Bayesian Survival Tree Ensembles with Submodel Shrinkage

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

March 25, 2021

Abstract

We consider Bayesian nonparametric estimation of a survival time subject to right-censoring in the presence of potentially high-dimensional predictors. We argue that several approaches, such as random survival forests and existing Bayesian nonparametric approaches, possess several drawbacks, including: computational difficulties; lack of known theoretical properties; and ineffectiveness at filtering out irrelevant predictors. We propose two models based on the Bayesian additive regression trees (BART) framework. The first, Modulated BART (MBART), is fully-nonparametric and models the failure time as the first occurrence of a non-homogeneous Poisson process. The second, CoxBART, uses a Bayesian implementation of Cox's partial likelihood. These models are adapted to high-dimensional predictors, have default prior specifications, and require simple modifications of existing BART methods to implement. We show the effectiveness of these methods on simulated and benchmark datasets. We also establish that, for a simplified variant of MBART, the posterior distribution contracts at a near-minimax optimal rate in a high-dimensional sparse asymptotic regime.

1 Introduction

We consider the nonparametric conditional survival analysis problem, where our goal is to assess the impact of P predictors $x = (x_1, \ldots, x_P)$ on the survival function $S(t \mid x) = \Pr(T > t \mid X = x)$ and the hazard function $h(t \mid x) = -\frac{d}{dt} \log S(t \mid x)$. The time-constant predictors x may include treatments and prognostic markers in a clinical study. A popular semiparametric model in survival analysis is the Cox proportional hazards model (Cox, 1972) $h(t \mid x) = \lambda(t)\, e^{x^\top \beta}$, where $\lambda(t)$ is a nonparametric baseline hazard model. Instead of this

restrictive assumption, methods based on *decision trees* (Breiman et al., 1984) construct a partition of the predictor space $\mathcal{X}$ and estimate $S(t \mid x)$ separately for each equivalence class. Decision trees are also used as building blocks for *ensemble methods*. For example, the random survival forests algorithm (Ishwaran et al., 2008) aggregates many decision trees together to obtain a flexible estimate of $S(t \mid x)$. Models based on Bayesian additive regression trees (BART) have also been proposed.

While Bayesian nonparametric methods for survival regression analysis exist, we argue that they often lack the following properties we feel to be desirable:

(i) An argument for using Bayesian nonparametrics is that it allows us to use a prior which centers on a specified parametric model (similar to a "prior guess"), effectively "shrinking" the fitted model towards the parametric structure while allowing the model to adapt to lack-of-fit when the parametric structure is incorrect. For example, parametric accelerated failure time (AFT) models and the semiparametric Cox models are two examples of such "prior guesses." This gives us the best of both worlds: the flexibility of a nonparametric model and (when the guess is supported by data) the stability of a (semi) parametric model. Arguments for the desirability of this property include: maintenance of interpretability when the prior guess is accurate (Müller and Mitra, 2013); increased stability of inference with small sample sizes; losing the minimal amount of predictive accuracy when the prior guess is accurate; and guarding oneself from under-fitting when the prior guess is inaccurate (Dunson, 2009).

(ii) The prior should be able to adapt to structure in the data, such as low-order interactions in the predictors, sparsity, or smoothness of the hazard as a function of time.

(iii) The posterior should be computationally tractable.

(iv) The prior should have "large support," with the posterior ideally concentrating at a near-minimax rate adaptively over a variety of function spaces.

We propose two models using the BART framework. The first approach, which we refer to as Modulated BART (MBART), is fully nonparametric and satisfies (i)–(iv) above. The MBART model sets $h(t \mid x) = \lambda(t \mid x, \theta)\, \Phi\{g(t, x)\}$ where $\lambda(t \mid x, \theta)$ is the hazard of a (semi) parametric model parameterized by θ that we wish to shrink towards, while $g(t, x)$ is a decision tree ensemble that controls deviations from the base model through the *link function* $\Phi : \mathbb{R} \to [0, 1]$. The second approach, which we refer to as CoxBART, is based on a Bayesian interpretation of the Cox partial likelihood. CoxBART is less computationally intensive than MBART and retains the simpler interpretation of the Cox proportional hazards model; CoxBART is also a useful point of comparison for MBART because we will often shrink the MBART model towards a proportional hazards model.

We provide the first theoretical guarantees for BART survival models. We show that a simplified version of our MBART model, when combined with the smooth decision trees used by Linero and Yang (2018), adapts to sparsity in $h(t \mid x)$ when $x \in \mathbb{R}^P$ is high dimensional but $h(t \mid x)$ depends on $D \ll P$ predictors. MBART also adapts to the smoothness level of $h(t \mid x)$. In both cases, we obtain near-minimax rates of convergence with respect to a type of Hellinger distance.

1.1 Related Methods

There are several existing approaches to using the BART framework with survival data. Sparapani et al. (2016) developed a fully nonparametric regression method for discrete survival data. This approach can be used for continuous survival data only after discretizing the N observed survival times to a grid of N_{grid} times, and the fitted model depends on both the number and location of the N_{grid} grid points. The Gibbs sampler developed by Sparapani et al. (2016) has computational complexity $\Omega(MNN_{\text{grid}})$ where M is the number of trees in the ensemble; this is substantially more computationally intensive than usual BART algorithms, which have complexity $\Omega(MN)$, and forces N_{grid} to be small. It also violates (i) by not being centered on any (semi)parametric submodel.

Bonato et al. (2010) proposed a variety of semiparametric models based on BART, including semiparametric accelerated failure time models of the form $\log T_i = g(X_i) + \epsilon_i$ and a Weibull regression model. Most important for our purposes, they considered the proportional hazards model $h(t \mid \omega_i) = \lambda(t)\exp(\omega_i)$ conditional on the latent variable $\omega_i \sim \text{Normal}\{g(X_i), \sigma^2_\omega\}$. This latent variable structure is imposed only to allow for existing Gibbs samplers to be used with ω_i as the response; the ω_i's can then be updated by Metropolis-Hastings during Gibbs sampling. This model is essentially a frailty model $h(t \mid x, z) = \lambda(t)\exp\{g(x) + z\}$ where $\exp(Z_i)$ is a log-normal frailty. Similar to the identifiability issues of the frailty distribution and marginal regression structure for semiparametric univariate survival models (Oakes, 1989; Hougaard, 2000), the distribution of ω_i is not identifiable and the marginal distribution of T_i given X_i does not preserve the proportional hazards structure (thus $g(x)$ fails to describe a proportional hazards relationship between T_i and X_i). By contrast, our CoxBART model induces a proportional hazards model *marginally*. Henderson et al. (2020) introduced an AFT model with large support in the class of all AFT models $\log T_i = g(X_i) + \epsilon_i$ by modeling the residual distribution $\epsilon_i \sim F$ as a Dirichlet process mixture; this accomplishes a similar goal as our CoxBART model by allowing for the use of BART in a large class of nonparametric survival models.

Our proposed modeling strategy is similar in spirit to recent work of Li et al. (2020) on nonparametric conditional distribution estimation. In concurrent work by the authors of this paper, the methodology is extended to the setting of interval-censored clustered survival times (Basak et al., 2020); the present work differs by incorporating targeted smoothing, centering the nonparametric model on a "prior guess," providing theoretical justification, and by introducing the CoxBART model. A similar data augmentation algorithm to the one described here is used with Gaussian processes to perform survival analysis (Fernández et al., 2016). More generally, there is a sizable literature on Bayesian nonparametric survival analysis based on Dirichlet processes and Gaussian processes. See, for example, De Iorio et al. (2009), who develop an ANOVA-DDP model to perform fully-nonparametric Bayesian

survival analysis.

1.2 Outline of the Paper

In Section 2 we describe the MBART and CoxBART models. In Section 3 we propose several base models that MBART can be shrunk towards. In Section 4 we study the theoretical properties of a special case of the MBART model, establishing posterior concentration rates. In Section 5 we illustrate MBART and CoxBART on simulated data and on publicly available data on liver disease. We conclude in Section 6 with a discussion. Additional computational details and proofs of the main theorems are in the appendix. Algorithms and proofs of lemmas are given in the Supplementary Material.

2 Survival Models Using Bayesian Tree Ensembles

Let (T_i, C_i) denote the survival and censoring times respectively for $i = 1, \ldots, N$. We observe data $\mathcal{D} = \{(Y_i, \delta_i, X_i) : i = 1, \ldots, N\}$ where $Y_i = \min(T_i, C_i)$ is the observed (right-censored) survival time, $\delta_i = I(T_i \leq C_i)$ is the censoring indicator, and $X_i \in \mathbb{R}^P$ is a vector of covariates. Our goal is to model the conditional survival function $S_0(t \mid x)$ of T_i conditional on X_i. Other relevant quantities include the corresponding cumulative hazard function $H_0(t \mid x) = -\log S_0(t \mid x)$ and the hazard function $h_0(t \mid x) = \frac{d}{dt} H_0(t \mid x)$. Throughout this paper we assume that the censoring time C_i is independent of T_i given X_i.

2.1 Review of Bayesian Additive Regression Trees

BART models an unknown function $g(x)$ as a sum of M decision trees $\sum_{m=1}^{M} \text{Tree}(x; \mathcal{T}_m, \mathcal{M}_m)$ where $\mathcal{T}_m$ denotes the tree topology and splitting rules of the tree and $\mathcal{M}_m$ denotes the predicted response for each leaf node. For detailed reviews of BART, see Linero (2017); Hill et al. (2019). The function $\text{Tree}(x; \mathcal{T}_m, \mathcal{M}_m)$ returns $\mu_{m\ell}$ if x is associated to leaf node ℓ of tree m. Each tree induces a partition of the predictor space $\mathcal{X}$ such that $g(x)$ is constant on

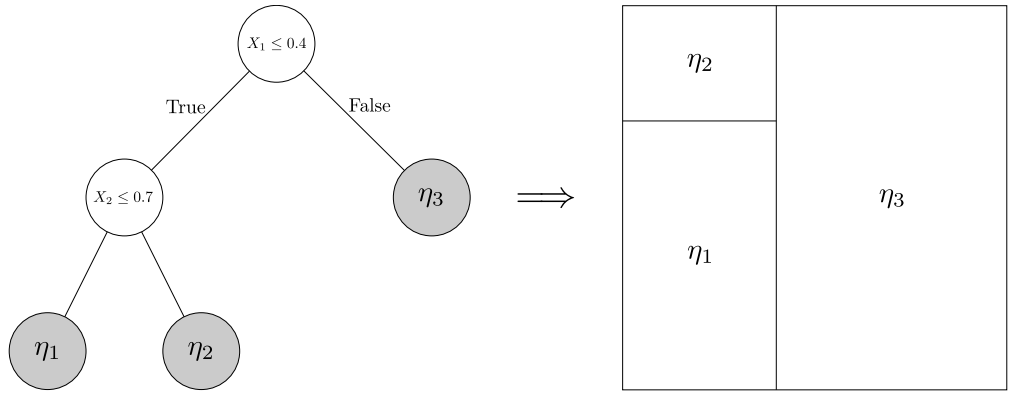

Figure 1: Left: an example of a decision tree with two variables $X = (X_1, X_2)$. Right: the piecewise-constant function induced from the decision tree, taking values η_1, η_2, and η_3 depending on the value of X.

each equivalence class of the partition. A schematic showing a particular decision tree with the induced partition over $\mathcal{X} = [0,1]^2$ is given in Figure 1. We divide the decision tree nodes into a collection of leaf nodes $\ell \in \mathcal{L}$ and branch nodes $b \in \mathcal{B}$, where $\mathcal{L}$ consists of the nodes with no children. Associated to each branch b is a *decision rule* of the form $[X_{j_b} \leq C_b]$, while each leaf ℓ is associated to a prediction $\mu_{m\ell}$.

We write $g \sim \text{BART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ for the BART prior with prior $\mathcal{T}_m \overset{\text{iid}}{\sim} \pi_{\mathcal{T}}$ and $\mu_{m\ell} \overset{\text{iid}}{\sim} \pi_{\mathcal{M}}$ given $\{\mathcal{T}_m\}$. It is standard to take $\pi_{\mathcal{M}}$ to be $\text{Normal}(0, \sigma_\mu^2)$ where $\sigma_\mu \propto M^{-1/2}$; this is conditionally conjugate, and ensures that $\text{Var}\{g(x)\} = M\sigma_\mu^2$ does not depend on the number of trees M. We take $\pi_{\mathcal{T}}$ to be a the prior described by Chipman et al. (1998), which can be sampled from by initializing $\mathcal{T}_m$ with a single root note of depth $d = 0$. This node is then made branch with probability $\gamma/(1+d)^\beta$ and a leaf node otherwise; if the node is a branch node, we add its two children at depth $d+1$. This process then iterates over all the nodes of depth $d = 1, 2, \ldots$ until all of the nodes at some depth are leaf nodes. We use the following prior on the splitting rules through this paper. First $j_b \sim \text{Categorical}(s)$ for some probability vector s. Then, given j_b and the values of $(j_{b'}, C_{b'})$ for the ancestor nodes of b, we set $C_b \sim \text{Uniform}(L_{j_b}, R_{j_b})$ where $\prod_{j=1}^P (L_j, R_j)$ is the hyperrectangle of x-values which are associated to branch b. Following Linero (2018), we set $s \sim \text{Dirichlet}(\alpha/P, \ldots, \alpha/P)$ and in our illustrations use $\alpha/(\alpha+P) \sim \text{Beta}(0.5, 1)$. This prior for s encourages the model to

concentrate on models with a small number of relevant predictors.

