi(f^K) - \Psi_n) \mid \boldsymbol{Y}^{(n)}] \tag{14}$$

converges to a normal distribution (in $\mathbb{P}_0^n$ probability) with mean zero and variance $V_0 = \|a\|_L^2$, where Ψ_n is a random centering, distributed according to a Gaussian distribution with mean $\Psi(f_0)$ and variance V_0.

Using the fact that convergence of Laplace transforms for all t in probability implies convergence in distribution in probability (Section 1 of Castillo and Rousseau (2015)) this BvM statement holds when $\forall t \in \mathbb{R}$

$$\mathbb{E}^{\Pi}[\,\mathrm{e}^{t\sqrt{n}(\Psi(f^K) - \widehat{\Psi}_K)} \mid \boldsymbol{Y}^{(n)}] \to \mathrm{e}^{\frac{t^2}{2}\|a\|_L^2}, \tag{15}$$

in $\mathbb{P}_0^n$ probability as $n \to \infty$, where $\widehat{\Psi}_K$ is a linear efficient estimator of $\Psi(f_0)$ such that

$$\sqrt{n}(\widehat{\Psi}_K - \Psi_n) = o_P(1).$$

In order to show (15), we first need to introduce some notation. Let a^K be the projection of a onto $\mathcal{F}[K]$, i.e.

$$a^K(\boldsymbol{x}) = \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{x} \in \Omega_k) a_k^K$$

with $a_k^K = \sum_{i=1}^n \frac{\mathbb{I}(\boldsymbol{x}_i \in \Omega_k)}{n} \frac{a(\boldsymbol{x}_i)}{\mu(\Omega_k)}$, where $\mu(\Omega_k)$ was defined above as $\mu(\Omega_k) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(\boldsymbol{x}_i \in \Omega_k)$. Similarly, we denote with $f_0^K = \sum_{k=1}^K \mathbb{I}(\boldsymbol{x} \in \Omega_k)\beta_k^K$ the projection of f_0 onto $\mathcal{F}[K]$ with jump sizes $\boldsymbol{\beta}^K = (\beta_1^K, \ldots, \beta_K^K)'$. Next, we define

$$\widehat{\Psi}_K = \Psi\left(f_0^K\right) + \frac{W_n(a^K)}{\sqrt{n}} \quad \text{and} \quad \Psi_n = \Psi\left(f_0\right) + \frac{W_n(a)}{\sqrt{n}}.$$

The following Theorem characterizes the BvM property when K is sufficiently large and when α is known.

Theorem 4.1. *Assume the model (1) with $p = 1$, where f_0 is endowed with a prior on $\mathcal{F}[K]$ in (13) induced by (11). Assume $f_0 \in \mathcal{H}_1^\alpha$ and $a \in \mathcal{H}_1^\gamma$ for $1/2 < \alpha \leq 1$ and $\gamma > 1/2$. With the choice $K = K_n = \lfloor (n/\log n)^{1/(2\alpha+1)} \rfloor$, we have*

$$\Pi\left(\sqrt{n}\left(\Psi(f^K) - \widehat{\Psi}_K\right) \mid \boldsymbol{Y}^{(n)}\right) \rightsquigarrow \mathcal{N}(0, \|a\|_L^2)$$

in $\mathbb{P}_0^n$-probability as $n \to \infty$.

Proof: Appendix Section 1.1.

Remark 4.1. *Theorem 4.1 applies to the so-called symmetric dyadic trees. In particular, when $n = 2^r$ for some $r > 0$, the equivalent blocks partition with $K = 2^L$ cells can be represented with a symmetric full binary tree which splits all the nodes at dyadic rationals up to a resolution L.*

Theorem 4.1 is related to Theorem 4.2 of Castillo and Rousseau (2015) for density estimation with non-adaptive histogram priors. The proof here also requires two key ingredients. The first one is the construction of a prior which does not change too much under the change of measures from f^K to f_t^K, where f_t^K is a step function with shifted heights $\beta_k^t = \beta_k - \frac{t a_k^K}{\sqrt{n}}$. This property holds for (correlated) Gaussian step heights and is safely satisfied by other prior distributions with heavier tails.

