Adaptive Conditional Distribution Estimation with Bayesian Decision Tree Ensembles

Yinpu Li[1], Antonio R. Linero[2,*], and Jared Murray[2]
[1]*Florida State University and* [2]*University of Texas at Austin*
[*]antonio.linero@austin.utexas.edu

April 16, 2021

Abstract

We present a Bayesian nonparametric model for conditional distribution estimation using Bayesian additive regression trees (BART). The generative model we use is based on rejection sampling from a base model. Like other BART models, our model is flexible, has a default prior specification, and is computationally convenient. To address the distinguished role of the response in our BART model, we further introduce an approach to targeted smoothing of BART models which is possibly of independent interest. We study the proposed model theoretically and provide sufficient conditions for the posterior distribution to concentrate at close to the minimax optimal rate adaptively over smoothness classes in the high-dimensional regime in which many predictors are irrelevant. To fit our model, we propose a data augmentation algorithm which allows for existing BART samplers to be extended with minimal effort. We illustrate the performance of our methodology on simulated data and use it to study the relationship between education and body mass index using data from the medical expenditure panel survey (MEPS).

1 Introduction

We consider here the Bayesian nonparametric estimation of the conditional distribution of a response Y_i based on predictors X_i. A common strategy is to introduce a latent variable b, and set $Y_i \sim h(y \mid X_i, b, \theta)$ given b, where $h(y \mid x, b, \theta)$ is a parametric model. This includes mixture models where b is a latent class indicator and the conditional density has the mixture form $f(y \mid x) = \sum_{k=1}^{\infty} \omega_k(x)\ h(y \mid x, \theta_k)$ (MacEachern, 1999; Dunson and Park, 2008; Chung and Dunson, 2009; Rodriguez and Dunson, 2011), as well as Gaussian process

latent variable/covariate models where b is continuous (Wang and Neal, 2012; Kundu and Dunson, 2014; Dutordoir et al., 2018).

A conceptually simpler approach models $f(y \mid x)$ by tilting a base model:

$$f(y \mid x) = \frac{h(y \mid x, \theta)\ \Phi\{r(y,x)\}}{\int h(\widetilde{y} \mid x, \theta)\ \Phi\{r(\widetilde{y},x)\}\ d\widetilde{y}}, \tag{1}$$

where $\{h(\cdot \mid \cdot, \theta) : \theta \in \Theta\}$ is a parametric family of conditional densities. We refer to $h(y \mid x, \theta)$ as the *base model* and $\Phi(\mu)$ as the *link function*. When $r(y,x)$ is a constant, this model reduces to the base model, allowing the user to center (1) on a desired parametric model. A special case of (1) sets $\Phi(\mu) = e^{\mu}$ and takes $r(y,x)$ to be a Gaussian process (Tokdar et al., 2010). In the context of (marginal) density estimation, Murray et al. (2009) proposed the Gaussian process density sampler (GP-DS), which sets $\Phi(\mu)$ to be a sigmoidal function such as the logistic function $\Phi(\mu) = (1 + e^{-\mu})^{-1}$. Methods based on Gaussian processes have elegant theoretical properties (van der Vaart and van Zanten, 2008) but are somewhat difficult to work with due to the integral in the denominator of (1) and the need to compute, store, and invert an $N \times N$ where N is the sample size. The goal of this paper is to propose a method based on (1) with the following desirable properties.

- Algorithms for posterior inference are straight-forward to implement.
- The posterior possesses strong theoretical properties, obtaining posterior convergence rates close to the best possible.
- For routine use, a default prior can be used which empirically obtains good practical performance.
- Shrinkage towards the base model $h(y \mid x, \theta)$ is easy to perform, so that the model naturally adapts to the complexity of the data.

We propose a modified variant of the Bayesian additive regression trees (BART) model of Chipman et al. (2010) which we refer to as the SBART density sampler (SBART-DS). We

choose $r(y, x)$ to be a *soft* decision tree (Linero and Yang, 2018; Irsoy et al., 2012) which smooths in a targeted fashion on the response variable y (Starling et al., 2018). A benefit of the BART framework is that we are able to develop default priors based on well-known heuristics and show that these default priors perform well in practice.

To construct inference algorithms, we restrict the choice of $\Phi(\mu)$ to the logit, probit, or t_ν-link functions. Our proposal is similar to the GP-DS, but is adapted to conditional distribution estimation. We construct an efficient Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution by combining a data augmentation scheme of Rao et al. (2016) with an additional layer of data augmentation. After performing this data augmentation, we can update the parameters of the model using the same Bayesian backfitting algorithm as Chipman et al. (2010). Given that one has the ability to perform Bayesian backfitting, the algorithms we construct are simple to implement.

We present theoretical results which show that suitably-specified SBART-DS priors attain convergence rates which are close to the best possible. Simplifying slightly, we show that in the high-dimensional sparse setting, where only $D - 1 \ll P$ of the P predictors are relevant, SBART-DS can obtain the oracle rate of convergence $\epsilon_n = n^{-2\alpha/(2\alpha+D)}$ up-to a logarithmic term, where α is related to the smoothness level of the true conditional density. In a simulation study we show that these theoretical results are suggestive of what occurs in practice, as the performance of SBART-DS is robust to the presence of irrelevant predictors.

Related Methods Starting with the works of Müller et al. (1996) and MacEachern (1999), there is a large literature which tackles the density regression problem from the perspective of infinite mixture modeling, with $f(y \mid x) = \sum_k \omega_k(x)\, h(y \mid x, \theta_k)$. This allows for shrinkage towards a base model $h(y \mid x, \theta)$ by choosing the prior so that $\omega_1(x) \approx 1$ for all x. The weights can be modeled implicitly by jointly modeling the predictors and responses (Müller et al., 1996; Shahbaba and Neal, 2009; Wade et al., 2014). Alternatively, the weights can be directly modeled. A popular approach to modeling the weights is to use various extensions of

the Sethurmann stick-breaking construction of the Dirichlet process $\omega_k(x) = V_k(x)\prod_{j=1}^{K-1}\{1-V_j(x)\}$ (Sethuraman, 1994) such as the kernel stick breaking process (Dunson and Park, 2008) or probit stick-breaking process (Rodriguez and Dunson, 2011). Alternatively, Antoniano-Villalobos et al. (2014) model $\omega_k(x) \propto w_k K(x \mid \lambda_k)$ for some choice of kernel function $K(\cdot \mid \cdot)$. In Chung and Dunson (2009) the probit stick breaking process is extended to account for potentially high-dimensional predictors. Beyond infinite mixture models, there is a rich literature on Gaussian process based density regressors, either based on the structure (1) with the exponential link (Riihimäki et al., 2014; Tokdar et al., 2010) or based on a latent variable structure $f(y \mid x) = \int_0^1 \text{Normal}\{y \mid r(x,b), \sigma^2\}\, db$ (Titsias and Lawrence, 2010; Wang and Neal, 2012; Kundu and Dunson, 2014). Each of the above approaches comes with challenges — implementations of these models may not be easy for non-experts, default priors may not be reliable, and our ability to shrink towards a base model may be limited.

Why Density Regression? As pointed out by an anonymous reviewer, there is an argument that direct analysis of the features of interest of the conditional density (such as estimating the conditional quantiles through quantile regression) is more appropriate than indirect analysis through a complicated density regression model. We find specification of a large Bayesian model to be attractive due to its ability to learn many complex features of the distribution of interest, in a principled fashion, without knowing a-priori which features will be present. For example, the shape/scale effects of the predictors in Section 5.3 were not known to us a-priori; we had no reason to expect them and would not have thought to look for them. By using a single Bayesian nonparametric model, we are able to examine the posterior for interesting features while still remaining principled from the Bayesian perspective; if we instead use separate models to analyze different features, the different results may be difficult to reconcile in a principled fashion. This feature of the logical coherence of Bayesian methods has recently been used to generate explainable Bayesian inference using loss functions (Hahn and Carvalho, 2015; Woody et al., 2020).

Outline In Section 2 we review BART and describe a naive version of SBART-DS; we then describe our approach for targeted smoothing which centers our prior for $r(\cdot, x)$ on a desired Gaussian process. In Section 3 we provide data augmentation algorithms for fitting (1) when the link function $\Phi(\mu)$ is a probit, logit, or Student's t_ν link. In Section 4 we present our theoretical results. In Section 5 we conduct a simulation study and then apply SBART-DS to data from the Medical Expenditure Panel Survey (MEPS) to study the relationship between educational attainment and body mass index in adult women. We conclude in Section 6 with a discussion.

2 Model Description

2.1 Review of Bayesian Additive Regression Trees

The Bayesian additive regression trees (BART) framework models a function $r(x)$ as a sum of regression trees $r(x) = \sum_{m=1}^{M} g(x; \mathcal{T}_m, \mathcal{M}_m)$ where $\mathcal{T}_m$ denotes the topology and splitting rules of a binary decision tree and $\mathcal{M}_m = \{\mu_{m1}, \ldots, \mu_{mL_m}\}$ gives a prediction for each of the L_m terminal (leaf) nodes of $\mathcal{T}_m$. Figure 1 gives a schematic which shows how predictions are obtained from a decision tree. Chipman et al. (2010) specify priors $\pi_{\mathcal{T}}$ and $\pi_{\mathcal{M}}$ for the tree topologies $\mathcal{T}_m$ and the $\mu_{m\ell}$'s given $\mathcal{T}_m$, respectively. We write $r \sim \text{BART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ to denote that r has the associated BART prior. Typically we set $\mu_{m\ell} \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_\mu^2/M)$ so that $\text{Var}\{r(x)\} = \sigma_\mu^2$ regardless of the number of trees used in the model.

A problem with methods based on decision trees is that realizations of $r \sim \text{BART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ will not be continuous in x. This is particularly problematic for density estimation, as we generally prefer estimates of the density to be smooth. A smooth variant of BART called soft BART (SBART) was introduced by Linero and Yang (2018). SBART takes the tree $\mathcal{T}_m$ to be a *smooth* decision tree, where observations are assigned a weight $\varphi_{m\ell}(x)$ to leaf node ℓ of tree m. As a point of comparison, non-soft decision trees use the weights $\varphi_{m\ell}(x) = \prod_{b \in \mathcal{A}_{m\ell}} I(x_{j_b} \leq C_b)^{1-R_b} \, I(x_{j_b} > C_b)^{R_b}$, where $\mathcal{A}_{m\ell}$ denotes the collection of branches which

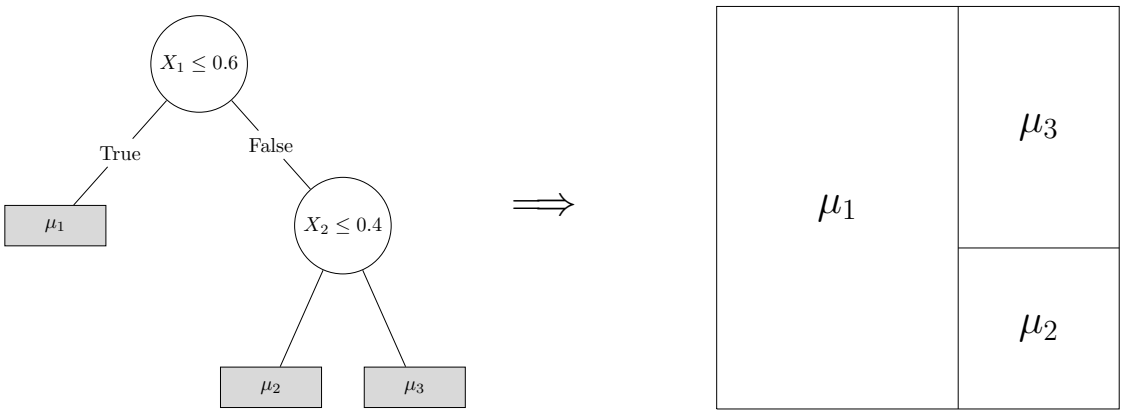

Figure 1: Schematic which shows how a decision tree induces a regression function. Associated to the decision tree is a partition of $[0,1]^2$ with the function $g(x;\mathcal{T},\mathcal{M})$ returning μ_1, μ_2, or μ_3.

are *ancestors* of leaf ℓ of tree m, j_b denotes the coordinate along which b splits, C_b denotes the cutpoint of branch b, and R_b is the indicator that the path from the root to the leaf goes right at b. A soft decision tree instead takes

$$\varphi_{m\ell}(x) = \prod_{b \in \mathcal{A}_{m\ell}} \psi(x; C_b, \tau_b)^{1-R_b} \left\{1 - \psi(x; C_b, \tau_b)\right\}^{R_b},$$

where $\psi(x; c, \tau)$ is the cumulative distribution function of a location-scale family with location c and scale τ. If $\psi(x) = I(x \leq 0)$ (or, equivalently, as $\tau \to 0$) we get a standard decision tree. If we instead take $\psi(x)$ to be a smooth function then $r(y, x)$ will also be smooth. The parameter τ is analogous to a bandwidth parameter, with larger values of τ giving smoother functions. Like Linero and Yang (2018) we will take $\psi(x) = (1+e^{-x})^{-1}$ and use tree-specific bandwidths τ_m. We write $r \sim \text{SBART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ to denote that r has an SBART prior, where $\pi_{\mathcal{T}}$ is now a prior over the *soft* trees $\mathcal{T}_m$.