We also review the *soft Bayesian additive regression trees* (SBART) prior defined by Linero and Yang (2018). Note that $g \sim \text{BART}(\pi_{\mathcal{T}}, \mathcal{M}_{\mathcal{T}})$ can equivalently be expressed as

$$g(x) = \sum_{m=1}^{M} \sum_{\ell \in \mathcal{L}_m} w_{m\ell}(x)\, \mu_{m\ell} \tag{1}$$

where $w_{m\ell}(x) = 1$ or 0 according as x is associated with leaf ℓ of tree m or not. The SBART model is of the form (1), but uses smooth weights $w_{m\ell}(x) = \prod_{b \in A(\ell)} \psi(x; C_b, \tau_b)^{1-R_b}\{1 - \psi(x; C_b, \tau_b)\}^{R_b}$ where R_b is the indicator that leaf ℓ of tree m is associated to the right path of branch b. The function $\psi(x; C, \tau)$ is the cumulative distribution function of a continuous symmetric random variable; throughout this work, we will take $\psi(x; C, \tau) = \text{expit}\{(x - C)/\tau\}$ where $\text{expit}(x) = (1 + e^{-x})^{-1}$. We note that, in the limit as $\tau \to 0$, we revert to the usual (non-soft) decision trees. Trees which use smooth weights $w_{m\ell}(x)$ are referred to as *soft decision trees*; see, for example, Irsoy et al. (2012). When $g(x)$ has an SBART prior we will write $g \sim \text{SBART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$. Using SBART makes the function $g(x)$ continuous in x, which leads to better theoretical and practical performance (Linero and Yang, 2018). A drawback of SBART is that the it is more computationally intensive to fit.

2.2 Modulated BART

Modulated BART (MBART) models hazard function $h_0(t \mid x)$ as the random function

$$h(t \mid x) = \lambda(t \mid x, \theta)\, \Phi\{g(t, x)\} \quad \text{where} \quad g(t, x) = \mathfrak{a} + \sum_{m=1}^{M} \psi_m(t)\ \text{Tree}(x; \mathcal{T}_m, \mathcal{M}_m), \tag{2}$$

where $\text{Tree}(x; \mathcal{T}_m, \mathcal{M}_m)$ is a smooth decision tree, and $\lambda(t \mid x, \theta)$ is either a parametric or semiparametric model which serves as the "best guess" at $h_0(t \mid x)$. Under (2), a random variable T can be generated by simulating a Poisson process with intensity $\lambda(t \mid x, \theta)$ on $(0, \infty)$, thinning the points with probability $\Phi\{g(t, x)\}$, and setting T to be the smallest

accepted point.

We treat t different in $g(t,x)$ in order to facilitate shrinkage towards a semiparametric proportional hazards model (See Section 3). In particular, we use the *targeted smoothing* framework of Starling et al. (2020). This also ensures that estimated survival functions will be smooth in time even if we replace SBART with BART for computational purposes, and removes the need to standardize T_i to be supported on $[0,1]$ (which may be difficult due to right-censoring).

We implement targeted smoothing over t using a random Fourier series $\psi_m(t) = \cos(w_m t + b_m)$ where $b_m \sim \text{Uniform}(0, 2\pi)$ and $w_m \sim \text{Normal}(0, \rho^{-2})$ (Li et al., 2020; Rahimi and Recht, 2008; Fernández et al., 2016). This approximates the ideal model proposed by Starling et al. (2020) where $\text{Tree}(x; \mathcal{T}_m, \mathcal{M}_m)$ is a Gaussian process with kernel $\Sigma(t,t') = e^{-(t-t')^2/(2\rho)}$. We cannot use the ideal model because it requires $O(N_T^3)$ computations where N_T is the unique number of observed failure times. By contrast, the random Fourier series imposes no additional burden on the original BART algorithm. For more details on using random Fourier series with BART, see Li et al. (2020).

Model (2) is "centered" on $\lambda(t \mid x, \theta)$ in two ways. If the prior encourages $|g(t,x)|$ to be small, then $\Phi\{g(t,x)\} \approx 0.5$ so that $h(t \mid x) \approx \lambda(t \mid x, \theta)/2$. If $|g(t,x)| \gg 0$, then $\Phi\{(g(t,x))\} \approx 1$ and the model becomes $h(t \mid x) \approx \lambda(t \mid x, \theta)$. Our preference is to choose a prior which encourages $\mathfrak{a} \gg 0$ in (2) so that the resulting prior puts a high weight around the event $\Phi\{g(t,x)\} \approx 1$.

2.2.1 Default Prior Specification of the Thinning Process

An advantage of the BART framework is that it is straight-forward to develop a default prior which works well in practice. For our survival model we must specify $\pi_{\mathcal{T}}$, $\pi_{\mathcal{M}}$, $\mathfrak{a}$, the link Φ, the base model $\lambda(t \mid x, \theta)$, and the associated prior for θ. In our illustrations we take $\Phi(\cdot)$ to be the probit link. We discuss specification of the base model in Section 3. We specify a Half-Normal$(1,1)$ prior for $\mathfrak{a}$ to encourage larger values of $g(t,x)$; as previously noted, this

shrinks our model towards $\lambda(t \mid x, \theta)$. Encouraging $g(t, x)$ to be large has the additional benefit of improving the efficiency the data augmentation scheme described in Section 2.2.2. We set $\sigma_\mu \sim \text{Half-Cauchy}(0, 1.5/\sqrt{M})$ and use the default prior for $\pi_{\mathcal{T}}$ described in Section 2.1.

2.2.2 Data Augmentation

The observed data likelihood of model (2) has the form

$$\prod_{i=1}^{N} \lambda(Y_i \mid X_i, \theta)^{\delta_i} \Phi\{g(Y_i, X_i)\}^{\delta_i} \exp\left\{ - \int_0^{Y_i} \lambda(t \mid X_i, \theta)\, \Phi\{g(t, X_i)\}\, dt \right\}. \tag{3}$$

This likelihood is inconvenient to work with both because of the analytically intractable integral and because it does not cleanly allow for the use of existing Bayesian backfitting algorithms for fitting BART models. To construct a Markov chain Monte Carlo (MCMC) algorithm for MBART we use two steps of data augmentation. The first step removes the intractable integral $\int_0^{Y_i} \lambda(t \mid X_i, \theta)\, \Phi\{g(t, X_i)\}\, dt$ from the likelihood by augmenting a Poisson process with intensity $\lambda(t \mid X_i, \theta)\, [1 - \Phi\{g(t, X_i)\}]$.

Proposition 1. *If $\{W_{ij} : 1 \leq j \leq J_i\}$ given $(Y_i, \delta_i, X_i, g, \theta)$ is sampled according to a non-homogeneous Poisson process on the interval $(0, Y_i)$ with intensity function $\lambda(t \mid X_i, \theta)\, [1 - \Phi\{g(t, X_i)\}]$, then the joint likelihood of (g, θ) given $(Y_i, \delta_i, X_i, \{W_{ij}\}_{j=1}^{J_i})$ for $i = 1, \ldots, N$ is*

$$e^{-\sum_{i=1}^{N} \Lambda(Y_i \mid X_i, \theta)} \left(\prod_{i=1}^{N} \lambda(Y_i \mid X_i, \theta)^{\delta_i} \prod_{j=1}^{J_i} \lambda(W_{ij} \mid X_i, \theta) \right) \left(\prod_{i=1}^{N} \Phi\{g(Y_i, X_i)\}^{\delta_i} \prod_{j=1}^{J_i} [1 - \Phi\{g(W_{ij}, X_i)\}] \right),$$

where $\Lambda(t \mid x, \theta) = \int_0^t \lambda(s \mid x, \theta)\, ds$ is the cumulative hazard of the base model.

A variant of Proposition 1 is used by Adams et al. (2009) to fit non-homogeneous Poisson processes. For completeness, we give a simple proof of this result.

Proof. The likelihood component of $(Y_1, \ldots, Y_N)$ is given by (3), whereas the likelihood component of the event times $\{W_{ij} : i = 1, \ldots, N, j = 1, \ldots, J_i\}$ simulated independently

from our non-homogeneous Poisson processes on $(0, Y_i)$ is

$$\prod_{ij} \lambda(W_{ij} \mid X_i, \theta) \left[1 - \Phi\{g(W_{ij} \mid X_i, \theta)\}\right] \prod_{i=1}^{N} \exp\left\{ - \int_0^{Y_i} \lambda(t \mid X_i, \theta)[1 - \Phi\{g(t, X_i)\}] \, dt \right\}.$$

Multiplying these quantities and noting that the exponential terms can be combined to give $\exp\{-\int_0^{Y_i} \lambda(t \mid X_i, \theta) \, dt\} = e^{-\Lambda(Y_i \mid X_i, \theta)}$ establishes the result. □

Proposition 1 eliminates the intractable integral $\int_0^{Y_i} \lambda(t \mid X_i, \theta) \, \Phi\{g(t, X_i) \, dt\}$ from the likelihood. When $\Phi(\mu)$ is the probit link, we can use the approach of Albert and Chib (1993) to simplify the likelihood further. We introduce $\{Z_{ij} : i = 1, \ldots, N, j = 0, \ldots, J_i\}$ where $Z_{ij} \sim \text{Normal}\{g(W_{ij}, X_i), 1\}$ truncated to $(-\infty, 0)$ for $j \geq 1$ and $Z_{ij} \sim \text{Normal}\{g(Y_i, X_i), 1\}$ truncated to $(0, \infty)$ for $j = 0$. This gives the joint likelihood

$$\begin{aligned} e^{-\sum_{i=1}^{N} \Lambda(Y_i \mid X_i, \theta)} \times \left(\prod_{i=1}^{N} \lambda(Y_i \mid X_i, \theta)^{\delta_i} \prod_{j=1}^{J_i} \lambda(W_{ij} \mid X_i, \theta) \right) \\ \times \left(\prod_{i=1}^{N} \text{Normal}(Z_{i0} \mid g(Y_i, X_i), 1)^{\delta_i} \prod_{j=1}^{J_i} \text{Normal}(Z_{ij} \mid g(W_{ij}, X_i), 1) \right). \end{aligned} \tag{4}$$

The usual Bayesian backfitting algorithm of Chipman et al. (2010) can now be applied by treating the Z_{ij}'s as the response. Similarly, the logistic link $\Phi(\mu) = (1 + e^{-\mu})^{-1}$ or Student's T link can be implemented using the Gaussian scale-mixture representation of the logistic and Student's T distribution (Holmes and Held, 2006). Detailed algorithms are given in Section S.4 of the Supplementary Material.

2.3 Proportional Hazards with CoxBART

We now consider the proportional hazards model $h(t \mid x) = \lambda(t) \exp\{g(x)\}$ where $\lambda(t)$ is the baseline hazard and $g(x)$ is unknown, with $g(x)$ estimated using the Cox partial likelihood $\text{PL}(g) = \prod_{i:\delta_i=1} \frac{\exp\{g(X_i)\}}{\sum_{j \in \mathscr{R}_i} \exp\{g(X_j)\}}$ where $\mathscr{R}_i = \{j : Y_j \geq Y_i\}$ is the set of subjects at-risk of

failure at time Y_i. Letting Π denote the prior, we define the pseudo-posterior

$$\Pi(dg \mid \mathcal{D}_N) = \frac{\mathrm{PL}(g)\ \Pi(dg)}{\int \mathrm{PL}(g)\,\Pi(dg)}. \tag{5}$$

Expression (5) arises under an improper prior for $\Lambda(t)$. Consider a discrete time proportional hazards model $S(t \mid x) = \exp\big\{ - e^{g(x)} \sum_{t_i \le t} \phi_i \big\}$. Sinha et al. (2003) show that $\mathrm{PL}(g)$ is the integrated likelihood of $g(x)$ when the parameters ϕ_i are given an improper data-dependent prior where the t_i's are set equal to the observed values of the Y_i's and $\pi(\phi_1, \ldots, \phi_N) \propto \prod_{i=1}^N \delta_i \phi_i^{-1} + (1 - \delta_i)\delta_0(\phi_i)$, where $\delta_0(\cdot)$ is a point-mass at 0. The likelihood of (ϕ, g) is

$$\prod_{i=1}^N \phi_i^{\delta_i} \exp\Big[\delta_i g(X_i) - \exp\{g(X_i)\} \sum_{j : Y_j \le Y_i} \phi_j \delta_j\Big] = \prod_{i:\delta_i=1} \phi_i \exp\Big[g(X_i) - \phi_i \sum_{j \in \mathscr{R}_i} \exp\{g(X_j)\}\Big].$$

The conditional distribution of $(\phi_1, \ldots, \phi_N, \mathcal{D})$ given g under this model is

$$\pi(\phi_1, \ldots, \phi_N, \mathcal{D} \mid g) = \prod_{i:\delta_i=1} \exp\Big[g(X_i) - \phi_i \sum_{j \in \mathscr{R}_i} \exp\{g(X_j)\}\Big]. \tag{6}$$

From (6), $\phi_i \overset{\text{ind}}{\sim} \mathrm{Gam}(1, \sum_{j \in \mathscr{R}_i} \exp\{g(X_j)\})$ given $(g, \mathcal{D})$. Integrating out ϕ, we obtain $\mathrm{PL}(g)$.