Remark 4.2. *Theorem 4.1 holds for Laplace product prior $\pi(\boldsymbol{\beta} \mid K) = \prod_{k=1}^K \psi(\beta_k, \lambda)$ (as shown in the Supplemental Material). It is interesting to note that under the Laplace prior, one can obviate the need for showing posterior concentration around a projection of f_0 onto $\mathcal{F}[K]$, which is needed for the Gaussian case.*

The second crucial ingredient (as in the Proposition 1 of Castillo and Rousseau (2015)) is the so-called *no bias condition*:

$$\sqrt{n}\langle a - a^K, f_0 - f_0^K \rangle_L = o(1). \tag{16}$$

This condition vaguely reads as follows: one should be able to approximate *simultaneously* $a(\cdot)$ and $f_0(\cdot)$ well enough using functions in the inferential class $\mathcal{F}[K]$. Using the Cauchy-Schwartz inequality and Lemma 3.2 of Ročková and van der Pas (2017), this condition will be satisfied when $\sqrt{n}K^{-(\alpha+\gamma)} \to 0$. Choosing $K = K_n = \lfloor (n/\log n)^{1/(2\alpha+1)} \rfloor$, (16) holds as long as $\gamma > 1/2$. Different choices of K_n, however, would imply different restrictions on α and γ. The no-bias condition thus enforces certain limitations on the regularities α and γ. This poses challenges for adaptive priors that only adapt to the regularity of f_0, which may not be necessarily similar to the regularity of a. This phenomenon has been coined as the *curse of adaptivity* (Castillo and Rousseau, 2015).

5. Overcoming the Curse of Adaptivity

The dependence of K_n on α makes the result in Theorem 4.1 somewhat theoretical. In practice, it is common to estimate the regularity from the data using, e.g., a hierarchical Bayes method, which treats both K and the partition $\mathcal{T}$ as unknown with a prior. This fully Bayes model brings us a step closer to the actual Bayesian CART and BART priors. Treating both K and $\mathcal{T}$ as random and assuming $T = 1$, the class of approximating functions now constitutes a union of shells

$$\mathcal{F} = \bigcup_{K=1}^{\infty} \mathcal{F}[K],$$

where each shell

$$\mathcal{F}[K] = \bigcup_{\mathcal{T} \in \mathcal{V}^K} \mathcal{F}[\mathcal{T}],$$

itself is a union of sets $\mathcal{F}[\mathcal{T}] = \{f_{\mathcal{T},\boldsymbol{\beta}}$ of the form (6) for some $\boldsymbol{\beta} \in \mathbb{R}^K\}$. The sets $\mathcal{F}[\mathcal{T}]$ collect all step functions that grow on the same tree partition $\mathcal{T} \in \mathcal{V}^K$.

As mentioned in Castillo and Rousseau (2015), obtaining BvM in the case of random K is case dependent. As the prior typically adapts to the regularity of f_0, the no-bias condition (16) may not be satisfied if the regularities of a and f_0 are too different. The adaptive prior can be detrimental in such scenarios, inducing a non-vanishing bias in the centering of the posterior distribution (see Castillo (2012a) or Rivoirard and Rousseau (2012)). Roughly speaking, one needs to make sure that the prior supports large enough K values and sufficiently regular partitions $\mathcal{T}$ so that f_0 and a can be both safely approximated. To ensure this behavior, we enforce a signal strength assumption through

self-similarity requiring that the function f_0 "does not appear smoother than it actually is" (Gine and Nickl, 2015). Such qualitative assumptions are natural and necessary for obtaining adaptation properties of confidence bands (Picard and Tribouley, 2000; Bull, 2012; Gine and Nickl, 2010; Nickl and Szabo, 2016; Ray, 2017).

5.1. Self-similarity

Various self-similarity conditions have been considered in various estimation settings, including the supremum-norm loss (Bull, 2012; Gine and Nickl, 2010; 2011) as well as the ℓ_2 loss (Nickl and Szabo, 2016; Szabo et al., 2015). In the multi-scale analysis, the term self-similar coins "desirable truths" f_0 that are not so difficult to estimate since their regularity looks similar at small and large scales. Gine and Nickl (2010) argue that such self-similarity is a reasonable modeling assumption and Bull (2012) shows the set of self-dissimilar Hölder functions (in the ℓ_∞ sense) is negligible. Szabo et al. (2015) provided an ℓ_2-style self-similarity restriction on a Sobolev parameter space. Nickl and Szabo (2016) weakened this condition and showed that it "cannot be improved upon" and that the statistical complexity of the estimation problem does not decrease quantitatively under self-similarity in Sobolev spaces.