For completeness, we describe the prior over the tree structures we will use. We assume that each coordinate x_j of the predictors has been scaled to lie in $[0,1]$. This can be done, for example, by applying the empirical quantile transform to a subset of the observed values for each predictor and then interpolating the remaining values. A tree $\mathcal{T}_m$ can be sampled from the prior as follows:

1. Initialize $\mathcal{T}_m$ with an single node of depth $D_m = 0$.

2. For all nodes of depth D_m, make that node a branch node, with a left and right child of depth $D_m + 1$, with probability $\alpha(1 + D_m)^{-\beta}$; otherwise, make the node a leaf node.

3. For all branch nodes b of depth D_m, sample the splitting coordinate $j_b \sim \text{Categorical}(s)$ and a splitting point $C_b \sim \text{Uniform}(L_{j_b}, U_{j_b})$ where $\prod_{j=1}^{P}[L_j, U_j]$ denotes the hyperrectangle of x-values which lead to node b.

4. If all nodes of depth D_m are leaf nodes, terminate; otherwise, set $D_m \leftarrow D_m + 1$ and return to Step 2.

The distribution of the splitting coordinate $j_b \sim \text{Categorical}(s)$ determines how relevant a-priori we expect each predictor to be; for example, if $s_1 = 0.99$ we expect most splitting rules to use x_1, whereas if $s_1 = 10^{-10}$ we expect none of the splitting rules to use x_1. Linero (2018) took advantage of this fact to perform automatic relevance determination (Neal, 1995) for BART models by using a sparsity-inducing Dirichlet prior $s \sim \text{Dirichlet}(a/P, \ldots, a/P)$ for some $a \ll P$. We also use this prior for the splitting proportion, which will allow us to perform automatic relevance determination in the density regression setting. This prior is crucial for proving that the posterior adapts to the presence of irrelevant predictors.

2.2 The Soft BART Density Sampler

Our modeling strategy is based on representation (1), where $\Phi(\mu)$ is a continuous, non-negative, monotonically increasing *link* function. Such an $r(y, x)$ is guaranteed to exist whenever $R(y, x) = f(y \mid x)/h(y \mid x, \theta)$ is non-zero and bounded for each x, e.g., we have $r(y, x) = \Phi^{-1}\{R(y, x)/\sup_y R(y, x)\}$. Because a BART model is bounded almost-surely, we note that $0 < \Phi\{r(y, x)\} < 1$ so that $0 < \int h(y \mid x, \theta)\, \Phi\{r(y, x)\}\, dy \leq \int h(y \mid x, \theta)\, dy = 1$ for all x almost-surely; hence the right-hand-side of (1) is defined almost-surely when $r(y, x)$ has a BART prior.

A naive approach is to set $r(y,x) \sim \text{BART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$. This specification has two problems. First, $r(y,x)$ will not be smooth in y so that draws from the prior and posterior of $f(y \mid x)$ will also not be smooth. The smoothness problem can be addressed by setting $r(y,x) \sim \text{SBART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ instead. It is this model that we study the theoretical properties of in Section 4.

Setting $r(y,x) \sim \text{SBART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ is still naive because of the way in which BART shrinks $r(y,x)$ towards additive models $r(y,x) = \sum_{v=1}^{V} r_v(y,x)$ where each $r_v(y,x)$ depends on a small subset of the coordinates of (y,x) (Linero and Yang, 2018; Rockova and van der Pas, 2017). In the regression setting, we often expect that an underlying regression function will have exactly this form; in the case of sparse additive models (Ravikumar et al., 2007) for example, each $r_v(y,x)$ would depend on exactly one coordinate. This type of shrinkage-towards-additivity is not appropriate for conditional density estimation due to the distinguished nature of the response y; we instead want the predictors to *interact* with y.

To see why we want to force interactions with y, consider the strictly additive function $r(y,x) = r_Y(y) + \sum_{p=1}^{P} r_p(x_p)$. If we take $\Phi(\mu) = e^{\mu}$, a massive cancellation occurs in (1) and the model reduces to $f(y \mid x) = h(y \mid x, \theta)\ \Phi\{r_Y(y)\} / \int h(\widetilde{y} \mid x, \theta)\ \Phi\{r_Y(\widetilde{y})\}\ d\widetilde{y}$, effectively eliminating the predictors from the model. More generally, any trees which do not split on Y_i will have no effect on $f(y \mid x)$. While exact cancellation is unique to the exponential link, it occurs in an approximate form for the logistic link as well. At the other extreme, SBART-DS uses a prior which favors utilizing a small number of coordinates. If y is eliminated, massive cancellation occurs irrespective of the link function, and gives $f(y \mid x) = h(y \mid x, \theta)$. Combined with a Dirichlet prior for s, this approach encodes prior information that $f(y \mid x)$ is exactly equal to $h(y \mid x, \theta)$ with high probability.

Why BART? At the point of writing down model (1), we might have chosen a different specification for $r(y,x)$, such as a Gaussian process. We find SBART or BART with targeted smoothing to be an attractive model for the following reasons. First, in our experience,

algorithms based on BART priors are surprisingly robust; we made little effort to tune the hyperparameters when designing our default prior, and have not yet run into any problems. Second, by using targeted smoothing, we ensure both that the estimated densities are smooth (this is true for both BART and SBART) and that we can control the interactions of x with y in a more transparent fashion. Third, as argued by Rockova and van der Pas (2017); Linero and Yang (2018), BART shrinks our fitted model towards a parsimonious structure consisting of low-order interactions, which are common in practice. Fourth, using the Dirichlet prior described in Section 2.4, it is particularly convenient for implementing sparsity. Lastly, computations for the BART and SBART models are straight-forward and scale favorably with the sample size; additionally, while we have not pursued this here, there is scope for very fast implementations of our model based on BART without using soft decision trees, and we see no obvious reason why such an implementation would be much slower than existing BART software for fitting semiparametric regression.

2.3 Targeted Smoothing via Random Basis Function Expansions

We use the "targeted smoothing" approach of Starling et al. (2018) to overcome the problems of the naive SBART-DS prior. They set $r(y,x) = \gamma + \sum_{m=1}^{M} g(y; x, \mathcal{T}_m, \mathcal{M}_m)$ where each leaf node is associated with a Gaussian process; that is, for fixed x, we have $g(\cdot; x; \mathcal{T}_m, \mathcal{M}_m) \sim \text{GP}\{0, \Sigma(\cdot,\cdot)\}$ where $\text{GP}\{0, \Sigma(\cdot,\cdot)\}$ denotes a mean-0 Gaussian process with covariance function $\Sigma(y, y')$.

Starling et al. (2018) consider the case where the number of unique values y takes, N_y, is small. When N_y is large this is no longer practical due to the need to store and invert an $N_y \times N_y$ matrix for all m trees. For SBART-DS we cannot guarantee that this is the case. Instead, we set $r(y,x) = \gamma + \sum_{m=1}^{M} \mathcal{B}_m(y)\ g(x; \mathcal{T}_m, \mathcal{M}_m)$ where each $\mathcal{B}_m$ is a random basis function.

To construct our approximation, consider the case where $\mathcal{T}_m$ is a non-soft decision tree. For fixed x we can write $r(y,x) = \gamma + M^{-1/2} \sum_{m=1}^{M} \mu_m\ \mathcal{B}_m(y)$ where the μ_m's are iid

Normal$(0, \sigma_\mu^2/M)$ random variables. Under mild regularity conditions on the distribution of the $\mathcal{B}_m$'s, as $M \to \infty$ a functional central limit theorem will hold and this will converge weakly to a Gaussian process with mean γ and covariance function

$$\Sigma(y, y') = \sigma_\mu^2 \mathbb{E}\{\mathcal{B}_1(y)\ \mathcal{B}_1(y')\}. \tag{2}$$

This is the same as the distribution of $r(\cdot, x)$ used by Starling et al. (2018); for fixed x, this approximation is likely to be accurate for the values of M we use due to the fact that y is one-dimensional. Rather than directly choose the basis functions $\mathcal{B}_m$, we specify $\Sigma(y, y')$ and derive a distribution for $\mathcal{B}_m$ which matches (2). We make use of the following proposition, which follows from Bochner's Theorem, and is stated for completeness.

Proposition 1. *Let $\Sigma(y, y') = \sigma_\mu^2 \delta(y - y')$ be a shift-invariant kernel with $\delta(0) = 1$. Then there exists a distribution $P(d\omega)$ such that $\Sigma(y, y') = \sigma_\mu^2 \mathbb{E}\{2\cos(\omega y + b)\cos(\omega y' + b)\}$ where $\omega \sim P(d\omega)$ and $b \sim$ Uniform$(0, 2\pi)$. Moreover, $\delta(\cdot)$ is the characteristic function of $P(d\omega)$, i.e., $\delta(t) = \int e^{it\omega}\ P(d\omega)$.*

The approach of using random Fourier features in this fashion was popularized by Rahimi and Recht (2008). It follows from Proposition 1 that we can take $\mathcal{B}_m(y) = \sqrt{2}\cos(\omega_m y + b_m)$ where $\omega_m \overset{\text{iid}}{\sim} P(d\omega)$ and $b_m \overset{\text{iid}}{\sim}$ Uniform$(0, 2\pi)$. We list some possible choices below.

- $\omega_m \sim$ Normal$(0, \rho^{-2})$ corresponds to the squared exponential covariance $\Sigma(y, y') = \sigma_\mu^2 \exp\{-(y - y')^2/(2\rho^2)\}$.
- Setting $\omega_m \sim t_\nu$ with location 0 and scale ρ^{-1} gives the Matern kernel

$$\Sigma(y, y') = \sigma_\mu^2 \frac{1}{2^{\nu/2-1}\Gamma(\nu/2)} \left(\frac{\sqrt{\nu}|y - y'|}{\rho}\right)^{v/2} K_{\nu/2}\left(\frac{\sqrt{\nu}|y - y'|}{\rho}\right)$$

  where $K_\nu(\cdot)$ is a modified Bessel function of the second kind. The exponential kernel $\Sigma(y, y') = \sigma_\mu^2 \exp\{-|y - y'|/\rho\}$ is a special case ($\nu = 1$).

- In general, $P(d\omega) = p(\omega)\ d\omega$ can be obtained from the inversion formula $p(\omega) = \frac{1}{2\pi}\int e^{-it\omega}\ \delta(t)\ dt$. For example, inverting a Cauchy kernel $\delta(t) = \{1 + t^2/\rho^2\}^{-1}$ shows that we can get a Cauchy kernel from the Laplace distribution $p(\omega) = \frac{\rho}{2}e^{-\rho|\omega|}$.

Other choices of random basis functins are also possible. For example, a referee suggested the use of local kernels as basis functions, with $\mathcal{B}_m(y; \mathfrak{s}, \mathfrak{c}) = \mathfrak{s}^{-1}\kappa\left(\frac{y-\mathfrak{c}}{\mathfrak{s}}\right)$ where κ is (say) a Gaussian kernel. While these basis functions make it more difficult to control the induced covariance function, the localization of the kernel may be more desirable than the oscillatory behavior of the Fourier basis functions. These basis functions are also easier to analyze theoretically, as the use of kernel functions dovetails nicely with the kernel-based theory used to establish results for SBART.

2.4 Default Prior Specification

In our illustrations we use the following default prior specification. Following Chipman et al. (2010), we fix the parameters $\alpha = 0.95$ and $\beta = 2$ in the prior for $\pi_{\mathcal{T}}$ and set $\mu_{m\ell} \sim \text{Normal}(0, \sigma_\mu^2/M)$. We fix $M = 50$; in general, we recommend trying multiple values of M. We set $\sigma_\mu \sim \text{Half-Cauchy}(0, 1.5)$ to learn an appropriate value of σ_μ from the data. By having mass near $\sigma_\mu = 0$, this also allows us to revert to the base model $h(y \mid x, \theta)$. To induce further shrinkage to the base model, we set $\gamma \sim \text{Normal}(1, 1)$. This has the additional benefit of making the prior prefer models for which $\Phi\{r(y, x)\}$ is close to 1, which reduces the number of latent variables we need to introduce when fitting the model by MCMC (see Section 3). We use tree-specific bandwidths τ_m which are exponentially distributed with mean 0.1. To perform variable selection we specify $s \sim \text{Dirichlet}(a/P, \ldots, a/P)$ and use a hyperprior $a/(a + P) \sim \text{Beta}(0.5, 1)$.

