# Robust Semiparametric Inference for Bayesian Additive Regression Trees*

Christoph Breunig† Ruixuan Liu‡ Zhengfei Yu§

October 20, 2025

**Abstract**

We develop a semiparametric framework for inference on the mean response in missing-data settings using a corrected posterior distribution. Our approach is tailored to Bayesian Additive Regression Trees (BART), which is a powerful predictive method but whose nonsmoothness complicates asymptotic theory with multi-dimensional covariates. When using BART combined with Bayesian bootstrap weights, we establish a new Bernstein–von Mises theorem and show that the limit distribution generally contains a bias term. To address this, we introduce RoBART, a posterior bias-correction that robustifies BART for valid inference on the mean response. Monte Carlo studies support our theory, demonstrating reduced bias and improved coverage relative to existing procedures using BART.

KEY WORDS: Causal inference, posterior correction, nonparametric Bayesian inference, Bernstein–von Mises theorem, BART.

*Breunig gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813. Yu gratefully acknowledges the support of JSPS KAKENHI Grant Number 25K05032.

†Department of Economics and Hausdorff Center for Mathematics, University of Bonn. Email: cbreunig@uni-bonn.de

‡Department of Decisions, Operations and Technology, Chinese University of Hong Kong. Email: ruixuanliu@cuhk.edu.hk

§Faculty of Humanities and Social Sciences, University of Tsukuba. Email: yu.zhengfei.gn@u.tsukuba.ac.jp

# 1 Introduction

The Bayesian additive regression trees (BART) is one of the most powerful statistical methods. Its empirical success in predictive modeling has been demonstrated in numerous studies that uncover complex nonlinear relationships in high-dimensional data. Additionally, in the context of nonparametric function estimation, the BART achieves the adaptive near rate-minimax posterior contraction, as shown by [27, 33, 31, 25]. Beyond prediction, BART has also shown strong performance in semiparametric inference for causal parameters of interest [14, 20]. However, a theoretical justification for semiparametric inference based on BART under reasonable regularity conditions is still lacking.

Semiparametric inference based on BART is challenging because tree- and forest-based methods build on piecewise constant functions. While BART can capture additive structures and perform variable selection, the standard Donsker property, which requires the unknown function to have smoothness exceeding one-half of the covariate dimension, is difficult to justify. In the context of missing data and causal inference, we impose the standard identifying assumption that potential outcomes and the response indicator are independent, conditional on the observed covariates. Within this framework, the main objective of the paper is to develop a robust inference procedure for the mean response based on BART.

We propose a novel robust Bayesian procedure that is asymptotically valid without imposing the Donsker property or the no-bias condition. For estimation of the mean response in the missing data problem or the average treatment effect (ATE) in program evaluation, we explore the role of propensity score by making use of the double robust functional. By combining a preliminary estimator of the propensity score, the BART induced posterior for the conditional mean function, and the Bayesian bootstrap weights, we establish a semiparametric Bernstein-von Mises (BvM) Theorem, which implies inference on the mean response at a $\sqrt{n}$-rate and asymptotic efficiency in the semiparametric sense. When the Donsker property fails, however, an additional bias term appears in the posterior distribution in the Bernstein-von Mises Theorem. Interestingly, this bias term is of the exact same form as that obtained by [4], who consider the prior correction approach of [28] based on least favorable directions of semiparametric submodels.

To adjust for the bias in the BvM statement, we propose a posterior correction approach based on pilot estimators for the propensity score and the conditional mean functions. As a technical device, we rely on sample splitting for the pilot estimation, in line with the recent double machine learning (DML) literature [11]. In doing so, our method differs from the proposal of [40], which places a prior on the propensity score or estimates it within a

one-step posterior correction. In contrast to their BvM results (given in their Theorems 3 and 5), we do not impose Donsker conditions in our paper. This is possible because our additional posterior correction removes the bias term in non-Donsker regimes for the mean response. Such an adjustment is crucial for BART, which builds on non-smooth base learners and thereby only adapts to conditional mean functions that are Lipschitz smooth [33]. As a result, the Donsker requirement is particularly stringent with multi-dimensional covariates. Our proposed robust BART (RoBART) achieves the BvM under more flexible smoothness conditions: the lack of smoothness of conditional mean functions can be compensated by high regularity of the propensity score.

Monte Carlo simulation results show that RoBART improves robustness over conventional BART while maintaining competitive credible interval lengths. Our method also shows finite-sample advantage over the one-step posterior correction. With increasing complexity of the underlying conditional mean and propensity score functions, one-step posterior correction is not sufficient to address the undercoverage of BART. By contrast, incorporating the additional debiasing term into the BART posterior allows our robust BART approach to achieve improved bias and coverage performance in complex model designs, without sacrificing estimation precision. We also showcase the practicality of our method by a real-data application.

*Related literature.* Our work is most closely related to active research areas. There has been considerable interest in developing semiparametric Bayesian inference that deviates from the plug-in principle. [28] pioneered the use of data-dependent priors to weaken the regularity conditions on the propensity score, for which they coined the term *single robustness*. Maintaining the same prior construction, [4] introduce an additional debiasing step that allows for double robustness. This is further extended to the popular difference-in-differences design by [3]. In this paper, we start from a different perspective by examining the one-step posterior of [40] and establish its asymptotic equivalence with the prior adjusted method. More importantly, we conduct another posterior bias correction to remove the bias term $b_{0,\eta}$. These are designed to the relax the Donsker property of the conditional mean, which is required by [40], but it is not met for the BART with high-dimensional covariates and nonadditive functions.

As a general nonparametric function estimation, the BART achieves the adaptive near rate-minimax posterior contraction, which has been the major breakthrough, obtained by [27, 33, 31, 25]. Compared with the extensive results on the posterior convergence for BART, the corresponding inferential theory is much scarce. [31] established a

semiparametric BvM theorem using BART type priors for a particular type of linear functional of the regression function, under the fixed design. Her result requires Donsker property of the conditional mean. In the presence of multi-dimensional covariates, she further requires the weight function to be uniform in this linear functional. For the Bayesian CART (for a single tree not the forest), [6] also establish general semiparametric BvM theorem within the Donsker regime.

The paper is structured as follows. Section 2 describes the missing data model together with the identifying assumptions and presents the robust semiparametric Bayesian procedure. In Section 3, we derive the BvM Theorem under high-level assumptions. Section 4 illustrates these results for BART. Section 5 provides numerical results via Monte Carlo simulations and an empirical illustration. Proofs of the main theoretical results are presented in Appendix A. Lemmas and auxiliary results are collected in Appendix B. Appendix C extends the framework to inference on average treatment effects. Appendix D provides additional implementation details of our methodology.

*Notation.* We adopt the standard empirical process notation as follows. For a function $h$ of a random vector $Z$ that follows distribution $P$, we let $P[h] = \int h(z)\,\mathrm{d}P(z), \mathbb{P}_n[h] = n^{-1}\sum_{i=1}^n h(Z_i)$, and $\mathbb{G}_n[h] = \sqrt{n}\,(\mathbb{P}_n - P)\,[h]$. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \lesssim b_n$ if $\limsup_{n\to\infty}(a_n/b_n) < \infty$, and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

# 2 Setup and Implementation

This section provides the main setup of our missing data framework and motivates the new Bayesian methodology. We provide the implementation of our robust Bayesian algorithm and discuss its connection and difference to the existing literature.

## 2.1 The Model Setup

We consider a family of probability distributions $\{P_\eta : \eta \in \mathcal{H}\}$ for some parameter space $\mathcal{H}$. The (possibly infinite dimensional) parameter $\eta$ characterizes the probability model. An independent and identically distributed (i.i.d.) sample $\{(R_iY_i, R_i, X_i^\top)^\top\}_{i=1}^n$ is available, where the dependent variable $Y_i$ is observed only if $R_i = 1$ and it is missing if $R_i = 0$. We assume the outcomes $Y_i$ are missing at random (MAR); that is, the outcome $Y_i$ and the missingness indicator $R_i$ are conditionally independent given $X_i$.

Throughout the paper, the distribution of the observable vector $(RY, R, X^\top)^\top$ is completely modeled by three components: (i) the marginal distribution $\mu_\eta(x)$ of the $p$-

dimensional covariates $X$; (ii) the propensity score $\pi_\eta(x) = P_\eta(R = 1 \mid X = x)$; and (iii) the conditional density function $f_\eta(\cdot \mid x)$ of the outcome $Y$ given $X = x$ and $R = 1$. Consequently, the conditional mean function of the outcome given covariates is

$$m_\eta(x) = \int y f_\eta(y \mid x) \, \mathrm{d}y.$$

Under the MAR assumption, the probability density function $p_\eta$ for $Z_i = (R_i Y_i, R_i, X_i^\top)^\top$ can be written as

$$p_\eta(z) = \mu_\eta(x) \pi_\eta^r(x) (1 - \pi_\eta(x))^{1-r} f_\eta^r(y \mid x), \tag{2.1}$$

Let $\eta_0$ be the true value of the parameter and denote $P_0 = P_{\eta_0}$, which corresponds to the frequentist distribution that generates the observed data. Accordingly, we denote $\mu_0$ as the true marginal covariates distribution, $\pi_0$ as the true propensity score, and $f_0(\cdot \mid x)$ as the true conditional density of $Y$ given $X = x$ and $R = 1$.

The key parameter of interest is the mean of the outcome variable:

$$\chi_0 := \mathbb{E}_0[Y_i],$$

where $\mathbb{E}_0[\cdot]$ denotes the expectation under $P_0$. Under the MAR assumption, $\chi_0$ can be identified using the conditional mean function $m_0(x) := \mathbb{E}_0[R_i Y_i \mid R_i = 1, X_i = x]$. Specifically, we have

$$\chi_0 = \int m_0(x) \, \mathrm{d}\mu_0(x).$$

For our approach, it is useful to consider an alternative expression of $\chi_0$, augmented by a term depending on the inverse propensity score $\pi_0$ as follows

$$\chi_0 = \int (m_0(x) + \gamma_0(r, x)(y - m_0(x))) \, \mathrm{d}P_0(y, r, x) \tag{2.2}$$

where the Riesz representer $\gamma_0$ is given by

$$\gamma_0(r, x) = \frac{r}{\pi_0(x)}, \tag{2.3}$$

which satisfies $\mathbb{E}_0[m_0(X)\gamma_0(R, X)] = \chi_0$; see [9].

Let $W^{(n)} := (W_{n1}, \ldots, W_{nn})$ denote the Bayesian bootstrap weights by [34], i.e., $W_{ni} = e_i / \sum_{j=1}^n e_j$ for $e_i \overset{iid}{\sim} \mathrm{Exp}(1)$, $i = 1, \ldots, n$. Referring to the expression (2.2), the natural

starting point for semiparametric Bayesian inference is

$$\chi_\eta = \sum_{i=1}^{n} W_{ni} \left( m_\eta(X_i) + \widehat{\gamma}(R_i, X_i)\big(Y_i - m_\eta(X_i)\big)\right), \tag{2.4}$$

where $m_\eta(\cdot)$ is some stochastic function with a given prior and $\widehat{\gamma}(r,x) = r/\widehat{\pi}(x)$ is the estimated Riesz representer with a pilot estimator $\widehat{\pi}$ of $\pi_0$ computed over some auxiliary dataset. The posterior law of $\chi_\eta$ or its conditional distribution given the observed data $Z^{(n)} = (Z_1, \ldots, Z_n)$ is determined jointly by the posterior law of $m_\eta$ denoted by $\Pi_m(\cdot \mid Z^{(n)})$, as well as the distribution of random vector $W^{(n)}$ involving Bayesian bootstrap weights. For the latter part, its probability law is known and easy to simulate. In accordance with the study on the bootstrap law [38], we adopt the notation $\Pi_W(\cdot \mid Z^{(n)})$. While our general results are applicable to a broad class of nonparametric priors for $m_\eta$, we particularly analyzes prior modeling under more basic conditions based on the BART. The posterior draws of the vector $(m_\eta(X_1), \ldots, m_\eta(X_n))$ can be easily obtained using the R package BART [36].