We consider a related notion of ℓ_2 self-similar classes within the context of fixed-design regression as opposed to the asymptotically equivalent white noise model. To this end, let us first formalize the notion of the cell size in terms of the local spread of the data and introduce the partition *diameter* (Verma et al., 2009; Ročková and van der Pas, 2017).

Definition 5.1. *(Diameter) Denote by $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ a partition of $[0,1]^p$ and by $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ a collection of data points in $[0,1]^p$. We define a diameter of Ω_k as*

$$\mathrm{diam}\,(\Omega_k) = \max_{\boldsymbol{x},\boldsymbol{y} \in \Omega_k \cap \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

and with $\mathrm{diam}\,(\mathcal{T}) = \sqrt{\sum_{k=1}^K \mu(\Omega_k)\mathrm{diam}^2\,(\Omega_k)}$ we define a diameter of the entire partition $\mathcal{T}$ where $\mu(\Omega_k) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(\boldsymbol{x}_i \in \Omega_k) = n(\Omega_k)/n$.

Here, we do not require that the design points are strictly on a grid as long as they are regular according to Definition 3.3 of Ročková and van der Pas (2017). We define regular datasets below. First we introduce the notion of the k-d tree (Bentley, 1975). Such a partition $\widehat{\mathcal{T}}$ is constructed by cycling over coordinate directions in $\mathcal{S} = \{1, \ldots, p\}$, 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 $\Omega_k^\star$, will be split at a median of the point set (along the j^{th} axis). This split will pass $\lfloor \mu(\Omega_k^\star)n/2 \rfloor$ and $\lceil \mu(\Omega_k^\star)n/2 \rceil$ 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 $K = 2^{s\,|\mathcal{S}|}$.

We now define regular datasets in terms of the k-d tree partition.

Definition 5.2. *Denote by $\widehat{\mathcal{T}} = \{\widehat{\Omega}_k\}_{k=1}^K \in \mathcal{V}^K$ the k-d tree where $K = 2^{s\,p}$. We say that a dataset $\mathcal{X} \equiv \{\boldsymbol{x}_i\}_{i=1}^n$ is regular if*

$$\max_{1 \le k \le K} \mathrm{diam}(\widehat{\Omega}_k) < M \sum_{k=1}^K \mu(\widehat{\Omega}_k)\mathrm{diam}(\widehat{\Omega}_k) \qquad (17)$$

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.

Definition 5.3. *We say that the function $f_0 \in \mathcal{H}_p^\alpha$ is self-similar, if there exists constant $M > 0$ and $D > 0$ such that*

$$\|f_{\mathcal{T}}^0 - f_0\|_L^2 \ge M\mathrm{diam}(\mathcal{T})^{2\alpha} \qquad (18)$$

for all $\mathcal{T} \in \mathcal{V}^K$ such that $\mathrm{diam}(\mathcal{T}) \le D$ where $f_{\mathcal{T}}^0$ is the the $\|\cdot\|_L$ projection of f_0 onto $\mathcal{F}[\mathcal{T}]$.

We can relate the assumption (18) to the notion of self-similarity in the Remark 3.4 of Szabo et al. (2015). To see this connection, assume for now the equivalent block partition $\mathcal{T}$ from Section 4, whose diameter $\mathrm{diam}(\mathcal{T})$ is roughly $1/K$ when the design points lie on a regular grid. The study of regression histograms with $K = 2^L$ equivalent blocks under a fixed regular design is statistically equivalent to the multi-scale analysis of Haar wavelet coefficients up to the resolution $L-1$ in the white noise model. The projected model onto the Haar basis can be written as

$$Y_{lk} = f_{lk}^0 + \frac{1}{\sqrt{n}}\varepsilon_{lk} \quad \text{for} \quad 0 \le l < L \quad \text{and} \quad 0 \le k < 2^l,$$