We refer to this model with $g \sim \mathrm{BART}(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ as CoxBART. CoxBART is a nonparametric model in the sense that $g(x)$ is nonparametric, but we do assume proportional hazards. It is more flexible than the usual Cox model, where $g(x) = x^\top \beta$, but is not as flexible as MBART. CoxBART is interesting both for providing a benchmark to measure MBART against and independently due to the common use of the proportional hazards assumption in practice. A related model is given by Bonato et al. (2010), however this model only satisfies the proportional hazards assumption conditional on latent variables.

Conveniently, (6) can be combined with the log-linear BART prior of Murray (2020) to construct a Bayesian backfitting algorithm for sampling $g(x)$ and ϕ. Details are given in the appendix, with $\mu_{m\ell} \sim \log \mathrm{Gam}(\alpha_\mu, \beta_\mu)$.

2.3.1 Default Prior for CoxBART

Using the default choice of $\pi_{\mathcal{T}}$ described in Section 2.1, the only remaining parameters to choose are (α_μ, β_μ) in the prior $\mu_{m\ell} \sim \log \text{Gam}(\alpha_\mu, \beta_\mu)$. To impose that $\mu_{m\ell}$ has mean 0, we set $\log(\beta_\mu) = \psi(\alpha_\mu)$ where $\psi(\alpha)$ is the digamma function $\frac{d}{d\alpha} \log \Gamma(\alpha)$. The variance of $\mu_{m\ell}$ is given by $\sigma_\mu^2 = \psi'(\alpha_\mu)$ where $\psi'(\alpha)$ is the trigamma function. We set $\sigma_\mu \sim \text{Half-Cauchy}(0, 1.5/\sqrt{M})$ and update σ_μ using slice sampling (Neal, 2003). We note that the required special functions ψ, ψ' and $(\psi')^{-1}$ are all straight-forward to calculate numerically. The simple approximation $\alpha_\mu = \sigma_\mu^{-2} + 1/2$ and $\beta_\mu = \sigma_\mu^{-2}$ given by Murray (2020) also works well when σ_μ is small.

3 Base Models for MBART

MBART can be used with essentially any base hazard $\lambda(t \mid x, \theta)$. It is ideal for $\lambda(t \mid x, \theta)$ to be a good approximation to the true hazard $h_0(t \mid x)$ for two reasons. First, a good specification of $\lambda(t \mid x, \theta)$ will encourage the model to disregard the nonparametric component, reducing the model complexity. Second, the efficiency of the MCMC algorithm depends on the sizes of the J_i's, which will be large if the base model fits poorly.

Parametric Weibull A commonly used parametric subclass of the proportional hazards model is the *Weibull* model $\lambda(t \mid x, \theta) = \kappa\, e^{b^\top x}\, t^{\kappa-1}$, where $\theta = (\kappa, b)$ (Ibrahim et al., 2001, Section 2.2). The full conditional of θ when using Proposition 1 is given by

$$\pi(\kappa, b \mid -) \propto \kappa^{\sum_i (J_i + \delta_i)} \left(\prod_i Y_i^{\delta_i} \prod_j W_{ij} \right)^{\kappa - 1} \exp\left\{ \sum_i (J_i + \delta_i) b^\top X_i - e^{b^\top X_i} Y_i^\kappa \right\} \pi(\kappa, b).$$

This full conditional can be sampled using Hamiltonian Monte Carlo.

Weibull-BART A semiparametric variant of the Weibull model can be developed using the log-linear BART model of Murray (2020). This model sets $\lambda(t \mid x, \theta) = \kappa\, e^{r(X_i)}\, t^{\kappa-1}$ where $r \sim \text{BART}(\pi_{\mathcal{T}}^r, \pi_{\mathcal{M}}^r)$, written $r(x) = \eta_0 + \sum_{m=1}^M \text{Tree}(x; \mathcal{T}_m^r, \mathcal{M}_m^r)$. Here, $\theta = (\kappa, r)$. Similar

to CoxBART, we set $\eta_{m\ell} \sim \log \text{Gam}(\alpha_\eta, \beta_\eta)$. In the appendix, we derive a simple Bayesian backfitting algorithm for this model.

CoxBART2 It is possible to shrink towards a *semiparametric* proportional hazards model using the Weibull-BART base model. In the special case where $g(t, x)$ does not depend on x, the Weibull-BART model simplifies to $h(t \mid x) = \lambda^\star(t)\, e^{r(x)}$ where $\lambda^\star(t) = \kappa t^{\kappa-1} \Phi\{g(t)\}$, which is a Cox proportional hazards model. Because $g(t)$ is modeled nonparametrically, the baseline hazard $\lambda^\star(t)$ has a very flexible prior. We can shrink towards $g(t, x) \equiv g(t)$ by using a prior which encourages the tree structures to consist of only a root node. To accomplish this, we simply choose the tree parameters (γ, β) so that most trees consist only of the root node (e.g., by taking γ small). A similar approach is used by Hahn et al. (2020) in the context of causal inference in order to shrink towards a homogeneous treatment effect.

As a default, we have used Weibull-BART in all of our illustrations and simulations, with $\kappa \equiv 1$; equivalently, we have used an exponential distribution for the base model. For some discussion of possible priors for κ, see Van Niekerk et al. (2020), who recommend a penalized complexity prior.

4 Theoretical Results

We now establish convergence rates for the MBART posterior. These results are similar in spirit to results of Linero and Yang (2018) for regression and Li et al. (2020) for conditional distribution estimation. We operate in the Frequentist setup, with uncoarsened data $(T_1, C_1, X_1), \ldots, (T_n, C_n, X_n)$ sampled iid from a joint distribution F_0. Throughout, we make the *independent censoring assumption*.

Condition R (random censoring) The true joint distribution F_0 of (T_i, C_i, X_i) is such that T_i and C_i are conditionally independent given X_i.

We assume the T_i's have conditional density $p_0(t \mid x)$ on $(0, \infty)$. The C_i's are assumed to

be bounded by a constant $C_i \leq C$ (typically the time period of the study), with conditional density $f_C(c \mid x) = p_C(c \mid x)I(c < C) + S_C(c \mid x)\, I(c = C)$ with respect to the sum of Lebesgue measure on $(0, C)$ and a point mass at C; we write $\nu(dt)$ for this measure, and we will use it throughout this section. Without loss of generality we assume that $C \leq 1$. The constant C represents the time of the end of the study, so that all observations with $T_i > C$ are censored. We define $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. We also define $\widetilde{Y}_i = \min(T_i, C)$, so that we can study the setting of fixed censoring within this framework as well. Let $\mathcal{D}_n = \{Y_i, \delta_i, X_i : i = 1, \ldots, n\}$ and let $\widetilde{\mathcal{D}}_n = \{\widetilde{Y}_i, X_i : i = 1, \ldots, n\}$.

For simplicity we assume that the baseline hazard $\lambda(t)$ is fixed, so that we need only consider a prior distribution Π on $g(t, x)$ and its associated hyperparameters. To each $u : [0,1]^{P+1} \to \mathbb{R}$ we associate a density function $p_u(t \mid x) = \lambda(t)\, \Phi\{u(t,x)\} \exp\{-H_u(t \mid x)\}\}$ where $H_u(t \mid x) = \int_0^t \lambda(s)\, \Phi\{u(s,x)\}\, ds$ is a cumulative hazard. We make the following assumption about the true hazard function $h_0(t \mid x)$, which guarantees that we can define the "true" value of $g_0(t, x)$ by $\Phi^{-1}(h_0/\lambda)$.

Condition H (on p_0) For $(t, x) \in [0, C] \times [0,1]^P$, the hazard ratio $R(t,x) = h_0(t \mid x)/\lambda(t)$ is bounded away from 0 and 1, and $R(\cdot, \cdot) \in C^{\alpha, \mathcal{R}}([0,1]^{P+1})$ where $C^{\alpha,\mathcal{R}}([0,1]^{P+1})$ is the ball of radius $\mathcal{R}$ of α-smooth Hölder functions (see Ghosal and van der Vaart, 2017, Appendix C). Additionally, $R(t,x)$ depends on $D \leq P + 1$ many coordinates of (t, x).

At face value Condition H is highly restrictive, as it requires an a-priori known upper bound for $h_0(t \mid x)$. We stress that this is a minor technical convenience, as if (say) $h_0(t \mid x)$ is bounded then the theory we develop goes through with $\lambda(t) \equiv \lambda$ with λ given an appropriate prior. Additionally, the restriction that $x \in [0,1]^P$ is not restrictive, as in practice we will always perform a quantile transformation to all predictors.

Condition P (on Π) The function g 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 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.

We also make the assumption that $\Phi(\mu)$ corresponds to the distribution function of a heavy-tailed distribution.

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 such that $\frac{d}{d\mu} \log \Phi(\mu) \leq \mathcal{K}$.

4.1 Fixed Censoring

Fixed censoring occurs if T_i is observed as long as $T_i \leq C$. We study this by conditioning on $\widetilde{\mathcal{D}}_n$. Given $g(t,x)$, the $\widetilde{Y}_i$'s have conditional density given by

$$f_g(y \mid x) = p_g(y \mid x)\, I(y < C) + \exp\{-H_g(C \mid x)\}\, I(y = C) \tag{7}$$

with respect to $\nu(dy)$. We measure the accuracy of the posterior distribution using the *integrated Hellinger metric* $\mathcal{H}(g_0, g)$ defined by

$$\begin{aligned}\mathcal{H}^2(g_0,g) &= \int\int \{\sqrt{f_{g_0}(y \mid x)} - \sqrt{f_g(y \mid x)}\}^2 \; \nu(dy) \; F_X(dx) \\ &= \int \left[\int_0^C \{\sqrt{p_{g_0}(y \mid x)} - \sqrt{p_g(y \mid x)}\}^2 \; dy\right] + (e^{-H_{g_0}(C|x)} - e^{-H_g(C|x)})^2 \; F_X(dx),\end{aligned}$$

where F_X is the true distribution of the X_i's. While not used as part of the model, integrating with respect to F_X gives a natural metric by which to judge estimation accuracy. Our goal is to establish $\Pi\{\mathcal{H}(g_0,g) \leq K\epsilon_n \mid \widetilde{\mathcal{D}}_n\} \stackrel{n\to\infty}{\longrightarrow}$ in F_0-probability for a sequence $\epsilon_n \downarrow 0$ and some fixed positive constant K; in this case, we say that the convergence rate of the posterior is faster than ϵ_n. Under the assumption that $g_0(t,x)$ is an α-Hölder smooth function depending on D coordinates, the oracle minimax estimation rate when α and the relevant coordinates of (t,x) are known is $\epsilon_n = n^{-\alpha/(2\alpha+D)}$. The following theorem, which is proved in the appendix, shows that we adaptively obtain this rate up-to a logarithmic term $\log(n)^t$ and a variable selection term $\sqrt{D\log(P+1)/n}$. This result allows P (but not D) to diverge, and permits consistent estimation even when $\log P$ grows nearly linearly with n.

Theorem 1. *Suppose that Condition H, Condition L, and Condition P hold. Let* $\epsilon_n = n^{-\alpha/(2\alpha+D)}\log(n)^t + \sqrt{D\log(P+1)/n}$ *where* $t = \alpha(D+1)/(2\alpha+D)$. *Then*

$$\Pi\{\mathcal{H}(g_0,g) \leq K\epsilon_n \mid \widetilde{\mathcal{D}}_n\} \stackrel{n\to\infty}{\longrightarrow} 0 \qquad \textit{in } F_0\textit{-probability}$$

for some sufficiently large constant K.

4.2 Random Censoring

Our results for a fixed censoring time C extend in a straight-forward fashion to the case where the C_i's are bounded with $C_i \leq C$. Under the independent censoring assumption, the joint density of (Y_i, δ_i) given $X_i = x$ and g is

$$q_g(y, \delta \mid x) = \begin{cases} S_g(y \mid x)\, p_C(y \mid x) & \text{if } \delta = 0 \text{ and } y < C, \\ S_g(y \mid x)\, S_C(y \mid x) & \text{if } \delta = 0 \text{ and } y = C, \\ p_g(y \mid x)\; S_C(y \mid x) & \text{if } \delta = 1, \\ 0 & \text{otherwise.} \end{cases}$$

We now study posterior concentration with respect to the integrated Hellinger distance

$$\mathcal{H}_q^2(g_0, g) = \int \int \sum_{\delta=0}^{1} \{\sqrt{q_{g_0}(y, \delta \mid x)} - \sqrt{q_g(y, \delta \mid x)}\}^2 \; \nu(dy) \; F_X(dx). \tag{8}$$

In addition to the covariate distribution F_X, $\mathcal{H}_q$ also depends on the distribution of C_i. In the appendix we prove the following result.

