On Theory for BART

Veronika Ročková and Enakshi Saha
University of Chicago

Abstract

Ensemble learning is a statistical paradigm built on the premise that many weak learners can perform exceptionally well when deployed collectively. The BART method of Chipman et al. (2010) is a prominent example of *Bayesian* ensemble learning, where each learner is a tree. Due to its impressive performance, BART has received a lot of attention from practitioners. Despite its wide popularity, however, theoretical studies of BART have begun emerging only very recently. Laying down foundation for the theoretical analysis of Bayesian forests, Rockova and van der Pas (2017) showed optimal posterior concentration under *conditionally uniform tree priors.* These priors deviate from the actual priors implemented in BART. Here, we study the exact BART prior and propose a simple modification so that it *also* enjoys optimality properties. To this end, we dive into the branching processes theory. We obtain tail bounds for the distribution of total progeny under heterogeneous Galton-Watson (GW) processes using their connection to random walks. We conclude with a result stating optimal rate of convergence for BART.

1 Bayesian Machine Learning

Bayesian Machine Learning and Bayesian Nonparametrics share the same objective: increasing flexibility necessary to address very complex problems using a Bayesian approach with minimal subjective input. While the two fields can be, to some extent, regarded as synonymous, their emphasis is quite different. Bayesian non-parametrics has subsumed a theoretical field focused on studying frequentist properties of posterior objects in inifinite-dimensional parameter spaces. Bayesian machine learning, on the other hand, has been primarily concerned with developing scalable tools for computing such posterior objects. In this work, we bridge these two fields by providing theoretical insights into one of the workhorses of Bayesian machine learning, the BART method.

Bayesian Additive Regression Trees (BART) are one of the more widely used Bayesian prediction tools and their popularity continues to grow. Compared to its competitors (e.g. Gaussian processes, random forests or neural networks) BART requires considerably less tuning, while maintaining robust and relatively scalable performance (`BART` R package of McCulloch (2017), `bartMachine` R package of Bleich et al. [2014], top down particle filtering of Lakshminarayanan et al. [2013]). BART has been successfully deployed in many prediction tasks, often outperforming its competitors (see predictive comparisons on 42 data sets in Chipman et al. [2010]). More recently, its flexibility and stellar prediction has been capitalized on in causal inference tasks for heterogeneous/average treatment effect estimation (Hill [2011], Hahn et al. [2017] and references therein). BART has also served as a springboard for various incarnations and extensions including: Monotone BART (Chipman et al. [2016]), Heteroscedastic BART (Pratola et al. [2017]), treed Gaussian processes (Gramacy and Lee [2008]) and dynamic trees (Taddy et al. [2011]), to list a few. Related non-parametric constructions based on recursive partitioning have proliferated in the Bayesian machine learning community for modeling relational data (Mondrian process of Roy and Teh [2008], Mondrian forests (Lakshminarayanan et al. [2014]). In short, BART continues to have a decided impact on the field of Bayesian non-parametrics/machine learning.

Despite its widespread popularity, however, the theory has not caught up with its applications. First theoretical results were obtained only very recently. As a precursor to these developments, Coram and Lalley [2006] obtained a consistency result for Bayesian histograms in binary regression with a single predictor. van der Pas and Rockova [2017] provided a posterior concen-

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

tration result for Bayesian regression histograms in Gaussian non-parametric regression, also with one predictor. Rockova and van der Pas [2017] (further referred to as RP17) then extended their study to trees and forests in a high-dimensional setup where $p > n$ and where variable selection uncertainty is present. They obtained the first theoretical results for Bayesian CART, showing optimal posterior concentration (up to a log factor) around a ν-Hölder continuous regression function (i.e. Hölder smooth with smoothness $0 < \nu \leq 1$). Going further, they also show optimal performance for Bayesian forests, both in additive and non-additive regression. Linero and Yang [2017] obtained similar results for Bayesian ensembles, but for *fractional* posteriors (raised to a power). The proof of RP17, on the other hand, relies on a careful construction of sieves and applies to regular posteriors. Building on RP17, Linero and Yang [2017] subsequently obtained also results for actual posteriors. In addition, Linero and Yang [2017] do not study step functions (the essence of Bayesian CART and BART) but aggregated smooth kernels, allowing for $\nu > 1$. Liu et al. [2018] obtained model selection consistency results (for variable and regularity selection) for Bayesian forests.

Albeit related, the tree priors studied in RP17 are *not* the actual priors deployed in BART. Here, we develop new tools for the analysis of the actual BART prior and obtain parallel results to those in RP17. To begin, we dive into branching process theory to characterize aspects of the distribution on total progeny under heterogeneous Galton-Watson processes. Revisiting several useful facts about Galton-Watson processes, including their connection to random walks, we derive a new prior tail bound for the tree size under the BART prior. With our proving strategy, the actual prior of Chipman et al. [2010] *does not* appear to penalize large trees aggressively enough. We suggest a very simple modification of the prior by altering the splitting probability. With this minor change, the prior is shown to induce the right amount of regularization and optimal speed of posterior convergence.

The paper is structured as follows. Section 2 revisits trees and forests in the context of non-parametric regression and discusses the BART prior. Section 3 reviews the notion of posterior concentration. Section 4 discusses Galton Watson processes and their connection to Bayesian CART. Section 5 is concerned with tail bounds on total progeny. Section 6 and 7 describe prior and concentration properties of BART. Section 7 wraps up with a discussion.