where $\varepsilon_{lk} \overset{iid}{\sim} \mathcal{N}(0,1)$ and where f_{lk}^0 are the wavelet coefficients indexed by the resolution level l and the shift index k. The speed of decay of f_{lk}^0 determines the statistical properties of f_0, where α-Hölder continuous functions satisfy $|\, f_{lk}^0 \,| \lesssim 2^{-l(\alpha+1/2)}$. Assuming the equivalent blocks partition $\mathcal{T}$ with $K = 2^L$ blocks, the condition in the Remark 3.4 of Szabo et al. (2015) writes as follows: there exists $K_0 \in \mathbb{N}$ such that $\forall K \ge K_0$ we have $\int_0^1 |\, f_{\mathcal{T}}^0(x) - f_0(x)\,|^2 \mathrm{d}x = \sum_{l \ge L} \sum_{0 \le k < 2^l} (f_{lk}^0)^2 \ge MK^{-2\alpha} \asymp \mathrm{diam}(\mathcal{T})^{2\alpha}$. The first equality above stems from the orthogonality of the Haar bases. In this vein, (18) can be regarded a generalization of this condition to imbalanced partitions and fixed design under the norm $\|\cdot\|_L$.

To get even more insight into (18) in fixed design regression, we take a closer look at the approximation gap. We have

$\|f_0 - f_{\mathcal{T}}^0\|_L^2 = \sum_{k=1}^K \mu(\Omega_k)\mathsf{Var}\left[f_0 \mid \Omega_k\right]$, where

$$\mathsf{Var}\left[f_0 \mid \Omega_k\right] = \frac{1}{n(\Omega_k)} \sum_{\boldsymbol{x}_i \in \Omega_k} \left(f_0(\boldsymbol{x}_i) - \frac{1}{n(\Omega_k)} \sum_{\boldsymbol{x}_i \in \Omega_k} f_0(\boldsymbol{x}_i) \right)^2$$

is the local variability of the function f_0 inside Ω_k. The function f_0 will be self-similar when the variability inside each cell Ω_k's is large enough to be detectable, i.e.

$$\inf_{1 \le k \le K} \mathsf{Var}\left[f_0 \mid \Omega_k\right] > M \mathrm{diam}^{2\alpha}(\mathcal{T})$$

for all $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}^K$ such that $\mathrm{diam}(\mathcal{T}) \le D$ for some $D > 0$. From the definition of the partition diameter, it turns out that this will be satisfied when $\mathsf{Var}\left[f_0 \mid \Omega_k\right] > \mathrm{diam}^{2\alpha}(\Omega_k)$ for all $1 \le k \le K$ and $\mathcal{T} = \{\Omega_k\}_{k=1}^K \in \mathcal{V}^K$. Functions that are nearly constant in long intervals will not satisfy this requirement. The premise of self-similar functions is that their signal should be detectable with partitions $\mathcal{T}$ that undersmooth the truth. In addition, it follows from the proof of Lemma 3.2 of Ročková and van der Pas (2017) that $\|f_{\mathcal{T}}^0 - f_0\|_L^2 \lesssim \mathrm{diam}(\mathcal{T})^{2\alpha}$. The lower bound in (18) thus matches the upper bound making the approximation error behave similarly across partitions with similar diameters, essentially identifying the smoothness. Based on these considerations, we introduce the notion of *regular* partitions that are not too rough in the sense that their diameters shrink sufficiently fast.

Definition 5.4. *For some $M > 0$ and for some arbitrarily slowly increasing sequence $M_n \to \infty$, we denote*

$$d_n(\alpha) \equiv (M_n/M)^{1/\alpha} n^{-1/(2\alpha+p)} \log^{1/2\alpha} n. \tag{19}$$

A tree partition $\mathcal{T} \in \mathcal{V}^K$ is said to be n-regular for a given $n \in \mathbb{N}$ if

$$\mathrm{diam}(\mathcal{T}) \le d_n(\alpha).$$

We denote the subset of all n-regular partitions with $\mathcal{R}_n$.

The following Lemma states that, when f_0 is self-similar, the posterior concentrates on partitions that are not too complex or irregular.

Lemma 5.1. *Assume that $f_0 \in \mathcal{H}_p^\alpha$ is self-similar, $p \lesssim \sqrt{\log n}$ and that the design $\mathcal{X} \equiv \{\boldsymbol{x}_i\}_{i=1}^n$ is regular. Under the Bayesian CART prior ((8) and (9) or (10)) with Gaussian or Laplace step heights ((11) or (12)) we have*

$$\Pi\left(\{\mathcal{T} \notin \mathcal{R}_n\} \cup \{\mathcal{T} \in \mathcal{V}^K : K > K_n\} \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

in $\mathbb{P}_0^n$-probability as $n, p \to \infty$, where $\mathcal{R}_n$ are all regular partitions and $K_n = M_2 \lfloor n\varepsilon_n^2 / \log n \rfloor \asymp (n/\log n)^{p/(2\alpha+p)}$ for some $M_2 > 0$.