Theorem 2. *Suppose that Condition R, Condition H, Condition L, and Condition P hold. Let $\epsilon_n = n^{-\alpha/(2\alpha+D)} \log(n)^t + \sqrt{D \log(P+1)/n}$ where $t = \alpha(D+1)/(2\alpha+D)$. Then*

$$\Pi\{\mathcal{H}_q(g_0, g) \leq K\epsilon_n \mid \mathcal{D}_n\} \overset{n\to\infty}{\longrightarrow} 0 \qquad \textit{in } F_0\textit{-probability}$$

for some sufficiently large constant K.

Remark 1. Our proof of Theorem 1 is based on checking the sufficient conditions for posterior convergence rates given by Ghosal et al. (2000). To prove Theorem 2 we use a monotonicity property of *f-divergences* (Ali and Silvey, 1966) to show that the sufficient conditions used to prove Theorem 1 also suffice to prove Theorem 2; for example, the fact that the Hellinger distance is an f-divergence can be used to show that $\mathcal{H}_q(g_0, g) \lesssim \mathcal{H}(g_0, g)$. This strategy

works because all divergences used in Theorem 2.1 of Ghosal et al. (2000) are f-divergences, or can be made equivalent to an f-divergence after applying some linear interpolation (see Lemma S.2 in the Supplementary Material).

Remark 2. We cannot make progress on divergences like $\mathcal{H}_p^2(g_0, g) = \int\int\{\sqrt{p_{g_0}(y \mid x)} - \sqrt{p_g(y \mid x)}\}^2 \, dy \; F_X(dx)$ because there is no information in the data about $g_0(t, x)$ for $t > C$. The best we can do is control $\int\int_0^C\{\sqrt{p_{g_0}(y \mid x)} - \sqrt{p_g(y \mid x)}\}^2 \, dy \; F_X(dx)$, which is accomplished by $\mathcal{H}^2(g_0, g)$, as well as $H_q^2(g_0, g)$ when $S_C(y \mid x)$ (the probability that an individual is not censored before the end of the study) is bounded away from 0.

5 Illustrations

5.1 Simulation Experiment

We conduct a simulation study to assess (i) the ability of MBART and CoxBART to capture nonlinear relationships and (ii) to assess what one loses when CoxBART is used when the proportional hazards assumption fails. The function $f(x) = \sin(\pi x_1 x_2) + 2(x_3 - 0.5)^2 + x_4 + 0.5x_2$ used in S2 and S3 is a nonlinear function described by Friedman (1991), having linear, nonlinear, and interaction effects. We consider the following models for the hazard.

- **S1, Cox:** $h(t \mid x) = \exp\{\sum_{j=1}^P x_j\} \, h_{\text{Gam}}(t)$ where $P = 10$.
- **S2, Semiparametric Exponential** $h(t \mid x) = \exp\{f(x)\}$, $P = 10$;
- **S3, Nonparametric Cox:** $h(t \mid x) = \exp\{f(x)\} \, h_{\text{Gam}}(t)$, $P = 5$;
- **S4, SLML:** T_i has a Weibull distribution with $\kappa(x) = 0.7 + 1.3x_7$ and scale parameter $1 + 0.25\sum_{j=1}^6 x_j + 2.5x_7$ with $P = 10$. This simulation setting is taken from Sparapani et al. (2016), and strongly violates the proportional hazards assumption.
- **S5, ZK:** $T_i \sim \text{Gam}\{\alpha(x), 0.5\}$ where $\alpha(x) = 0.5 + 0.3|\sum_{j=11}^{15} x_j|$ and $P = 25$. This simulation is taken from Zhu and Kosorok (2012) and strongly violates the proportional

hazards assumption.

Covariates are simulated independently from a Uniform$(0, 1)$ distribution, with the exception of S5 where the covariates X_i are multivariate normal with mean 0 and covariance matrix V with entries $V_{jk} = (0.75)^{|j-k|}$. The function $h_{\text{Gam}}(t)$ used in S1 and S3 as a baseline hazard denotes the hazard function of a Gam$(3, 1)$ random variable. For each simulation setting we took $N = 500$, except for the ZK setting where we took $N = 300$ to match Zhu and Kosorok (2012).

The purpose of S1 is to determine how much is lost by the nonparametric models when a simple semiparametric model holds. The purpose of S2 and S3 is to assess how much is lost if we use MBART instead of CoxBART when the proportional hazards assumption holds. Settings S4 and S5 are designed to assess how well MBART performs relative to other nonparametric techniques like random survival forests (Ishwaran et al., 2008), as well as how much is lost using CoxBART when the proportional hazards assumption fails.

Figure 2 displays randomly sampled survival curves for the settings S3, S4, and S5. Setting S3 obeys the proportional hazards assumption; consequently, the shape of each survival curve is similar and there is no curve crossing. Setting S5 also has no curve crossing because the Gam$(\alpha, 0.5)$ distribution is stochastically increasing in α; the shape of the survival curve, however, varies substantially. Depending on the value of α, both decreasing and increasing hazard functions are possible under S5. Setting S4 allows for both a variety of shapes for the survival curves and the possibility for the curves to cross. We compare the following methods.

- **MBART-Light:** The MBART model with mild shrinkage towards a proportional hazards model. The baseline model is an exponential BART model (i.e., Weibull-BART with $\kappa = 1$). This is the default prior described in Section 2.2.1.

- **MBART-Heavy:** The same as the default MBART model, but with $\gamma = 0.3$ so that most trees in the ensemble a-priori do not split on any covariates. This encourages the

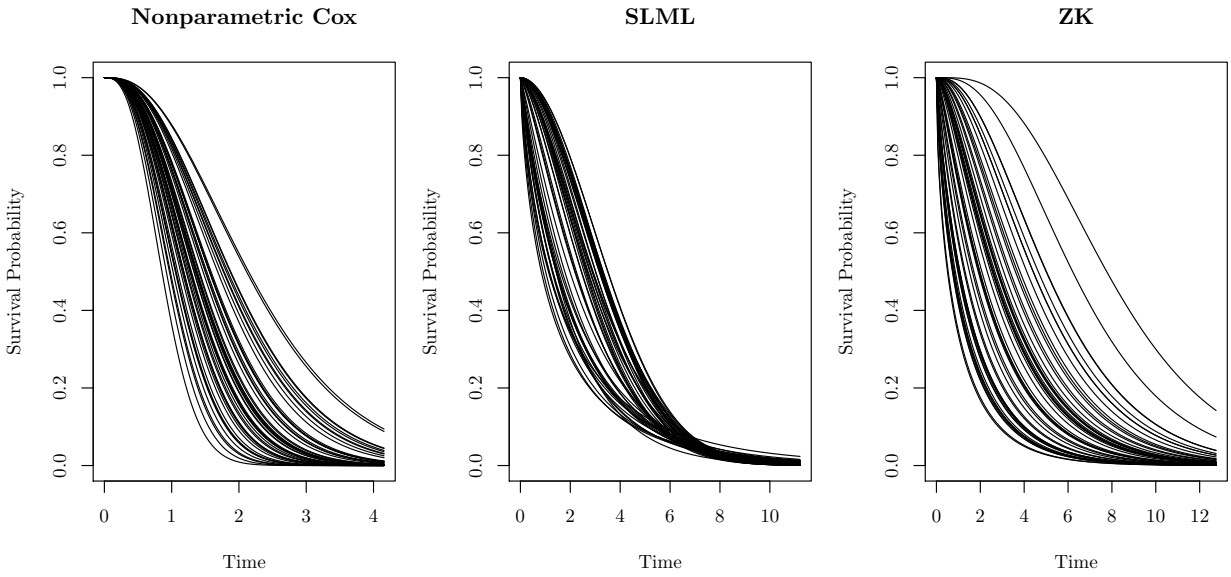

Figure 2: Samples of survival functions $S(t \mid X_i)$ for 50 samples of X_i for each model.

model to resemble the CoxBART2 base model described in Section 3.

- **CoxBART:** The CoxBART model with the default prior.
- **Cox-Linear:** The semiparametric Cox proportional hazards model fit with the log-linear link $h(t \mid x) = \lambda(t) \exp(x^\top \beta)$. This is fit with the `coxph` function in the `survival` package in `R`.
- **Random Survival Forests:** The random survival forests algorithm (Ishwaran et al., 2008) fit with the `randomForestSRC` package in `R`. This fits a fully-nonparametric model to the survival function, and does not invoke the proportional hazards assumption.
- **Exp-BART:** the exponential BART base model used with our MBART models. This is included to determine if the base model is adequate by itself.

We attempted to fit the `surv.bart` function in the `BART` package in `R`. Unfortunately, because we do not coarsen the time scale, the algorithm was too memory intensive.

We assess the performance of each method according to how well they estimate the conditional function $S(t \mid x)$. As a measure of accuracy, we consider the average integrated

Setting	Cox-Linear	CoxBART	RSF	MBART-Light	MBART-Heavy	Exp-BART
S1	**1.00**	3.13	10.2	2.91	2.88	28.2
S2	1.22	**1.00**	1.25	1.29	1.31	1.24
S3	1.33	**1.00**	1.50	1.13	1.06	24.9
S4	1.26	1.37	**1.00**	**1.00**	**1.00**	2.32
S5	2.89	1.05	1.08	**1.00**	1.02	1.68

Table 1: Results for the simulation study described in Section 5.1 for the settings S1–S5. Each entry is the integrated mean squared error in estimating the survival function. All results are relative to the best performing method, i.e., the best performing method always has the result 1.00.

squared distance between the estimated survival function $\widehat{S}(t \mid x)$ and the true survival function $S_0(t \mid x)$, MSE $= \int \int_0^{T_{\max}} \{\widehat{S}(t \mid x) - S_0(t \mid x)\}^2 \, dt \; F_X(dx)$ where F_X denotes the true distribution of the predictors and $T_{\max}$ is the maximum of all observed survival times in the sample. Because MSE cannot be computed in closed form, we approximate the integral numerically as $N_\star^{-1} \sum_{i=1}^{N_\star} \int_0^{T_{\max}} \{\widehat{S}(t \mid X_i^\star) - S_0(t \mid X_i^\star)\}^2 \; dt$ where the dt integral is computed numerically and the $X_1^\star, \ldots, X_{N_\star}^\star$ are held-out covariates sampled independently from F_X. Results are averaged over 200 simulated datasets. We used $N_* = 100$ held-out covariates and approximated the dt integral using a grid of size 500.

Results are given in Table 1. Except for S1, either CoxBART or MBART-Light performs the best. Predictably, CoxBART performs the best when the nonparametric proportional hazards model holds, while MBART performs best when it does not. Curiously, MBART has mediocre performance under S2, but does manage to outperform CoxBART under S1 despite not making the proportional hazards assumption. The performance of Exp-BART indicates that the baseline model alone is typically not adequate, and performs very poorly in S1 and S3 where the baseline hazard is far away from an exponential.

To better understand how the presence of irrelevant predictors affects the resulting estimators, we conducted an additional simulation under S4 with the number of predictors P being increased. To ensure that the comparison with Cox-Linear is fair, we use a lasso penalty implemented with the `glmnet` package in `R` (Tibshirani, 1997). We consider only

P	CoxNet	CoxBART	RSF	MBART	Exp-BART
25	1.67	1.31	1.17	**1.00**	2.33
50	1.56	1.24	1.25	**1.00**	2.17
100	1.53	1.27	1.42	**1.00**	2.20
200	1.46	1.27	1.47	**1.00**	2.22
400	1.37	1.23	1.50	**1.00**	2.09

Table 2: Results for the simulation experiment in Section 5.1 for setting S4 with $P \in \{25, 50, 100, 200, 400\}$. CoxNet refers to the Cox-Linear approach using the lasso. Each entry is the integrated mean squared error in estimating the survival function. All results are relative to the best performing method; if the result is 1.00, i.e., the best performing method always has the result 1.00.

MBART-Light, as the performance of MBART-Light and MBART-Heavy is similar in Table 1. Results are given in Table 2 for $P \in \{25, 50, 100, 200, 400\}$. Across all settings, we see that MBART outperforms the other methods, and that all methods except for random survival forests are robust to the inclusion of irrelevant predictors. The random survival forests algorithm performs poorly as the dimensionality of the problem increases; at $P = 50$, random survival forests performs the same as CoxBART, but by $P = 400$ it performs worse than all methods except for Exp-BART. The lack of robustness of random forests to irrelevant predictors has been noted in other works (Zhu et al., 2015; Linero, 2018) as well, and does not depend on the selection of tuning parameters.

We conclude that both MBART and CoxBART are generally effective for situations when the semiparametric Cox proportional hazards model $h(t \mid x) = \lambda(t) \exp(x^\top \beta)$ fails. For situations where the nonparametric proportional hazards model $h(t \mid x) = \lambda(t) \exp\{f(x)\}$ holds CoxBART performs best, while for situations where the nonparametric proportional hazards model fails MBART performs best. Both methods are highly robust to the inclusion of irrelevant predictors, which is not true of random survival forests.