For targeted smoothing, we approximate the squared exponential kernel by setting $\omega_m \sim \text{Normal}(0, \rho^{-2})$. We set $\rho^2 \sim \text{Gamma}(\alpha_\rho, \beta_\rho)$ to allow the length scale to be learned from the data. As a default, we set $\alpha_\rho = 1$ and $\beta_\rho = \pi^2/4$ after scaling the Y_i's to have unit variance. This choice is based on the number of times a Gaussian process with length-scale

ρ is expected to cross 0 on the interval $(-1, 1)$: by Rice's formula, the expected number of crossings is $2/(\pi\rho)$ so that if $\rho^2 = 4/\pi^2$ the expected number of crossings is 1 (Adler et al., 2015, Exercise 2.8.23). Smaller values of ρ^2 correspond to more wiggly functions. Because our prior has positive density at 0, setting $\alpha_\rho = 1$ allows for the possibility that the function is very wiggly while defaulting to the prior belief that it is not.

In this paper, we consider two possible choices for the base model. The first is the Gaussian linear model with $h(y \mid x, \theta) = \text{Normal}(y \mid \alpha_\theta + x^\top \beta_\theta, \sigma_\theta)$ where $\text{Normal}(y \mid \mu, \sigma)$ is the density of a Gaussian random variable with mean μ and variance σ^2. We use this model with the MEPS dataset. We recommend a weakly-informative prior such as $(\alpha_\theta, \beta_\theta) \sim \text{Normal}(0, 5\text{I})$ and $\sigma_\theta \sim \text{Half-Cauchy}(0, 1)$ after the response and predictors. We can also shrink towards a semiparametric Gaussian model by setting $h(y \mid x, \theta) = \text{Normal}\{y \mid r_\theta(x), \sigma_\theta\}$ where $r_\theta(x) \sim \text{SBART}(\pi^\theta_{\mathcal{T}}, \pi^\theta_{\mathcal{M}})$ and the default prior of Linero and Yang (2018) is specified for $(\pi^\theta_{\mathcal{T}}, \pi^\theta_{\mathcal{M}})$. While the SBART base model does not add any flexibility beyond SBART-DS, it can result in faster computations if the true data generating process has Gaussian errors but a non-linear mean. By modeling the nonlinearity in the base model, fewer augmented data points are required by our MCMC algorithm. We use the SBART base model in the simulation study of Section 5.1.

3 Posterior Computation

3.1 Rejection Sampling Data Augmentation

We use a two-layer data augmentation scheme which removes both the intractable integral in the denominator of (1) and the link function $\Phi(\mu)$ from the likelihood. Our approach is based on the following method for sampling from $f(y \mid x)$.

Proposition 2. *Suppose that we sample* $Y_1, Y_2, Y_3, \ldots \overset{\text{iid}}{\sim} h(y \mid x, \theta)$ *and independently sample* $A_j \overset{\text{indep}}{\sim} \text{Bernoulli}[\Phi\{r(Y_j, x)\}]$. *Let* Z *denote the* Y_j *associated with the smallest index* $J+1$ *for which* $A_{J+1} = 1$. *Then conditional on* $\{J, A_j : 1 \leq j \leq J+1\}$, Z *is a draw from* $f(y \mid x) \propto$

$h(y \mid x, \theta)\Phi\{r(y,x)\}$ *and* $Y_1, \ldots, Y_J$ *are draws from* $\bar{f}(y \mid x) \propto h(y \mid x, \theta)[1 - \Phi\{r(y,x)\}]$.

We make use of Proposition 2 by augmenting the latent index J and the sequence of rejected points. Associated to each observation $Y_i = Y_{i0}$ we independently sample $Y_{ij} \overset{\text{iid}}{\sim} h(y \mid X_i, \theta)$ and $A_{ij} \overset{\text{indep}}{\sim} \text{Bernoulli}[\Phi\{r(Y_{ij}, X_i)\}]$ until we reach the first iteration $J_i + 1$ such that $A_{i(J_i+1)} = 1$. We then work with the augmented state $\{Y_{ij} : 1 \leq i \leq N, 0 \leq j \leq J_i\}$, which has likelihood

$$\prod_{i=1}^{N} \prod_{j=0}^{J_i} h(Y_{ij} \mid X_i, \theta) \times \prod_{i=1}^{N} \left(\Phi\{r(Y_{i0}, X_i)\} \prod_{j=1}^{J_i} [1 - \Phi\{r(Y_{ij}, X_i)\}] \right). \tag{3}$$

For more details on the derivation of this expression, see Rao et al. (2016), who consider the GP-DS model. At this stage Rao et al. (2016) propose the use of Hamiltonian Monte Carlo to sample from the posterior distribution. This is not an option for us, as the $\mathcal{T}_m$'s are discrete parameters.

3.2 Bayesian Backfitting for Probit, Logit, and Student's t_ν Links

We now apply the data augmentation strategy of Albert and Chib (1993). Suppose that $\Phi(\mu)$ is either the probit, logit, or Student's t_ν link. We can then associate to each A_{ij} from Section 3.1 a random variable

$$Z_{ij} = r(Y_{ij}, X_i) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \text{Normal}(0, \lambda_{ij}^{-1}), \qquad \lambda_{ij} \sim g(\lambda).$$

Setting $A_{ij} = I(Z_{ij} \geq 0)$ recovers the $\text{Bernoulli}[\Phi\{r(Y_{ij}, X_i)\}]$ model. This model captures the three links we consider:

- For the probit link, λ_{ij}^{-1} has a point-mass distribution at 1.
- When $\Phi(\mu) = T_\nu(\mu)$ is the Student's t link with ν degrees of freedom, $\lambda_{ij} \sim \text{Gamma}(\nu/2, \nu/2)$.
- When $\Phi(\mu) = (1 + e^{-\mu})^{-1}$ is the logistic link, $\lambda_{ij}^{-1/2}/2$ has a Kolmogorov-Smirnov

distribution (Holmes and Held, 2006).

Compared to (3), introducing the latent variables (Z_{ij}, λ_{ij}) leads to a more tractable likelihood:

$$\prod_{i=1}^{N} \prod_{j=0}^{J_i} h(Y_{ij} \mid X_i, \theta) \times \text{Normal}\{Z_{ij} \mid r(Y_{ij}, X_i), \lambda_{ij}^{-1}\} \times g(\lambda_{ij}). \tag{4}$$

After reaching expression (4) we can apply a Bayesian backfitting algorithm to update $r(y, x)$. While the Bayesian backfitting algorithm originally proposed by Chipman et al. (2010) does not account for heteroskedasticity in the Z_{ij}'s, several recent works have shown how to accommodate this (Bleich and Kapelner, 2014; Pratola et al., 2017; Linero et al., 2018). Consider the prior $\gamma \sim \text{Normal}(\mu_\gamma, \lambda_\gamma^{-1})$ and let $R_{ij} = Z_{ij} - \gamma - \sum_{m \neq k} \mathcal{B}_m(Y_{ij})\, g(X_i; \mathcal{T}_m, \mathcal{M}_m)$. When updating $\mathcal{T}_k$ it suffices to consider the backfit model

$$R_{ij} = \mathcal{B}_k(Y_{ij})\; g(X_i; \mathcal{T}_k, \mathcal{M}_k) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \text{Normal}(0, \lambda_{ij}^{-1}). \tag{5}$$

Recall that $\varphi_{k\ell}(X_i)$ is the weight associated to leaf ℓ of tree k at X_i. Let $\varphi_k(X_i)$ be a vector with ℓ^{th} entry $\varphi_{k\ell}(X_i)$. Then we can rewrite (5) as $R_{ij} = \mathcal{B}_k(Y_{ij})\; \varphi_k(X_i)^\top \mu_k + \epsilon_{ij}$ or, in multivariate form, $\boldsymbol{R} \sim \text{Normal}(\mathcal{B}_k \mu, \Lambda^{-1})$ where the rows of $\mathcal{B}_k$ correspond to $\mathcal{B}_k(Y_{ij})\varphi_k(X_i)^\top$ and Λ is diagonal with entries λ_{ij}. If $\mu_k \sim \text{Normal}(0, \lambda_\mu^{-1})$ where $\lambda_\mu = M/\sigma_\mu^2$ then it follows from standard properties of the multivariate Gaussian distribution that

$$\begin{aligned} \boldsymbol{R} &\sim \text{Normal}(0, \Lambda^{-1} + \mathcal{B}_k \mathcal{B}_k^\top / \lambda_\mu)] \qquad \text{and} \\ [\mu_k \mid \boldsymbol{R}] &\sim \text{Normal}(V \mathcal{B}_k^\top \Lambda \boldsymbol{R}, V) \qquad \text{where } V = (\mathcal{B}_k^\top \Lambda \mathcal{B}_k + \lambda_\mu \text{I})^{-1}. \end{aligned} \tag{6}$$

After applying the Woodbury matrix identity and the matrix determinant lemma (Brookes, 2011, matrix identities), the likelihood of $\mathcal{T}_k$ after integrating out μ_k is given by

$$(2\pi)^{-N/2} \prod_{i=1}^{N} \lambda_i^{1/2} \det(\text{I} + \mathcal{B}_k^\top \Lambda \mathcal{B}_k / \lambda_\mu)\; \exp\left[-\frac{1}{2}\left\{\boldsymbol{R}^\top \Lambda \boldsymbol{R} - \delta^\top (\text{I} + \mathcal{B}_k^\top \Lambda \mathcal{B}_k / \lambda_\mu)^{-1} \delta\right\}\right] \tag{7}$$

where $\delta = \mathcal{B}_k^\top \Lambda \boldsymbol{R}$. The value of (7) is that it avoids taking the determinant of and inverting

Algorithm 1 An iteration of the data augmentation algorithm for SBART-DS

1. For $i = 1, \ldots, N$, set $Y_{i0} = Y_i$ and sample $Y_{i1}, Y_{i2}, \ldots \sim h(y \mid X_i, \theta)$ and $A_{i1}, A_{i2}, \ldots \sim \text{Bernoulli}[\Phi\{r(Y_{ij}, X_i)\}]$ until $A_{i(J_i+1)} = 1$. Retain the samples $Y_{i0}, \ldots, Y_{iJ_i}$.
2. Make an update to θ which leaves the full conditional $\pi(\theta \mid -) \propto \pi(\theta) \prod_{i,j} h(Y_{ij} \mid X_i, \theta)$ invariant.
3. Sample $Z_{ij} \sim f(z \mid \mu_{ij})$ truncated to $(0, \infty)$ for $j = 0$ and $(-\infty, 0)$ for $j > 0$ where $\mu_{ij} = r(Y_{ij}, X_i)$ and $f(z \mid \mu_{ij})$ is a normal, logistic, or Student's t_ν distribution with location μ_{ij} and scale 1 for the probit, logit, and T_ν links respectively.
4. Sample λ_{ij} from its full conditional given Z_{ij} for all $1 \leq i \leq N$ and $0 \leq j \leq J_i$.
   - For the probit link, $\lambda_{ij} \equiv 1$.
   - For the Student's t_ν link, $\lambda_{ij} \sim \text{Gam}\{(\nu+1)/2, (\nu + [Z_{ij} - r(Y_{ij}, X_i)]^2)/2\}$.
   - For the logit link, sample λ_{ij}^{-1} using the rejection sampling algorithm of Holmes and Held (2006).
5. For $m = 1, \ldots, M$ update $(\mathcal{T}_m, \mathcal{M}_m, \mathcal{B}_m)$ using the Metropolis-Hastings algorithm given in Algorithm 2.

the $N \times N$ matrix $\Lambda^{-1} + \mathcal{B}_k \mathcal{B}_k^\top / \lambda_\mu$. The marginal likelihood $L_k(\mathcal{T}, \mathcal{B})$ given by (7) is used to update both the tree topology $\mathcal{T}_k$ and the random basis function $\mathcal{B}_k(y)$ using Metropolis-Hastings.

Our final MCMC scheme is summarized in Algorithm 1, which calls Algorithm 2 to update $(\mathcal{T}_k, \mathcal{B}_k, \mathcal{M}_k)$. The Markov transition function $Q(\mathcal{T}_k \to \mathcal{T}')$ used to propose new tree topologies is a mixture of the `BIRTH`, `DEATH`, and `CHANGE` proposals described by Chipman et al. (1998) and a `PRIOR` proposal which samples $\mathcal{T}'$ from the prior.