To motivate the approach from a Bayesian perspective, one can view the conditional distribution of (2.4) as approximating the posterior of

$$\int \left( m_\eta(x) + \frac{r}{\pi_\eta(x)}(y - m_\eta(x)) \right) \mathrm{d}P_\eta(y, r, x), \tag{2.5}$$

where we use a degenerate posterior for the propensity score from the pilot estimation. When it comes to the prior for $P_\eta$, we consider the Dirichlet process prior with its base measure taken to be zero. It coincides with the Bayesian bootstrap process as in (2.4). Note that the original notation in [40] defines the one-step posterior of the mean functional $\chi_\eta$ given $Z^{(n)}$ as the conditional measure of

$$\tilde{P}\left[ m_\eta(X) + \frac{R}{\pi_\eta(X)}(Y - m_\eta(X)) \right],$$

where the probability model $P_\eta$ is parameterized by $(m_\eta, \pi_\eta)$, and the augmented part $\tilde{P}$ denotes the Bayesian bootstrap law. The posterior law of $(m_\eta, \pi_\eta)$ is also assumed to be conditionally independent of the bootstrap law $\tilde{P}$. Both angles lead to the same approach because the Bayesian bootstrap law is known and does not depend on the unknown data generating process for the data. The main contribution of the current paper is that we show the one-step posterior still contains a bias term when the conditional mean function does not satisfy the Donsker property. This is particularly relevant for the BART with

multi-dimensional covariates.

## 2.2 A Robust Semiparametric Bayesian Procedure

Our Bayesian approach connects to the least favorable direction as specified by the efficient influence function

$$\widetilde{\chi}_\eta(Z) = m_\eta(X) + \gamma_\eta(R, X)(Y - m_\eta(X)) - \chi_\eta, \tag{2.6}$$

for some Riesz representer $\gamma_\eta$ given by

$$\gamma_\eta(R, X) = \frac{R}{\pi_\eta(X)}. \tag{2.7}$$

A pilot estimator for the propensity score $\pi_0$ is denoted by $\widehat{\pi}$ based on an auxiliary sample, so that $\widehat{\gamma}(r, x) = r/\widehat{\pi}(x)$ is an estimator of the Riesz representer $\gamma_0$. We also make use of a pilot estimator $\widehat{m}$ for the conditional mean function $m_0$ in the debiasing step. The use of an auxiliary data for the estimation of unknown functional parameters simplifies the technical analysis and is common in the related Bayesian literature; see [28] for propensity score adjusted priors in the case of missing data. It is also inspired by the recent literature on the DML [9], which yield negligibly of certain smaller order terms. This technique also dates back to the early development in semiparametric estimation [35]. In practice, we use the full data twice and do not split the sample, as we have not observed any over-fitting or loss of coverage thereby.

We are now in a position to describe our robust Bayesian procedure. The inputs to the algorithm are the observable data $\{(R_iY_i, R_i, X_i^\top)^\top\}_{i=1}^n$, the pilot estimators $\widehat{\gamma}$ and $\widehat{m}$ as discussed above, and the number of posterior draws $S$. The posterior distribution of $(m_\eta(X_i))_{i=1}^n := (m_\eta(X_1), \ldots, m_\eta(X_n))$ is generated by applying BART to estimate $m_\eta(\cdot)$ using the data $\{(R_iY_i, X_i^\top) : R_i = 1\}_{i=1}^n$.

**Algorithm 1** Posterior Computation

**for** $s = 1, \ldots, S$ **do**

(a) Obtain one draw from the posterior of $(m_\eta(X_i))_{i=1}^n$, denote it as $(m_\eta^s(X_i))_{i=1}^n$.

(b) Draw Bayesian bootstrap weights $W_{ni}^s = e_i^s / \sum_{j=1}^n e_j^s$ with $e_i^s \overset{iid}{\sim} \text{Exp}(1)$, $1 \leqslant i \leqslant n$.

(c) Calculate $\check{\chi}_\eta^s = \chi_\eta^s - \widehat{b}_\eta^s$ where

$$\chi_\eta^s = \sum_{i=1}^n W_{ni}^s \left( m_\eta^s(X_i) + \widehat{\gamma}(R_i, X_i)\big(Y_i - m_\eta^s(X_i)\big)\right) \tag{2.8}$$

and the debiasing term is

$$\widehat{b}_\eta^s = \frac{1}{n}\sum_{i=1}^n \left(\widehat{\gamma}(R_i, X_i) - 1\right)\left(\widehat{m}(X_i) - m_\eta^s(X_i)\right). \tag{2.9}$$

**end for**

**Output:** $\{\check{\chi}_\eta^s : s = 1, \ldots, S\}$

Given the posterior simulation draws, the $100 \cdot (1-\alpha)\%$ credible set $\mathcal{C}_n(\alpha)$ for the parameter of interest $\chi_0$ is computed by

$$\mathcal{C}_n(\alpha) = \big\{\tau : q(\alpha/2) \leqslant \chi \leqslant q(1-\alpha/2)\big\}, \tag{2.10}$$

where $q(a)$ denotes the $a$-quantile of $\{\check{\chi}_\eta^s : s = 1, \ldots, S\}$. We take the Bayesian point estimator (the posterior mean) by averaging the simulation draws: $\overline{\chi}_\eta = S^{-1}\sum_{s=1}^S \check{\chi}_\eta^s$.

Given some pilot estimators $\widehat{m}$ for the conditional mean $m_0$ and $\widehat{\pi}$ for the propensity score $\pi_0$, the well-known doubly robust estimator takes the form:

$$\widehat{\chi} = \frac{1}{n}\sum_{i=1}^n \left(\widehat{m}(X_i) + \frac{R_i}{\widehat{\pi}(X_i)}(Y_i - \widehat{m}(X_i))\right),$$

This has been extensively studied in the frequentist literature, starting with the augmented inverse propensity score weighting (AIPW) by [30]. Its extension to using pilot machine learning estimators combined with cross-fitting by [8] has generated considerable interest in the literature. The formulation in Step (2.8) of our algorithm mimics the frequentist construction and it agrees with the one-step posterior proposed by [40].

**Remark 2.1** (Estimation of the Riesz Representer)**.** *Our proposal builds on a plug-in approach that employs pilot estimators of the Riesz representer. Alternatively, one could*

*assign some prior to the propensity score function, as in the main main proposal of [40]*[1]*. Our approach in comparison is motivated by the following considerations. First, by adopting sample splitting for the pilot estimators, it becomes feasible to relax the Donsker condition by our proposal. Sample splitting has long been used in the semiparametric models, [35, 37] and regained popularity in the DML literature [8, 9]. Second, in more complicated semiparametric models, the Riesz representer may lack a tractable analytical form, but one can still construct a pilot estimator following the approach of [10]. In such cases, designing a feasible algorithm to obtain its posterior can be challenging. Finally, compared to the conditional mean function, the Riesz representer is often of secondary interest, one can view the plug-in estimator as a particular type of degenerate posterior for the propensity score [28].*

# 3 BvM Theorems under High-level Assumptions

Although our primary interest lies in establishing the BvM theorem for the BART, we first present the theory under high-level conditions. The robustness achieved through our additional debiasing step is thus applicable to other types of priors beyond the BART framework.

The one-step posterior of $\chi_\eta$ is determined jointly by the posterior of the conditional mean and the Bayesian bootstrap law. Recall that the conditional probability density function associated with the conditional mean function by $f_\eta(\cdot \mid x)$. For simplicity, we work with the conditional density class which only depends on the unknown conditional mean function. Such a class includes the standard Gaussian, Bernoulli, and Poisson outcomes, as discussed in [4]. Accordingly, we denote it by $f_{m_\eta}(\cdot \mid x)$ thereafter. The posterior distribution is formally given by

$$\begin{aligned}\Pi\big(m_\eta \in A, W^{(n)} \in B \;\big|\; Z^{(n)}\big) &= \Pi_m\big(m_\eta \in A \;\big|\; Z^{(n)}\big) \times \Pi_W\big(W^{(n)} \in B \;\big|\; Z^{(n)}\big) \\ &= \int_B \frac{\int_A \prod_{i=1}^n f_{m_\eta}^{R_i}(Y_i \mid X_i)\,\mathrm{d}\Pi_m(m_\eta)}{\int \prod_{i=1}^n f_{m_\eta}^{R_i}(Y_i \mid X_i)\,\mathrm{d}\Pi_m(m_\eta)}\,\mathrm{d}\Pi_W\big(W^{(n)} \;\big|\; Z^{(n)}\big).\end{aligned}$$

Regarding the centering point in the asymptotic normal approximation, we can consider

[1] [40] also consider the version that plugs in an estimator for the propensity score considered within a discussion of the distinction to the prior adjustment approach of [28].

any asymptotically efficient estimator $\widehat{\chi}$ with the following linear representation:

$$\widehat{\chi} = \chi_0 + \frac{1}{n}\sum_{i=1}^{n} \widetilde{\chi}_0(Z_i) + o_{P_0}(n^{-1/2}), \tag{3.1}$$

where $\widetilde{\chi}_0 = \widetilde{\chi}_{\eta_0}$ is the efficient influence function given in (2.6) under $\eta = \eta_0$. Denote its variance by $\mathrm{v}_0 = \mathbb{E}_0[\widetilde{\chi}_0^2(Z)]$. We write $\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta - \widehat{\chi}) \mid Z^{(n)})$ for the marginal posterior distribution of $\sqrt{n}(\chi_\eta - \widehat{\chi})$.

Below, we consider some measurable sets $\mathcal{H}_n^m$ of functions $m_\eta$ such that $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z) \to_{P_0} 1$. To abuse the notation for convenience, we also denote $\mathcal{H}_n = \{\eta : m_\eta \in \mathcal{H}_n^m\}$ where we index the conditional mean function $m_\eta$ by its subscript $\eta$. We introduce the notation $\|\phi\|_{2,\mu_0} := \sqrt{\int \phi^2(x)\,\mathrm{d}\mu_0(x)}$ for all $\phi \in L^2(\mu_0) := \{\phi : \|\phi\|_{2,\mu_0} < \infty\}$, as well as the supremum norm $\|\cdot\|_\infty$. We introduce the notation for the conditional variance $\sigma_0^2(x) = \mathbb{E}_0[(RY - m_0(X))^2 \mid R = 1, X = x]$ and assume throughout the paper that $\sigma_0 \in L^2(\mu_0)$. We denote the support of a random variable $V$ by $\mathcal{V}$.

**Assumption 1.** [Identification] We observe an *i.i.d.* sample $\{(R_iY_i, R_i, X_i)\}_{i=1}^n$, where the propensity score $\pi_0$ satisfies $\pi_0(x) = \mathbb{P}(R_i = 1 \mid Y_i = y, X_i = x)$ for all $(y, x) \in \mathcal{Y} \times \mathcal{X}$ and $\inf_{x\in\mathcal{X}} \pi_0(x) \geqslant c > 0$ for some constant $c > 0$.

Assumption 1 imposes a missing-at-random (MAR) assumption and an overlap assumption, both of which are sufficient for identifying the conditional mean function $m_0(x) = \mathbb{E}_0[Y_i \mid X_i = x]$.

**Assumption 2.** [Rates of Convergence] The pilot estimators $\widehat{\pi}$ and $\widehat{m}$, which are based on an auxiliary sample independent of $Z^{(n)}$, satisfy $\|\widehat{\gamma} - \gamma_0\|_{2,\mu_0} = O_{P_0}(r_n)$ and for $d \in \{0, 1\}$:

$$\|\widehat{m} - m_0\|_{2,\mu_0} = O_{P_0}(\varepsilon_n) \quad \text{and} \quad \sup_{\eta\in\mathcal{H}_n} \|m_\eta - m_0\|_{2,\mu_0} \lesssim \varepsilon_n,$$

where $\max\{\varepsilon_n, r_n\} \to 0$ and $\sqrt{n}\,\varepsilon_n r_n \to 0$. Further, $\|\widehat{\gamma}\|_\infty = O_{P_0}(1)$.

Assumption 2 imposes sufficiently fast convergence rates for the estimators for the conditional mean function $m_0$ and the propensity score $\pi_0$. In practice, one can explore the recent proposals from [12] and [23]. The posterior concentration rate for the conditional mean can be derived by verifying high-level assumptions of [16].