Proof: Appendix Section 1.5.

5.2. Adaptive BvM for Smooth Functionals when $p = 1$

It is known that signal-strength conditions enforced through self-similarity allow for the construction of honest adaptive credible balls (Gine and Nickl, 2010). Our notion of self-similarity is sufficient for obtaining the adaptive semi-parametric BvM phenomenon for smooth linear functionals. Denote with

$$\mathcal{R}(K_n) = \{\mathcal{T} \in \mathcal{R}_n \cap \mathcal{V}^K \quad \text{for} \quad K \le K_n\}.$$

Lemma 5.1 shows that the posterior concentrates on this set so that we can perform the analysis locally. For any $\mathcal{T} \in \mathcal{R}(K_n)$, we write

$$\widehat{\Psi}_{\mathcal{T}} = \Psi(f_0^{\mathcal{T}}) + \frac{W_n(a_{\mathcal{T}})}{\sqrt{n}},$$

where $f_0^{\mathcal{T}}$ and $a_{\mathcal{T}}$ are the $\|\cdot\|_L$ projections onto $\mathcal{F}[\mathcal{T}]$. Under the adaptive prior (treating the partitions as random with a prior) the posterior is asymptotically close to a *mixture* of normals indexed by $\mathcal{T} \in \mathcal{R}(K_n)$ with weights $\pi(\mathcal{T} \mid \boldsymbol{Y}^{(n)}, \mathcal{R}(K_n))$. When

$$\max_{\mathcal{T} \in \mathcal{R}(K_n)} \mid \|a_{\mathcal{T}}\|_L - \|a\|_L \mid = o(1) \tag{20}$$

and

$$\max_{\mathcal{T} \in \mathcal{R}(K_n)} \sqrt{n}(\Psi_n - \widehat{\Psi}_{\mathcal{T}}) = o_P(1) \tag{21}$$

this mixture boils down to the target law $\mathcal{N}(0, \|a\|_L^2)$. The first condition (20) holds owing to the fact that $\|a - a_{\mathcal{T}}\|_L \lesssim d_n(\alpha)^\gamma \to 0$ (Lemma 3.2 of Rockova and van der Pas (2017)). The second condition (21) will be satisfied as long as

$$\sqrt{n}\, d_n(\alpha)^{\alpha+\gamma} \to 0. \tag{22}$$

For $\mathcal{T} \in \mathcal{R}(K_n)$ and assuming $p = 1$, we have for $M_n \lesssim \sqrt{\log n}$

$$\sqrt{n}\, d_n(\alpha)^{\alpha+\gamma} \lesssim n^{1/2 - \frac{\alpha+\gamma}{2\alpha+1}} (\log n)^{\frac{\alpha+\gamma}{\alpha}} \to 0$$

for $\gamma > 1/2$. We can now state an adaptive variant of Theorem 4.1 for random partitions.

Theorem 5.1. *Assume model (1) with $p = 1$, where $f_0 \in \mathcal{H}_1^\alpha$ and $a \in \mathcal{H}_1^\gamma$ for $1/2 < \alpha \le 1$ and $\gamma > 1/2$. Assume that f_0 is self-similar. Under the Bayesian CART prior ((8) and (9) or (10)) with Gaussian or Laplace step heights ((11) or (12)), we have*

$$\Pi\left(\sqrt{n}\left(\Psi(f_{\mathcal{T},\boldsymbol{\beta}}) - \widehat{\Psi}_{\mathcal{T}}\right) \mid \boldsymbol{Y}^{(n)}\right) \rightsquigarrow \mathcal{N}(0, \|a\|_L^2)$$

in $\mathbb{P}_0^n$-probability as $n \to \infty$.

Proof: Appendix Section 1.3.

Theorem 5.1 can be regarded as an adaptive extension of the regular density histogram result of Castillo and Rousseau (2015). Here, we instead focus on irregular and adaptive regression histograms underpinned by tree priors and treat both K and $\mathcal{T}$ as random under self-similarity. The change of measure argument is performed locally for each regular partition.