2 The Appeal of Trees/Forests

The data setup under consideration consists of $Y_i \in \mathbb{R}$, a set of one-dimensional outputs, and $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})' \in [0,1]^p$, a set of high dimensional inputs for $1 \leq i \leq n$. Our statistical framework is nonparametric regression, which characterizes the input-output relationship through

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

where $f_0 : [0,1]^p \to \mathbb{R}$ is an unknown regression function. A regression tree can be used to reconstruct f_0 via a mapping $f_{\mathcal{T},\boldsymbol{\beta}} : [0,1]^p \to \mathbb{R}$ so that $f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) \approx f_0(\boldsymbol{x})$ for $\boldsymbol{x} \notin \{\boldsymbol{x}_i\}_{i=1}^n$. Each such mapping is essentially a step function

$$f_{\mathcal{T},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \beta_k \mathbb{I}(\boldsymbol{x} \in \Omega_k) \tag{2}$$

underpinned by a tree-shaped partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ and a vector of step heights $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)'$. The vector $\boldsymbol{\beta}$ represents quantitative guesses of the average outcome inside each cell. Each partition $\mathcal{T}$ consists of rectangles obtained by recursively applying a splitting rule (an axis-parallel bisection of the predictor space). We focus on *binary* tree partitions, where each internal node (box) is split into two children (formal definition below).

Definition 2.1. *(A Binary Tree Partition) A binary tree partition $\mathcal{T} = \{\Omega_k\}_{k=1}^K$ consists of K rectangular cells Ω_k obtained with $K-1$ successive recursive binary splits of the form $\{x_j \leq c\}$ vs $\{x_j > c\}$ for some $j \in \{1, \dots, p\}$, where the splitting value c is chosen from observed values $\{x_{ij}\}_{i=1}^n$.*

Partitioning is intended to increase within-node homogeneity of outcomes. In the traditional CART method (Breiman et al. [1984]), the tree is obtained by "greedy growing" (i.e. sequential optimization of some impurity criterion) until homogeneity cannot be substantially improved. The tree growing process is often followed by "optimal pruning" to increase generalizability. Prediction is then determined by terminal nodes of the pruned tree and takes the form either of a class level in classification problems, or the average of the response variable in least squares regression problems (Breiman et al. [1984]).

In tree *ensemble* learning, each constituent is designed to be a weak learner, addressing a slightly different aspect of the prediction problem. These trees are intended to be shallow and are woven into a forest mapping

$$f_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}) = \sum_{t=1}^{T} f_{\mathcal{T}_t,\boldsymbol{\beta}_t}(\boldsymbol{x}), \tag{3}$$

where each $f_{\mathcal{T}_t,\boldsymbol{\beta}_t}(\boldsymbol{x})$ is of the form (2), $\mathcal{E} = \{\mathcal{T}_1,\dots,\mathcal{T}_T\}$ is an ensemble of trees and $\boldsymbol{B} = \{\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_T\}'$ is a collection of jump sizes for the T trees. Random forests obtain each tree learner from a bootstrapped version of the data. Here, we consider a Bayesian variant, the BART method of Chipman et al. [2010], which relies on the posterior distribution over $f_{\mathcal{E},\boldsymbol{B}}$ to reconstruct the unknown regression function f_0.

2.1 Bayesian Trees and Forests

Bayesian CART was introduced as a Bayesian alternative to CART, where regularization/stabilization is obtained with a prior rather than with pruning (Chipman et al. [1998], Denison et al. [1998]). The prior distribution is assigned over a class of step functions

$$\mathcal{F} = \{f_{\mathcal{E},\boldsymbol{B}}(\boldsymbol{x}) \quad \text{of the form (3) for some } \mathcal{E} \text{ and } \boldsymbol{B}\}$$

in a hierarchical manner.

The BART prior by Chipman et al. [2010] assumes that the number of trees T is fixed. The authors recommend a default choice $T = 200$ which was seen to provide good results. Next, the tree components $(\mathcal{T}_t, \boldsymbol{\beta}_t)$ are a-priori independent of each other in the sense that

$$\pi(\mathcal{E},\boldsymbol{B}) = \prod_{t=1}^{T} \pi(\mathcal{T}_t)\pi(\boldsymbol{\beta}_t \mid \mathcal{T}_t), \tag{4}$$

where $\pi(\mathcal{T}_t)$ is the prior probability of a partition $\mathcal{T}_t$ and $\pi(\boldsymbol{\beta}_t \mid \mathcal{T}_t)$ is the prior distribution over the jump sizes.

2.1.1 Prior on Partitions $\pi(\mathcal{T})$

In BART and Bayesian CART of Chipman et al. [1998], the prior over trees is specified implicitly as a tree generating stochastic process, described as follows:

1. Start with a single leave (a root node) $[0,1]^p$.
2. Split a terminal node, say Ω_t, with a probability
$$p_{split}(\Omega_t) = \frac{\alpha}{(1+d(\Omega_t))^{\gamma}} \tag{5}$$
for some $\alpha \in (0,1)$ and $\gamma \geq 0$, where $d(\Omega_t)$ is the depth of the node Ω_t in the tree architecture.
3. If the node Ω_t splits, assign a splitting rule and create left and right children nodes. The splitting rule consists of picking a split variable j uniformly from available directions $\{1,\dots,p\}$ and picking a split point c uniformly from available data values $x_{1j},\dots,x_{nj}$. Non-uniform priors can also be used to favor splitting values that are thought to be more important. For example, splitting values can be given more weight towards the center and less weight towards the edges.