5.2 Liver Disease Data

We analyze a publicly available dataset on time to death (scaled to have standard deviation 1) for patients of primary biliary cirrhosis (PBC), a chronic disease in which the bile ducts of the

liver are slowly destroyed (available as the `pbc` dataset in `randomForestSRC`). The survival times were subject to right censoring due to either liver transplant or survival beyond the end of the study. We select this study for re-analysis to demonstrate the capability of our methods (MBART and CoxBART) to accommodate unknown non-linear and time-varying effects of the covariates on the hazard function, because it has observed in prior analyses that at least two covariates (Bilirubin and Protime) have such effects. We first fit the usual semiparametric Cox proportional hazards model $\lambda(t \mid x) = \lambda_0(t)e^{x^\top \beta}$ using the following covariates: age, baseline bilirubin level (Bilirubin), baseline albumin level (Albumin), presence of edema, and prothrombin time (Protime). An analysis of the martingale residuals (Fleming and Harrington, 2011, Section 4.6) of the predictors show that some covariates have non-linear and possibly time-varying effects on the hazard rate; for example, the top left panel of Figure 3 suggests a non-linear effect of bilirubin which can be accommodated by using a log transformation (top right panel). CoxBART detects these effects automatically, eliminating the need to look for appropriate transformations, and also outperforms a similar generalized additive model based on natural cubic splines. To demonstrate this, we performed 5-fold cross-validation and computed the held-out deviance $-2 \log \mathrm{PL}^\star = -2 \sum_{k=1}^{5} \log \mathrm{PL}_{-k}$ where PL_{-k} denotes the Cox partial likelihood obtained by regressing (Y_i, δ_i) in the k^{th} fold on the estimated risk $\widehat{g}(X_i)$, where CoxBART used the posterior mean of $g(X_i)$ for $\widehat{g}(X_i)$. This was replicated across 10 different splits into 5-folds and averaged over the 10 splits. Results are given in Table 3. The best performing proportional hazards model uses hand-selected transformations for the continuous predictors — in particular, taking the logarithm of the bilirubin level — but CoxBART performs nearly as well (and much better than the linear and GAM models).

While CoxBART is able to automatically find an appropriate transformation of the predictors, the Schoenfeld residuals in Figure 3 (middle left) suggest a time-varying effect of prothrombin time (Protime) which cannot be accommodated using a non-linear transformation, implying the proportional hazards assumption does not hold. A formal test for the violation

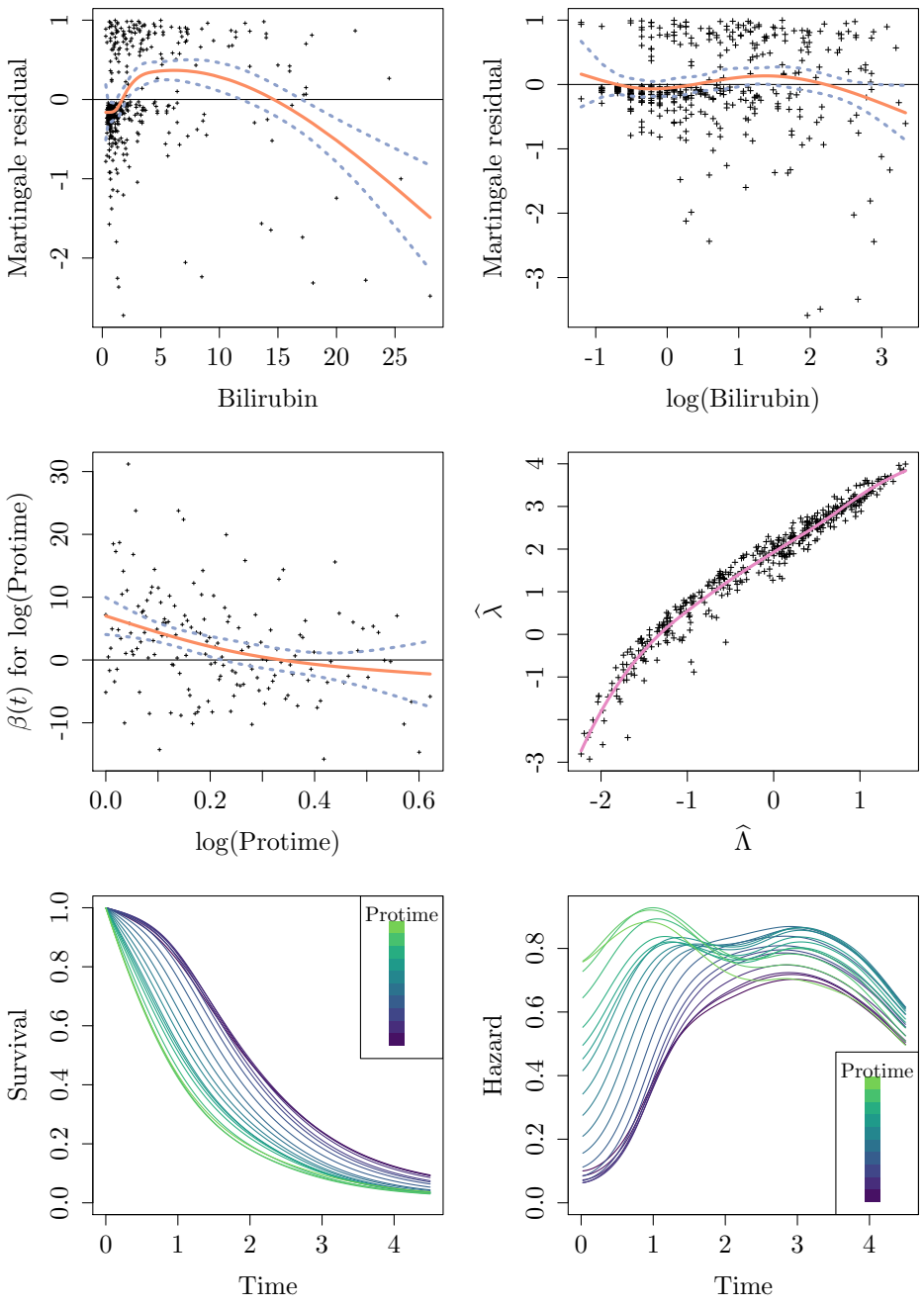

Figure 3: Top left: plot of bilirubin against martingale residuals from the model with linear effects. Top right: plot of log-bilirubin against martingale residuals from the model using log-bilirubin. Middle left: plot of log-protime against an estimate of the time-varying effect of protime obtained from the `cox.zph` function in `R`. Middle right: cumulative risk from the MBART model against the estimated risk from the CoxBART model. Bottom left: mean MBART survival functions for different levels of protime (darker means lower protime). Bottom right: estimated MBART hazard functions for different levels of protime. Smooth solid lines for the top and middle plots are spline estimates of the mean relationship between the variables, while dashed lines are 95% confidence bands for the splines.

Method	CoxBART	CoxLM	CoxLog	CoxGAM
Heldout Deviance	1014	1058	1008	1034

Table 3: The five-fold cross-validated estimate $-2 \log \mathrm{PL}^\star$ for the different proportional hazards models considered. CoxLM, CoxLog, and CoxGAM denote the semiparametric Cox proportional hazards obtained with no transformations, log transformations, and natural cubic splines, respectively.

of the proportional hazards assumption using the `cox.zph` function in `R` gives a significant global test, primarily driven by the effect of protime (P-value ≈ 0.003). To show that MBART is able to accommodate this failure of the proportional hazards assumption, we use 5-fold cross validation to compare the fit of MBART to the CoxBART2 model described in Section 3; we use CoxBART2 here because it admits a density and so can be readily compared to MBART using likelihood-based criteria. We measure predictive accuracy using held-out deviance of the data $\mathrm{dev} = -2 \sum_{i=1}^{N} \delta_i \log \widehat{f}_{\rho(i)}(Y_i \mid X_i) + (1 - \delta_i) \log \widehat{S}_{\rho(i)}(Y_i \mid X_i)$ where $\widehat{f}_{\rho(i)}$ and $\widehat{S}_{\rho(i)}$ denote the estimated density and survival function computed on the fold which excluded observation i. This experiment was replicated on all 10 splits. MBART outperformed CoxBART2 over all 10 splits, with the average difference in deviance between the two models being $\overline{\mathrm{dev}}_{\mathrm{Cox}} - \overline{\mathrm{dev}}_{\mathrm{M}} = 7.81$. Hence MBART is able to obtain a better fit by weakening the proportional hazards assumption.

Interestingly, MBART and CoxBART give similar risk measures for each individual; see Figure 3 (middle right), where $\widehat{\lambda}$ is the posterior mean of $g(X_i)$ for the CoxBART model and $\widehat{\Lambda}$ is the posterior mean of the cumulative hazard $\int_0^{\max(Y_i)} \lambda(t \mid X_i, \theta)\, \Phi\{g(t, X_i)\}\, dt$. We see that MBART detects the same risk structure as CoxBART. The bottom panels of Figure 3 show the estimated survival function and hazard function as a function of protime when all other variables are fixed at their sample medians. We see that the estimated hazard functions are clearly not proportional; in particular, individuals with high values of protime have relatively constant risk, while individuals with low values of protime have a small risk at the beginning of the study and large risk at the end. The corresponding survival function estimates for high values of protime are similar to what was observed under the ZK simulation

setting, which violated the proportional hazards assumption.

We conclude that MBART is able to effectively model the PBC data. Importantly, MBART finds the same structure as CoxBART, while also being able to account for failure of the proportional hazards assumption. This allows MBART to be used without needing to check assumptions regarding proportional hazards or lack of functional form fit.

6 Discussion

In this paper we introduced two survival models for right-censored data using Bayesian additive regression trees. The first model is based on the observation that a response T_i with hazard $h(t \mid x)$ can be modeled as the first occurrence of a non-homogeneous Poisson process with intensity function $h(t \mid x)$; by taking $h(t \mid x)$ to correspond to a thinned Poisson process $\lambda(t \mid x, \theta)\, \Phi\{g(t, x)\}$ where $\lambda(t \mid x, \theta)$ is chosen to be tractable, we were able to construct a two-layer data augmentation scheme for updating the function $g(t, x)$. A single-layer strategy using this idea has been used for Gaussian processes (Fernández et al., 2016), but is not applicable to our model due to our use of discrete parameters and only shrinks towards marginal models $\lambda(t)$. We also establish theoretically that our thinned Poisson process model obtains near-minimax optimal rates of posterior contraction in the high-dimensional setting.

The second model we introduced is a nonparametric variant of the Cox proportional hazards model, which takes $h(t \mid x) = \lambda(t) \exp\{g(x)\}$ with the baseline hazard function $\lambda(t)$ modeled nonparametrically. Inference in this model proceeds from the Cox partial likelihood $\mathrm{PL}(g)$. By using a Bayesian justification of the Cox proportional hazards model (Sinha et al., 2003) we were able to both construct a tractable Gibbs sampling algorithm and obtain a nonparametric estimate of the cumulative hazard function.

An interesting avenue for future research is to develop methods for summarizing the BART posteriors for these models. A key advantage that classical semiparametric models have is that they reduce the effect of covariates to interpretable scalar parameters. One possibility in

this direction is to develop tools which systematically project the BART posterior onto more interpretable models (Woody et al., 2020).

A Bayesian Backfitting for CoxBART

Let $\mathcal{T}_{-m}$ and $\mathcal{M}_{-m}$ denote the tree topologies and leaf node parameters for all trees except for the m^{th}, and let $x \rightsquigarrow (\ell, m)$ mean that x is associated to leaf ℓ of $\mathcal{T}_m$. In order to implement the generalized Bayesian backfitting approach of Hill et al. (2019) we require $\pi(\mathcal{T}_m \mid \mathcal{T}_{-m}, \mathcal{M}_{-m}, \phi, \mathcal{D})$ and $\pi(\mathcal{M}_m \mid \mathcal{T}_m, \mathcal{T}_{-m}, \mathcal{M}_{-m}, \phi, \mathcal{D})$.