4 Theoretical Results

4.1 Rates of Convergence

We show that SBART-DS attains close to the minimax-optimal concentration rate for $(P+1)$-dimensional functions $r(y, x)$ in the high-dimensional sparse setting. All proofs are deferred to the appendix. We consider the case where $r(y, x)$ depends on only D co-

Algorithm 2 Metropolis-Hastings update for $(\mathcal{T}_k, \mathcal{M}_k, \mathcal{B}_k)$

1. Compute $\boldsymbol{R}$ as in (5).
2. Propose a tree $\mathcal{T}'$ from a Markov transition kernel $Q(\mathcal{T}_k \to \mathcal{T}')$.
3. Set $\mathcal{T}_k = \mathcal{T}'$ with probability
$$\min\left\{\frac{\pi_{\mathcal{T}}(\mathcal{T}')\ L_k(\mathcal{T}', \mathcal{B}_k)\ Q(\mathcal{T}' \to \mathcal{T}_k)}{\pi_{\mathcal{T}}(\mathcal{T}_k)\ L_k(\mathcal{T}_k, \mathcal{B}_k)\ Q(\mathcal{T}_k \to \mathcal{T}')}, 1\right\}.$$
Otherwise, do not change $\mathcal{T}_k$.
4. Sample a basis function $\mathcal{B}'(y) = \sqrt{2}\cos(\omega' y + b')$ by sampling $\omega' \sim P(d\omega)$ and $b' \sim \text{Uniform}(0, 2\pi)$. Then set $\mathcal{B}_k = \mathcal{B}'$ with probability
$$\min\left\{\frac{L_k(\mathcal{T}_k, \mathcal{B}')}{L_k(\mathcal{T}_k, \mathcal{B}_k)}, 1\right\}.$$
Otherwise, do not change $\mathcal{B}_k$.
5. Sample $\mu_k \sim \text{Normal}(V\mathcal{B}_k^\top \Lambda \boldsymbol{R}, V)$ where $V = (\mathcal{B}_k^\top \Lambda \mathcal{B}_k + \mathrm{I}/\lambda_\mu)^{-1}$ and $\lambda_\mu = M/\sigma_\mu^2$.

ordinates of $(y, x)^\top$ where the relevant coordinates are unknown and must be learned from the data. Following Pati et al. (2013) we study concentration with respect to the integrated Hellinger distance. Let $H(f, f_0) = [\int\{\sqrt{f_0(y \mid x)} - \sqrt{f(y \mid x)}\}^2\ dy\ F_X(dx)]^{1/2}$ denote the *F_X-integrated Hellinger distance* between $f_0(y \mid x)$ and $f(y \mid x)$. The covariates X_i are assumed to be iid from F_X, which is not assumed to be known. We similarly define an *F_X-integrated Kullback-Leibler* neighborhood. Define $K(f_0, f) = \int f_0 \log \frac{f_0}{f}\ dy\ dF_X$ and $V(f_0, f) = \int f_0 \left(\log \frac{f_0}{f}\right)^2\ dy\ dF_X$. Then the integrated Kullback-Leibler neighborhood is given by $K(\epsilon) = \{f : K(f_0, f) \leq \epsilon^2 \text{ and } V(f_0, f) \leq \epsilon^2\}$. Let $\mathcal{D}_n$ denote the data $\{X_i, Y_i : i = 1, \ldots, n\}$ and let Π denote a prior distribution on r and additional hyperparameters. We say that the posterior has a convergence rate of at least ϵ_n if there exists a constant $C > 0$ such that $\Pi\{H(f_r, f_0) \geq C\epsilon_n \mid \mathcal{D}_n\} \to 0$ in probability. To simplify the theoretical results, we assume that X_i and Y_i take values in $[0, 1]^{P+1}$. We additionally make the following assumptions about the true data generating process F_0.

Condition F (on F_0): The true conditional density $f_0(y \mid x)$ can be written as $f_{r_0}(y \mid x)$ for some $r_0 \in C^{\alpha,R}([0,1]^{P+1})$ where $C^{\alpha,R}([0,1]^{P+1})$ is the ball of radius R in the space of α-Hölder smooth functions on $[0,1]^{P+1}$, where $f_r(y \mid x)$ is defined as

$$f_r(y \mid x) = \frac{h(y)\ \Phi\{r(y,x)\}}{\int h(\widetilde{y})\ \Phi\{r(\widetilde{y},x)\}\ d\widetilde{y}}$$

for some density $h(y)$ on $[0,1]$. Additionally, we can write $r_0(y,x) = \widetilde{r}(y, x_{\mathcal{S}})$ where $x_{\mathcal{S}} = \{x_j : j \in \mathcal{S}\}$ and $\mathcal{S}$ is a subset of $\{1,\ldots,P\}$ of cardinality $D-1$. That is, $r_0(y,x)$ depends on at most D coordinates of $(y,x)^\top$. The number of predictors $P \equiv P_n$ depends on n but is such that $\log(P+1) \leq C_\eta n^\eta$ for some $\eta \in (0,1)$.

Remark 1. For simplicity, we consider $\widetilde{r}$ and $\mathcal{S}$ to be independent of n; in particular, we do not consider D diverging with n. As noted in Section 2, there will exist some r_0 such that $f_0 = f_{r_0}$ provided that $R(y,x) = f(y \mid x)/h(y)$ is non-zero and $\sup_y R(y,x) < \infty$ for all x. When $h(y) = 1$, r_0 being continuous on $[0,1]^{P+1}$ implies that $C^{-1} \leq f_0(y \mid x) \leq C$ for some constant $C > 0$, i.e., $f_0(y \mid x)$ is bounded and bounded away from 0.

Condition L (on Φ): The link function $\Phi(\mu)$ is strictly increasing and is the cumulative distribution function of a random variable Z which is symmetric about 0 and has density $\phi(\mu)$ satisfying $\phi(\mu)/\Phi(\mu) \leq \mathcal{K}$ for all μ and some constant $\mathcal{K}$.

Remark 2. We show in the appendix that Condition L holds for the logit ($\mathcal{K} = 1$) and t_ν ($\mathcal{K} = \sqrt{\nu}$) links, but fails for the probit link.

Condition P (on Π): The function r is given an SBART$(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ prior with M trees conditional on $(\pi_{\mathcal{T}}, \pi_{\mathcal{M}}, M)$. Additionally, the prior Π satisfies the following conditions.

(P1) There exists positive constants (C_{M1}, C_{M2}) such that the prior on the number of trees M in the ensemble is $\Pi(M = t) = C_{M1} \exp\{-C_{M2} t \log t\}$.

(P2) A single bandwidth $\tau_m \equiv \tau$ is used and its prior satisfies $\Pi(\tau \geq x) \leq C_{\tau 1} \exp(-x^{C_{\tau 2}})$

and $\Pi(\tau^{-1} \geq x) \leq C_{\tau 3} \exp(-x^{C_{\tau 4}})$ for some positive constants $C_{\tau 1}, \ldots, C_{\tau 4}$ for all sufficiently large x, with $C_{\tau 2}, C_{\tau 4} < 1$. Moreover, the density of τ^{-1} satisfies $\pi_{\tau^{-1}}(x) \geq C_{\tau 5} e^{-C_{\tau 6} x}$ for large enough x and some positive constants $C_{\tau 5}$ and $C_{\tau 6}$.

(P3) The prior on the splitting proportions is $s \sim \text{Dirichlet}(a/P^{\xi}, \ldots, a/P^{\xi})$ for some $\xi > 1$ and $a > 0$.

(P4) The $\mu_{m\ell}$'s are iid from a density $\pi_{\mu}(\mu)$ such that $\pi_{\mu}(\mu) \geq C_{\mu 1} e^{-C_{\mu 2}|\mu|}$ for some coefficients $C_{\mu 1}, C_{\mu 2}$. Additionally, there exists constants $C_{\mu 3}, C_{\mu 4}$ such that $\Pi(|\mu_{m\ell}| \geq t) \leq C_{\mu 3} \exp\{-t^{C_{\mu 4}}\}$ for all t.

(P5) Let D_m denote the depth of tree $\mathcal{T}_m$. Then $\Pi(D_m = k) > 0$ for all $k = 0, 1, \ldots, 2D$ and $\Pi(D_m > d_0) = 0$ for some $d_0 \geq D$.

(P6) The gating function $\psi : \mathbb{R} \to [0, 1]$ of the SBART prior is such that $\sup_x |\psi'(x)| < \infty$ and the function $\rho(x) = \psi(x)\{1 - \psi(x)\}$ is such that $\int \rho(x)\, dx > 0$, $\int |x|^m \rho(x)\, dx < \infty$ for all integers $m \geq 0$, and $\rho(x)$ can be analytically extended to some strip $\{z : |\Im(z)| \leq U\}$ in the complex plane.

Remark 3. Conditions other than Condition P might also be used. Recent work of Rockova and van der Pas (2017), for example, studies concentration results for BART using different sets of conditions, and the conditions overall are weaker than the conditions presented here. A downside of these results is they apply only when non-smooth decision trees are used, which induces non-smooth densities. Condition P2 holds when τ is given an inverse-gamma prior truncated from above, while Condition P4 holds when the $\mu_{m\ell}$'s are given a Laplace prior, although as noted by Linero and Yang (2018) this could potentially be weakened to allow a Gaussian prior with a hyperprior on σ_{μ} (we do not pursue this here). Condition P6 holds for the logistic gating function $\psi(x) = \{1 + \exp(-x)\}^{-1}$, which is used by default. Condition P5 holds if we truncate the prior of Chipman et al. (2010) at some large d_0, which is extremely unlikely to affect the MCMC in practice. Hence, satisfying P5 is not a practical

concern. Condition P1 is problematic because BART implementations do not use a prior on M. In practice, we find that selecting M by cross validation is more reliable than using a prior; we recommend either (i) using cross validation to select M or (ii) fixing M at a default value such as $M = 200$ (recommended by Chipman et al., 2010) or $M = 50$ (used here).

Theorem 1. *Suppose that Condition L, Condition F, and Condition P hold. Then there exists a positive constant C such that $\Pi\{H(f_0, f_r) \geq C\epsilon_n \mid \mathcal{D}_n\} \to 0$ in probability, where $\epsilon_n = n^{-\alpha/(2\alpha+D)}(\log n)^t + \sqrt{\frac{D\log(P+1)}{n}}$ and $t = \alpha(D+1)/(2\alpha+D)$.*

We prove Theorem 1 by checking (a)—(c) in Proposition 3, which are analogous to conditions of Ghosal et al. (2000).

Proposition 3. *Let Π denote a prior for a conditional density $f(y \mid x)$ and let ϵ_n and $\bar{\epsilon}_n$ be sequences of positive numbers such that $\bar{\epsilon}_n, \epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, and $\epsilon_n \leq \bar{\epsilon}_n$. Let $N(\epsilon, \mathcal{F}, H)$ denote the ϵ-covering number of $\mathcal{F}$ with respect to H (i.e., the number of balls of radius ϵ required to cover $\mathcal{F}$). Suppose that there exist positive constants C, C_N such that for all sufficiently large n there exist sets of conditional densities $\mathcal{F}_n$ satisfying the following conditions:*

(a) Entropy Bound: $\log N(\bar{\epsilon}_n, \mathcal{F}_n, H) \leq C_N n\bar{\epsilon}_n^2$.

(b) Support Condition: $\Pi(\mathcal{F}_n^c) \leq \exp\{-(C+4)n\epsilon_n^2\}$.

(c) Prior Thickness: $\Pi\{f \in K(\epsilon_n)\} \geq \exp(-Cn\epsilon_n^2)$.

Then $\Pi\{H(f_0, f) \geq A\bar{\epsilon}_n \mid \mathcal{D}_n\} \to 0$ in probability for some constant $A > 0$.

The proof that our SBART prior satisfies these conditions is similar to the proof of Theorem 3.1 of van der Vaart and van Zanten (2008), who established posterior convergence rates for density estimation using logistic Gaussian processes. We use a collection of results of Linero and Yang (2018), who established results similar to (a)—(c) for a regression function $r \sim \text{SBART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ with respect to the supremum norm $\|r - r_0\|_\infty = \sup_{x,y} |r(y,x) -$

$r_0(y,x)|$. We then use the following lemma, which links the supremum-norm neighborhoods of r_0 with the integrated Hellinger and Kullback-Leibler neighborhoods of f_0; this allows us to convert results about the $\|\cdot\|_\infty$-norm neighborhoods to results about the integrated neighborhoods. This lemma is similar to Lemma 3.1 of van der Vaart and van Zanten (2008), but with exponential link $\Phi(\mu) = e^\mu$ replaced with a link satisfying Condition L.

Lemma 1. *Let $\Phi(\mu)$ be a link function satisfying Condition L. Then for any measurable functions $u, v : [0,1]^{P+1} \to \mathbb{R}$ we have the following:*

- $H^2(f_u, f_v) \leq \mathcal{K}^2 \|u - v\|_\infty^2 \exp(\mathcal{K}\|u - v\|_\infty)$;
- $K(f_u, f_v) \lesssim \|u - v\|_\infty^2 \exp(\mathcal{K}\|u - v\|_\infty)(1 + 2\mathcal{K}\|u - v\|_\infty)$; *and*
- $V(f_u, f_v) \lesssim \|u - v\|_\infty^2 \exp(\mathcal{K}\|u - v\|_\infty)(1 + 2\mathcal{K}\|u - v\|_\infty)^2$.

The expression $a \lesssim b$ here denotes that $a \leq Cb$ for some constant C depending only on $\mathcal{K}$.