2.1.2 Prior on Step Heights $\pi(\boldsymbol{\beta} \mid \mathcal{T})$

Given a tree partition $\mathcal{T}_t$ with K_t steps, we consider iid Gaussian jumps

$$\pi(\boldsymbol{\beta}_t \mid \mathcal{T}_t) = \prod_{k=1}^{K_t} \phi(\beta_{tj};\ 0, 1/T),$$

where $\phi(x;\ 0,\sigma^2)$ is a Gaussian density with mean 0 and variance σ^2. Chipman et al. [2010] recommend first shifting and rescaling Y_i's so that the observed transformed values range from -0.5 to 0.5. Then they assign a conjugate normal prior $\beta_{tj} \sim N(0,\sigma^2)$, where $\sigma = 0.5/k\sqrt{T}$ for some suitable value of k. This is to ensure that the prior assigns substantial probability to the range of the Y_i's.

The BART prior also involves an inverse chi-squared distribution on residual variance in (1). While the case of random variance can be incorporated in our analysis (de Jonge and van Zanten [2013]), we will for simplicity assume that the residual variance is fixed and equal to one.

Existing theoretical work for Bayesian forests (RP17) is available for a different prior on tree partitions $\mathcal{T}$. Their analysis assumes a hierarchical prior consisting of (a) a prior on the size of a tree K and (b) a uniform prior over trees of size K. This prior is equalitarian in the sense that trees with the same number of leaves are a-priori equally likely regardless of their topology. The prior on the number of leaves K is a very important ingredient for regularization. We will study aspects of its distribution under the actual BART prior in later sections.

3 Bayesian Non-parametrics Lense

One way of assessing the quality of a Bayesian procedure is by studying the learning rate of its posterior, i.e. the speed at which the posterior distribution shrinks around the truth as $n \to \infty$. These statements are ultimately framed in a frequentist way, describing the typical behavior of the posterior under the true generative model $\mathbb{P}^{(n)}_{f_0}$. Posterior concentration rate results have been valuable for the proposal and calibration of priors. In infinite-dimensional parameter spaces, such as the one considered here, seemingly innocuous priors can lead to inconsistencies (Cox [1993], Diaconis and Freedman [1986]) and far more care has to be exercised to come up with well-behaved priors.

The Bayesian approach requires placing a prior measure $\Pi(\cdot)$ on $\mathcal{F}$, the parameter space consisting of qualitative guesses of f_0. Given observed data $\boldsymbol{Y}^{(n)} = (Y_1, \dots, Y_n)'$, inference about f_0 is then carried out via the posterior distribution

$$\Pi(A \mid \boldsymbol{Y}^{(n)}) = \frac{\int_A \prod_{i=1}^n \Pi_f(Y_i \mid \boldsymbol{x}_i) \mathrm{d}\Pi(f)}{\int \prod_{i=1}^n \Pi_f(Y_i \mid \boldsymbol{x}_i) \mathrm{d}\Pi(f)} \quad \forall A \in \mathcal{B}$$

where $\mathcal{B}$ is a σ-field on $\mathcal{F}$ and where $\Pi_f(Y_i \mid \boldsymbol{x}_i)$ is the likelihood function for the output Y_i under f.

In Bayesian non-parametrics, one of the usual goals is determining *how fast the posterior probability measure concentrates around* f_0 as $n \to \infty$. This speed can be assessed by inspecting the size of the smallest $\|\cdot\|_n$-neighborhoods around f_0 that contain most of the posterior probability (Ghosal and van Der Vaart [2007]), where $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)^2$. For a diameter $\varepsilon > 0$ and some $M > 0$, we denote with

$$A_{\varepsilon,M} = \{f_{\mathcal{E},\boldsymbol{B}} \in \mathcal{F} : \|f_{\mathcal{E},\boldsymbol{B}} - f_0\|_n \le M\,\varepsilon\}$$

the $M\varepsilon$-neighborhood centered around f_0. We say that the posterior distribution concentrates at speed $\varepsilon_n \to 0$ such that $n\,\varepsilon_n^2 \to \infty$ when

$$\Pi(A^c_{\varepsilon_n,M_n} \mid \boldsymbol{Y}^{(n)}) \to 0 \quad \text{in } \mathbb{P}_{f_0}^{(n)}\text{-probability as } n \to \infty \tag{6}$$

for any $M_n \to \infty$, where A^c stands for the complement of the set A. Posterior consistency statements are a bit weaker, where ε_n in (6) is replaced with a fixed neighborhood $\varepsilon > 0$. We will position our results using $\varepsilon_n = n^{-\nu/(2\nu+p)} \log^{1/2} n$, the near-minimax rate for estimating a p-dimensional ν-smooth function. We will also assume that f_0 is Hölder continuous, i.e. ν-Hölder smooth with $0 < \nu \le 1$. The limitation $\nu \le 1$ is an unavoidable consequence of using step functions to approximate smooth f_0 and can be avoided with smooth kernel methods (Linero and Yang [2017]).

The statement (6) can be proved by verifying the following three conditions (suitably adapted from Theorem 4 of Ghosal and van Der Vaart [2007]):

$$\sup_{\varepsilon > \varepsilon_n} \log N\left(\tfrac{\varepsilon}{36};\ A_{\varepsilon,1} \cap \mathcal{F}_n;\ \|.\|_n\right) \le n\,\varepsilon_n^2 \tag{7}$$

$$\Pi(A_{\varepsilon_n,1}) \ge \mathrm{e}^{-d\,n\,\varepsilon_n^2} \tag{8}$$

$$\Pi(\mathcal{F}\backslash\mathcal{F}_n) = o(\,\mathrm{e}^{-(d+2)\,n\,\varepsilon_n^2}) \tag{9}$$

for some $d > 2$. In (7), $N(\varepsilon;\ \Omega;\ d)$ is the ε-covering number of a set Ω for a semimetric d, i.e. the minimal number of d-balls of radius ε needed to cover a set Ω. A few remarks are in place. The condition (9) 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 (7) is known as "the entropy condition" and controls the combinatorial richness of the approximating sets $\mathcal{F}_n$. Finally, condition (8) requires that the prior charges an ε_n neighborhood of the true function. The results of type (6) quantify not only the typical distance between a point estimator (posterior mean/median) and the truth, but also the typical spread of the posterior around the truth. These results are typically the first step towards further uncertainty quantification statements.