**Assumption 3.** [Complexity] (i) For $\mathcal{G}_n = \{m_\eta(\cdot) : \eta \in \mathcal{H}_n\}$ it holds that

$$\sup_{m_\eta\in\mathcal{G}_n} |(\mathbb{P}_n - P_0)m_\eta| = o_{P_0}(1). \tag{3.2}$$

Let $G_n$ be the envelope function of the functional class $\mathcal{G}_n$ satisfying

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{E}_0[G_n^2 \mathbb{1}\{G_n > C\}] = 0, \quad \mathbb{E}_0[G_n^4] = o(n). \tag{3.3}$$

(ii) Furthermore, we assume

$$\sup_{\eta\in\mathcal{H}_n} |\mathbb{G}_n [(\gamma_0 - \widehat{\gamma}) (m_\eta - m_0)]| = o_{P_0}(1). \tag{3.4}$$

Assumption 3 (i) restricts the functional class $\mathcal{G}_n$ to form a $P_0$-Glivenko-Cantelli class; see Section 2.4 of [38]. The moment conditions on the envelope functions follow from [40], which address the uniformity issues when showing the convergence of the conditional Laplace transform. Assumption 3 (ii) imposes a new stochastic equicontinuity condition on the product structure involving $\widehat{\gamma}$ and $m_\eta$, which sets us apart from the existing literature. In contrast to this assumption, [28] and [40] both require a stochastic equicontinuity condition $\sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n [m_\eta - m_0]| = o_{P_0}(1)$[2]. As demonstrated by [4], this is weaker than directly imposing the Donsker property on $(m_\eta - m_0)$. For the Hölder smooth class, [4] show that the low-order differentiability of the conditional mean can be compensated by exploring the high-order smoothness of the propensity score, and vice versa. Note that $\sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n [m_\eta - m_0]|$ diverges for the non-Donsker class, by the Sudakov's inequality [38].

We now present a semiparametric Bernstein–von Mises theorem, which establishes asymptotic normality of the posterior distribution, modulo a bias term. This asymptotic equivalence result is established using the so called *bounded Lipschitz distance* on probability distributions on $\mathbb{R}$, see (see Chapter 11 of [15]). As the one-step posterior is built from mimicking the frequentist double robust estimand, it is surprising at first sight that this does not fully remove the bias term, hence the technical proof of the BvM theorems in [40] still relies on the Donsker property. Note that one crucial difference to the frequentist method is that we need to characterize the weak convergence for the entire posterior distribution (not just its center as the point estimator), there is this crucial bias term that we identify. This shows the distinctive feature of deriving the frequentist validity of semiparametric Bayesian inference.

**Theorem 3.1.** *Let Assumptions 1, 2 and 3 (i) hold. Then, with the one-step posterior*

[2]Because a nonparametric prior is also assigned to the propensity score, [40] further requires the propensity score function to belong to the Donsker class.

*correction, we have*

$$d_{BL}\left(\mathcal{L}_{\Pi}(\sqrt{n}(\chi_\eta - \widehat{\chi} - b_{0,\eta}) \mid Z^{(n)}), N(0, \mathrm{v}_0)\right) \to_{P_0} 0,$$

*where* $b_{0,\eta} = \mathbb{P}_n[(\gamma_0 - 1)(m_0 - m_\eta)]$.

**Remark 3.1** (Bias Equivalence for Prior Correction)**.** *Under the double robust smoothness conditions imposed in Assumption 2, [4] show that the prior adjustment approach of [28], denoted by* $\chi_\eta^{PA}$*, also yields a semiparametric BvM theorem up to the exact same bias term* $b_{0,\eta}$*. Specifically, as a consequence of the triangular inequality for the bounded Lipschitz distance, we have the asymptotic equivalence*

$$d_{BL}\left(\mathcal{L}_{\Pi}(\sqrt{n}(\chi_\eta^{PA} - \widehat{\chi} - b_{0,\eta}) \mid Z^{(n)}), \mathcal{L}_{\Pi}(\sqrt{n}(\chi_\eta - \widehat{\chi} - b_{0,\eta}) \mid Z^{(n)})\right) \to_{P_0} 0,$$

*by Theorem 3.1 above and Theorem 3.1 of [4]. Surprisingly, the one-step updated posterior exhibits the exact same bias term as the plug-in version using adjusted prior of [28]. Therefore, we follow the strategy of [4] to carry out a feasible bias correction.*

The previous result shows that, under double robust smoothness conditions, the posterior of the one-step corrected mean only satisfies the BvM result within a bias term $b_{0,\eta}$. We emphasize that the Bayesian procedure that achieves the BvM equivalence in Theorem 3.1 is not feasible, because it depends on the term $b_{0,\eta}$, which is a function of the unknown conditional mean $m_0$. This bias term vanishes, if we impose additional smoothness restriction on the conditional mean function $m$ satisfying Donsker property.

Aimed at relaxing the Donsker property, our objective is to maintain double robust smoothness conditions while considering pilot estimators for the unknown functional parameters in $b_{0,\eta}$. We explicitly correct the posterior distribution, following our proposed methodology in Algorithm 1. Recall the definition of the bias correction term $\widehat{b}_\eta$ given in Algorithm 1:

$$\widehat{b}_\eta = \mathbb{P}_n[(\widehat{\gamma} - 1)(\widehat{m} - m_\eta)]. \tag{3.5}$$

The next theorem is an immediate consequence of [4].

**Theorem 3.2.** *Let Assumptions 1, 2 and 3 hold. Then, we have*

$$d_{BL}\left(\mathcal{L}_{\Pi}(\sqrt{n}(\chi_\eta - \widehat{\chi}_n - \widehat{b}_\eta) \mid Z^{(n)}), N(0, \mathrm{v}_0)\right) \to_{P_0} 0.$$

As importance consequences of the BvM theorems, we hence provide the frequentist validity of the Bayesian credible set built from the corrected posterior. For any $\alpha \in (0,1)$ and Bayesian credible set $\mathcal{C}_n(\alpha)$ for $\chi_\eta - \widehat{b}_{0,\eta}$, then we have $P_0\big(\chi_0 \in \mathcal{C}_n(\alpha)\big) \to 1-\alpha$.

# 4 BvM Theorems under Primitive Conditions

We turn to our main task of establishing a new BvM theorem where the conditional mean function is modeled by the BART. The BART prior expresses the unknown function as a sum of many binary regression trees. Each individual tree consists of a set of internal decision nodes which define a partition of the covariate space, as well as a set of terminal nodes or leaves. Those partitions are generated by recursively applying some binary split rules of the covariate space as the form $\{x_j \leqslant \tau\}$ versus $\{x_j > \tau\}$ where $x_j$ signifies that $j$-th coordinated is chosen for the splitting and $\tau$ is the split point. The good approximation property of tree-based learners relies on finding a fine partition scheme that divides the data into more homogeneous groups and learning a piecewise constant function on the partition. For this purpose, we introduce the notion about the split-net from [33]. Given an increasing integer sequence $b_n$, a split-net $\mathcal{X} = \{x_j \in [0,1]^p, j = 1, \ldots, b_n\}$ is a discrete collection of $b_n$ points $x_j$ at which possible splits occur along coordinates. For a set $\Omega \subset \mathbb{R}^p$, we denote the $j$–th projection mapping of $\Omega$ by $[\Omega]_j = \{x_j \in \mathbb{R} : (x_1, \ldots, x_p)^\top \in \Omega\}$.

The construction of each individual tree starts with some root node in $[0,1]^p$ at depth $l = 0$, where the depth of a node means the number of nodes along the path from the root node down to that node. Any binary tree can be characterized by the (i) the probability that a node at depth $l$ is nonterminal, (ii) the distribution on the splitting variable assignments at each splitting node, and (iii) the rules with which the splits are made. Referring to point (i), each node at depth $l \in \{0, 1, 2, \ldots\}$ is split (hence nonterminal) with some prior probability depending on the depth. When it comes to point (ii), we follow [19, 27] using a sparse Dirichlet prior. This approach chooses a splitting covariate $j$ from a proportion vector $\vartheta = (\vartheta_1, \ldots, \vartheta_p)^\top$ belonging to the $p$-dimensional simplex. Finally for point (iii), if a node corresponding to a box $\Omega$ is split, a splitting coordinate is drawn from the proportion vector and a split-point $\tau_j$ is chosen randomly from $[\mathcal{X}]_j \bigcap \mathsf{int}([\Omega]_j)$ for a given split-net $\mathcal{X}$. The procedure ends until all nodes become terminal. We further denote the cardinality of the split points in the split-net $\mathcal{X}$ projected onto the $j$-th coordinate by $b_j(\mathcal{X})$ for $j = 1, \ldots, p$.

Throughout the paper, we pick a fixed number of trees $T$ and assume an independent product prior for the tree ensemble, $\Pi(\mathcal{E}) = \prod_{t=1}^{T} \Pi(\mathcal{T}^t)$. [13] recommend taking $T = 200$

based on extensive numerical evidence. For a given split-net $\mathcal{X}$ and for each $1 \leqslant t \leqslant T$, we denote with $\mathcal{T}^t = (\Omega_1^t, \ldots, \Omega_{K^t}^t)$ a $\mathcal{X}$-tree partition of size $K^t$ and with step-heights as $\boldsymbol{\beta}^t = (\beta_1^t, \ldots, \beta_{K^t}^t) \in \mathbb{R}^{K^t}$. An additive tree-based learner is fully described by a tree ensemble $\mathcal{E} = \{\mathcal{T}^1, \ldots, \mathcal{T}^T\}$ and terminal node parameters $\mathcal{B} = (\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^T)^\top \in \mathbb{R}^{\sum_{t=1}^T K^t}$ as follows

$$m_{\mathcal{E},\mathcal{B}}(x) = \sum_{t=1}^{T} \sum_{k=1}^{K^t} \beta_k^t \mathbb{1}\{x \in \Omega_k^t\}. \tag{4.1}$$

Given an ensemble $\mathcal{E}$ of trees, we denote $\mathcal{M}_\mathcal{E} := \{m_{\mathcal{E},\mathcal{B}}(x) : \mathcal{B} \in \mathbb{R}^{\sum_{t=1}^T K^t}\}$. If $\mathcal{E}$ consists of a single tree $\mathcal{T}$, we denote $\mathcal{M}_\mathcal{E}$ by $\mathcal{M}_\mathcal{T}$. Binary decision trees or forests offer good approximation properties to the class of Hölder *continuous class*, which is indexed by an exponent $\alpha \in (0, 1]$. This parameter $\alpha$ does not exceed one, which is standard in the literature studying the piecewise constant estimators and priors. Another notable feature of BART is that one can explore the sparsity of relevant regressors and perform variable selection. The true conditional mean function is assumed to belong to the following class of functions. Below we introduce the norm $\|\phi\|_{\mathcal{H}^\alpha} := \sup_{x,y\in[0,1]^p} |\phi(x) - \phi(y)|/\|x - y\|_2^\alpha$, for functions $\phi : [0, 1]^p \mapsto \mathbb{R}$, with some $\alpha > 0$, where $\|\cdot\|_2$ denotes the Euclidean norm.

**Definition 4.1.** *We denote the space of uniformly $\alpha$-Hölder continuous functions that only depend on some subset of covariates $\mathcal{S}$ as follows,*

$$\mathcal{F}_p(\alpha, \mathcal{S}) := \Big\{ m : [0,1]^p \mapsto \mathbb{R} : \|m\|_{\mathcal{H}^\alpha} < \infty \text{ and } m \text{ is constant in the directions } \{1, \ldots, p\} \backslash \mathcal{S} \Big\}$$

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

For the set $\mathcal{S}_0$ with its cardinality $|\mathcal{S}_0| = q_0 < p$, we assume $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$. In addition, we also explore the case where $m_0$ is additively separable in terms of low-dimensional covariates. Consider the following additive class with $T_0$ components

$$\mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}}) := \left\{ m_0(x) = \sum_{t=1}^{T_0} m_0^t(x), \quad \text{such that} \quad m_0^t \in \mathcal{F}_p(\alpha^t, \mathcal{S}^t) \right\},$$

where $\boldsymbol{\alpha} := (\alpha^t)_{t=1}^{T_0}$ and $\boldsymbol{\mathcal{S}} := (\mathcal{S}^t)_{t=1}^{T_0}$. We denote $q_0^t = |\mathcal{S}_0^t|$ for the subset $\mathcal{S}_0^t \subset \{1, \ldots, p\}$. We define

$$\varepsilon_n := n^{-\alpha/(2\alpha+q_0)} \sqrt{\log n},$$

related to the rate of posterior contraction for $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S})$. When considering the additive case $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}})$, we define $\varepsilon_n^{add} = \sqrt{\sum_{t=1}^{T_0} \varepsilon_{n,t}^2}$, where $\varepsilon_{n,t} := n^{-\alpha^t/(2\alpha^t+q_0^t)} \sqrt{\log n}$.