For fixed m, let $\zeta_i = g(X_i) - \text{Tree}(X_i; \mathcal{T}_m, \mathcal{M}_m)$ and let $r_i = \sum_{j: i \in \mathscr{R}_j} \phi_j \, \delta_j$. Marginalizing the joint distribution of $(\mathcal{D}, \phi, g)$ from Section 2.3 over $\mathcal{M}_m$ gives

$$
\begin{aligned}
&\pi(\mathcal{T}_m \mid \mathcal{T}_{-m}, \mathcal{M}_{-m}, \phi, \mathcal{D}) \\
&\quad \propto \pi_{\mathcal{T}}(\mathcal{T}_m) \prod_{\ell} \int \prod_{i \rightsquigarrow (\ell,m)} \exp\left\{\delta_i(\zeta_i + \mu) - r_i e^{\zeta_i} e^{\mu}\right\} \times \frac{\beta_\mu^{\alpha_\mu}}{\Gamma(\alpha_\mu)} \exp\{\alpha_\mu \mu - \beta_\mu e^{\mu}\} \, d\mu \\
&\quad = \pi_{\mathcal{T}}(\mathcal{T}_m) \prod_{\ell} \int \frac{\beta_\mu^{\alpha_\mu}}{\Gamma(\alpha_\mu)} \exp\left\{Z_\ell + \mu(\alpha_\mu + S_\ell) - e^{\mu}(\beta_\mu + E_\ell)\right\} \, d\mu,
\end{aligned}
\tag{9}
$$

where $S_\ell = \sum_{i \rightsquigarrow (\ell,m)} \delta_i$, $Z_\ell = \sum_{i \rightsquigarrow (\ell,m)} \delta_i \zeta_i$, and $E_\ell = \sum_{i \rightsquigarrow (\ell,m)} r_i e^{\zeta_i}$. The integrand above is the kernel of a gamma distribution, giving $\pi_{\mathcal{T}}(\mathcal{T}_m) \prod_\ell e^{Z_\ell} \frac{\beta_\mu^{\alpha_\mu}}{\Gamma(\alpha_\mu)} \frac{\Gamma(\alpha_\ell)}{\beta_\ell^{\alpha_\ell}}$ where $\beta_\ell = \beta_\mu + E_\ell$ and $\alpha_\ell = \alpha_\mu + S_\ell$. Additionally, by computations identical to those in (9), it follows that the full conditional of $\mathcal{M}_m$ is $\pi(\mathcal{M}_m \mid \mathcal{T}_m, \mathcal{T}_{-m}, \mathcal{M}_{-m}, \phi, \mathcal{D}) \propto \prod_\ell \exp\{\mu_{m\ell}\alpha_\ell - e^{\mu_{m\ell}}\beta_\ell\}$, i.e., $\mu_{m\ell} \overset{\text{ind}}{\sim} \log \text{Gam}(\alpha_\ell, \beta_\ell)$.

B Bayesian Backfitting for Weibull-BART

After performing the data augmentation in Proposition 1, the likelihood becomes

$$
\prod_{i=1}^{N} \left\{ (\kappa \, e^{r(X_i)} \, Y_i^{\kappa-1})^{\delta_i} \prod_{j=1}^{J_i} \kappa \, e^{r(X_i)} \, W_{ij}^{\kappa-1} \right\} \times \prod_i \exp\{-e^{r(X_i)} \, Y_i^{\kappa}\}.
$$

Hold $\mathcal{T}_m^r$ fixed, set $\zeta_i = r(X_i) - \text{Tree}(X_i; \mathcal{T}_m^r, \mathcal{M}_m^r)$, and define the leaf node sufficient statistics $N_\ell = \sum_{i \rightsquigarrow (\ell,m)} J_i + \delta_i$, $E_\ell = \sum_{i \rightsquigarrow (\ell,m)} e^{\zeta_i} Y_i^\kappa$, $S_\ell = \sum_{i \rightsquigarrow (\ell,m)} \delta_i \log Y_i + \sum_{j=1}^{J_i} \log W_{ij}$, and $Z_\ell = \sum_{i \rightsquigarrow (\ell,m)} \zeta_i (J_i + \delta_i)$, where we write $i \rightsquigarrow (\ell, m)$ if X_i is associated to leaf ℓ of tree m. Then the likelihood of $\mathcal{M}_m^r$ holding all other quantities fixed is

$$\prod_\ell \frac{\kappa^{N_\ell} \beta_\eta^{\alpha_\eta}}{\Gamma(\alpha_\eta)} \exp\left\{\eta_{m\ell}(\alpha_\eta + N_\ell) - e^{\eta_{m\ell}}(\beta_\eta + E_\ell)\right\} \exp\{Z_\ell + (\kappa - 1) S_\ell\}.$$

This is proportional to a product of $\log \text{Gam}(\alpha_\eta + N_\ell, \beta + E_\ell)$ densities. Integrating out the $\eta_{m\ell}$'s, we obtain the conditional distribution

$$\pi(\mathcal{T}_m^r \mid \mathcal{T}_{-m}^r, \mathcal{M}_{-m}^r, \eta_0, g, \mathcal{D}) \propto \prod_\ell \frac{\kappa^{N_\ell} \beta_\eta^{\alpha_\eta}}{\Gamma(\alpha_\eta)} \times \frac{\Gamma(\alpha_\ell)}{\beta_\ell^{\alpha_\ell}} \exp\{Z_\ell + (\kappa - 1) S_\ell\}, \tag{10}$$

where $\alpha_\ell = \alpha_\eta + N_\ell$ and $\beta_\ell = \beta_\eta + E_\ell$. Similarly, the full conditional of $\eta_{m\ell}$ is $\log \text{Gam}(\alpha_\ell, \beta_\ell)$. We can now update $(\mathcal{T}_m^r, \mathcal{M}_m^r)$ by first updating $\mathcal{T}_m^r$ using Metropolis-Hastings and then sampling $\mathcal{M}_m^r$ from its full conditional.

C Proof of Theorem 1

Our results are proved by verifying sufficient conditions similar to those of Ghosal et al. (2000). Let u_0 be an α-smooth function depending on $D \leq P + 1$ many coordinates of (t, x) such that $f_0 = f_{u_0}$; for example, we can take $u_0(t, x) = \Phi^{-1}\{R(t, x)\}$, which is well-defined because $R(t, x)$ is bounded away from 0 and 1 by Condition H. In addition to the integrated Hellinger divergence $\mathcal{H}(u, u_0)$, we introduce two other Kullback-Leibler type divergences

$$\text{KL}(u_0 \| u) = \int \int f_{u_0}(y \mid x) \log \frac{f_{u_0}(y \mid x)}{f_u(y \mid x)} \, \nu(dy) \, F_X(dx) \qquad \text{and}$$
$$V(u_0 \| u) = \int \int f_{u_0}(y \mid x) \left(\log \frac{f_{u_0}(y \mid x)}{f_u(y \mid x)} \right)^2 \nu(dy) \, F_X(dx).$$

We define the *integrated Kullback-Leibler neighborhood* of u_0 as

$$\mathrm{KL}(u_0, \epsilon) = \left\{u : \max\{\mathrm{KL}(u_0 \| u), V(u_0 \| u)\} \leq \epsilon^2\right\}$$

Theorem 1 is proved by verifying the following sufficient conditions; the fact that these conditions are sufficient is proved in Section S.3 of the Supplementary Material.

(A1) **Prior thickness:** $\Pi\{u \in \mathrm{KL}(u_0, \epsilon_n)\} \geq \exp(-K_1 n\epsilon_n^2)$ for some constant K_1.

(A2) **High Mass Sieve:** for every n there exists a set $\mathcal{F}_n$ of $u : [0,1]^{P+1} \to \mathbb{R}$ such that $\Pi(\mathcal{F}_n^c) \leq \exp\{-(K_1 + 4)n\epsilon_n^2\}$.

(A3) **Low Entropy Sieve:** The $\mathcal{F}_n$ from (A2) satisfies $\log N(\bar{\epsilon}_n, \mathcal{F}_n, \mathcal{H}) \leq K_2 n\bar{\epsilon}_n$.

If (A1)–(A3) for some $\epsilon_n \leq \bar{\epsilon}_n \downarrow 0$ such that $n\epsilon_n^2 \to \infty$, then the convergence rate is at least $\bar{\epsilon}_n$. The quantity $\log N(\epsilon, \mathcal{F}, d)$ denotes the log of the ϵ-covering number of $\mathcal{F}$ with respect to a metric d (see, for example, Van Der Vaart and Wellner, 1996). To establish A1—A3, we use the following results about SBART priors with respect to the supremum norm $\|u\|_\infty = \sup_{t,x} |u(t,x)|$; these follow from Condition P and the fact that u_0 is bounded, α-Hölder smooth, and depends on only $D \leq P + 1$ coordinates. These statements are proved in the appendix of Li et al. (2020) and are similar to Theorem 3 and Lemma S.1 of Linero and Yang (2018).

(B1) $\Pi\{\|u - u_0\|_\infty \leq A_1\epsilon_n\} \geq \exp\{-A_2 n\epsilon_n^2\}$ for some positive constants A_1, A_2.

(B2) For any constant A_3, we can find a $\mathcal{G}_n$ and a constant A_4 such that $\Pi(u \notin \mathcal{G}_n) \leq \exp\{-A_3 n\epsilon_n^2\}$ and $\log N(\bar{\epsilon}_n, \mathcal{G}_n, \|\cdot\|_\infty) \leq A_4 n\bar{\epsilon}_n^2$ and $\bar{\epsilon}_n$ is a constant multiple of ϵ_n.

We remark that it is in establishing that (B1) and (B2) hold that we make use of the fact that $\epsilon_n = n^{-\alpha/(2\alpha+D)} \log(n)^t + \sqrt{D\log(P+1)/n}$, as this is required by Theorem 3 and Lemma S.1 of Linero and Yang (2018). To leverage (B1)–(B2) to prove (A1)–(A3), we prove the following lemma in the Supplementary Material.

Lemma 1. *There exist constants $C_{\mathcal{H}}$ and C_K depending only on the link function $\Phi(\mu)$ and the quantity $\Lambda(C) = \int_0^C \lambda(t)\, d$ such that, for any measurable functions u and v from $[0,1]^{P+1}$ to $\mathbb{R}$,*

$$\begin{aligned}
\mathcal{H}^2(u,v) &\le C_{\mathcal{H}}^2 \|u-v\|_\infty^2 \exp(C_{\mathcal{H}}\|u-v\|_\infty), \\
\mathrm{KL}(u\|v) &\le C_K^2 \|u-v\|_\infty^2 \exp(C_K\|u-v\|_\infty)\,(1+\|u-v\|_\infty), \\
V(u\|v) &\le C_K^2 \|u-v\|_\infty^2 \exp(C_K\|u-v\|_\infty)\,(1+\|u-v\|_\infty)^2.
\end{aligned}$$

To establish (A1), we note by Lemma 1 that, for sufficiently small ϵ_n,

$$\Pi\{\mathrm{KL}(u_0, 2A_1C_K\epsilon_n)\} \ge \Pi(\|u-u_0\|_\infty \le A_1\epsilon_n) \ge \exp(-A_2 n\epsilon_n^2).$$

To simplify notation, we redefine ϵ_n to $2A_1C_K\epsilon_n$, so that this is equivalent to (A1) with $K_1 = A_2/(4C_K^2A_1^2)$; we note that replacing ϵ_n with a constant multiple does not affect (B2), as the constants can be absorbed into A_3 and A_4.

To establish (A2) and (A3), we set $\mathcal{F}_n$ equal to $\mathcal{G}_n$ from (B2) with $A_3 = (K_1+4)$. Using Lemma 1, for sufficiently large n, any $\bar{\epsilon}_n$-net of $\mathcal{G}_n$ with respect to the uniform metric $d(u,v) = \|u-v\|_\infty$ is also a $2C_{\mathcal{H}}\bar{\epsilon}_n$-net of $\mathcal{G}_n$ with respect to $\mathcal{H}(u,v)$. Therefore $\log N(2C_{\mathcal{H}}\bar{\epsilon}_n, \mathcal{G}_n, \mathcal{H}) \le A_4 n\epsilon_n^2$, establishing (A3) with $2C_{\mathcal{H}}\bar{\epsilon}_n$ in the role of $\bar{\epsilon}_n$ and $K_2 = A_4/(4C_{\mathcal{H}}^2)$.

D Proof of Theorem 2

It is again sufficient to check conditions (A1)–(A3), with the modification that the divergences $\mathcal{H}$, KL, and V are replaced with $\mathcal{H}_q(q_0,q)$, $\mathrm{KL}_q(q_0\|q) = \int\int \sum_\delta q_{u_0}(y,\delta) \log \frac{q_{u_0}(y,\delta)}{q(y,\delta)}\, \nu(dy)\, F_X(dx)$ and $V_q(q_0\|q) = \int\int \sum_\delta q_{u_0}(y,\delta) \left(\log \frac{q_{u_0}(y,\delta)}{q(y,\delta)}\right)^2 \nu(dy)\, F_X(dx)$. Let G_u denote the joint distribution of $(\widetilde{Y}_i, \delta_i, X_i)$ when $\widetilde{Y}_i \sim f_u(y \mid x)$,with (δ_i, X_i) having their true distributions. Then it is easy to verify by definition that $\mathcal{H}(G_u, G_v) = \mathcal{H}(u,v)$, $\mathrm{KL}(G_u\|G_v) = \mathrm{KL}(u\|v)$, and

$V(G_u\|G_v) = V(_u|\ v)$ where $\mathcal{H}^2(G_u, G_v) = \int(\sqrt{dG_u} - \sqrt{dG_v})^2$, $\mathrm{KL}(G_u\|G_v) = \int \log \frac{dG_u}{dG_v}\ dG_u$, and $V(G_u\|G_v) = \int \left(\log \frac{dG_u}{dG_v}\right)^2\ dG_u$ denote the usual Hellinger and Kullback-Leibler divergences between two distributions. Similarly, by definition we have $\mathcal{H}(\mathscr{T}^{-1}G_u, \mathscr{T}^{-1}G_v) = H_q(u, v)$, $\mathrm{KL}(\mathscr{T}^{-1}G_u\|\mathscr{T}^{-1}G_v) = \mathrm{KL}_q(u\|v)$, and $V(\mathscr{T}^{-1}G_u\|\mathscr{T}^{-1}G_v) = V_q(u\|v)$, where $\mathscr{T}$ is the mapping $(\widetilde{Y}_i, \delta_i, X_i) \mapsto (Y_i, \delta_i, X_i)$ and $\mathscr{T}^{-1}G$ denotes the distribution on (Y_i, δ_i, X_i) induced by $(\widetilde{Y}_i, \delta_i, X_i) \sim G$.