4 The Galton-Watson Process Prior

The Galton-Watson (GW) process provides a mathematical representation of an evolving population of individuals who reproduce and die subject to laws of chance. Binary tree partitions $\mathcal{T}$ under the prior (5) can be thought of as realizations of such a branching process. Below, we review some terminology of branching processes and link them to Bayesian CART.

We denote with Z_t the population size at time t (i.e. the number of nodes in the t^{th} layer of the tree). The process starts at time $t = 0$ with a single individual, i.e. $Z_0 = 1$. At time t, each member is split *independently* of one another into a random number of offsprings. Let Y_{ti} denote the number of offsprings produced by the i^{th} individual of the t^{th} generation and let $g_t(s)$ be the associated probability generating function. A binary tree is obtained when each node has either *zero* or *two* offsprings, as characterized by

$$g_t(s) = s^0\mathbb{P}(Y_{t1} = 0) + s^2\mathbb{P}(Y_{t1} = 2), \quad 0 \le s \le 1. \tag{10}$$

Homogeneous GW process is obtained when all Y_{ti}'s are iid. A *heterogeneous* GW process is a generalization where the offspring distribution is allowed to vary according to the generations, i.e. the variables Y_{ti} are independent but *non-identical*. The Bayesian CART prior of Chipman et al. [1998] can be framed as a heterogeneous GW process, where the probability of splitting a node (generating offsprings) depends on the depth t of the node in the tree. In particular, using (5) one obtains for $0 < \alpha < 1$ and $\gamma > 0$

$$\mathbb{P}(Y_{t1} = 2) = 1 - \mathbb{P}(Y_{t1} = 0) = \frac{\alpha}{(1+t)^\gamma}. \tag{11}$$

The population size at time t satisfies $Z_t = \sum_{i=1}^{Z_{t-1}} Y_{ti}$ and its expectation can be written as

$$\mathbb{E}Z_t = \mathbb{E}[\mathbb{E}(Z_t \mid Z_{t-1})] = (2\alpha)^t[(t+1)!]^{-\gamma}.$$

Since $\mathbb{E}Z_1 < 1$ under (11), the process is subcritical and thereby it dies out with probability one. This means that the random sequence $\{Z_t\}$ consists of zeros for all but a finite number of t's. The overall number of nodes in the tree (all ancestors in the family pedigree)

$$X=\sum_{t=0}^{\infty} Z_t \tag{12}$$

is thus finite with probability one. The number of *leaves* (bottom nodes) K can be related to X through

$$K=(X+1)/2 \tag{13}$$

and satisfies

$$T_{ex}+1\leq K\leq 2^{T_{ex}}, \tag{14}$$

where $T_{ex}=\min\{t: Z_t=0\}$ is the time of extinction. In (14), we have used the fact that $T_{ex}-1$ is the depth of the tree, where the lower bound is obtained with asymmetric trees with only one node split at each level and the upper bound is obtained with symmetric full binary trees (all nodes are split at each level).

Regularization is an essential remedy against overfitting and Bayesian procedures have a natural way of doing so through a prior. In the context of trees, the key regularization element is the prior on the number of bottom leaves K, which is completely characterized by the distribution of total progeny X via (13). Using this connection, in the next section we study the tail bounds of the distribution $\pi(K)$ implied by the Galton-Watson process.

5 Bayesian Tree Regularization

If we knew ν, the optimal (rate-minimax) choice of the number of tree leaves would be $K\asymp K_\nu=n^{p/(2\nu+p)}$ (RP17). When ν is unknown, one can do almost as well (sacrificing only a log factor in the convergence rate) using a suitable prior $\pi(K)$. As noted by Coram and Lalley [2006], the tail behavior of $\pi(K)$ is critical for controlling the vulnerability/resilience to overfitting. The anticipation is that with 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, Denison et al. [1998] suggest the Poisson distribution (constrained to $\mathbb{N}\backslash\{0\}$), which satisfies

$$\mathbb{P}(K>k)\lesssim \mathrm{e}^{-C_K\,k\log k}\quad\text{for some } C_K>0. \tag{15}$$

Under this prior, one can show that $\mathbb{P}(K>C\,K_\nu\mid \boldsymbol{Y}^{(n)})\rightarrow 0$ in $\mathbb{P}_{f_0}^{(n)}$ probability (RP17). The posterior thus does not overshoot the oracle K_ν too much.

In the BART prior, the distribution $\pi(K)$ is implicitly defined through the GW process rather than directly through (15). In order to see whether BART induces a sufficient amount of regularization, we first need to obtain a tail bound of $\pi(K)$ under the GW process and show that it decays fast enough. One seemingly simple remedy would be to set $\gamma=0$ (which coincides with the homogeneous GW case) and $\alpha=c/n$ with some $c>0$. Standard branching process theory then implies $\Pi(K>k)\lesssim \mathrm{e}^{-C_K\,k\log n}$. This prior is more aggressive than (15). Moreover, letting the split probability $p_{split}(\Omega_k)$ decay with sample size is counterintuitive. By choosing $\alpha=c$, on the other hand, one obtains $\Pi(K>k)\lesssim \mathrm{e}^{-C_K\,k}$ which is not aggressive enough.