The case where $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$ provides a clean illustration of our theory, which shows that our additional debiasing step is essential if $q_0 > 1$. The class $\mathcal{F}_p(\alpha, \mathcal{S}_0)$ fails to satisfy the Donsker property whenever $\alpha < q_0/2$. A subtle consequence, is that the size of the resulting sieve set used to approximate the Hölder continuous function becomes too large, so that the standard maximal inequality via the entropy bound fail to deliver the stochastic equicontinuity. The same issue occurs to the additive class whenever $\alpha^t < q_0^t/2$ for any $1 \leqslant t \leqslant T_0$.

Next, we list assumptions of covariates, error term and the prior specifications, followed by some remark.

**Assumption 4** (Model Specification)**.** Under $P_0$, the conditional density of $Y$, given $(R = 1, X = x)$ is standard Gaussian, i.e.,

$$f_0(y \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - m_0(x))^2}{2}\right).$$

Moreover, let $\log p \lesssim n^{q_0/(2\alpha+q_0)}$ if $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$, and $\log p \lesssim \min_{1\leqslant t\leqslant T_0} n^{q_0^t/(2\alpha^t+q_0^t)}$ if $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}})$.

**Assumption 5** (Tree-based Partition)**.** The split-net $\mathcal{X}$ satisfies the following conditions: (i) $\max_{1\leqslant j\leqslant p} \log b_j(\mathcal{X}) \lesssim \log n$. (ii) One can construct a $\mathcal{X}$-tree partition $\widehat{\mathcal{T}}$ such that there exists $m_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}$ with its step heights $\widehat{\boldsymbol{\beta}}$, satisfying $\|m_0 - m_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_\infty \lesssim \varepsilon_n$ if $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$. For $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}}_0)$, we assume this tree-based approximation exists for each individual component in the additive function, that is, $\|m_0 - m_{0,\widehat{\mathcal{T}}^t,\widehat{\boldsymbol{\beta}}^t}\|_\infty \lesssim \varepsilon_n^t$ for $t = 1, \ldots, T_0$.

**Assumption 6** (Prior)**.** (i) For a fixed $T > 0$, each tree $\mathcal{T}^t$, $t = 1, \ldots, T$ is independently assigned a tree prior with the following Dirichlet sparsity, that is, the $j$-the covariate is chosen for splitting the nonterminal nodes from a proportion vector $\vartheta := (\vartheta_1, \ldots, \vartheta_p)^\top$, s.t. $(\vartheta_1, \ldots, \vartheta_p)^\top \sim \mathrm{Dir}(\zeta/p^\xi, \ldots, \zeta/p^\xi)$ with $\zeta > 0$ and $\xi > 1$. (ii) Given any tree, each node at depth $l \in \{0, 1, 2, \ldots\}$ is split with some prior probability $\nu^{l+1}$ for some $\nu \in (0, 1/2)$. (iii) Given $K^1, \ldots, K^T$ induced by $\mathcal{E}$, we consider the independent priors for the step-heights:

$$\mathrm{d}\Pi(\mathcal{B} \mid K^1, \ldots, K^T) = \prod_{t=1}^{T} \prod_{k=1}^{K^t} \phi_T(\beta_k^t),$$

where $\phi_T(\cdot)$ is some bounded and compactly supported probability density function that can depend on the tree size $T$.

**Remark 4.1** (Discussion of Assumptions)**.** *The Gaussian likelihood with unit variance in Assumption 4 is standard in the literature. One can also incorporate the unknown*

*variance term and assign a prior as in [39] and [25]. The asymptotic analysis allows for a large ambient dimension p. The requirement about the growth of* $\log p$ *simplifies the presentation of the contraction rate. Otherwise, one needs to incorporate an additional term as* $\sqrt{q_0(\log p)/n}$ *in the high dimensional setup. The construction of the piecewise constant approximation with the tree partition, which satisfies Assumption 5, can be found in Section 4 of [25].*

*Assumption 6 lists the specification of priors in the BART. The control of the unique elements in the split-net along each coordinate is crucial to establish the posterior rate of contraction. The original algorithm in [13] does not take into account of the sparsity of the true regression model. [27] suggests the sparse Dirichlet prior on splitting coordinates to perform the variable selection. Following [32], the splitting probabilities decay exponentially with respect to the depth l, which gives rise to the desired exponential tail of the total tree sizes. The compact support condition on each step height is assumed for technical reasons when verifying high-level conditions in [16]; see [39] and [25].*

**Theorem 4.1.** *Let Assumptions 1 and 4–6 hold, together with* $\|1/\widehat{\pi}\|_\infty = O_{P_0}(1)$. *If* $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$ *and* $\sqrt{n}\,\varepsilon_n\|\widehat{\pi} - \pi_0\|_\infty \to_{P_0} 0$, *then*

$$d_{BL}\left(\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta - \widehat{\chi} - \widehat{b}_{0,\eta}) \mid Z^{(n)}), N(0, \mathrm{v}_0)\right) \to_{P_0} 0.$$

*The same conclusion holds if* $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}}_0)$ *and* $\sqrt{n}\,\varepsilon_n^{add}\|\widehat{\pi} - \pi_0\|_\infty \to_{P_0} 0$.

Theorem 4.1 establishes a BvM result for BART without requiring the Donsker property for either the propensity score or the conditional mean function. This result follows from verifying the high-level assumption from the previous section. The proof is based on constructing a sieve set that receives posterior mass with probability approaching one. This sieve set is defined by piecewise constant functions over increasingly fine grids and with a growing number of covariates. Due to the discontinuity of such piecewise constant functions, the complexity with increasing number of covariates exceeds the threshold of the Donsker class. Hence, the standard stochastic equicontinuity as imposed in the semiparametric Bayesian literature is not satisfied.

**Remark 4.2** (Donsker Class)**.** *One can establish the BvM theorem*

$$d_{BL}\left(\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta - \widehat{\chi}) \mid Z^{(n)}), N(0, \mathrm{v}_0)\right) \to_{P_0} 0,$$

*under the Donsker property of* $\mathcal{H}_n$ *so that* $b_{0,\eta}$ *becomes asymptotically negligible itself. This condition holds if* $\alpha > q_0/2$ *for* $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$. *With the additive structure, the above*

*property holds for* $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}}_0)$ *if* $\alpha^t > q_0^t/2$ *for all* $1 \leqslant t \leqslant T_0$. *For semiparametric Bayesian inference, such Donsker properties are imposed in Assumption 2(c) in [40] or Condition (3.12) in [28]. Our results exploit the regularity of the propensity score, expressed in terms of new stochastic equicontinuity to restore the frequentist validity of our robust procedure.*

**Remark 4.3** (Smoothed BART)**.** *Given the non-smooth feature of the BART, it is natural to consider the smoothed BART (see [27]), because it explores the smoothness of the conditional mean. Our main message remains unchanged as one can trade-off the orders of smoothness for the conditional mean and propensity score by incorporating the additional bias correction step. Consider functions in the Hölder smooth class, i.e., the space of functions on* $[0,1]^p$ *with bounded partial derivatives up to order* $\lfloor\beta\rfloor$, *where* $\lfloor\beta\rfloor$ *is the largest integer strictly less than* $\beta$ *and such that the partial derivatives of order* $\lfloor\beta\rfloor$ *are Hölder continuous of order* $\beta - \lfloor\beta\rfloor$. *If the true function only depends on at most* $q_0$ *covariates, Theorem 2 in [27] states that the posterior rate of contraction is of the order* $O(n^{\beta/(2\beta+q_0)}\log(n))$. *If the conditional mean function and propensity score function are in such Hölder smooth classes with smoothness indices* $(\beta_m, \beta_\pi)$ *and they only depends on* $q_0$ *regressors, our robust approach is asymptotically normal when* $\sqrt{\beta_m\beta_\pi} > q_0/2$. *In comparison, the Donsker properties in [40] will force* $\min\{\beta_m, \beta_\pi\} > q_0/2$.

# 5 Numerical Studies

In this section, we first examine the finite-sample performance of BART inference for the mean response $\chi_0 = \mathbb{E}_0[Y_i]$ in missing data models. To illustrate the practical relevance of the theoretical discussion in Remark 4.2, we consider simulation designs that include interactions. As Appendix C extends the proposed robust BART inference to the average treatment effect (ATE) framework, we then apply it to the well-known National Health and Nutrition Examination Survey data to revisit the average effect of participation in meal programs on students' body mass index.

## 5.1 Monte Carlo Simulation

The data-generating process for i.i.d. observations is specified as follows. For each unit $i = 1, \ldots, n$, we generate the observed variables $(R_iY_i, R_i, X_i^\top)$ by $X_i = (X_{i1}, \ldots, X_{i5})^\top$ where $X_{i1}, X_{i2}, X_{i3} \sim N(0,1)$, $X_{i4} \sim$ Bernoulli(0.5), $X_{i5}$ is a categorical variable taking

values $\{1, 2, 3\}$ with equal probability. The distributions of $R_i$ and $Y_i$ are given by

$$R_i \mid X_i \sim \text{Bernoulli}\left(\Psi\left[e(X_i)\right]\right), \quad Y_i \mid X_i \sim N\left(m(X_i), 1\right),$$

where $\Psi(t) = 1/(1 + e^{-t})$. We analyze the finite-sample effects of varying the complexity of the conditional mean function $m$ and the propensity scores for sample sizes $n \in \{125, 250, 500\}$. Throughout our simulations, the number of Monte Carlo replication is set to 1000.

We consider four designs for the function $e(\cdot)$ that determines the propensity score by $\pi(x) = \Psi(e(x))$ and the conditional mean function $m(\cdot)$. In Design I, $e(x)$ includes an interaction term $x_1x_3$, whereas $m(\cdot)$ contains no interaction terms. In Designs II–IV, both $e(\cdot)$ and $m(\cdot)$ include interaction terms, with increasing complexity (i.e., more interaction terms) across designs:

$$\text{Design I:} \quad e(x) = x_1(2x_3 - 1)/5,\ m(x) = 1 - 2x_1 - x_1^2/2 + x_2 + x_3 + x_4 + h(x_5),$$

$$\text{Design II:} \quad e(x) = x_1(2x_3 - 1)/5,\ m(x) = 1 + x_1(x_2 + x_3) + x_2 + x_4 + h(x_5),$$

$$\text{Design III:} \quad e(x) = [2x_3(x_1 + x_2) - x_1]/5,\ m(x) = 1 + x_1(1 + x_3) + x_2x_3 + x_4 + h(x_5),$$

$$\text{Design IV:} \quad e(x) = [2x_3(x_1 + x_2) - x_1]/5,\ m(x) = 1 + x_1x_3 + x_2x_3 + x_2x_4 + h(x_5),$$

where $h(x_5) = 2\mathbb{1}\{x_5 = 1\} - \mathbb{1}\{x_5 = 2\} - \mathbb{1}\{x_5 = 3\}/2$ with $\mathbb{1}\{\cdot\}$ denoting the indicator function.

We evaluate three inference methods. Standard **BART** obtains the posterior of the conditional mean function $m_\eta(\cdot)$ using BART, and then averages over the covariates using Bayesian bootstrap weights. Implementation relies on the R package BART [36]: the posterior of $m_\eta(\cdot)$ is computed using the function gbart with argument type = wbart. We draw 2000 posterior samples after discarding a burn-in of 500, with the number of trees set to $T = 200$. Default priors from the package are used, see Appendix D for details. **One-step BART** applies the posterior correction of [40], which uses posteriors of both $m(\cdot)$ and $\pi(\cdot)$ obtained via BART or its logistic variant.[3] **RoBART**, the robust BART method described in Algorithm 1, combines the BART posterior of $m_\eta(\cdot)$, a plug-in estimator for $\pi(\cdot)$, and the debiasing term $\widehat{b}_\eta$ in (2.9). We consider two estimators for propensity score: **Logit**, a quadratic logistic regression; and **SL**, a SuperLearner combining logistic regression and a generalized additive model.[4] Both include pairwise interactions of covariates.