As $\mathcal{H}(G_u, G_v)$ and $\mathrm{KL}(G_u\|G_v)$ are f-divergences (Ali and Silvey, 1966) and $(\widetilde{Y}_i, \delta_i, X_i) \mapsto (Y_i, \delta_i, X_i)$ is a measurable transformation, we immediately have $\mathcal{H}_q(u_0, u) \leq \mathcal{H}(u_0, u)$ and $\mathrm{KL}_q(u_0\|u) \leq \mathrm{KL}(u_0\|u)$. In Lemma S.2, we use the properties of f-divergences to prove that, while $V(G_u\|G_v)$ is not an f-divergence, we can still obtain a similar inequality

$$V_q(u\|v) = V(\mathscr{T}^{-1}G_u\|\mathscr{T}^{-1}G_v) \leq \psi\left(\left\|\frac{dG_u}{dG_v}\right\|_\infty\right) V(G_u\|G_v) = \psi\left(\left\|\frac{f_u}{f_v}\right\|_\infty\right) V(u\|v),$$

where $\psi(x) = \max\{\frac{2x-e}{e\log(x)^2}, 1\}$. In the proof of Lemma 1 it is shown that $\|f_{u_0}/f_u\|_\infty$ can be made arbitrarily close to 1 when $\|u_0 - u\|_\infty$ is sufficiently small. Hence, for sufficiently large n, $\|u_0 - u\|_\infty \leq \epsilon_n$ implies $V_q(u_0\|u) \leq V(u_0\|u)$.

Given these results, we have $\Pi\{u \in \mathrm{KL}_q(u_0, \epsilon_n)\} \geq \Pi(u \in \mathrm{KL}(u_0, \epsilon_n))$ for sufficiently large n, and the same set $\mathcal{F}_n$ can be chosen to satisfy conditions (A2) and (A3) as in the proof of Theorem 2 because $\mathcal{H}_q(u_0, u) \leq \mathcal{H}(u_0, u)$ implies $\log N(\bar{\epsilon}_n, \mathcal{F}_n, \mathcal{H}_q) \leq \log N(\bar{\epsilon}_n, \mathcal{F}_n, \mathcal{H})$. Because (A1)–(A3) have been shown to hold, we obtain the desired rate of convergence.

References

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response

data. *Journal of the American Statistical Association*, 88:669–679.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B*, pages 131–142.

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*.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2010). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont, CA.

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.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771.

Dunson, D. B. (2009). Bayesian nonparametric hierarchical modeling. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):273–284.

Fernández, T., Rivera, N., and Teh, Y. W. (2016). Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems*, pages 5021–5029.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and Survival Analysis*, volume 169. John Wiley & Sons.

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

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. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68.

Hill, J., Linero, A. R., and Murray, J. (2019). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7.

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

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag New York, 1 edition.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag New York, 1 edition.

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

Ishwaran, H., Kogalur, U. B., Blackson, E. H., and Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3):841–860.

Li, Y., Linero, A. R., and Murray, J. S. (2020). Adaptive conditional distribution estimation with Bayesian decision tree ensembles. *arXiv preprint arXiv:2005.02490*.

Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559.

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

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

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference–why and how. *Bayesian Analysis*, 8(2).

Murray, J. S. (2020). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*. To appear.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31:705–767.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493.

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.

Sinha, D., Ibrahim, J. G., and Cen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90(3):629–641.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in medicine*.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., and Scott, J. G. (2020). BART with targeted smoothing: An analysis of patient-specific stillbirth risk. *Annals of Applied Statistics*, 14(1):28–50.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer, New York.

Van Niekerk, J., Bakka, H., and Rue, H. (2020). A principled distance-based prior for the shape of the Weibull model. *arXiv preprint arXiv:2002.06519*.

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.

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

Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.

Supplementary Material for Bayesian Survival Tree Ensembles with Submodel Shrinkage

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

March 25, 2021

S.1 Auxiliary Results

Lemma S.1. *Let u and v be measurable mappings from $[0,1]^{P+1}$ to $\mathbb{R}$. Define*

$$H_u(t \mid x) = \int_0^t \lambda(s)\, \Phi\{u(s,x)\}\, ds,$$

where $\lambda(s)$ is a hazard function and $\Phi(\mu)$ is a link function satisfying Condition L. Then $\sup_{t\leq C} |H_u(t \mid x) - H_v(t \mid x)| \leq \|\phi\|_\infty \Lambda(C)\, \|u - v\|_\infty$, where $\Lambda(t) = \int_0^t \lambda(s)\, ds$. Further,

$$\sup_{t\leq C} \left| 1 - \sqrt{\frac{p_u(t \mid x)}{p_v(t \mid x)}} \right| \leq \frac{\{\mathcal{K} + \|\phi\|_\infty \Lambda(C)\}\|u - v\|_\infty}{2} \exp\left\{ \frac{\{\mathcal{K} + \|\phi\|_\infty \Lambda(C)\}\|u - v\|_\infty}{2} \right\}.$$

where $p_u(t \mid x) = \lambda(t)\, \Phi\{u(t,x)\}\, \exp\{-H_u(t \mid x)\}$.

Proof. Note that Condition L implies that $\|\phi\|_\infty \leq \mathcal{K}$. By Taylor's theorem, for $t \leq C$,

$$\begin{aligned}
|H_u(t \mid x) - H_v(t \mid x)| &\leq \int_0^t \lambda(s) |\Phi\{u(s,x)\} - \Phi\{v(s,x)\}|\, ds \\
&\leq \int_0^t \lambda(s) |u(s,x) - v(s,x)| \|\phi\|_\infty\, ds \\
&\leq \|\phi\|_\infty \|u - v\|_\infty \int_0^t \lambda(s) \leq \Lambda(C) \|\phi\|_\infty \|u - v\|_\infty.
\end{aligned}$$

This proves the first result. Next, let $\ell_u = \log p_u$ and $\ell_v = \log p_v$. Using the fact that $|1 - \exp(x)| \leq |x|e^{|x|}$ we have

$$\left|1 - \sqrt{\frac{p_u(t \mid x)}{p_v(t \mid x)}}\right| \leq \frac{|\ell_u - \ell_v|}{2} \exp\left(\frac{|\ell_u - \ell_v|}{2}\right).$$

By the triangle inequality we have $|\ell_u - \ell_v| \leq |\log \frac{\Phi(u)}{\Phi(v)}| + |H_u - H_v|$. By integrating Condition L we can show that $\log \frac{\Phi(u)}{\Phi(v)} \in [-\mathcal{K}\|u - v\|_\infty, \mathcal{K}\|u - v\|_\infty]$ so that $|\log \frac{\Phi(u)}{\Phi(v)}| \leq \mathcal{K}\|u - v\|_\infty$. Combining this with the previous bound on $|H_u - H_v|$ gives

$$\left|1 - \sqrt{\frac{p_u(t \mid x)}{p_v(t \mid x)}}\right| \leq \frac{\{\mathcal{K} + \|\phi\|_\infty \Lambda(C)\}\|u - v\|_\infty}{2} \exp\left\{\frac{\{\mathcal{K} + \|\phi\|_\infty \Lambda(C)\}\|u - v\|_\infty}{2}\right\}.$$

□

Lemma S.2. *Let $\mathscr{T}$ denote any measurable mapping between measurable spaces $\mathscr{X}$ to $\mathscr{Y}$. Let P and Q be probability distributions on $\mathscr{X}$ such that $\left\|\frac{dP}{dQ}\right\|_\infty < \infty$. Define $V(P, Q) = \int(\log \frac{dP}{dQ})^2 \, dP$. Then*

$$V(\mathscr{T}^{-1}P, \mathscr{T}^{-1}Q) \leq \psi(\|dP/dQ\|_\infty)V(P, Q)$$

where $\psi(x) = \max\{\frac{2x-e}{e\log(x)^2}, 1\}$; in particular, if $\|dP/dQ\|_\infty < e$, then $V(\mathscr{T}^{-1}P, \mathscr{T}^{-1}Q) \leq V(P, Q)$.

Proof. Let $g(x) = \log(x)^2$ and upper-bound $g(x)$ by the function $f(x) = \log(x)^2 I(x \leq e) + \left(\frac{2x}{e} - 1\right) I(x > e)$. It can be verified that $f(x)$ is a convex function with $f(1) = 0$, and hence f defines an f-divergence which we denote by $D_f(P, Q)$. It follows that

$$V(\mathscr{T}^{-1}P, \mathscr{T}^{-1}Q) \leq D_f(\mathscr{T}^{-1}P, \mathscr{T}^{-1}Q) \leq D_f(P, Q).$$

It can further be shown that $f(dP/dQ) = \psi(dP/dQ)g(dP/dQ)$ and that $\psi(x)$ is monotonically

increasing. Hence,

$$D_f(P,Q) = \int \psi(dP/dQ)\, g(dP/dQ)\; dQ$$
$$\leq \psi(\|dP/dQ\|_\infty) \int g(dP/dQ)\; dQ = \psi(\|dP/dQ\|_\infty) V(P,Q).$$

Hence $V(\mathscr{T}^{-1}P, \mathscr{T}^{-1}Q) \leq \psi(\|dP/dQ\|_\infty)\, V(P,Q)$, completing the proof. □

S.2 Proof of Lemma 1

We first bound $\mathcal{H}(u,v)$ in terms of $\|u-v\|_\infty$. By Lemma S.1

$$\int_0^C \{\sqrt{p_u(t \mid x)} - \sqrt{p_v(t \mid x)}\}^2\; dt = \int_0^C \left|1 - \sqrt{\frac{p_v(t \mid x)}{p_u(t \mid x)}}\right|^2 \; p_u(t \mid x)\; dt$$
$$\leq \frac{(\{\mathcal{K} + \|\phi\|_\infty\; \Lambda(C)\}\; \|u-v\|_\infty)^2}{4} \exp\left\{(\mathcal{K} + \|\phi\|_\infty\; \Lambda(C))\; \|u-v\|_\infty\right\}. \tag{S.1}$$

Next,

$$(e^{-H_u(C|x)} - e^{-H_v(C|x)})^2 = e^{-2H_v(C|x)} \left(\exp\left\{-H_u(C \mid x) + H_v(C \mid x)\right\} - 1\right)^2.$$
$$\leq e^{-2H_v(C|x)} \left(e^{\|\phi\|_\infty\; \Lambda(C)\; \|u-v\|_\infty} - 1\right)^2 \tag{S.2}$$
$$\leq (\|\phi\|_\infty \Lambda(C)\; \|u-v\|_\infty)^2\; e^{2\|\phi\|_\infty \Lambda(C)\; \|u-v\|_\infty}.$$

Integrating (S.1) and (S.2) with respect to $F_X(dx)$ and adding the results together gives

$$\mathcal{H}^2(u,v) \leq \frac{(\{\mathcal{K} + \|\phi\|_\infty \Lambda(C)\}\; \|u-v\|_\infty)^2}{4} e^{\{\mathcal{K} + \|\phi\|_\infty \Lambda(C)\}\; \|u-v\|_\infty}$$
$$+ (\|\phi\|_\infty \Lambda(C)\; \|u-v\|_\infty)^2\; e^{2\|\phi\|_\infty\; \Lambda(C)\; \|u-v\|_\infty}$$
$$\leq C_{\mathcal{H}}^2\; \|u-v\|_\infty^2\; \exp(C_{\mathcal{H}} \|u-v\|_\infty)$$

for some choice of $C_{\mathcal{H}}$ depending on $\mathcal{K}$, $\|\phi\|_\infty$, and $\Lambda(C)$. This proves the bound on $\mathcal{H}^2(u,v)$. To prove the bounds for $\mathrm{KL}(u\|v)$ and $V(u\|v)$, Lemma 8 of Ghosal and Van Der Vaart (2007)

(the proof of which does not depend on the dominating measure of the densities) gives the bounds

$$\begin{aligned}
\mathrm{KL}(u\|v) &\lesssim_{\Phi,\Lambda} \|u-v\|_\infty^2 \ \exp(C_{\mathcal{H}}\|u-v\|_\infty) \ \left[1+\log\left\|\frac{f_u(y\mid x)}{f_v(y\mid x)}\right\|_\infty\right] \\
V(u\|v) &\lesssim_{\Phi,\Lambda} \|u-v\|_\infty^2 \ \exp(C_{\mathcal{H}}\|u-v\|_\infty) \ \left[1+\log\left\|\frac{f_u(y\mid x)}{f_v(y\mid x)}\right\|_\infty\right]^2.
\end{aligned}$$

For $t \le C$, the proof of Lemma S.1 gives $\|p_u(t\mid x)/p_v(t\mid x)\|_\infty \le \exp\{(\mathcal{K}+\|\phi\|_\infty\,\Lambda(C))\|u-v\|_\infty\}$ and $\exp\{-H_u(C\mid x)+H_v(C\mid x)\} \le \exp\{\|\phi\|_\infty\Lambda(C)\|u-v\|_\infty\}$, so that $\|f_u(y\mid x)/f_v(y\mid x)\|_\infty \le \exp\{(\mathcal{K}+\|\phi\|_\infty\,\Lambda(C))\|u-v\|_\infty\}$. It follows that

$$\begin{aligned}
\mathrm{KL}(u\|v) &\lesssim_{\Phi,\Lambda} \|u-v\|_\infty^2 \ \exp(C_{\mathcal{H}}\|u-v\|_\infty) \ (1+\|u-v\|_\infty) \\
V(u\|v) &\lesssim_{\Phi,\Lambda} \|u-v\|_\infty^2 \ \exp(C_{\mathcal{H}}\|u-v\|_\infty) \ (1+\|u-v\|)_\infty^2,
\end{aligned}$$

proving the result for C_K sufficiently large.