While the homogeneous GW processes have been studied quite extensively, the literature on tail bounds for *heterogeneous* GW processes (for when $\gamma\neq 0$) has been relatively deserted. We first review one interesting approach in the next section and then come up with a new bound in Section 5.2.

5.1 Tail Bounds à la Agresti

Agresti [1975] obtained bounds for the extinction time distribution of branching processes with independent non-identically distributed environmental random variables Y_{ti}.

Theorem 5.1. *[Agresti, 1975] Consider the heterogeneous Galton-Watson branching process with offspring p.g.f.'s $\{g_j(s);\ j\geq 0\}$ satisfying $g_j^{''}(1)<\infty$ for $j\geq 0$. Denote $P_t=\prod_{j=0}^{t-1}g_j'(1)$. Then*

$$\mathbb{P}(T_{ex}>t)\leq\left[P_t^{-1}+\frac{1}{2}\sum_{j=0}^{t-1}(g_j^{''}(0)/g_j'(1)P_{j+1})\right]^{-1}. \tag{16}$$

Using this result, we can obtain a tail bound on the extinction time under the Bayesian CART prior.

Corollary 5.1. *For the heterogeneous Galton-Watson branching process with offspring p.g.f.'s* (10) *with* (11) *we have*

$$\mathbb{P}(T_{ex}>t)<C_0\left(\frac{t^\gamma}{2\alpha\,\mathrm{e}^\gamma}\right)^{-t} \tag{17}$$

for a positive constant C_0 that depends on α and γ.

Proof. We have $g_0(s)=s$ and for $j\geq 1$

$$\begin{aligned}
g_j(s)&=1-\alpha(1+j)^{-\gamma}+s^2\alpha(1+j)^{-\gamma},\\
g_j^{'}(s)&=2s\alpha(1+j)^{-\gamma},\\
g_j^{''}(s)&=2\alpha(1+j)^{-\gamma}.
\end{aligned}$$

Thus we have $g_0'(1)=1$ and $g_j^{'}(1)=g_j^{''}(0)=2\alpha(1+j)^{-\gamma}$ for $j\geq 1$. Then we can write

$$P_t^{-1}=\frac{\prod_{i=0}^{t-1}(1+i)^\gamma}{(2\alpha)^t}=\frac{(t!)^\gamma}{(2\alpha)^t} \tag{18}$$

and

$$\sum_{j=0}^{t-1}(g_j^{''}(0)/g_j^{'}(1)P_{j+1}) = \sum_{j=0}^{t-1}\frac{1}{P_{j+1}} = \sum_{j=1}^{t}\frac{(j)!^{\gamma}}{(2\alpha)^{j}} > \frac{(t!)^{\gamma}}{(2\alpha)^{t}}.$$

Using (18) and the fact that $t! > (t/\mathrm{e})^t\,\mathrm{e}$, we can upper-bound the right hand side of (16) with $C_0[t^\gamma/(\mathrm{e}^\gamma 2\alpha)]^{-t}$. ■

Remark 5.1. *A simpler bound on the extinction time can be obtained using Markov's inequality as follows:* $\mathbb{P}(T_{ex} > t) = \mathbb{P}(Z_t \geq 1) \leq \mathbb{E}Z_t \leq (2\alpha)^t[(t+1)!]^{-\gamma}$.

Using the upper bound in (14) we immediately conclude that

$$\mathbb{P}(K > k) < \mathbb{P}(T_{ex} > \log_2 k) \leq C_0\left(\frac{\log_2^\gamma k}{2\alpha\,\mathrm{e}^\gamma}\right)^{-\log_2 k}.$$

This decay, however, is not fast enough as we would ideally like to show (15). We try a different approach in the next section.

5.2 Trees as Random Walks

There is a curious connection between branching processes and random walks (see e.g. Dwass [1969]). Suppose that a binary tree $\mathcal{T}$ is revealed in the following node-by-node exploration process: one exhausts all nodes in generation d before revealing nodes in generation $d+1$. Namely, nodes are implicitly numbered (and explored) according to their priority and this is done in a top/down manner according to their layer and a left-to-right manner within each layer (i.e. Ω_0 is the root node and, if split, Ω_1 and Ω_2 are the two children (left and right) etc.)

Nodes that are waiting to be explored can be organized in a queue Q. We say that a node is *active* at time t if it resides in a queue. Starting with one active node at $t = 0$ (the root node), at each time t we deactivate (remove from Q) the node with the highest priority (lowest index) and add its children to Q. Letting S_t be the number of active nodes at time t, one finds that $\{S_t\}$ satisfies

$$S_t = S_{t-1} - 1 + Y_t, \quad t \geq 1,$$

and $S_0 = 1$, where Y_t are sampled from the offspring distribution. For the *homogeneous* GW process, S_t is an actual random walk where Y_t are iid with a probability generating function (10). For the *heterogeneous* GW process, S_t is not strictly a random walk in the sense that $Y_t's$ are not iid. Nevertheless, using this construction one can see that the total population X equals the first time the queue is empty:

$$X = \min\{t \geq 0 : S_t = 0\}.$$

Linking Galton-Watson trees to random walk excursions in this way, one can obtain a useful tail bound of the distribution of the population size X. While perhaps not surprising, we believe that this bound is new, as we could not find any equivalent in the literature.

Lemma 5.1. *Denote by X the total population size* (12) *arising from the heterogeneous Galton-Watson process. Then we have for any $c > 0$*

$$\mathbb{P}(X > k) \leq \mathrm{e}^{-k\,c + (\,\mathrm{e}^{2c}-1)\mu}, \tag{19}$$

where $\mu = \sum_{i=1}^k p_i$ and $p_i = p_{split}(\Omega_i)$, where nodes Ω_i are ordered in a top-down left-to-right fashion.