[3] We implement logistic BART by calling gbart with type = lbart.

[4] We use the R package SuperLearner with the library of prediction algorithms including SL.glm and SL.gam for implementation.

Table 1: Finite-sample performance of BART-based inference methods for the mean response with missing data.

| $n$ | Methods | Design I | | | Design II | | | Design III | | | Design IV | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Bias | CP | CIL | Bias | CP | CIL | Bias | CP | CIL | Bias | CP | CIL |
| 125 | BART | 0.086 | 0.918 | 1.114 | 0.187 | 0.885 | 1.039 | 0.279 | 0.821 | 1.041 | 0.343 | 0.724 | 1.006 |
| | One-step BART | 0.073 | 0.946 | 1.284 | 0.196 | 0.959 | 1.401 | 0.308 | 0.891 | 1.360 | 0.355 | 0.841 | 1.358 |
| | $\text{RoBART}_{\text{Logit}}$ | 0.082 | 0.916 | 1.192 | 0.048 | 0.963 | 1.215 | 0.059 | 0.938 | 1.207 | 0.083 | 0.934 | 1.209 |
| | $\text{RoBART}_{\text{SL}}$ | 0.050 | 0.916 | 1.416 | 0.037 | 0.949 | 1.477 | 0.072 | 0.916 | 1.228 | 0.075 | 0.918 | 1.345 |
| 250 | BART | 0.051 | 0.935 | 0.786 | 0.171 | 0.848 | 0.724 | 0.270 | 0.673 | 0.724 | 0.320 | 0.577 | 0.705 |
| | One-step BART | 0.040 | 0.960 | 0.866 | 0.163 | 0.939 | 0.953 | 0.273 | 0.823 | 0.922 | 0.305 | 0.770 | 0.931 |
| | $\text{RoBART}_{\text{Logit}}$ | 0.039 | 0.935 | 0.822 | 0.019 | 0.965 | 0.816 | 0.011 | 0.958 | 0.875 | 0.028 | 0.963 | 0.875 |
| | $\text{RoBART}_{\text{SL}}$ | 0.081 | 0.932 | 0.954 | 0.028 | 0.964 | 0.802 | 0.047 | 0.953 | 0.898 | 0.045 | 0.959 | 0.849 |
| 500 | BART | 0.046 | 0.924 | 0.565 | 0.094 | 0.866 | 0.469 | 0.142 | 0.761 | 0.475 | 0.189 | 0.607 | 0.451 |
| | One-step BART | 0.040 | 0.948 | 0.599 | 0.077 | 0.932 | 0.535 | 0.115 | 0.888 | 0.542 | 0.154 | 0.814 | 0.533 |
| | $\text{RoBART}_{\text{Logit}}$ | 0.040 | 0.927 | 0.570 | 0.030 | 0.944 | 0.493 | 0.026 | 0.947 | 0.519 | 0.053 | 0.930 | 0.511 |
| | $\text{RoBART}_{\text{SL}}$ | 0.042 | 0.926 | 0.566 | 0.034 | 0.941 | 0.486 | 0.033 | 0.945 | 0.503 | 0.061 | 0.925 | 0.494 |

Table 1 presents the bias of the posterior mean, coverage probability (CP) and the average length (CIL) of the 95% credible interval formed by the (corrected) posteriors. In Design I, where $m(\cdot)$ is linear, all methods perform well including standard BART. In Design II, where interaction terms appear in both $e(\cdot)$ and $m(\cdot)$, BART tends to undercover, while one-step BART and RoBART restore CP close to nominal levels.[5] Between the two, RoBART exhibits smaller bias, especially in small samples. In Designs III and IV, as more interaction terms appear in $e(\cdot)$ and/or $m(\cdot)$, one-step BART produces larger bias and undercoverage, whereas RoBART continues to yield small bias and improved coverage. These results align with our theory. As the additive components of $\pi(\cdot)$ or $m(\cdot)$ depend on more than one covariate (see Remark 4.2), the Donsker property is not guaranteed. Evidently, our RoBART demonstrates robust performance in such complex designs. In addition to its improved bias and coverage performance, RoBART (with a logit estimator for the propensity score) yields shorter credible intervals than one-step BART for all designs and sample sizes, while RoBART (with a SuperLearner for the propensity score) does so in 9 out of 12 cases.

[5]This complements the simulations in [40], which show improved coverage for one-step BART in a design where both $\pi(\cdot)$ and $m(\cdot)$ are one-dimensional but discontinuous, see their Table 2.

Table 2: ATE estimation of the school meal programs on children's and youths' BMI, sample trimmed based on estimated propensity score (using Logit) within $[t, 1-t]$, $\bar{n}$= effective sample size after trimming.

| $t$ | $\bar{n}$ | Methods | ATE | 95% CI | CIL |
|---|---|---|---|---|---|
| 0 | 2330 | BART | 0.160 | [-0.176, 0.675] | 0.851 |
| | | One-step BART | -0.164 | [-1.440, 0.984] | 2.425 |
| | | $\text{RoBART}_{\text{Logit}}$ | 0.088 | [-0.356, 0.519] | 0.875 |
| | | $\text{RoBART}_{\text{SL}}$ | 0.062 | [-0.354, 0.471] | 0.826 |
| 0.05 | 2326 | BART | 0.195 | [-0.258, 0.688] | 0.946 |
| | | One-step BART | -0.109 | [-1.235, 0.994] | 2.229 |
| | | $\text{RoBART}_{\text{Logit}}$ | 0.094 | [-0.348, 0.521] | 0.870 |
| | | $\text{RoBART}_{\text{SL}}$ | 0.066 | [-0.343, 0.467] | 0.810 |
| 0.10 | 2136 | BART | 0.207 | [-0.131, 0.681] | 0.813 |
| | | One-step BART | 0.140 | [-0.775, 1.061] | 1.836 |
| | | $\text{RoBART}_{\text{Logit}}$ | 0.215 | [-0.230, 0.660] | 0.890 |
| | | $\text{RoBART}_{\text{SL}}$ | 0.200 | [-0.225, 0.611] | 0.835 |

## 5.2 Empirical Application

We apply the BART inference methods to a subsample of data from the National Health and Nutrition Examination Survey (NHANES) 2007–2008, previously analyzed by [7]. The parameter of interest is the average treatment effect (ATE), defined as $\mathbb{E}[Y_i(1) - Y_i(0)]$, of participating in school meal programs ($D$) on body mass index (BMI) for children and youths aged 4–17 years ($Y$). The covariates $X$ include eleven variables: child age, child gender, race dummies (Black and Hispanic), an indicator for family income above 200% of the federal poverty level, indicators for participation in the special supplemental nutrition program and in the food stamp program, an indicator for childhood food security, an insurance coverage dummy, and the age and gender of the survey respondent (an adult in the family). The sample size is 2330.

We adapt the BART inference methods from the simulation study to the ATE setting (see Appendix C for the robust BART (RoBART) method for ATE), and present the results in Table 2. Table 2 shows that all ATE estimates of school meal programs on BMI are small relatively to the sample average BMI 20.11 (with the standard deviation 5.42). All 95% credible intervals include zero, suggesting that the average effect may go either direction when uncertainty is taken into account. One-step BART produces more dispersed posterior and thus more volatile point estimates than other methods. The estimates tend to increase when observations with propensity score near the boundary are discarded. We compare

the BART-based estimates in Table 2 with the frequentist results reported in Table 1 of [7] using the full sample ($\bar{n} = 2330$): Horvitz-Thompson (HT) estimate of $-1.48$ with a 95% confidence interval $[-2.50, -0.46]$, the inverse probability weighting (IPW) estimate $-0.41$ with confidence interval $[-0.62, 0.34]$, and their preferred calibration estimate of $-0.04$ with the confidence interval $[-0.48, 0.40]$.[6] While point estimates differ in sign across methods, all but the HT estimator indicate that participation in school meal programs had no statistically significant effect on children's and youths' BMI. In particular, both our proposed RoBART and the calibration estimator of [7] yield small point estimates when using the full sample. And the credible intervals generated by RoBART are similar in magnitude to the confidence intervals of the calibration estimator.

[6]We cite the calibration estimator with exponential tilting as representative; other calibration estimates in [7] are similar in magnitude and interval length.

# A Proofs of Main Results

In the following, we denote the log-likelihood of the conditional probability density function as

$$\ell_n(m_\eta) = \sum_{i=1}^{n} R_i \log f_{m_\eta}(Y_i \mid X_i),$$

which depends on the conditional mean function $m_\eta(\cdot)$. We use the empirical process and bootstrap process notations by writing $\mathbb{P}_n[h] = n^{-1}\sum_{i=1}^n h(Z_i)$ and $\mathbb{P}_n^*[h] = \sum_{i=1}^n W_{ni} h(Z_i)$. Recall the definition of the measurable sets $\mathcal{H}_n^m$ of functions $m_\eta$ such that $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \to_{P_0} 1$. We introduce the conditional prior $\Pi_n(\cdot) := \Pi(\cdot \bigcap \mathcal{H}_n^m)/\Pi(\mathcal{H}_n^m)$. Our BvM theorem makes use of a frequentist estimator $\widehat{\chi}_n$ as the centering term. Note that this estimator itself is not needed in constructing the Bayesian point estimator or the credible set. All we require is that it admits the linear representation with the efficient influence function. Thus, we have the freedom to choose the following one to simplify our subsequent asymptotic analysis:

$$\widehat{\chi} = \mathbb{P}_n\left[m_0(X) + \widehat{\gamma}(R, X)(Y - m_0(X))\right]. \tag{A.1}$$

For simplicity of notation, we introduce

$$\rho^m(y, x) = y - m(x),$$

which is used in the proofs presented below.

*Proof of Theorem 3.1.* In the same spirit of Theorem 2 in [28], we work that the estimated least favorable direction $\widehat{\gamma}$ is based on observations that are independent of $Z^{(n)}$ by Assumption 2. In the sequel, we denote it as a deterministic sequence of $\gamma_n$. Consequently, we can write $\chi_\eta = \mathbb{P}_n^*[m_\eta + \gamma_n \rho^{m_\eta}]$

By Theorem 1.13.1 in [38], we can show this by establishing the convergence of the conditional Laplace transform

$$I_n(t) := \frac{\iint_{\eta\in\mathcal{H}_n} \exp\left(t\sqrt{n}(\chi_\eta - \widehat{\chi} - b_{0,\eta})\right) \mathrm{d}\Pi_W(W^{(n)}|Z^{(n)}) e^{\ell_n(m_\eta)} \mathrm{d}\Pi(m_\eta)}{\int_{\eta\in\mathcal{H}_n} e^{\ell_n(m_\eta)} \mathrm{d}\Pi(m_\eta)}.$$

We proceed with the following decomposition:

$$\begin{aligned}\chi_\eta - \widehat{\chi} =& \mathbb{P}_n^* \left[m_0 + \gamma_n \rho^{m_0}\right] - \mathbb{P}_n \left[m_0 + \gamma_n \rho^{m_0}\right] + \mathbb{P}_n^* \left[m_\eta - m_0 - \gamma_n (m_\eta - m_0)\right] \\ =& (\mathbb{P}_n^* - \mathbb{P}_n) \left[m_0 + \gamma_0 \rho^{m_0}\right] + (\mathbb{P}_n^* - \mathbb{P}_n) \underbrace{\left[m_\eta - m_0 - \gamma_n (m_\eta - m_0) + (\gamma_n - \gamma_0)\rho^{m_0}\right]}_{=:\vartheta_{n,\eta}} \\ &+ \underbrace{\mathbb{P}_n \left[m_\eta - m_0 - \gamma_n (m_\eta - m_0)\right]}_{=b_{0,\eta}}.\end{aligned}$$