S.3 Additional Details of Proof of Theorem 1

We show in this section that conditions (A1)–(A3) are sufficient to prove Theorem 1. Consider a joint prior $\widetilde{\Pi}$ on (u,F) where F is a distribution on $[0,1]^P$, and consider the joint distribution for $(\widetilde{Y}_i, X_i)$ given by $f_u(\widetilde{y}\mid x)\,\nu(d\widetilde{y})\,F(dx)$. We consider in particular a point-mass prior for F where $F = F_X$ with prior probability 1. Let $G_{u,F}$ denote the random joint distribution of $(\widetilde{Y}_i, X_i)$ under $(u,F) \sim \widetilde{\Pi}$ and G_0 represent the true joint distribution. Then it is easy to show that $\mathcal{H}^2(u_0,u)$ corresponds to the usual Hellinger distance between $G_{u,F}$ and G_0, and similarly $\mathrm{KL}(u_0\|u)$ and $V(u_0\|u)$ correspond to the usual Kullback-Leibler-type divergences from G_0 to $G_{u,F}$ given by $\mathrm{KL}(G_0\|G) = \int \log\frac{dG_0}{dG}\,dG_0$ and $V(G_0\|G) = \int \left(\log\frac{dG_0}{dG}\right)^2\,dG_0$, respectively. It is then easy to verify that conditions (A1)–(A3), by the variant of Theorem 2.1 of Ghosal and Van Der Vaart (2007) given by Shen et al. (2013, page 627), are sufficient

Algorithm 1 One Iteration of MBART Bayesian Backfitting

Input: $\{\mathcal{T}_m, \mathcal{M}_m\}_{m=1}^M$, $\{Y_i, \delta_i, X_i\}_{i=1}^N$, $Q(\cdot \mid \cdot)$, $\pi_{\mathcal{T}}$, σ_μ, Φ, θ, π_θ, π_σ

for $i = 1, \ldots, N$ **do**
  Sample $q_i \leftarrow \text{Poisson}\{\Lambda(Y_i \mid X_i, \theta)\}$
  Sample $U_{ij} \leftarrow \text{Uniform}(0, \Lambda(Y_i \mid X_i, \theta))$ for $j = 1, \ldots, q_i$
  $\widetilde{G}_{ij} \leftarrow \Lambda^{-1}(U_{ij} \mid X_i, \theta)$
  Sample $V_{ij} \leftarrow \text{Uniform}(0, 1)$ for $j = 1, \ldots, q_i$
  $\{W_{ij}\} \leftarrow \{\widetilde{G}_{ij} : V_{ij} \leq 1 - \Phi\{g(\widetilde{G}_{ij}, X_i)\}\}$
  $J_i \leftarrow \sum_{j=1}^{q_i} I(V_{ij} \leq 1 - \Phi\{g(\widetilde{G}_{ij}, X_i)\})$
  Sample Z_{ij} according to (S.3) for $j = 0, \ldots, J_i$.
  Sample ω_{ij} according to (S.3) depending on the choice of link $\Phi(\mu)$
end for
Update g using Algorithm 2 of Li et al. (2020) (treating $\mathfrak{a}$ as a tree with fixed depth 0), omitting Z_{i0} if $\delta_i = 0$
Make an update to θ which leaves invariant the unnormalized density

$$\pi(\theta) e^{-\sum_i \Lambda(Y_i \mid X_i, \theta)} \prod_{i=1}^N \lambda(Y_i \mid X_i, \theta)^{\delta_i} \prod_{j=1}^J \lambda(W_{ij} \mid X_i, \theta)$$

Make an update to σ_μ which leaves $\pi_\sigma(\sigma_\mu) \prod_m \prod_{\ell \in \mathcal{L}_m} \text{Normal}(\mu_{m\ell} \mid 0, \sigma_\mu^2)$ invariant

to establish the following result: there exists a K such that

$$\widetilde{\Pi}\{\mathcal{H}(G_0, G_{u,F}) \geq K\epsilon_n \mid \widetilde{\mathcal{D}}\} \to 0 \qquad \text{in } G_0\text{-probability.}$$

However, because $\widetilde{\Pi}$ places a point-mass prior on F at F_X, and because $\mathcal{H}(G_0, G_{u,F_X}) = \mathcal{H}(u_0, u)$, this is equivalent to the statement $\Pi\{\mathcal{H}(u, u_0) \geq K\epsilon_n \mid \widetilde{\mathcal{D}}\} \to 0$ in G_0-probability, as desired.

Algorithm 2 One iteration of Weibull-BART Bayesian Backfitting

Input: $\{\mathcal{T}_m^r, \mathcal{M}_m^r\}_{m=1}^M, \{Y_i, \delta_i, X_i\}_{i=1}^N, \{W_{ij} : i = 1, \ldots, N, j = 1, \ldots, J_i\}$, $Q(\cdot \mid \cdot)$

for $m = 1, \ldots, M$ **do**
  for $i = 1, \ldots, N$ **do**
    $\zeta_i \leftarrow r(X_i) - \text{Tree}(X_i; \mathcal{T}_m^r, \mathcal{M}_m^r)$
  end for
  for $\ell \in \mathcal{L}_m$ **do**
    $N_\ell \leftarrow \sum_{i \rightsquigarrow (\ell, m)} J_i + \delta_i$
    $E_\ell \leftarrow \sum_{i \rightsquigarrow (\ell, m)} e^{\zeta_i} Y_i^\kappa$
    $S_\ell \leftarrow \sum_{i \rightsquigarrow (\ell, m)} \delta_i \log Y_i + \sum_{j=1}^{J_i} \log W_i$
    $Z_\ell \leftarrow \sum_{i \rightsquigarrow (\ell, m)} \zeta_i (J_i + \delta_i)$
  end for
  Sample $\mathcal{T}' \leftarrow Q(\cdot \mid \mathcal{T}_m^r)$
  Compute the acceptance probability according to (10)

$$A \leftarrow \min \left\{ 1, \frac{\pi(\mathcal{T}' \mid \mathcal{T}_{-m}^r, \mathcal{M}_{-m}^r, \mathcal{D})\, Q(\mathcal{T}_m^r \mid \mathcal{T}')}{\pi(\mathcal{T}_m^r \mid \mathcal{T}_{-m}^r, \mathcal{M}_{-m}, \mathcal{D})\, Q(\mathcal{T}' \mid \mathcal{T}_m^r)} \right\}$$

  Set $\mathcal{T}_m^r \leftarrow \mathcal{T}'$ with probability A
  For $\ell \in \mathcal{L}_m^r$, sample $\eta_{m\ell} \sim \log \text{Gam}(\alpha_\eta + N_\ell, \beta_\eta + E_\ell)$
end for

S.4 Algorithms

For MBART, we restrict attention to the normal, logistic, or Student's T_ν links. In each case, we can express the model for Z_{ij} as a scale mixture of normal distributions

$$Z_{ij} \sim \text{Normal}\{g(W_{ij}, X_i), \omega_{ij}\},$$
$$\omega_{ij} \sim H,$$

where the distribution H varies depending on the choice of the link function. For the probit link, for example, $\omega_{ij} \equiv 1$, while for the Student's T_ν link we have $\omega_{ij} \sim \text{Gam}(\nu/2, \nu/2)$, and for the logistic link $\sqrt{\omega_{ij}}/2$ has a KS distribution (Holmes and Held, 2006). Set $W_{i0} = Y_i$ for

simplicity. In each case, one can sample (ω_{ij}, Z_{ij}) as a block by sampling

$$Z_{ij} \sim \begin{cases} \Phi\{(\cdot - g(W_{ij}, X_i))\}\, I(-\infty, 0) & \text{if } j \neq 0, \\ \Phi\{(\cdot - g(W_{ij}, X_i))\}\, I(0, \infty) & \text{if } j = 0. \end{cases} \tag{S.3}$$

$$\omega_{ij} \sim H(\omega \mid Z_{ij}, W_{ij}, X_i, g), \tag{S.4}$$

In the case of the Student's T_ν link, it can be shown that the distribution $H(\omega \mid Z_{ij}, W_{ij}, X_i, g)$ is given by a $\text{Gam}\{(\nu+1)/2, (\nu+[Z_{ij}-g(W_{ij}, X_i)])/2\}$, while for the logistic link we can sample ω_{ij} using the rejection sampling algorithm of Holmes and Held (2006). After augmenting the variables Z_{ij} and ω_{ij} the likelihood simplifies to

$$\begin{aligned} e^{-\sum_{i=1}^N \Lambda(Y_i|X_i,\theta)} \times \left(\prod_{i=1}^N \lambda(Y_i \mid X_i, \theta)^{\delta_i} \prod_{j=1}^{J_i} \lambda(W_{ij} \mid X_i, \theta) \right) \times \prod_{i=1}^N \prod_{j=0}^{J_i} H(\eta_{ij}) \\ \times \left(\prod_{i=1}^N \text{Normal}(Z_{i0} \mid g(Y_i, X_i), \omega_{i0}) \right\}^{\delta_i} \prod_{j=1}^{J_i} \text{Normal}(Z_{ij} \mid g(W_{ij}, X_i), \omega_{ij}) \Bigg). \end{aligned} \tag{S.5}$$

Hence the usual Bayesian backfitting approach can be used to update g by treating $Z_{ij} \sim \text{Normal}(g(W_{ij}, X_i), \omega_{ij})$ as the response; the only difference here is that we require that the Bayesian backfitting algorithm be able to handle heteroskedastic errors and the targeted smoothing. An algorithm for this is given by Algorithm 2 of Li et al. (2020) for SBART. Algorithm 1 gives our algorithm for a Gibbs sampling update for MBART.

The update of θ in Algorithm 1 requires updating the parameter θ of the base model. When we take the base model to be the Weibull hazard model, another round of Bayesian backfitting is used to update θ. This is given in Algorithm 2. In order to describe the algorithm, $Q(\cdot \mid \cdot)$ in Algorithm 2 denotes a *transition kernel* on the space of possible trees to be used as part of a Metropolis-Hastings step. For BART, one typically uses a transition kernel which is a mixture of the birth, death, and change rules introduced by Chipman et al. (1998); see Kapelner and Bleich (2016) for a detailed description of a possible transition

Algorithm 3 One Iteration of CoxBART Bayesian Backfitting

Input: $\{\mathcal{T}_m, \mathcal{M}_m\}_{m=1}^M$, $\{Y_i, \delta_i, X_i\}_{i=1}^N$, $Q(\cdot \mid \cdot)$, $\pi_{\mathcal{T}}$, α_μ, β_μ

1: **for** $i = 1, \ldots, N$ **do**
2: Sample $\phi_i \leftarrow \text{Gam}(1, \sum_{j \in \mathscr{R}_i} e^{g(X_i)})$ if $\delta_i = 1$, otherwise $\phi_i \leftarrow 0$.
3: **end for**
4: **for** $m = 1, \ldots, M$ **do**
5: Sample $\mathcal{T}' \leftarrow Q(\cdot \mid \mathcal{T}_m)$
6: Compute the acceptance probability according to (9)

$$A \leftarrow \min\left\{1, \frac{\pi(\mathcal{T}' \mid \mathcal{T}_{-m}, \mathcal{M}_{-m}, \phi, \mathcal{D})\, Q(\mathcal{T}_m \mid \mathcal{T}')}{\pi(\mathcal{T}_m \mid \mathcal{T}_{-m}, \mathcal{M}_{-m}, \phi, \mathcal{D})\, Q(\mathcal{T}' \mid \mathcal{T}_m)}\right\}$$

7: Set $\mathcal{T}_m \leftarrow \mathcal{T}'$ with probability A
8: For $\ell \in \mathcal{L}_m$, sample $\mu_{m\ell} \leftarrow \log \text{Gam}(\alpha_\ell, \beta_\ell)$ with $(\alpha_\ell, \beta_\ell)$ defined as in Section A
9: **end for**

kernel.

Finally, Algorithm 3 gives our Bayesian backfitting algorithm for CoxBART. This algorithm is substantially simpler than the MBART algorithm, and is computationally more efficient.

References

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.

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.

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

Kapelner, A. and Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40.

Li, Y., Linero, A. R., and Murray, J. S. (2020). Adaptive conditional distribution estimation with Bayesian decision tree ensembles. *arXiv preprint arXiv:2005.02490*.

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