More generally, one expects that Theorem 1 can be improved to allow for additive decompositions $r_0(y,x) = \sum_{j=1}^J r_{0j}(y,x)$ where the r_{0j}'s are functions which are D_j-sparse and α_j-Hölder continuous. Results in this framework (Linero and Yang, 2018; Rockova and van der Pas, 2017; Yang and Tokdar, 2015) suggest that we should be able to obtain a rate $\epsilon_n = \sum_{j=1}^J n^{-\alpha_j/(2\alpha_j + D_j)} \log(n)^{t_j} + \sqrt{n^{-1} D_j \log(P+1)}$, which is a substantial improvement on Theorem 1. One difficulty with extending these results is that Condition P2 only allows a single bandwidth, while different τ's will be optimal for different α_j's. Unlike the nonparametric regression setting, however, it is unclear how one would interpret the additivity assumption for SBART-DS. We leave examining the additive framework to future work.

4.2 Topological Support

Condition F implies the restrictive condition $\sup_y |\log f_0(y \mid x)/h(y)| < \infty$ for all x. For example, when $h(y)$ is a uniform distribution, this implies that $C^{-1} \leq f_0(y \mid x) \leq C$ for some constant C, prohibiting for example the Beta(α, α) distribution for all $\alpha \neq 1$.

Given the restrictiveness of Condition F, it is desirable to understand which densities can be estimated consistently (at any rate). This is intimately connected with the the Kullback-Leibler (KL) support $K(\Pi) = \{f_0 : \Pi[K(f_0, f_r) < \epsilon] > 0 \text{ for all } \epsilon > 0\}$ (Ghosal et al., 1999). Roughly speaking, f_0 will be consistently estimable if (i) $f_0 \in K(\Pi)$ and (ii) there exists a sieve satisfying the conditions of Proposition 3 for some sequence $\epsilon_n \downarrow 0$ with $n\epsilon_n^2 \to \infty$. As (ii) is established in the proof of Theorem 1, it suffices to characterize $K(\Pi)$.

In the supplementary material we establish the following result, which shows that $f_0 \in K(\Pi)$ for a substantially wider class of densities than suggested by Condition F. In particular, the KL support contains all uniformly bounded smooth densities, and a very large class of unbounded densities as well.

Theorem 2. *Suppose that $f_0(y \mid x)$ is α-Hölder smooth on $(0,1) \times [0,1]^P$ for some $\alpha > 0$. Suppose that there exists a constant B such that, for every $\delta \in (0,1)$, there exists an $a \leq \delta$ and $b \geq 1 - \delta$ such that*

(C1) $\sup_x a f_0(a \mid x) \leq \delta$ *and* $\sup_x (1-b) f_0(b \mid x) \leq \delta$;

(C2) $\sup_x |f_0(a \mid x) - \inf_{y \leq a} f_0(y \mid x)| \leq B$ *and* $\sup_x |f_0(b \mid x) - \inf_{y \geq b} f_0(y \mid x)| \leq B$; *and*

(C3) $\iint f_0(y \mid x) \, |\log f_0(y \mid x)| \, dy \, F_X(dx) < \infty$.

Then $f_0 \in K(\Pi)$ if Π satisfies Condition L and Condition P with $h(y) \equiv 1$.

Remark 4. Conditions C1, C2, and C3 are all mild. For example, all bounded continuous densities on $(0,1) \times [0,1]^P$ can be shown to satisfy C1 and C2. A large class of unbounded densities can also be shown to satisfy C1 and C2; for example, we show in the supplementary material that the conditional density $\text{Beta}\{y \mid \alpha(x), \beta(x)\}$ satisfies C1, C2, and C3 for all continuous choices of $\log \alpha(x)$ and $\log \beta(x)$. Additionally, the proof technique used to prove Theorem 2 can be used to show that $f_0 \in K(\Pi)$ for many discontinuous densities as well; for example, conditional densities which are simple functions $f_0(y \mid x) = \sum_{j=1}^{J} \pi_j \, I\{(y,x) \in A_j\}$ can be shown to lie in $K(\Pi)$.

4.3 Beyond the Naive Model

For simplicity, Section 4.1 and Section 4.2 consider a naive SBART-DS model rather than our targeted smoothing variant. Beyond the need to establish analogs of Theorem 2 and Lemma S.1 of Linero and Yang (2018) to accommodate targeted smoothing, there is no fundamental theoretical difference between the naive and targeted smoothing variants of SBART-DS. To illustrate this, we prove the following result in the Supplemental Material. This result takes the random basis functions to be a collection of logistic kernels $\mathcal{B}_m(y) = \exp\{(y - \mathfrak{c}_m)/\mathfrak{s}\}/[1 + \exp\{(y - \mathfrak{c}_m)/\mathfrak{s}\}]^2$ with $\mathfrak{c}_m \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$ and a prior on $\mathfrak{s}$ satisfying the same restrictions that Condition P2 places on τ.

Theorem 3. *Suppose that Condition F, Condition L, and Condition P′ (in the Supplementary Material) hold. Then the conclusion of Theorem 1 holds.*

5 Illustrations

5.1 A Simple Simulation Illustration

We now assess the performance of the SBART-DS using the simulation example described by Dunson et al. (2007). The response Y_i is sampled from a mixture model

$$Y_i \sim e^{-2x}\text{Normal}(x, 0.1^2) + (1 - e^{-2x})\text{Normal}(x^4, 0.2^2) \qquad \text{given } X_{i1} = x.$$

We set $N = 500$ and have $P - 1$ additional predictors which do not influence the response. The marginal density of the X_i's is uniform on $[0, 1]^P$. For SBART-DS we use the default prior with $M \equiv 50$ and the probit link. We do not make any attempt to tune the hyperparameters $(a, \sigma_\mu, \rho, \alpha, \beta, \gamma)$ beyond this. We take the base model to be the normal linear regression model $h(y \mid x, \theta) = \text{Normal}(y \mid \alpha_\theta + \beta_\theta^\top x, \sigma_\theta^2)$. We consider moderate dimensions P for illustrative purposes, but in higher dimensions one might wish to induce sparsity β_θ.

Figure 2 displays a scatterplot of the relationship between Y_i and X_{i1} as well as the

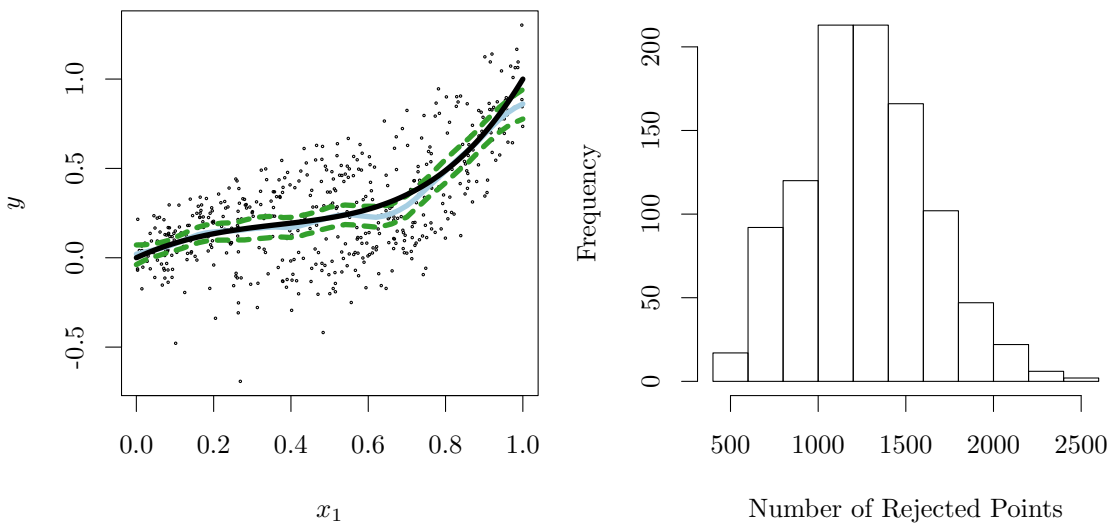

Figure 2: (Left) Plot of realized values of X_{i1} against Y_i for a single replication of the experiment, with solid black line indicating the true mean, light blue line indicating the estimated posterior mean, and the dashed green lines indicating 95% credible bands for the mean function. (Right) The posterior distribution of the number of rejected points estimated via Markov chain Monte Carlo.

posterior mean and credible band for the function $r(x) = \mathbb{E}(Y_i \mid X_i = x)$ with $P = 5$. We compare SBART-DS to a Dirichlet process mixture model described by Jara et al. (2011) as implemented in the function `DPcdensity` in the `DPpackage` package in `R`. This model uses the joint specification

$$(X_i, Y_i)^\top \sim \int \text{Normal}\left\{ \begin{pmatrix} x \\ y \end{pmatrix} \mid \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \right\} \, dG(\mu, \Sigma),$$

where $G \sim \text{DP}(\alpha G_0)$ is a Dirichlet process with a normal-inverse-Wishart base measure $G_0 \equiv \text{Normal}(\mu \mid m, \kappa_0 \Sigma)\ \text{IW}(\Sigma \mid \nu, \Psi)$. The conditional density of $[Y_i \mid X_i = x]$ can be estimated from an infinite mixture model as

$$f(y \mid x) = \sum_{k=1}^{\infty} \omega_k(x)\ \text{Normal}(y \mid \mu_{y|x}, \Sigma_{y|x})$$

where $\omega_k(x) \propto \pi_k \text{Normal}(x \mid \mu_x, \Sigma_{xx})$, $\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$, and $\Sigma_{y|x} = \Sigma_{yy} -$

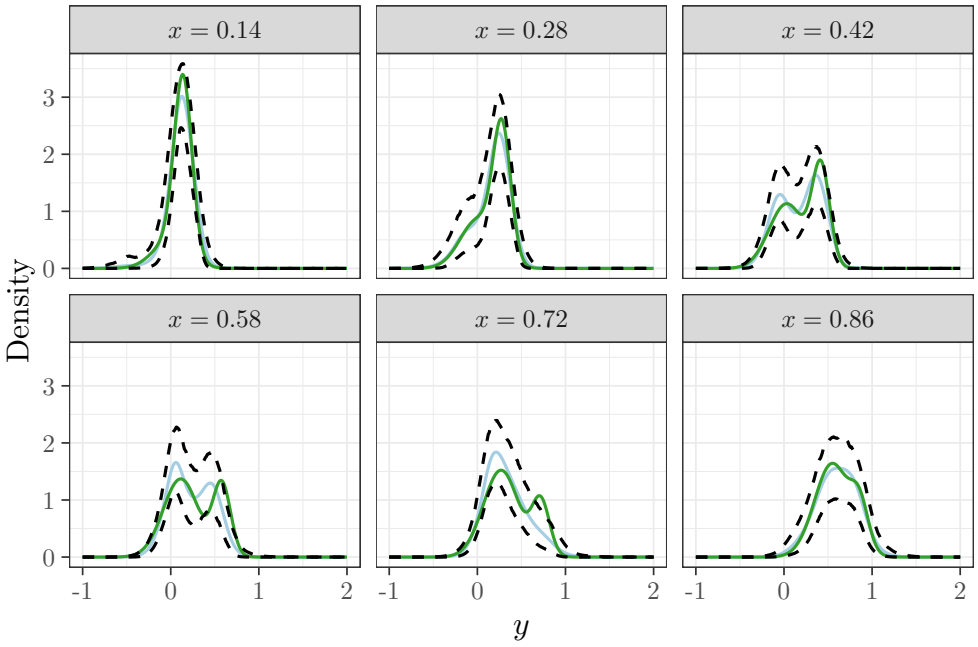

Figure 3: Posterior mean (blue), 95% credible bands for the density (dashed black) and true density function (green) for the simulated data, for the values $X_{i1} \in \{0.14, 0.28, 0.42, 0.58, 0.72, 0.86\}$.

$\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$. We use the same prior specification as Jara et al. (2011) but with a larger value of ν to accommodate the fact that $\nu > P - 1$ is required.

Figure 3 shows the fitted density for several fixed values of X_{i1} with all other predictors frozen at the value $X_{ij} = 0.5$ (as these predictors were correctly filtered out of the model, their particular value is irrelevant). To compute the density, we evaluated the numerator $h(y \mid x, \theta)\ \Phi\{r(y, x)\}$ on a grid of y values and applied the trapezoidal rule with these evaluations to approximate the denominator $\int h(y \mid x, \theta)\ \Phi\{r(y, x)\}\ dy$. We see SBART-DS successfully captures variability in the location, shape, and scale of the densities, and produces 95% credible bands which accurately account for uncertainty in the estimates. SBART-DS also captures the mean response accurately (left panel of Figure 2). Additionally, the number of rejected points is not prohibitively large, and fitting SBART-DS was faster than fitting the Dirichlet process mixture model using `DPcdensity`.

5.2 Comparison to Competing Methods

We now conduct an in-depth simulation experiment to assess the merits of SBART-DS relative to competing methods. In this simulation we took the baseline model $h(y \mid x, \theta)$ to itself correspond to an SBART model with the default prior described by Linero and Yang (2018).

Simulation Settings We consider a variety of different ground truths for comparison. All settings take $X_i \sim \text{Uniform}([0,1]^P)$, except for the ZK setting which has correlated predictors.