By definition, we can express the first term as $(\mathbb{P}_n^* - \mathbb{P}_n)\left[m_0 + \gamma_0 \rho^{m_0}\right] = (\mathbb{P}_n^* - \mathbb{P}_n)\widetilde{\chi}_0$. Using the notation of the influence function at $P_0$ given by $\widetilde{\chi}_0(z) = m_0(x) + \gamma_0(r,x)\rho^{m_0}(y,x) - \chi_0$, we may write

$$\chi_\eta - \widehat{\chi} - b_{0,\eta} = (\mathbb{P}_n^* - \mathbb{P}_n)(\widetilde{\chi}_0 + \vartheta_{n,\eta}).$$

Thus, the conditional Laplace transform becomes

$$I_n(t) = \frac{\iint_{\eta \in \mathcal{H}_n} \exp\left(t\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)(\widetilde{\chi}_0 + \vartheta_{n,\eta})\right) \mathrm{d}\Pi_W(W^{(n)} \mid Z^{(n)}) e^{\ell_n(m_\eta)} \mathrm{d}\Pi(m_\eta)}{\int_{\eta \in \mathcal{H}_n} e^{\ell_n(m_\eta)} \mathrm{d}\Pi(m_\eta)}.$$

Under Assumption 3 (i), we apply Lemma B.2 to the functional class $\{\widetilde{\chi}_0 + \vartheta_{n,\eta} : \eta \in \mathcal{H}_n\}$. Given that the estimated propensity score is bounded away from zero in Assumption 2, the Glivenko-Cantelli property as well as the required moment conditions on the envelope function are satisfied by the corresponding assumptions on the class $\{m_\eta : \eta \in \mathcal{H}_n\}$ in Assumption 3 (i). The conclusion from Lemma B.2 gives us

$$\sup_{\eta \in \mathcal{H}_n} \left| \mathbb{E}\left[ e^{t\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)(\widetilde{\chi}_0 + \vartheta_{n,\eta})} \;\middle|\; Z^{(n)} \right] - e^{t^2 Var_0(\widetilde{\chi}_0(Z) + \vartheta_{n,\eta}(Z))/2} \right| = o_{P_0}(1).$$

Under Assumption 2 we show in Lemma B.4 that

$$\sup_{\eta \in \mathcal{H}_n} \left| Var_0\left(\widetilde{\chi}_0(Z) + \vartheta_{n,\eta}(Z)\right) - Var_0\left(\widetilde{\chi}_0(Z)\right) \right| \to 0. \tag{A.2}$$

We emphasize that the above uniform convergence only requires the influence function $\widetilde{\chi}_0 + \vartheta_{n,\eta}$ to be in the $P_0$-Glivenko-Cantelli class, not necessarily the $P_0$-Donsker class. Therefore, we have

$$I_n(t) = e^{t^2 Var_0(\widetilde{\chi}_0(Z))/2} e^{o_{P_0}(1)} \frac{\int_{\eta \in \mathcal{H}_n} e^{\ell_n(m_\eta)} \, \mathrm{d}\Pi(m_\eta)}{\int_{\eta \in \mathcal{H}_n} e^{\ell_n(m_\eta)} \, \mathrm{d}\Pi(m_\eta)} = e^{t^2 Var_0(\widetilde{\chi}_0(Z))/2} e^{o_{P_0}(1)},$$

which concludes the proof. □

The original setup in [4] concerns the average treatment effect estimation. Our missing data example can be viewed as applying the proposal of [4] to the treated (or control) arm only. For concreteness, we adapt the proof of their Theorem 3.2 to show the feasible estimator $\widehat{b}_\eta$ approximates $b_{0,\eta}$ uniformly well in the missing data problem.

*Proof of Theorem 3.2.* It is sufficient to show that

$$\sup_{\eta\in\mathcal{H}_n}\left|b_{0,\eta}-\widehat{b}_\eta\right| = o_{P_0}(n^{-1/2}),$$

where $b_{0,\eta} = \mathbb{P}_n[(\gamma_0-1)(m_0-m_\eta)]$ and $\widehat{b}_\eta = \mathbb{P}_n[(\widehat{\gamma}-1)(\widehat{m}-m_\eta)]$. We make use of the decomposition

$$b_{0,\eta}-\widehat{b}_\eta = \mathbb{P}_n[(\gamma_0-\widehat{\gamma})(m_0-m_\eta)] + \mathbb{P}_n[(\widehat{\gamma}-1)(m_0-\widehat{m})]. \tag{A.3}$$

Consider the first summand on the right hand side of the previous equation. From Assumption 3 we infer

$$\begin{aligned}\sqrt{n}\sup_{\eta\in\mathcal{H}_n}|\mathbb{P}_n[(\gamma_0-\widehat{\gamma})(m_0-m_\eta)]| &\leqslant \sup_{\eta\in\mathcal{H}_n}|\mathbb{G}_n[(\gamma_0-\widehat{\gamma})(m_0-m_\eta)]| \\ &\quad + \sqrt{n}\sup_{\eta\in\mathcal{H}_n}|P_0[(\gamma_0-\widehat{\gamma})(m_0-m_\eta)]| \\ &\leqslant o_{P_0}(1) + O_{P_0}(1)\times\sqrt{n}\|\pi_0-\widehat{\pi}\|_{2,\mu_0}\sup_{\eta\in\mathcal{H}_n}\|m_\eta-m_0\|_{2,\mu_0} = o_{P_0}(1),\end{aligned}$$

using Assumption 3 (ii), the Cauchy-Schwarz inequality and Assumption 2. Consider the second summand on the right hand side of (A.3). Another application of the Cauchy-Schwarz inequality and Assumption 2 yields

$$\begin{aligned}\mathbb{P}_n[(\widehat{\gamma}-1)(m_0-\widehat{m})] &= \mathbb{P}_n[(\gamma_0-1)(m_0-\widehat{m})] + o_{P_0}(n^{-1/2}) \\ &= \mathbb{P}_n\left[\frac{R-\pi_0(X)}{\pi_0(X)}(m_0(X)-\widehat{m}(X))\right] + o_{P_0}(n^{-1/2}).\end{aligned}$$

Under Assumption 2, we apply Lemma B.3 to get $\mathbb{P}_n\left[\frac{R-\pi_0(X)}{\pi_0(X)}(m_0(X)-\widehat{m}(X))\right] + o_{P_0}(n^{-1/2})$. Hence, we obtain

$$\mathbb{P}_n[(\gamma_0-1)(m_0-\widehat{m})] = o_{P_0}(n^{-1/2}),$$

which completes the proof. □

Referring to BvM theorems under primitive conditions for the Bayesian regression trees (Bayesian CART) or forests (BART), we provide the detailed proof for Bayesian CART by verifying the high-level assumptions and highlight the modifications for the BART afterwards. Below, $a \vee b$ denotes the maximum of $a$ and $b$.

*Proof of Theorem 4.1.* First, we provide the construction of measurable sets $\mathcal{H}_n$ such that $\Pi(\eta \in \mathcal{H}_n \mid Z^{(n)}) \to_{P_0} 1$ in Assumption 2. As we impose the BART type priors over the conditional mean function, it suffices to find the proper sets $\mathcal{H}_n^m$ for $m_\eta$ such that $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \to_{P_0} 1$. Referring to the sieve set defined by equation (B.2), we work with

$$\mathcal{H}_n^m := \{m_\eta : m_\eta \in \mathcal{M}_n, \|m_\eta - m_0\|_{2,\mu_0} \lesssim \varepsilon_n\}. \tag{A.4}$$

We prove the following posterior contraction in Lemma B.5:

$$\Pi(m_\eta \in \mathcal{M}_n : \|m - m_0\|_{2,\mu_0} > C_n \varepsilon_n \mid Z^{(n)}) \to_{P_0} 0, \tag{A.5}$$

with $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)}\sqrt{\log n}$ for any $C_n \to \infty$ as $n, p \to \infty$ in Regime 1. In Lemma B.5, we have shown that $\Pi(m_\eta \notin \mathcal{M}_{\mathcal{T}}^n) \to_{P_0} 0$. Therefore, one also obtains $\Pi(m_\eta \notin \mathcal{M}_{\mathcal{T}}^n \mid Z^{(n)}) \to_{P_0} 0$, combined with Lemma 1 in [17]. Taken together, we have shown that $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \to_{P_0} 1$.

Under the uniformly boundedness assumption of the conditional mean, the requirement for the envelope functions in (3.3) are satisfied. Regarding the $P_0$-Glivenko-Cantelli property of $\mathcal{G}_n$, it is sufficient to verify this for the sieve set $\mathcal{M}^n$, because the functions in $\mathcal{G}_n$ correspond to the shifted versions by subtracting the true $m_0$. Let $\|\cdot\|_n$ denote the empirical $L^2(\mathbb{P}_n)$–norm. For $\mathcal{M}_n$ with $\bar{K}_n \sim n\varepsilon_n^2/\log n$ and $\bar{s}_n \sim n\varepsilon_n^2/\log(p \vee n)$. In the proof of Lemma B.5, we have $\log N(\epsilon, \mathcal{M}^n, \|\cdot\|_n) \lesssim n^{q_0/(2\alpha+q_0)} \log n = o(n)$ for any given $\epsilon > 0$. This satisfies the requirement about the growth of the entropy number in Theorem 2.4.3 of [38], which verifies the $P_0$-Glivenko-Cantelli property.

For the stochastic equicontinuity condition, we apply the multiplier inequality as stated

in Lemma B.1:

$$\begin{aligned}
&\sup_{\eta\in\mathcal{H}_n} |\mathbb{G}_n\left[(\gamma_0-\gamma_n)\,(m_\eta-m_0)\right]| \\
&\leqslant 4\|\gamma_n-\gamma_0\|_\infty \mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n} |\mathbb{G}_n(m_\eta-m_0)| + \sqrt{P_0(\gamma_n-\gamma_0)^2}\sup_{\eta\in\mathcal{H}_n}|P_0(m_\eta-m_0)| \\
&\lesssim \|\pi_n-\pi_0\|_\infty \mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n} |\mathbb{G}_n(m_\eta-m_0)| + \|\pi_n-\pi_0\|_{2,\mu_0}\sup_{\eta\in\mathcal{H}_n}\|m_\eta-m_0\|_{2,\mu_0} \\
&\lesssim \|\pi_n-\pi_0\|_\infty\sqrt{n}\varepsilon_n + \|\pi_n-\pi_0\|_{2,\mu_0}\varepsilon_n = o_{P_0}(1).
\end{aligned}$$

In the last inequality, we have applied the upper bound $\mathbb{E}_0 \sup_{m_\eta\in\mathcal{M}(\varepsilon_n)} |\mathbb{G}_n(m_\eta-m_0)| \lesssim \sqrt{n}\varepsilon_n$, as given in Lemma B.6.

When it comes to the functional class with the additive structure, we modify the posterior rate of contraction by $\varepsilon_n^{add} = \sqrt{\sum_{t=1}^{T_0} n^{-2\alpha^t/(2\alpha^t+q_0^t)}\log n}$ for a fixed $T_0 < T$. Following the second part of Lemma B.5, we have

$$\Pi(m_\eta \in \mathcal{M}_\mathcal{E} : \|m_\eta - m_0\|_{2,\mu_0} > C_n\varepsilon_n^{add} \mid Z^{(n)}) \to_{P_0} 0,$$

for any $C_n \to \infty$ as $n, p \to \infty$. The rest of the proof follows similarly to the Hölder continuous case. □

# B Useful Lemmas and Auxiliary Results

## B.1 Useful Lemmas

The following lemma is in the same spirit of Lemma 9 in [28] with one important difference. That is, we do not restrict the range of the function $\varphi$ to $[0, 1]$. This is important, as we apply this lemma with $\varphi = \gamma_n - \gamma_0$, which can take on negative values. Accordingly, we use the more general contraction principle from Theorem 4.12 of [26] instead of Proposition A.1.10 of [38]. This allows us to relax the positive range restriction in [28].