Theorem 5.1 can be extended to tree ensembles. The self-similarity assumption would be instead formulated in terms of a *global partition*, which is obtained by super-imposing all tree partitions inside $\mathcal{E}$ and by merging empty cells with one of their neighbors. Since tree ensembles also concentrate at near-minimax speed (Ročková and van der Pas, 2017; Ročková and Saha, 2019), one obtains that the posterior concentrates on regular ensembles (where the diameter is small). The analysis is then performed locally on regular ensembles in the same spirit as for single trees.

5.3. Average Regression Surface when $p > 1$

One of the main limitations of tree/forest methods is that they cannot estimate optimally functions that are smoother than Lipschitz (Scricciolo, 2007). The reason for this limitation is that step functions are relatively rough; e.g. the approximation error of histograms for functions that are smoother than Lipschitz is at least of the order $1/K$, where K is the number of bins (Ročková and van der Pas, 2017). The number of steps required to approximate a smooth function well is thus too large, creating a costly bias-variance tradeoff. When $p > 1$, the no-bias condition (16) would be satisfied if $\gamma > p/2$ which, from the Hölder continuity, holds when $a(\boldsymbol{x}_i)$ is a constant function.

Focusing on the actual BART method when $p > 1$, we now rephrase Theorem 5.1 for the average regression functional (5) obtained with $a(\cdot) = 1$. When $a(\cdot)$ is a constant function, the no-bias condition (16) is automatically satisfied. Recall that the second requirement for BvM pertains to the shift of measures. It turns out that the Gaussian prior (11) may induce too much bias when the variance is too small (fixed as $n \to \infty$). We thereby deploy an additional assumption in the prior to make sure that the variance increases suitably with the number of steps K. For the BART prior on step heights $\boldsymbol{\beta}^t$ of each tree $\mathcal{T}^t \in \mathcal{E}$, we assume either a Gaussian prior $\boldsymbol{\beta}^t \sim \mathcal{N}_{K^t}(\mathbf{0}, K^t \times I_{K^t})$ or the Laplace prior with $\lambda_t \asymp 1/\sqrt{K^t}$. Having the variance scale with the number of steps is generic in the multi-scale analysis of Haar wavelets.

The following theorem is formulated for a few variants of the BART prior. This prior consists of (a) either fixed number of trees T (as recommended by (Chipman et al., 2010)), (b) the Galton-Watson prior (10) or the conditionally uniform tree prior (8) and (9), independently for each tree, and (c) the Gaussian prior $\boldsymbol{\beta}^t \sim \mathcal{N}_{K^t}(\mathbf{0}, K^t \times I_{K^t})$ or the Laplace prior with $\lambda_t \asymp 1/\sqrt{K^t}$, where K^t is the number of bottom nodes of a tree $\mathcal{T}^t$. Below, we denote with $\widehat{\Psi}_{\mathcal{E}} = \Psi(f^0_{\mathcal{E}}) + W_n(a)/\sqrt{n}$, where $f^0_{\mathcal{E}}$ is a projection of f_0 onto $\mathcal{F}[\mathcal{E}]$, a set of all forest mappings (7) supported on the tree ensemble $\mathcal{E}$.

Theorem 5.2. *Assume model (1) with $p \geq 1$, where $f_0 \in \mathcal{H}^\alpha_p$ is endowed with the BART prior (as stated above) and where $\log p \lesssim n$ and $1/2 \leq \alpha < 1$. Assume that $a(\cdot) = 1$ in (5). When the design $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ is regular, we have $\Pi\left(\sqrt{n}\left(\Psi(f_{\mathcal{E},\mathcal{B}}) - \widehat{\Psi}_{\mathcal{E}}\right) \mid \boldsymbol{Y}^{(n)}\right) \rightsquigarrow \mathcal{N}(0, \|a\|^2_L)$ in $\mathbb{P}^n_0$-probability as $n \to \infty$.*

Proof: Appendix Section 1.4.

6. Discussion

This paper focuses on the important problem of uncertainty quantification and inference for machine learning. We provide frequentist justifications for Bayesian inference with BART based on (smooth) linear functionals of the regression function. These results can be used for testing hypotheses pertaining to exceedance of (weighted) average level, i.e. $\sum_{i=1}^n a_i f_0(\boldsymbol{x}_i) > c$ for some $c \in \mathbb{R}$, or for causal inference (Hill, 2011; Hahn et al., 2017). Indeed, embedding our development within the missing data framework of Ray (2017) will provide asymptotic normality results for average treatment estimation. This work will be reported elsewhere.