- **KD** We use a modified version of the simulation experiment of Kundu and Dunson (2014). Specifically, we take $Y_i \sim \text{Normal}(3e^{-Z_i}, Z_i + 0.05)$ where $Z_i = \sqrt{X_{i1}X_{i2}}$.

- **SN** We set $[Y_i \mid X_i = x] \sim \text{Skew-Normal}\{\xi(x), \omega(x), \alpha(x)\}$ where $\text{Skew-Normal}(\xi, \omega, \alpha)$ refers to the *skewed normal distribution* with location ξ, scale ω, and *slant parameter* α (Azzalini, 2013, Section 2.1). We take $\omega(x) \equiv 2$, $\xi(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ to correspond to Friedman's test function (Friedman, 1991; Chipman et al., 2010). We set $\alpha(x) = \{\xi(x) - 14.7\}/5.10$, which standardizes $\alpha(X_i)$ to have mean 0 and unit variance. This choice of $\alpha(x)$ allows for both positive and negative skewness, with the skewness increasing as a function of the mean.

- **DP** The simulation setting from Dunson et al. (2007) considered in Section 5.1.

- **ZK** We modify the simulation experiment of Zhu and Kosorok (2012), taking $[e^{Y_i} \mid X_i = x] \sim \text{Gamma}\{\alpha(x), \beta(x)\}$ with $\beta(x) \equiv 1$ and $\alpha(x) = 0.5 + 0.3\sum_{j=1}^{5} |x_j|$. We take $X_i \sim \text{Normal}(0, \Sigma)$ with $\Sigma_{jk} = \rho^{|i-j|}$ with $\rho = 0.75$ assess whether SBART-DS is robust to correlation structure in the predictors.

We consider differing values of P to assess robustness to the number of covariates, with $P \in \{5, 20, 100\}$. All cases use $N = 500$.

Competing Methods We compare SBART-DS to the probit stick-breaking (PSBP) approach of Chung and Dunson (2009), the joint Dirichlet process mixture model (Joint-DP) described by Jara et al. (2011), and a random forest conditional density estimation (RFCDE) algorithm proposed by Pospisil and Lee (2018) which constructs a conditional density estimate using random forests (Breiman, 2001). All methods use default hyperprior/hyperparameter settings. PSBP is implemented using `Matlab` code available at `github.com/david-dunson/probit-stick-breaking`, Joint-DP is implemented with the `DPcdensity` function in `DPpackage` in `R`, and RFCDE is implemented using the `RFCDE` package in `R`.

Evaluation Criteria Methods are compared according to the following criteria.

- **TV** The integrated total variation distance of the fitted density from the truth, $\text{TV}(f_0, \widehat{f}) = \int\int |f_0(y \mid x) - \widehat{f}(y \mid x)| \, dy \, F_X(dx)$. This integral is approximated numerically as $N^{\star -1} \sum_{i=1}^{N^\star} \int |f_0(y \mid X_i^\star) - \widehat{f}(y \mid X_i^\star)| \, dy$, where $\{X_i^\star\}_{i=1}^{N^\star}$ consists of heldout covariates sampled from the model and the dy integral is approximated numerically. We also consider nTV, a normalized version of TV such that the best performing method has an average TV of 1.

- **Coverage** For PSBP and SBART-DS, we examine the average coverage of posterior credible intervals for the quantiles $Q_{0.25}(X_i)$, $Q_{0.5}(X_i)$, and $Q_{0.75}(X_i)$, where $Q_\alpha(x)$ denotes the α^{th} quantile of Y_i given $X_i = x$. Quantiles are computed numerically from the density estimates at each MCMC iteration. Credible intervals are not available for RFCDE, and while they are in principle available for Joint-DP they are not readily accessible from the output of `DPpackage`.

- **RMSE** For $\alpha \in \{0.25, 0.5, 0.75\}$ we examine the root mean-squared error in estimating the quantiles $Q_\alpha(x)$, $\text{RMSE}_\alpha = \left\{ \int |Q_\alpha(x) - \widehat{Q}_\alpha(x)|^2 \, F_X(dx) \right\}^{1/2}$. For PSBP and SBART-DS, we take $\widehat{Q}_\alpha(x)$ to be the posterior mean of $Q_\alpha(x)$, while Joint-DP and RFCDE construct $\widehat{Q}_\alpha(x)$ from the point estimator of the conditional densities.

P	Method	nTV	$Q_{0.25}$		$Q_{0.5}$		$Q_{0.75}$	
			Coverage	RMSE	Coverage	RMSE	Coverage	RMSE
5	Joint DP	1.34	—	0.08	—	0.07	—	0.06
	PSBP	**1.00**	0.75	0.06	0.72	0.05	0.84	**0.04**
	RFCDE	1.89	—	0.08	—	0.07	—	0.06
	SBART-DS	1.09	**0.92**	**0.04**	**0.95**	**0.04**	**0.92**	0.04
20	Joint DP	2.31	—	0.15	—	0.12	—	0.08
	PSBP	**1.00**	0.73	0.06	0.71	0.05	0.85	**0.03**
	RFCDE	2.12	—	0.14	—	0.09	—	0.08
	SBART-DS	1.07	**0.93**	**0.04**	**0.96**	**0.04**	**0.93**	0.04
100	RFCDE	2.36	—	0.21	—	0.16	—	0.15
	SBART-DS	**1.00**	**0.92**	**0.04**	**0.95**	**0.04**	**0.93**	**0.04**

Table 1: Simulation results for the DP setting.

Other Details The simulation was replicated 200 times for each P, setting, and method. Results for $P = 100$ are missing for Joint-DP and PSBP; Joint-DP was impractically slow for this large P while the PSBP software produced errors. As the software for Joint-DP was quite slow, we performed less than 200 simulations, and replicated until its performance relative to the other approaches was clear.

Results We present partial results here, with complete results deferred to the Supplementary Material. Figure 4 gives the TV and coverage for each method, setting and P, and the average coverage of credible intervals for $Q_{0.75}(X_i^\star)$. For all settings and P, the best performing method by TV is SBART-DS by a considerable margin, with the single exception of DP where PSBP performs best; this is expected since the ground truth of DP is a mixture model. In terms of TV and RMSE in predicting the conditional quantiles, we also see that SBART-DS is highly robust to the inclusion of irrelevant predictors, with the performance varying negligibly as P is increased. SBART-DS also performs very well in terms of coverage, attaining close to the nominal coverage rate for all settings. By comparison, PSBP does not attain close to nominal coverage even on DP.

Detailed results for DP and SN are given in Table 1 and Table 2, with similar tables

P	Method	nTV	$Q_{0.25}$		$Q_{0.5}$		$Q_{0.75}$	
			Coverage	RMSE	Coverage	RMSE	Coverage	RMSE
5	Joint DP	2.05	—	1.15	—	1.10	—	1.32
	PSBP	3.71	0.37	2.55	0.33	2.43	0.27	2.85
	RFCDE	4.41	—	2.87	—	2.33	—	3.05
	SBART-DS	**1.00**	**0.97**	**0.53**	**0.98**	**0.52**	**0.97**	**0.53**
20	Joint DP	4.03	—	3.12	—	2.88	—	3.09
	PSBP	3.73	0.36	2.52	0.31	2.38	0.23	2.85
	RFCDE	5.11	—	4.32	—	3.29	—	4.46
	SBART-DS	**1.00**	**0.97**	**0.54**	**0.98**	**0.52**	**0.98**	**0.53**
100	RFCDE	5.58	—	5.71	—	4.55	—	5.84
	SBART-DS	**1.00**	**0.98**	**0.54**	**0.98**	**0.52**	**0.98**	**0.53**

Table 2: Simulation results for the SN setting.

for ZK and KD given in the Supplementary Material. Surprisingly, for most choices of α SBART-DS outperforms PSBP on DP; otherwise, the overall trends we observe in Figure 4 are present.

Computationally, the RFCDE method was by far the fastest, taking roughly 4 seconds to fit to the ZK ($P = 5$) setting. PSBP with 2000 MCMC iterations and our SBART-DS sampler with 10000 MCMC iterations each took roughly 30 minutes to run; we remark that both approaches used unoptimized code, and could both likely be sped up considerably. The slowest procedure was the Joint-DP, which took roughly 2 hours to fit 10000 iterations.

5.3 Analysis of MEPS Data

We apply SBART-DS to data from the Medical Expenditure Panel Survey (MEPS) from the year 2015. MEPS is an ongoing survey in the United States which collects data on families/individuals, their medical providers, and employers, with a focus on the cost and use of health care.

There is a large literature which has considered the relationship between socioeconomic status, education, and obesity. Educational attainment relates to obesity in a complex fashion, with the effect modified by the overall income of a region, gender, and other factors

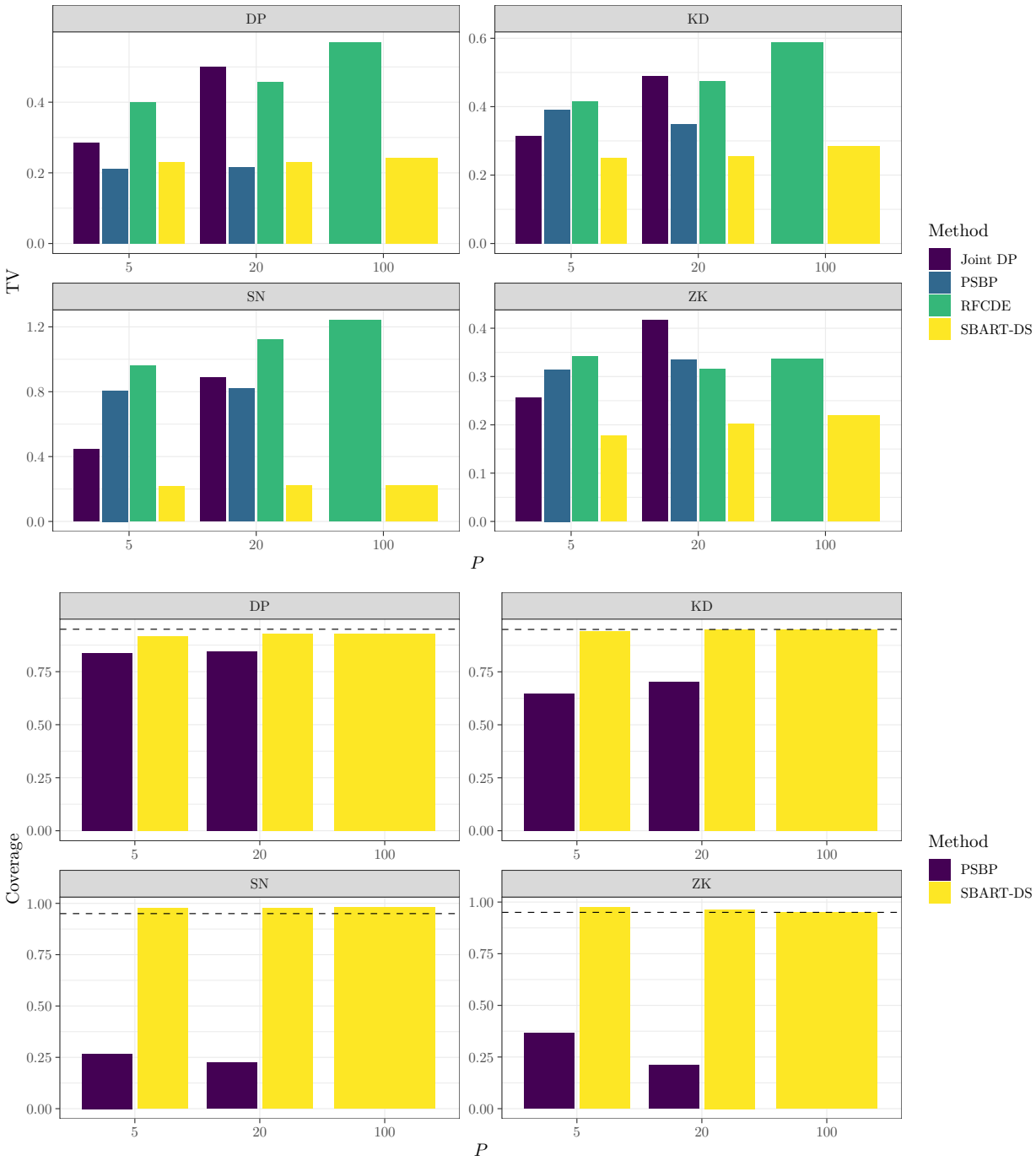

Figure 4: Simulation results for TV and coverage of $Q_{0.75}(X_i)$ for all P across all settings and methods.