Proof. For $k > 0$, we can write

$$\mathbb{P}(X > k) \leq \mathbb{P}(S_k > 0) = \mathbb{P}\left(\sum_{i=1}^k Y_i > k-1\right),$$

where X is the number of all nodes (internal and external) in the tree and Y_i has a two-point distribution characterized by $\mathbb{P}(Y_i = 2) = 1 - \mathbb{P}(Y_i = 0) = p_i$. Using the Chernoff bound, one deduces that for any $c > 0$

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^k Y_i > k-1\right) &\leq \mathrm{e}^{-k\,c}\,\mathbb{E}\,\mathrm{e}^{c\,\sum_{i=1}^k Y_i} \\ &= \mathrm{e}^{-k\,c}\prod_{i=1}^k[p_i\,\mathrm{e}^{2c} + 1 - p_i] \leq \mathrm{e}^{-k\,c + (\,\mathrm{e}^{2c}-1)\mu}\end{aligned}$$

where $\mu = \sum_{i=1}^k p_i$. ■

The goal throughout this section has been to understand whether the Bayesian CART prior of Chipman et al. [1998] yields (15) for some $C_K > 0$. The prior assumes $p_i = \alpha/(1 + d(\Omega_i))^\gamma$. Choosing $c = (\log k)/2$ in (19), the right hand side will be smaller than $\mathrm{e}^{-a\,k\log k}$, for some suitable $0 < a < 1/2$, as long as $\mu \leq (1/2 - a)\log k$. We note that

$$\mu = \sum_{i=1}^k p_i < \sum_{d=1}^{\lceil\log_2 k\rceil}\frac{\alpha}{(1+d)^\gamma}2^d.$$

Because the split probability p_i decreases only polynomially in depth of Ω_i, this *is not enough* to ensure $\mu < (1/2 - a)\log(k)$. The optimal decay, however, will be guaranteed if we instead choose

$$p_{split}(\Omega) \propto \alpha^{d(\Omega)} \quad \text{for some } 0 < \alpha < 1/2. \tag{20}$$

To conclude, from our considerations it is not clear that the Bayesian CART prior of Chipman et al. [1998] has the optimal tail-bound decay. The following Corollary certifies that the optimal tail behavior can be obtained with a suitable modification of $p_{split}(\Omega)$.

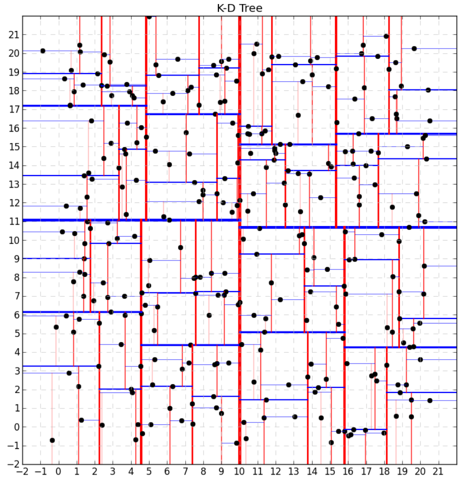

Figure 1: The k-d trees in two dimensions at various resolution levels.

Corollary 5.2. *Under the Bayesian CART prior of Chipman et al. [1998] with* (20)*, we obtain* (15)*.*

Proof. Follows from the considerations above and from (13).

6 Prior Concentration for BART

One of the prerequisites for optimal posterior concentration (6) is optimal prior concentration (Condition (8)). This condition ensures that there is enough prior support around the truth. It can be verified by constructing one approximating tree and by showing that it has enough prior mass. RP17 use the k-d approximating tree (Remark 3.1), which is a balanced full binary tree which partitions $[0,1]^p$ into nearly identical rectangles (in sufficiently regular designs). This tree can be regarded as the most regular partition that can be obtained by splitting at observed values and has been studied rigorously in the literature (see e.g. Verma et al. [2009]). A formal definition of the k-d tree is below and a few two-dimensional examples[1] (at various resolution levels) are in Figure 1.

Definition 6.1. *(k-d tree partition) The k-d tree partition with $p \times s$ layers is constructed by cycling s times over coordinate directions $\{1, \dots, p\}$. At each tree level, all nodes are split along the same axis. For a given direction $j \in \{1, \dots, p\}$, each internal node will be split at a median of the point set (along the j^{th} axis). Each split thus roughly halves the number of points inside the cell.*

After s rounds of splits on each variable, all K terminal nodes have at least $\lfloor n/K \rfloor$ observations, where $K = 2^{s\,p}$. The k-d tree partitions are thus balanced in light of Definition 2.4 of Rockova and van der Pas [2017] (i.e. have roughly the same number of observations inside). The k-d tree construction is instrumental in establishing optimal prior/posterior concentration. Lemma 3.2 of RP17 shows that there *exists* a step function supported by a k-d partition that safely approximates f_0 with an error smaller than a constant multiple of the minimax rate. The approximating k-d tree partition, denoted with $\widehat{\mathcal{T}}$, has $\widehat{K}$ steps where $\widehat{K} \asymp n\varepsilon_n^2/\log n$ when $p \lesssim \log^{1/2} n$ (as shown in Section 8.3 of RP17 and detailed in the proof of Theorem 7.1).

In order to complete the proof of posterior concentration for the Bayesian CART under the Galton-Watson process prior, we need to show that $\pi(\widehat{\mathcal{T}}) \geq \mathrm{e}^{-c_1 n \varepsilon_n^2}$ for some $c_1 > 0$. This is verified in the next lemma.