**Lemma B.1.** *Consider a set $\mathcal{H}$ of measurable functions $h : \mathcal{Z} \mapsto \mathbb{R}$ and a bounded measurable function $\varphi$. We have*

$$\mathbb{E}\sup_{h\in\mathcal{H}}|\mathbb{G}_n(\varphi h)| \leqslant 4\|\varphi\|_\infty \mathbb{E}\sup_{h\in\mathcal{H}}|\mathbb{G}_n(h)| + \sqrt{P_0\varphi^2}\sup_{h\in\mathcal{H}}|P_0 h|.$$

We now state the following generalization of Theorem 1 from [29], where the functional class $\mathcal{G}_n$ containing $g(\cdot)$ can vary with the sample size. We strengthen it following similar

moment conditions as in Lemma 11 of [40]. As discussed by [29] on Page 2225 therein, this uniformity refers to the marginal posterior distributions of the process $\widetilde{\chi}_0 + \vartheta_{n,\eta}$ for $\eta \in \mathcal{H}_n$. It is not about the distributional convergence of this process as a random element in the set of $\ell^\infty(\mathcal{H})$.

**Lemma B.2.** *Suppose $\mathcal{G}_n$ is a sequence of separable classes of measurable functions with envelope functions $G_n$, such that*

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}_0[g(Z)]\right| \to_{P_0} 0.$$

*In addition,* $\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{E}_0[G_n^2 \mathbb{I}\{G_n^2 > C\}] = 0$, *and* $\mathbb{E}_0[G_n^4] = o(n)$. *Then for every $t$ in a sufficiently small neighborhood of* 0,

$$\sup_{g\in\mathcal{G}_n}\left|\mathbb{E}_0\left[\exp\left(t\sqrt{n}\sum_{i=1}^{n}(W_{ni} - 1/n)g(Z_i)\right)\,\middle|\, Z^{(n)}\right] - \exp\left(t^2 Var_0(g(Z))/2\right)\right| \to_{P_0} 0.$$

## B.2 Smaller Order Terms

The next lemma verifies the asymptotic negligible term that appear in the debiasing step. Its proof is in the same spirit of bounding the $R_{n,2}$ term in Lemma C.8 of [5].

**Lemma B.3.** *Suppose that the pilot estimator computed over some external independent sample converges to the true conditional mean in the $L^2(\mu_0)$–norm, i.e., $\|\widehat{m} - m_0\|_{2,\mu_0}^2 \to 0$. Also, the true propensity score is uniformly bounded away from zero, then we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i - \pi_0(X_i)}{\pi_0(X_i)}(m_0 - \widehat{m})(X_i) = o_{P_0}(1). \tag{B.1}$$

*Proof.* We condition on $(X_1, \ldots, X_n)$, as well as the pilot estimator $\widehat{m}$, which is computed over the external independent sample. We use the fact that $(R_i - \pi_0(X_i))$ has a conditional zero mean. Specifically, this leads to

$$\begin{aligned}
&\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i - \pi_0(X_i)}{\pi_0(X_i)}(\widehat{m}(X_i) - m_0(X_i))\Big)^2 \,\Big|\, X_1, \ldots, X_n, \widehat{m}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{m}(X_i) - m_0(X_i)\big)^2 \frac{Var_0(R_i \mid X_i)}{\pi_0^2(X_i)}
\end{aligned}$$

using that $Var_0(R_i|X_i) = \pi_0(X_i)(1 - \pi_0(X_i))$. Consequently, by the overlapping condition

$1 \lesssim \pi_0(X_i)$ for all $1 \leqslant i \leqslant n$, we obtain

$$\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i-\pi_0(X_i)}{\pi_0(X_i)}(m_0-\widehat{m})(X_i)\Big)^2\right] \lesssim \|\widehat{m}-m_0\|_{2,\mu_0}^2 = o(1),$$

where the last equation is due to the $L^2(\mu_0)$-convergence for the pilot estimator $\widehat{m}$. Consequently, the result follows from using the Markov inequality. □

For the next result, recall the definition of $\vartheta_{n,\eta} = [m_\eta - m_0 - \gamma_n(m_\eta - m_0) + (\gamma_n - \gamma_0)\rho^{m_0}]$ introduced in the proof of Theorem 3.1.

**Lemma B.4.** *Let $\pi_n(\cdot)$ be a sequence of functions uniformly bounded away from zero and $\|\pi_n - \pi_0\|_{2,\mu_0} \to 0$. Uniformly in $\eta \in \mathcal{H}_n$, assume $\|m_\eta - m_0\|_{2,\mu_0} \to 0$. Then we have*

$$\sup_{\eta\in\mathcal{H}_n} |Var_0(\widetilde{\chi}_0 + \vartheta_{n,\eta}) - Var_0(\widetilde{\chi}_0)| \to 0.$$

*Proof.* Since $\widetilde{\chi}_0(Z)$ is centered under $P_0$, we obtain by elementary calculation and the Cauchy-Schwartz inequality

$$\begin{aligned}
&|Var_0(\widetilde{\chi}_0(Z) + \vartheta_{n,\eta}(Z)) - Var_0(\widetilde{\chi}_0(Z))| \\
&\quad\leqslant \left|\mathbb{E}_0(\widetilde{\chi}_0(Z) + \vartheta_{n,\eta}(Z))^2 - \mathbb{E}_0(\widetilde{\chi}_0(Z))^2\right| + (\mathbb{E}_0\vartheta_{n,\eta}(Z))^2 \\
&\quad\leqslant 2\sqrt{\mathbb{E}_0\widetilde{\chi}_0^2(Z)}\sqrt{\mathbb{E}_0\vartheta_{n,\eta}^2(Z)} + 2\mathbb{E}_0\vartheta_{n,\eta}^2(Z) \\
&\quad= 2\left(\sqrt{\mathrm{V}_0}\|\vartheta_{n,\eta}\|_{2,\mu_0} + \|\vartheta_{n,\eta}\|_{2,\mu_0}^2\right).
\end{aligned}$$

A close inspection of the function $\vartheta_{n,\eta}$ shows that we can proceed by

$$\begin{aligned}
\vartheta_{n,\eta}(Z) &= (1 - \frac{R}{\pi_n(X)})(m_\eta(X) - m_0(X)) + (\gamma_n(R,X) - \gamma_0(R,X))(Y - m_0(X)) \\
&= \frac{\pi_n(X) - \pi_0(X)}{\pi_n(X)}(m_\eta(X) - m_0(X)) + \frac{\pi_0(X) - R}{\pi_n(X)}(m_\eta(X) - m_0(X)) \\
&\quad + (\gamma_n(R,X) - \gamma_0(R,X))(Y - m_0(X)).
\end{aligned}$$

We now make use of the assumption that $\pi_n(\cdot)$ is uniformly bounded away from zero, as well as the conditional variance of $R_i(Y_i - m_0(X_i))$ is in $L^2(\mu_0)$, i.e., $\int \sigma_0^2(x)\,\mathrm{d}\mu_0(x) < \infty$. Thus, we obtain

$$\|\vartheta_{n,\eta}\|_{2,\mu_0} \lesssim \|m_\eta - m_0\|_{2,\mu_0} + \|\pi_n - \pi_0\|_{2,\mu_0} + \|m_\eta - m_0\|_{2,\mu_0}\|\pi_n - \pi_0\|_{2,\mu_0},$$

after some simple algebraic steps. The desired conclusion follows from the convergence of the pilot estimator of the propensity score and $\sup_{\eta\in\mathcal{H}_n}\|m_\eta - m_0\|_{2,\mu_0} \to 0$. □

## B.3 Results Related to Regression Trees

In tree-structured models, the idea is to recursively apply binary splitting rules to partition the support of covariates. Although tree-based partitioning allows splits to occur anywhere in the domain, it is often preferable to select split-points from a discrete set [13]. Recall that for the Cartesian product of $p$ subsets of $\mathbb{R}$, i.e., $\Omega \subset \mathbb{R}^p$, we denote the $j$–th projection mapping of $\Omega$ by $[\Omega]_j = \{x_j \in \mathbb{R} : (x_1, \ldots, x_p)^\top \in \Omega\}$. For a given split-net $\mathcal{X}$, we call each point $x_i = (x_{i1}, \ldots, x_{ip})^\top$ a split-candidate. For a given splitting coordinate $j$ and a split-net $\mathcal{X}$, a split-point will be chosen from $[\mathcal{X}]_j \bigcap \mathsf{int}([\Omega]_j)$ to dichotomize a box $\Omega$. Note that $[\mathcal{X}]_j = \{x_{ij} \in [0,1], i = 1 \ldots, b_n\}$ may have fewer elements than $\mathcal{X}$ due to duplication. We denote by $b_j(\mathcal{X})$ the cardinality of $[\mathcal{X}]_j$, i.e., the number of unique elements in the $b_n$-tuple $(x_{1j}, \ldots, x_{b_n j})$.

In this part, we prove the posterior rate of contraction for the BART. The first rigorous analysis covers the random design case appears [25] for a more general piecewise heterogeneous anisotropic Hölder class. Our setup can be seen as a special case of [25] by restricting to the isotropic Hölder continuous class. For two generic probability densities $p$ and $q$, we denote the Kullback-Leibler (KL) divergence by $K(p,q)$ and the square KL variation by $V(p,q)$; see Appendix B in [18]. Furthermore, denote the Hellinger distance by $\rho_H(\cdot,\cdot)$. Our derivation of the posterior contraction rate builds on considering sieve sets on which the prior concentrates on.

With given $\mathcal{E}$ and $0 < \bar{M} < \infty$, we define the function space

$$\mathcal{M}_{\mathcal{E},\bar{M}} := \left\{ m_{\mathcal{E},\mathcal{B}} \in \mathcal{M}_{\mathcal{E}} : \|\mathcal{B}\|_\infty \leqslant \bar{M} \right\},$$

where $\|\cdot\|_\infty$ denotes the supnorm of the vector $\mathcal{B}$. We consider the collection of all $\mathcal{M}_{\mathcal{E},\bar{M}}$ such that the the number of terminal nodes is upper bounded by $\bar{K}_n$ and the number of active splitting variables is upper bounded by $\bar{s}_n$ for each individual trees in the following sieve set

$$\mathcal{M}_n := \mathcal{M}_{\bar{s}_n,\bar{K}_n,\bar{M}} := \bigcup_{\mathcal{E}:|S|\leqslant\bar{s}_n, K^t\leqslant\bar{K}_n, 1\leqslant t\leqslant T} \mathcal{M}_{\mathcal{E},\bar{M}}. \tag{B.2}$$

where $\bar{K}_n \sim n\varepsilon_n^2/\log n$, $\bar{s}_n \sim n\varepsilon_n^2/\log(p \vee n)$, and $\bar{M}$ is the upper bound for individual step height in its prior distribution. Because the prior for each step height has a bounded density with compact support, we work with a fixed $\bar{M}$, rather than a growing upper bound

for the normal density in [25].

**Lemma B.5.** *Under the assumptions stated in Theorem 4.1 for $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$, we have*

$$\Pi\left(m \in \mathcal{M}_n : \|m - m_0\|_{2,\mu_0} > C_n \varepsilon_n \mid Z^{(n)}\right) \to_{P_0} 0,$$

*with $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)}\sqrt{\log n}$ for any $C_n \to \infty$ as $n, p \to \infty$. For $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \boldsymbol{\mathcal{S}}_0)$, the posterior rate of contraction holds with $\varepsilon_n^{add}$.*

*Proof.* We prove the result for $m_0$ in the Hölder continuous class and outline the necessary changes needed for the additive class. Let $m_1, m_2$ denote two generic conditional mean functions in the Gaussian model for the outcome variable with associated densities $f_{m_1}, f_{m_2}$. Then, the Hellinger distance $\rho_H(\cdot, \cdot)$ satisfies

$$\|m_1 - m_2\|_{2,\mu_0}^2 \lesssim \rho_H^2(f_{m_1}, f_{m_2}) \lesssim \|m_1 - m_2\|_{2,\mu_0},$$

by Lemma B.2 of [39] under our Assumption 6 (iii). Hence, it is sufficient to check the convergence in terms of $\rho_H$. In addition, we have

$$\max\{K(f_{m_0}, f_{m_\eta}), V(f_{m_0}, f_{m_\eta})\} \lesssim \|m_0 - m_\eta\|_{2,\mu_0}^2.$$

Define $B_n := \{m_\eta : \max\{K(f_{m_0}, f_{m_\eta}), V(f_{m_0}, f_{m_\eta})\} \leqslant \varepsilon_n^2\}$. To check the posterior rate of contraction, we verify the high-level assumptions from Theorem 2.1 in [16]:

$$\Pi(m_\eta \in B_n) \geqslant c_1 \exp(-c_2 n\varepsilon_n^2), \tag{B.3}$$

$$\log N(\varepsilon_n, \mathcal{M}_n, \rho_H) \leqslant c_3 n\varepsilon_n^2, \tag{B.4}$$

$$\Pi(m_\eta \notin \mathcal{M}_n) \leqslant \exp(-c_4 n\varepsilon_n^2), \tag{B.5}$$

for positive constant terms $c_1, \ldots, c_4$.