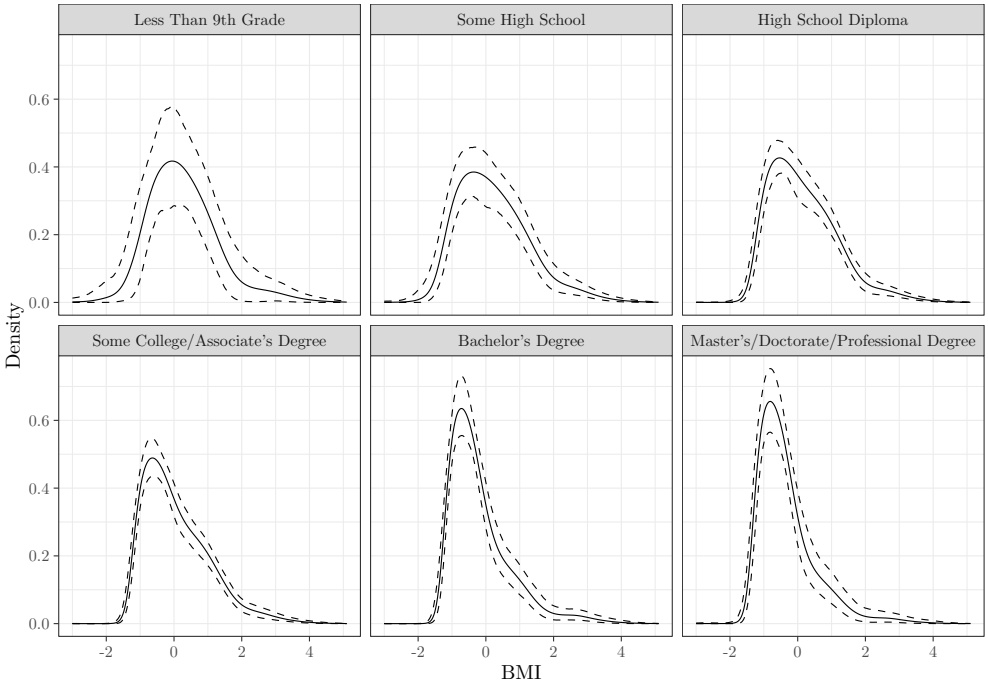

Figure 5: Density estimates and 95% credible bands for $f(y \mid x)$ for different educational levels for white women aged between 25 and 35, fixing log-income and age at their median values.

(Cohen et al., 2013). We examine this relationship on a subset MEPS consisting of responses from 1452 women aged between 25 and 35 years old, controlling for log-income (measured as a percentage of the poverty line), age, and race. Existing research predicts that higher educational attainment will be associated with lower obesity levels in this group.

In Figure 5 we display the estimated density as the level of educational attainment is varied from less-than-high-school to graduate degree for white women with all other covariates frozen at their median value. We see that as educational attainment increases the bulk of the distribution remains concentrated near 0 (the overall mean level of BMI) but goes from roughly symmetric to being highly right-skewed. The nature of this relationship is that, while the modal value of BMI is fairly stable as education level changes, highly educated women are less likely to be highly obese.

Each predictor j is associated to two coefficients: the base model coefficient $\beta_{\theta j}$ and the splitting proportion s_j. The posterior median, density, and (66%, 95%)-credible intervals

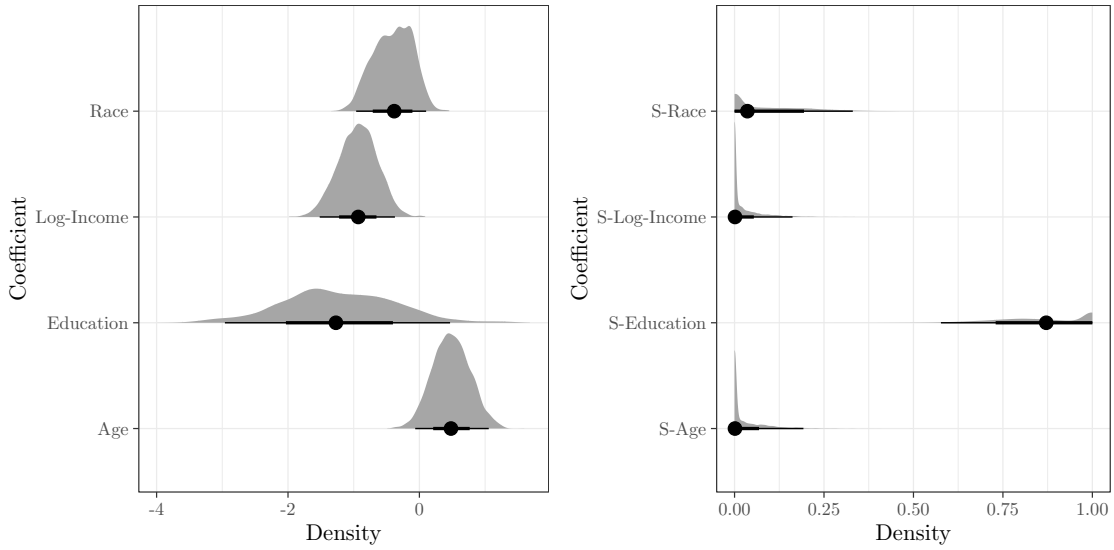

Figure 6: Left: Posterior medians, (66%, 95%)-credible intervals, and density estimates for the regression coefficients of the base model β_θ. Right: posterior medians, credible intervals, and density estimates for the splitting proportions s_j for each predictor.

are given for each coefficient in Figure 6. Interestingly, education level is the only relevant predictor in the nonparametric component $\Phi\{r(y, x)\}$ so that the overall shape of the density is primarily determined by education. Intuitively, one might expect that education is only relevant through its indirect effect on income, however our results suggest this is not the case. Log-income has a strong presence in the base model as well, while race and age have weaker effects.

6 Discussion

In this paper we proposed a new method for density regression based on Bayesian additive regression trees. SBART-DS is suitable for routine use — it has a simple default specification, strong theoretical properties, and can be fit using a tuning-parameter-free Gibbs sampling algorithm. On simulated data we illustrated how SBART-DS is robust to the presence of irrelevant variables, giving empirical support to the theoretical results supporting a faster posterior concentration rate when the rejection model $\Phi\{r(y, x)\}$ is sparse. Using data from

MEPS we showed how SBART-DS can capture the effect of education level on the conditional distribution of body mass index.

The general strategy of defining a prior using a rejection sampling model can be used to extend that SBART-DS model to other domains. For example, in subsequent work we have extended this approach to survival analysis by modeling the hazard function $\lambda(y \mid x)$ as the hazard of a thinned Poisson process $\lambda(y \mid x) = \lambda_0(y \mid x)\, \Phi\{r(y,x)\}$ (Basak et al., 2020).

A Proof of Auxiliary Results

Proof of Lemma 1. For posterity, we note that Condition L implies $\frac{d}{d\mu} \log \Phi(\mu) \leq \mathcal{K}$; integrating both sides on an interval $[L, U]$ gives

$$e^{-\mathcal{K}(U-L)} \leq \frac{\Phi(L)}{\Phi(U)} \leq \frac{\Phi(U)}{\Phi(L)} \leq e^{\mathcal{K}(U-L)}. \tag{8}$$

Let $a = \sqrt{\Phi\{u(y,x)\}}$ and $b = \sqrt{\Phi\{v(y,x)\}}$ and let $\|a\|_h^2$ denote the squared L_2-norm $\int a^2(y,x)\; h(y)\, dy$ (which implicitly depends on x). Then

$$H^2(f_u, f_v) = \int \left\| \frac{a}{\|a\|_h} - \frac{b}{\|b\|_h} \right\|_h^2 \; F_X(dx).$$

Two applications of the triangle inequality gives

$$H^2(f_u, f_v) \leq \int \left(\frac{2\|a-b\|_h}{\|a\|_h} \right)^2 \; F_X(dx) \leq 4\|1 - b/a\|_\infty^2. \tag{9}$$

The first inequality follows from the triangle inequality while the second follows from the inequality $\|a-b\|_h^2 = \int a^2(y,x)\{1 - b(y,x)/a(y,x)\}^2\; h(y) \leq \|a\|_h^2 \cdot \|1 - b/a\|_\infty^2$. Next, write $v(y,x) = u(y,x) + \Delta(y,x)$. Applying (8) and Taylor expanding the function $g(x) =$

$\sqrt{\Phi(\mu+x)/\Phi(\mu)}$ we get

$$\left|1-\sqrt{\frac{\Phi\{v(y,x)\}}{\Phi\{u(y,x)\}}}\right| \le |\Delta(y,x)|\frac{\phi\{u(y,x)+\Delta_1(y,x)\}}{2\Phi\{u(y,x)+\Delta_1(y,x)\}}\sqrt{\frac{\Phi\{u(y,x)+\Delta_1(y,x)\}}{\Phi\{u(y,x)\}}}$$
$$\le \frac{\mathcal{K}}{2}\|u-v\|_\infty \exp\left[\frac{\mathcal{K}\|u-v\|_\infty}{2}\right]$$

where $\Delta_1(y,x)$ is between 0 and $\Delta(y,x)$. Combining this with (9), $H(f_u,f_v) \le \mathcal{K}\|u-v\|_\infty \exp(\mathcal{K}\|u-v\|/2)$. By Lemma 8 of Ghosal and Van Der Vaart (2007), we have

$$K(f_u,f_v) \lesssim H^2(f_u,f_v)\left(1+\log\left\|\frac{f_u}{f_v}\right\|_\infty\right) \qquad \text{and}$$
$$V(f_u,f_v) \lesssim H^2(f_u,f_v)\left(1+\log\left\|\frac{f_u}{f_v}\right\|_\infty\right)^2.$$

Using (8) we have

$$\frac{f_u(y\mid x)}{f_v(y\mid x)} = \frac{\Phi\{u(y,x)\}\int h(\widetilde{y})\ \frac{\Phi\{v(\widetilde{y},x)\}}{\Phi\{u(\widetilde{y},x)\}}\ \Phi\{u(\widetilde{y},x)\}\ d\widetilde{y}}{\Phi\{v(y,x)\}\int h(\widetilde{y})\ \Phi\{u(\widetilde{y},x)\}\ d\widetilde{y}} \le \exp(2\mathcal{K}\|u-v\|_\infty). \tag{10}$$

Hence $\log\|f_u/f_v\|_\infty \le 2\mathcal{K}\|u-v\|_\infty$. □

Proof of Proposition 3. Consider an extended prior $\widetilde{\Pi}$ on the joint distribution of (X_i,Y_i) which places a point mass at F_X. We now have that f_r is contained in the integrated Hellinger and Kullback-Leibler neighborhoods whenever (f_r,F_X) are in the usual Hellinger and Kullback-Leibler neighborhoods of $F_0(dx,dy) = f_0(y\mid x)\,dy\,F_X(dx)$, so that the problem reduces to the setting of iid random vectors. The conditions (a)—(c) match one-to-one with the conditions of the variant of Theorem 2.1 of Ghosal et al. (2000) used by Shen et al. (2013, page 627), and hence suffice to establish the desired rate of convergence. □

We now prove that Condition L holds for the logit and t_ν links.

Proposition 4. *If $\Phi(\mu) = e^\mu/(1+e^\mu)$ then Condition L holds with $\mathcal{K}=1$. If $\Phi(\mu)=T_\nu(\mu)$ where T_ν is the distribution function of a t_ν random variable then Condition L holds with*

$\mathcal{K} = \sqrt{\nu}$. *Conversely, suppose* Z *is symmetric and has distribution function* $\Phi(\mu)$ *and* Z *is light-tailed in the sense that for all* $\mathcal{K}$ *we have* $\Pr(Z > z) < e^{-\mathcal{K}z}$ *for sufficiently large* z. *Then Condition L fails for* $\Phi(\mu)$. *In particular, Condition L fails for the probit link.*

Proof. For the logistic link it is straight-forward to check that $\phi(\mu)/\Phi(\mu) = 1 - \Phi(\mu) \leq 1$. To prove the result for the T_ν link we begin by deriving a lower bound for the survival function $\bar{T}_\nu(\mu) = \int_\mu^\infty t_\nu(x)\ dx$ for $\mu > 0$. Note that the density is $t_\nu(\mu) = c/(1 + \mu^2/\nu)^p$ where $p = (\nu + 1)/2$ and c is a normalizing constant. Define $z_0 = 1 + \mu^2/\nu$, $z = 1 + x^2/\nu$, $\theta_0 = \arcsin(1/\sqrt{z_0})$ and $\theta = \arcsin(1/\sqrt{z})$. Then after routine substitutions we have

$$\begin{aligned}\bar{T}_\nu(\mu) &= \int_{z_0}^\infty \frac{c\sqrt{\nu}}{2\sqrt{z-1}} z^{-p}\ dz = c\sqrt{\nu}\int_0^{\theta_0} \sin^{2p-2}(\theta)\ d\theta \\ &\geq c\sqrt{\nu}\int_0^{\theta_0} \sin^{2p-2}(\theta)\ \cos(\theta)\ d\theta = \frac{c\sqrt{\nu}}{2p-1} z_0^{-p+1/2}.\end{aligned}$$

Using this, the symmetry of the t_ν distribution, and substituting $2p - 1 = \nu$, we have

$$\frac{t_\nu(\mu)}{T_\nu(\mu)} \leq \frac{t_\nu(|\mu|)}{\bar{T}_\nu(|\mu|)} \leq \frac{c}{z_0^p} \cdot \frac{\sqrt{\nu}}{c z_0^{-p+1/2}} = \sqrt{\nu} z_0^{-1/2} \leq \sqrt{\nu}.$$

For the converse, set $U = 0$ and $L = -z$ in (8) to get $\Pr(Z > z) \geq \Phi(0) e^{-\mathcal{K}z}$ for some $\mathcal{K}$. In particular, $\Pr(Z > z) \geq \exp(-2\mathcal{K}z)$ for large enough z. In the case of the probit link, no such $\mathcal{K}$ can exist because $\Pr(Z > z) \leq e^{-z^2/2}$. □

The proof of Theorem 1 also requires a tail probability bound for the number of trees in the ensemble.