Lemma 6.1. *Denote with $\widehat{\mathcal{T}}$ the k-d tree partition described above. Assume the heterogeneous Galton-Watson process tree prior with $p_{split}(\Omega_k) \propto \alpha^{d(\Omega_k)}$ for some suitable $1/n \leq \alpha < 1/2$. Assume $p \lesssim \log^{1/2} n$. Then we have for some suitable $c_1 > 0$*

$$\pi(\widehat{\mathcal{T}}) \geq \mathrm{e}^{-c_1\, n\, \varepsilon_n^2}.$$

Proof. By construction, the k-d tree $\widehat{\mathcal{T}}$ has $\widehat{K} = 2^{p \times s}$ leaves and $p \times s$ layers for some $s \in \mathbb{N}$ where p is the number of predictors. In addition, the k-d tree is complete and balanced (i.e. every layer d, including the last one, has the maximal number 2^d of nodes). Since there are $\widehat{K}-1$ internal nodes and at least $1/(p\,n)$ splitting rules for each internal node, we have

$$\pi(\widehat{\mathcal{T}}) \geq \frac{(1-\alpha^{s\,p})^{\widehat{K}}}{(p\,n)^{\widehat{K}-1}} \prod_{d=0}^{\log_2 \widehat{K}-1} \alpha^{2^d} \geq \frac{(1-\alpha^{s\,p})^{\widehat{K}}}{(p\,n)^{\widehat{K}-1}} \alpha^{\widehat{K}-1}$$

$$\geq [\alpha(1-\alpha)]^{\widehat{K}} \left(\frac{1}{p\,n}\right)^{\widehat{K}-1}$$

$$> \mathrm{e}^{-\widehat{K}\log(2n) - (\widehat{K}-1)\log(p\,n)}.$$

Since $p \lesssim \log^{1/2} n$ and $\widehat{K} \asymp n\,\varepsilon_n^2/\log n$ we can lower-bound the above with $\mathrm{e}^{-c_1\, n \varepsilon_n^2}$ for some $c_1 > 0$. ■

For the actual BART method (similarly as in Theorem 5.1 of RP17), one needs to find an approximating *tree ensemble* and show that it has enough prior support. The approximating ensemble can be found in Lemma 10.1 of RP17 and consists of $\widehat{\mathcal{E}} = \{\widehat{\mathcal{T}}_1, \dots, \widehat{\mathcal{T}}_T\}$ tree partitions obtained by chopping of branches of $\widehat{\mathcal{T}}$. The number of trees T is fixed and the trees $\mathcal{T}_t$ will not overlap much when $1 \leq T \leq \widehat{K}/2$. The default BART choice $T = 200$ safely satisfies this as long as $p > 9$. The little trees $\widehat{\mathcal{T}}_t$ have $\widehat{K}^t$ leaves and satisfy $\log_2 \widehat{K} +$

[1] Source: https://salzis.wordpress.com/2014/06/28/

$1 \leq \widehat{K}^t \leq \widehat{K}$ (depending on the choice of T). Using Lemma 6.1 and the fact that the trees are independent a-priori (from (4)) and that T is fixed, we then obtain

$$\pi(\widehat{\mathcal{E}}) \geq \mathrm{e}^{-\sum_{t=1}^{T}[\widehat{K}^t \log 2n + (\widehat{K}^t - 1)\log(p\,n)]}$$
$$> \mathrm{e}^{-T\widehat{K}\log 2n - T(\widehat{K}-1)\log(p\,n)} > \mathrm{e}^{-c_2\,n\varepsilon_n^2}$$

for some $c_2 > 0$. The BART prior thus concentrates enough mass around the truth. Condition (8) also requires verification that the prior on jump sizes concentrates around the forest sitting on $\widehat{\mathcal{E}}$. This follows directly from Section 9.2 of RP17. We detail the steps in the proof of Theorem 7.1.

7 Posterior Concentration for BART

We now have all the ingredients needed to state the posterior concentration result for BART. The result is *different* from Theorem 5.1 of RP17 because here we (a) assume that T is fixed, (b) assume the branching process prior on $\mathcal{T}$ and (c) we do not have subset selection uncertainty. We will treat the design as fixed and *regular* according to Definition 3.3 of RP17.

Definition 7.1. *Denote by $\widehat{\mathcal{T}} = \{\widehat{\Omega}_k\}_{k=1}$ the k-d tree with $K = 2^{s\times p}$ bottom nodes. We say that a dataset $\{\boldsymbol{x}_i\}_{i=1}^n$ is regular if*

$$\max_{1\leq k\leq K} \mathrm{diam}(\widehat{\Omega}_k) < M \sum_{k=1}^{K} \mu(\widehat{\Omega}_k)\mathrm{diam}(\widehat{\Omega}_k) \qquad (21)$$

for all $s \in \mathbb{N}$ and for some $M > 0$, where

$$\mathrm{diam}\left(\widehat{\Omega}_k\right) = \max_{\boldsymbol{x},\boldsymbol{y}\in\widehat{\Omega}_k\cap\{\boldsymbol{x}_i\}_{j=1}^n} \|\boldsymbol{x}-\boldsymbol{y}\|_2.$$

The design regularity assumption translates as follows: the data points inside the cells of a *k-d* tree (Figure 1) have similar diameters (i.e. maximal interpoint distances inside the cell). The *k-d* tree is the most regular partition one can obtain by splitting at data points. If the cell diameters of the *k-d* tree are similar, the dataset is without outliers and/or isolated clouds of points. This is a mild and reasonable requirement (see RP17 for more discussion). Moreover, this assumption allows for correlated x's and holds trivially for fixed designs on a regular grid.