Both Theorem 4.1 and Theorem 5.1 impose restrictions on α and γ to make sure that they are compatible. These theorems can be obtained without assuming $\alpha > 1/2$ when $a(\cdot)$ is constant. The self-similarity assumption (a very typical assumption for uncertainty quantification of densities and functions) makes it possible to identify smoothness so that α does not need to be known in Theorem 5.1. Variants of this assumption are pervasive in the literature on adaptive confidence sets construction. This assumption is, again, not needed when $a(\cdot)$ is constant.

We focus on semi-parametric BvM's for linear functionals of the infinite-dimensional regression function parameters. This semi-parametric setup already poses nontrivial challenges on hierarchical Bayes. We have reiterated and highlighted some of the challenges here and addressed them with self-similarity identification. Our results serve as a step towards obtaining fully non-parametric BvM for the actual function f_0, as opposed to just its low-dimensional summaries. These results will be reported in follow-up work. Finally, the limitation $p = 1$ for smooth functions $a(\cdot)$ is due to the fact that BART cannot optimally approximate functions smoother than Lipschitz. This can be overcome by considering smoother versions of BART (Linero and Yang, 2017).

References

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

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

Bickel, P. and B. Kleijn (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics 27*, 1119–1140.

Bickel, P. and B. Kleijn (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics 40*, 206–237.

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

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

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

Castillo, I. (2012a). Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya 74*, 194–221.

Castillo, I. (2012b). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields 152*, 53–99.

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

Castillo, I. and R. Nickl (2014). On the Bernstein–von Mises theorem for nonparametric Bayes proceduress. *The Annals of Statistics 27*, 1941–1969.

Castillo, I. and J. Rousseau (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics 43*, 2353–2383.

Castillo, I. and V. Ročková (2019). Multiscale analysis of BART. *ArXiv*.

Cheng, G. and M. R. Kosorok (2008). General frequentist properties of the posterior profile distribution. *The Annals of Statistics 36*, 1819–1853.

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

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

Chipman, H., E. I. George, and R. E. McCulloch (2000). Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing 10*, 17–24.

Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics 21*, 903–923.

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

Diaconis, P. and D. Freedman (1986). On consistency of Bayes estimates. *The Annals of Statistics 14*, 1–26.

Diaconis, P. and D. Freedman (1998). Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli 4*, 411–444.

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

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

Gine, E. and R. Nickl (2010). Confidence bands in density estimation. *The Annals of Statistics 38*, 1122–1170.

Gine, E. and R. Nickl (2011). On adaptive inference and confidence bands. *The Annals of Statistics 39*, 2383–2409.

Gine, E. and R. Nickl (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Hahn, R. and Murray, J. and Carvalho, C. (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Journal of Computational and Graphical Statistics 29*, 217–240.

He, J. and Yalov, S. and Hahn, R. XBART: Accelerated Bayesian Additive Regression Trees. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 89*, 217–240.

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 29*, 217–240.

Johnstone, I. (2010). High dimensional Bernstein–von Mises: Simple examples. In *Borrowing Strength: Theory Powering Applications?A Festschrift for Lawrence D. Brown*.

Jonge, R. and H. van Zanten (2013). Semiparametric Bernstein?von Mises for the error standard deviation. *Electronic Journal of Statistics 7*, 217–243.

Kleijn, B. and A. van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics 34*, 837–877.

Laplace, P. S. (1810). Memoire sur les formules qui sont fonctions de tres grands nombres et sur leurs applications aux probabilites. *Oeuvres de Laplace 12*, 301–349.

Leahu, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat 4*, 373–404.

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

Linero, A. and Yang, Y. (2017). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Association (to appear)*.

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

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

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

Ray, K. and A. van der Vaart (2018). Semiparametric Bayesian causal inference using Gaussian process priors. *ArXiv*.

Rivoirard, V. and J. Rousseau (2012). Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics 40*, 1489–1523.

Ročková, V. and E. Saha (2019). On theory for BART. 22^{nd} *International Conference on Artificial Intelligence and Statistics*.

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

Scricciolo, C. (2007). On rates of convergence for Bayesian density estimation. *Scandinavian Journal of Statistics 34*, 626–642.

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association 97*, 222–235.

Szabo, B., A. van der Vaart, and J. van Zanten (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics 43*, 1391–1428.

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

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

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