Proposition 5. *Let* $\Pi(M = t)$ *satisfy Condition P1. Then there exist constants* C'_{M1} *and* C'_{M2} *such that* $\Pi(M \geq t) \leq C'_{M1} \exp\{-C'_{M2} t \log t\}$.

Proof. By the geometric series formula we have

$$\Pi(M \geq t) = C_{M1} \sum_{k=t}^\infty e^{-C_{M2} k \log k} \leq \frac{C_{M1} \exp\{-C_{M2} t \log t\}}{1 - t^{-C_{M2}}} = \frac{C_{M1} \exp\{-C_{M2}(t-1) \log t\}}{t^{C_{M2}} - 1}.$$

For $t > 2 \vee 2^{1/C_{M2}}$ this gives $\Pi(M \geq t) \leq C_{M1} \exp\{-(C_{M2}/2)t \log t\}$. The result follows by taking $C'_{M2} = C_{M2}/2$ and C'_{M1} to be the maximum of C_{M1} and $\exp\{(C_{M2}/2)t \log t\}$ for $t \leq 2 \vee 2^{1/C_{M2}}$. □

B Proof of Theorem 1

For completeness, we state two results of Linero and Yang (2018) which will be used in the proof. These two propositions capture the features of SBART that make it useful in high-dimensional sparse settings with smooth regression functions.

Proposition 6. *Suppose that Condition F and Condition P are satisfied and* $t \geq \alpha(D+1)/(2\alpha + D)$. *Then there exist constants* B *and* C *independent of* (n, P) *such that for all sufficiently large* n *the prior satisfies*

$$\Pi(\|r - r_0\| \leq B\epsilon_n) \geq e^{-Cn\epsilon_n^2},$$

where $\epsilon_n = n^{-\alpha/(2\alpha+D)} \log(n)^t + \sqrt{D \log(P+1)/n}$.

Proof. This is implied by Theorem 2 of Linero and Yang (2018); the only modification required is that Condition P1 and Condition P2 are modified from Linero and Yang (2018), but these modifications do not change the proof strategy. □

Proposition 7. *For fixed positive constants* $\epsilon, \sigma_1, \sigma_2, T, A$ *and integers* n, H, d *define the set*

$$\begin{aligned}\mathcal{G} = \Big\{ f(\cdot) = \sum_{t=1}^{T} g(\cdot; \mathcal{T}_t, \mathcal{M}_t) :\ & T \leq An\epsilon^2,\ \textit{each tree has depth at most } H, \\ & \textit{the common bandwidth parameter } \tau \textit{ satisfies } \sigma_1 \leq \tau^{-1} \leq \sigma_2, \\ & \textit{the total number of splitting directions is at most } d \textit{ out of } P+1, \\ & \textit{for each } (t,\ell),\ \mu_{t\ell} \in [-U, U] \Big\}.\end{aligned}$$

Then there exists a constant C_ψ depending only the gating function ψ of the SBART prior satisfying Condition P such that the following holds:

1. *Covering entropy control:* $\log N(\mathcal{G}, C_\psi\epsilon, \|\cdot\|_\infty) \le d\log(P+1)+3An\epsilon^2\, 2^H \log\left(d\,\sigma_1^{-1}\sigma_2^2 An\epsilon\, 2^H U\right)$; *and*

2. *Complement probability bound: if* $H \ge d_0$, *then* $\Pi(\mathcal{G}^c) \le C'_{M1}\exp\{-C'_{M2}\, An\epsilon^2\log(An\epsilon^2)\}+ 2^H\, An\epsilon^2\cdot\left[\exp\{-E\,d\log(P+1)\}+C_{\mu 1}\,\exp\{-U^{C_{\mu 2}}\}\right]+C_{\tau 1}\,\exp\{-\sigma_1^{-C_{\tau 2}}\}+C_{\tau 3}\exp\{-\sigma_2^{C_{\tau 4}}\}$ *for some constant* $E > 0$ *depending only on hyperparameter* $\xi > 1$ *in the Dirichlet prior.*

Proof. The proof is the same as the proof of Lemma 1 of the supplementary material of Linero and Yang (2018), except that Proposition 5 is used in the complementary probability bound. □

Proof of Theorem 1. Let B and C be chosen as in Proposition 6. By Lemma 1, note that for sufficiently large n, we have $\{f_r : \|r - r_0\| \le B\epsilon_n\} \subseteq K(BC_{\mathcal{K}}\epsilon_n)$ and $\{f_r : \|r - r_0\| \le \epsilon_n\} \subseteq \{f_r : H(f_r, f_{r_0}) \le C_{\mathcal{K}}\epsilon_n\}$ where $C_{\mathcal{K}}$ is a constant depending only on $\mathcal{K}$. Hence

$$\Pi\{f \in K(BC_{\mathcal{K}}\epsilon_n)\} \ge e^{-Cn\epsilon_n^2}.$$

To lighten notation, we redefine ϵ_n throughout the rest of the proof to be $\epsilon_n BC_{\mathcal{K}}$ and C to be $C/(BC_{\mathcal{K}})^2$ so that we have $\Pi\{f \in K(\epsilon_n)\} \ge e^{-Cn\epsilon_n^2}$. This verifies (c) of Proposition 3 using the modified choice of ϵ_n.

Next, for a large constant κ to be chosen later, set $A = \kappa/\log n$, $\sigma_1^{-C_{\tau 2}} = \sigma_2^{C_{\tau 4}} = U^{C_{\mu 2}} = \kappa n\epsilon_n^2$, $H = d_0$, and $d = \lfloor \kappa n\epsilon_n^2/\log(P+1)\rfloor$ for the set $\mathcal{G}_n$ in Proposition 7. Plugging these constants into the covering entropy bound, for sufficiently large n this implies that for $p_1 = 2 + C_{\mu 2}^{-1} + C_{\tau 2}^{-1} + 2C_{\tau 4}^{-1}$ and $p_2 = 2p_1 - 1$ we have

$$\log N(\mathcal{G}_n, C_\psi\epsilon_n, \|\cdot\|_\infty) \le \kappa n\epsilon_n^2\left\{1 + \frac{3\cdot 2^{d_0}}{\log n}\log\left(\frac{2^{d_0}(\kappa n)^{p_1}\epsilon_n^{p_2}}{\log n}\right)\right\} \le \kappa' n\epsilon_n^2$$

for some κ' larger than κ depending only on κ and the constants in Condition P. Define

$\mathcal{F}_n = \{f_r : r \in \mathcal{G}_n\}$. By Lemma 1, for large enough n any $C_\psi \epsilon_n$-net $\mathcal{G}_{\text{net}}$ for $\mathcal{G}_n$ can be converted into a $C_\mathcal{K} C_\psi \epsilon_n$-net $\mathcal{F}_{\text{net}} = \{f_r : r \in \mathcal{G}_{\text{net}}\}$ for $\mathcal{F}_n$. Hence we also have the bound

$$\log N(\mathcal{F}_n, C_\mathcal{K} C_\psi \epsilon_n, H) \leq \log N(\mathcal{G}_n, C_\psi \epsilon_n, \|\cdot\|_\infty) \leq \kappa' n \epsilon_n^2$$

which establishes condition (a) of Proposition 3 with $\bar{\epsilon}_n = C_\mathcal{K} C_\psi \epsilon_n$ and $C_N = \kappa'/(C_\mathcal{K} C_\psi)^2$. Finally, we show condition (b) holds. Applying the complementary probability bound we can make $\Pi(\mathcal{F}_n^c) \leq \exp\{-(C+4)n\epsilon_n^2\}$ for any choice of C by taking κ sufficiently large. To see why, note for example that $n\epsilon_n^2 \geq an^b$ for some positive constants (a, b) so that for large n we have

$$An\epsilon_n^2 \log(An\epsilon_n^2) \geq \kappa n \epsilon_n^2 \left\{ \frac{\log(\kappa/\log n) + \log a}{\log n} + b \right\} \geq \frac{\kappa b}{2} n \epsilon_n^2.$$

Using similar arguments, for large n we can bound each term of the complementary probability bound by $\exp\{-\kappa\delta n\epsilon_n^2/2\}$ for some δ depending only on the constants in Condition P. Taking κ sufficiently large we can make the total bound less than $\exp\{-(C+4)n\epsilon_n^2\}$ for arbitrary C. This proves condition (b). □

References

Adler, R. J., Taylor, J. E., and Worsley, K. J. (2015). *Applications of Random Fields and Geometry: Foundations and Case Studies.* Preprint, retrived from https://web.stanford.edu/class/stats317/hrf.pdf on March 29th, 2021.

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

Antoniano-Villalobos, I., Wade, S., and Walker, S. G. (2014). A Bayesian nonparametric regression model with normalized weights: a study of hippocampal atrophy in alzheimer's disease. *Journal of the American Statistical Association*, 109(506):477–490.

Azzalini, A. (2013). *The Skew-Normal and Related Families*, volume 3. Cambridge University Press.

Basak, P., Linero, A. R., Sinha, D., and Lipsitz, S. (2020). Semiparametric analysis of clustered interval-censored survival datau sing soft Bayesian additive regression trees (SBART). *arXiv preprint arXiv:02509*.

Bleich, J. and Kapelner, A. (2014). Bayesian additive regression trees with parametric models of heteroskedasticity. *arXiv preprint arXiv:1402.5397*.

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

Brookes, M. (2011). The matrix reference manual. Online. Accessed December 2019 at http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html.

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

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

Chung, Y. and Dunson, D. B. (2009). Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660.

Cohen, A. K., Rai, M., Rehkopf, D. H., and Abrams, B. (2013). Educational attainment and obesity: A systematic review. *Obesity Reviews*, 14:989–1005.

Dunson, D. B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307–323.

Dunson, D. B., Pillai, N., and Park, J. H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B*, 69:163–183.

Dutordoir, V., Salimbeni, H., Hensman, J., and Deisenroth, M. (2018). Gaussian process conditional density estimation. In *Advances in Neural Information Processing Systems*, pages 2385–2395.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.

Ghosal, S. and Van Der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168.

Irsoy, O., Yildiz, O. T., and Alpaydin, E. (2012). Soft decision trees. In *Proceedings of the International Conference on Pattern Recognition*.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011). Dppackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40:1–30.

Kundu, S. and Dunson, D. B. (2014). Latent factor models for density estimation. *Biometrika*, 101(3):641–654.

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

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2018). Semiparametric Mixed-Scale Models Using Shared Bayesian Forests. *arXiv e-prints arXiv:1809.08521*.

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

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79.

Murray, I., MacKay, D., and Adams, R. P. (2009). The gaussian process density sampler. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 9–16. Curran Associates, Inc.

Neal, R. M. (1995). *Bayesian Learning For Neural Networks*. PhD thesis, University of Toronto.

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472.

Pospisil, T. and Lee, A. B. (2018). RFCDE: Random forests for conditional density estimation. *arXiv preprint arXiv:1804.05753*.

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

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Rao, V., Lin, L., and Dunson, D. B. (2016). Data augmentation for models based on rejection sampling. *Biometrika*, 103(2):319–335.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007). SPAM: Sparse additive models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1201–1208.

Riihimäki, J., Vehtari, A., et al. (2014). Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448.

Rockova, V. and van der Pas, S. (2017). Posterior concentration for Bayesian regression trees and their ensembles. *arXiv preprint arXiv:1078.08734*.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian analysis (Online)*, 6(1).

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shahbaba, B. and Neal, R. M. (2009). Nonlineaer models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., and Scott, J. G. (2018). BART with targeted smoothing: An analysis of patient-specific stillbirth risk. arXiv:1805.07656.

Titsias, M. and Lawrence, N. D. (2010). Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851. JMLR Workshop and Conference Proceedings.

Tokdar, S. T., Zhu, Y. M., Ghosh, J. K., et al. (2010). Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian analysis*, 5(2):319–344.

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

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014). Improving prediction from dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research*, 15(1):1041–1071.

Wang, C. and Neal, R. M. (2012). Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*.

Woody, S., Carvalho, C. M., and Murray, J. S. (2020). Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics*, pages 1–9.

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

Zhu, R. and Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340.