Theorem 7.1. *(Posterior Concentration for BART) Assume that f_0 is ν-Hölder continuous with $0 < \nu \leq 1$ where $\|f_0\|_\infty \lesssim \log^{1/2} n$. Assume a regular design $\{\boldsymbol{x}_i\}_{i=1}^n$ where $p \lesssim \log^{1/2} n$. Assume the BART prior with T fixed and with $p_{split}(\Omega_t) = \alpha^{d(\Omega_t)}$ for $1/n \leq \alpha < 1/2$. With $\varepsilon_n = n^{-\alpha/(2\alpha+p)}\log^{1/2} n$ we have*

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

for any $M_n \to \infty$ in $\mathbb{P}_{f_0}^{(n)}$-probability, as $n, p \to \infty$.

Proof. Appendix. ■

Theorem 7.1 has very important implications. It provides a frequentist theoretical justification for BART claiming that the posterior is wrapped around the truth and its learning rate is near-optimal. As a by-product, one also obtains a statement which supports the empirical observation that BART is resilient to overfitting.

Corollary 7.1. *Under the assumptions of Theorem 7.1 we have*

$$\Pi\left(\bigcup_{t=1}^{T}\{K^t > C\,n^{p/(2\nu+p)}\} \mid \boldsymbol{Y}^{(n)}\right) \to 0$$

in $\mathbb{P}_{f_0}^{(n)}$-probability, as $n, p \to \infty$, for a suitable constant $C > 0$.

Proof. The proof follows from the proof of Theorem 7.1 and Lemma 1 of Ghosal and van Der Vaart [2007]. ■

In other words, the posterior distribution rewards ensembles that consist of small trees whose size does not overshoot the optimal number of steps $K_\nu = n^{p/(2\nu+p)}$ by much. In this way, the posterior is fully adaptive to unknown smoothness, not overfitting in the sense of split overuse.

8 Discussion

In this work, we propose a variant of the BART prior by modifying the split probability. This new prior (a) is shown to yield optimal posterior concentration (in the $\|\cdot\|_n$ sense) , (b) can be just as easily implemented and (c) might enhance the performance of BART in practice.

We have built on Rockova and van der Pas [2017] to show optimal posterior convergence rate of the BART method. Similar results have been obtained for other Bayesian non-parametric constructions such as Polya trees (Castillo [2017]), Gaussian processes (van der Vaart and van Zanten [2008], Castillo [2008]) and deep ReLU neural networks [Polson and Rockova, 2018]. Up to now, the increasing popularity of BART has relied on its practical performance across a wide variety of problems. The goal of this and future theoretical developments is to establish BART as a rigorous statistical tool with solid theoretical guarantees. Similar guarantees have been obtained for variants of the traditional forests/trees by multiple authors including Gordon and Olshen [1980, 1984], Donoho [1997], Biau et al. [2008], Scornet et al. [2015], Wager and Guenther [2015].

References

A. Agresti. On the extinction times of varying and random environment branching processes. *Journal of Applied Probability*, 12(1):39–46, 1975.

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.

J. Bleich, A. Kapelner, E.I. George, and S.T. Jensen. Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, 4(3): 1750–1781, 2014.

L. Breiman, J. Friedman, C.J. Stone, and R. A. Olshen. *Classification and Regression Trees (Wadsworth Statistics/Probability)*. Chapman and Hall/CRC, 1984.

I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

I. Castillo. Pólya tree posterior distributions on densities. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 2074–2102. Institut Henri Poincaré, 2017.

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

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

H.A. Chipman, E.I. George, R.E. McCulloch, and T.S. Shively. High-dimensional nonparametric monotone function estimation using BART. *arXiv preprint arXiv:1612.01619*, 2016.

M. Coram and S.P. Lalley. Consistency of Bayes estimators of a binary regression function. *The Annals of Statistics*, 34(3):1233–1269, 2006.

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

R. de Jonge and J.H. van Zanten. Semiparametric Bernstein-von Mises for the error standard deviation. *Electronic Journal of Statistics*, 7(1):217–243, 2013.

D. Denison, B. Mallick, and A. Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

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

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

M. Dwass. The total progeny in a branching process and a related random walk. *Journal of Applied Probability*, 6(3):682–686, 1969.

S. Ghosal and A.W. van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

L. Gordon and R. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627, 1980.

L. Gordon and R. Olshen. Almost sure consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 15: 147–163, 1984.

R.B. Gramacy and H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

P.R. Hahn, J.S. Murray, and C.M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. 2017.

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

B. Lakshminarayanan, D. Roy, and Y.W. Teh. Top-down particle filtering for Bayesian decision trees. In *International Conference on Machine Learning*, 2013.

B. Lakshminarayanan, D.M. Roy, and Y.W. Teh. Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

A.R. Linero and Y. Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *arXiv preprint arXiv:1707.09461*, 2017.

Y. Liu, V. Rockova, and Y. Wang. ABC variable selection with Bayesian forests. *arXiv preprint arXiv:1806.02304*, 2018.

N. Polson and V. Rockova. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

M. Pratola, H.A. Chipman, E.I. George, and R.E. McCulloch. Heteroscedastic BART using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*, 2017.

V. Rockova and S.L. van der Pas. Posterior concentration for Bayesian regression trees and their ensembles. *arXiv preprint arXiv:1708.08734*, 2017.

D.M. Roy and Y.W. Teh. The Mondrian process. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

E. Scornet, G. Biau, and J.P. Vert. Consistency of random forests. *Annals of Statistics*, 43:1716–1741, 2015.

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

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

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

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.

S. Wager and W. Guenther. Adaptive concentration of regression trees with application to random forests. *Manuscript*, 2015.