Referring to the first inequality (B.3), it follows by the direct calculation of the normal likelihood:

$$B_n \supset \{m_\eta : \|m_\eta - m_0\|_{2,\mu_0} \leqslant C_1 \varepsilon_n\}, \tag{B.6}$$

under Assumption 6 (iii). Then we construct an approximating ensemble denoted by $\widehat{\mathcal{E}} = (\widehat{\mathcal{T}}_1, \ldots, \widehat{\mathcal{T}}_T)$. By restricting the function space to the one constructed by $\widehat{\mathcal{E}}$, we have

$$\Pi(m_\eta : \|m_\eta - m_0\|_{2,\mu_0} \leqslant C_1\varepsilon_n) \geqslant \underbrace{\Pi(\widehat{\mathcal{E}})}_{①} \underbrace{\Pi(m_\eta \in \mathcal{M}_{\widehat{\mathcal{E}}} : \|m_\eta - m_0\|_{2,\mu_0} \leqslant C_1\varepsilon_n)}_{②}. \tag{B.7}$$

We provide the lower bound for the two prior probabilities separately. For a given split-net $\mathcal{X}$, there exists by Assumption 5 a tree partition $\widehat{\mathcal{T}}$ generating an approximating function $m_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}$ such that $\|m_0 - m_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_\infty \lesssim \varepsilon_n$. An approximating ensemble $\widehat{\mathcal{E}}$ can be constructed by setting $\widehat{\mathcal{T}}^1$ to be $\widehat{\mathcal{T}}$ and $\widehat{\mathcal{T}}^t = [0,1]^p$ for $t = 2, \ldots, T$ so that the rest of trees are root nodes with no splits. Under Assumption 6 (i)(ii), we have

$$\log \Pi(\widehat{\mathcal{E}}) = \sum_{t=1}^{T} \log \Pi(\widehat{\mathcal{T}}^t) = \log \Pi(\widehat{\mathcal{T}}^1) + (T-1)\log(1-\nu) \geqslant -C_2 n\varepsilon_n^2,$$

where the last inequality follows from Lemma 4 in [25].

Referring to the second part ②, we have $\|m_\eta - m_0\|_{2,\mu_0} \lesssim \|m_\eta - m_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_\infty + \varepsilon_n$ for some $m_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}} \in \mathcal{M}_{\widehat{\mathcal{T}}}$. We set all trees in $\widehat{\mathcal{E}}$ the root nodes except for the first one $\widehat{\mathcal{T}}^1 = \widehat{\mathcal{T}}$. Every step-heights vector $B$ for $\widehat{\mathcal{E}}$ has the form $B = (\beta^1, \beta^2, \ldots, \beta^T)^\top \in \mathbb{R}^{\widehat{K}+T-1}$ with $\beta^1 \in \mathbb{R}^{\widehat{K}}$ and $\beta^t \in \mathbb{R}$ with $t = 2, \ldots, T$, where $\widehat{K}$ is the size of $\widehat{\mathcal{T}}$. By letting $\widehat{B} = (\widehat{\beta}, 0, \ldots, 0)^\top$, we can write $m_{0,\widehat{\mathcal{T}},\widehat{\beta}} = m_{0,\widehat{\mathcal{E}},\widehat{B}}$. Thereafter, we obtain

$$\Pi(m_\eta \in \mathcal{M}_{\widehat{\mathcal{E}}} : \|m_\eta - m_0\|_{2,\mu_0} \leqslant C_1\varepsilon_n) \geqslant \Pi(m_\eta \in \mathcal{M}_{\widehat{\mathcal{E}}} : \|m_\eta - m_{0,\widehat{\mathcal{E}},\widehat{B}}\|_\infty \leqslant C_2\varepsilon_n).$$

A similar calculation as in the proof of Lemma 5 in [25] leads to the lower bound of the right hand side of the above inequality. By Assumption 6 (iii) where the priors step-height are independent and compactly supported, we can simply replace the Gaussian small ball probability bounds with the volume of order $C\nu_n^{\widehat{K}_*}$ with $\nu_n = \varepsilon_n \sigma_{\max}^{-1}(A^{-1})/\sqrt{\widehat{K}_*}$ for $\widehat{K}_* = \widehat{K} + T - 1$, by lower bounding the volume of $\{B \in \mathbb{R}^{\widehat{K}_*} : \|AB\|_2 \leqslant C\nu_n\}$ for some non-singular matrix $A$. Therein, we have

$$\log \Pi(\widehat{\mathcal{E}}) \geqslant -C_2 n\varepsilon_n^2, \quad \log \Pi(m_\eta \in \mathcal{M}_{\widehat{\mathcal{E}}} : \|m_\eta - m_0\|_\infty \leqslant C_3\varepsilon_n) \geqslant -C_4 n\varepsilon_n^2.$$

The above two inequalities combined with (B.6) and (B.7) gives us the desired bound in (B.3).

Next, we bound the entropy number in (B.4). Under the unit Gaussian error assumption, we can bound the Hellinger distance between $f_{m_1}$ and $f_{m_2}$ by the $L^2(\mu_0)$–distance $\|m_1 - m_2\|_{2,\mu_0}$ by Lemma 2.7 (i) in [18]. We proceed to upper bound $\log N(\varepsilon, \mathcal{M}_n, \|\cdot\|_\infty)$ for the stronger supnorm. Define $\boldsymbol{\mathcal{E}}_{S,K^1,\ldots,K^T}$ as the collection of all tree ensembles $\mathcal{E}$ with given $S, K^1, \ldots, K^T$. By construction, we have the upper bound

$$N(\varepsilon_n, \mathcal{M}_n, \|\cdot\|_\infty) \leqslant \sum_{S:|S|\leqslant \bar{s}_n} \sum_{(K^1,\ldots,K^T):K^t\leqslant \bar{K}_n} \sum_{\mathcal{E}\in\boldsymbol{\mathcal{E}}_{S,K^1,\ldots,K^T}} N(\varepsilon_n, \mathcal{M}_{\mathcal{E},\bar{M}}, \|\cdot\|_\infty),$$

where we recall that $\bar{K}_n \sim n\varepsilon_n^2/\log n$ and $\bar{s}_n \sim n\varepsilon_n^2/\log(p \vee n)$. For any given $\mathcal{E}$ and $\mathcal{B}_1, \mathcal{B}_2 \in \mathbb{R}^{\sum_{t=1}^T K^t}$, we have

$$\|m_{\mathcal{E},\mathcal{B}_1} - m_{\mathcal{E},\mathcal{B}_2}\|_\infty = \sup_{x\in[0,1]^p} \left| \sum_{t=1}^T \sum_{k=1}^{K^t} (\beta_{1k}^t - \beta_{2k}^t) \mathbb{1}\{x \in \Omega_k^t\} \right| \leqslant \sum_{t=1}^T K^t |\mathcal{B}_1 - \mathcal{B}_2|_\infty.$$

Following the proof of Lemma 6 of [25], the covering number of the sieve set is thus upper bounded by

$$\begin{aligned}
&N(\varepsilon_n, \mathcal{M}_n, \|\cdot\|_\infty) \\
&\quad \leqslant \bar{K}_n^T \sum_{s=1}^{\bar{s}_n} \binom{p}{s} \left( s \max_{1\leqslant j\leqslant p} b_j(\mathcal{X}) \right)^{T\bar{K}_n} N\left( \varepsilon_n/T\bar{K}_n, \{\mathcal{B} \in \mathbb{R}^{T\bar{K}_n} : \|\mathcal{B}\|_\infty \leqslant \bar{M}\}, \|\cdot\|_\infty \right) \\
&\quad \leqslant \bar{K}_n^T \bar{s}_n p^{\bar{s}_n} \left( \bar{s}_n \max_{1\leqslant j\leqslant p} b_j(\mathcal{X}) \right)^{T\bar{K}_n} \left( 3T\bar{K}_n\bar{M}/\varepsilon_n \right)^{T\bar{K}_n}.
\end{aligned}$$

Under Assumption 5 (i), that is, $\max_{1\leqslant j\leqslant p} b_j(\mathcal{X}) \lesssim \log n$, logarithm of the above quantity satisfies

$$\begin{aligned}
\log N(\varepsilon_n, \mathcal{M}_n, \|\cdot\|_\infty) &\lesssim \log \bar{K}_n + \log \bar{s}_n + \bar{s}_n \log p + \bar{K}_n \left( \log \bar{K}_n + \log \bar{s}_n + \log\log n - \log \varepsilon_n \right) \\
&\lesssim \bar{s}_n \log p + \bar{K}_n \log n
\end{aligned}$$

using $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)}\sqrt{\log n}$ and the choices of $\bar{K}_n$ and $\bar{s}_n$ given below the sieve set (B.2). These choices immediately imply $\log N(\varepsilon_n, \mathcal{M}_n, \|\cdot\|_\infty) \lesssim n\varepsilon_n^2$.

When it comes to the prior mass outside the sieve set, we follow the proof of the condition (2.3) in [33]. It boils down to check:

$$\sum_{t=1}^T \Pi\left(K^t > \bar{K}_n\right) + \Pi\left(S : s > \bar{s}_n \mid K^t \leqslant \bar{K}_n, t = 1, \ldots, T\right) \lesssim e^{-n\varepsilon_n^2},$$

with proper choices of $\bar{K}_n$ and $\bar{s}_n$ to be made later. Under Assumptions 6 (i)(ii), one can apply Corollary 7 of [32] to get

$$\log \Pi\left(K^t > \bar{K}_n\right) \lesssim -\bar{K}_n \log \bar{K}_n \lesssim -\bar{K}_n \log n,$$

for $\bar{K}_n = \lfloor Cn\varepsilon_n^2/\log n \rfloor$. Under Assumption 6 (i) with $\bar{s}_n = \lfloor Cn\varepsilon_n^2/\log(p \vee n) \rfloor$, it follows

from the proof on Page 50 of [25] that

$$\Pi\left(S : s > \bar{s}_n \mid K^t \leqslant \bar{K}_n, t = 1, \ldots, T\right) \lesssim e^{-n\varepsilon_n^2}.$$

This leads to the bound in (B.5) and completes the proof of the posterior rate of contraction for the Hölder continuous class.

We outline the modifications needed for the additive model. The first difference occurs to the construction of the approximating ensemble $\widehat{\mathcal{E}}$. For each $1 \leqslant t \leqslant T_0$, there exists a tree partition $\widehat{\mathcal{T}}^t$ generating an approximating function $m^t_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}$ such that $\|m_0^t - m^t_{0,\widehat{\mathcal{T}},\widehat{\boldsymbol{\beta}}}\|_\infty \lesssim \varepsilon_{n,t}$. An approximating ensemble $\widehat{\mathcal{E}}$ can be constructed by setting $\widehat{\mathcal{T}}^t$ as aforementioned with $1 \leqslant t \leqslant T_0$, and taking the rest of trees to be root nodes with no splits. The remaining proof about the posterior rate of contraction is similar to the proof of Theorem 7 of [25], and hence omitted. □

The next lemma bounds the local empirical process term without assuming Donsker property. Compared with more refined analysis in [21], the following bound is not sharp. Nonetheless, it is sufficient to show the validity of our debiasing step, under the rate condition $\sqrt{n}\varepsilon_n r_n \to 0$.

**Lemma B.6.** *Given the sieve set $\mathcal{M}_n$ over the tree ensembles in B.2 and the posterior rate of contraction $\varepsilon_n$, we have*

$$\mathbb{E}_0 \sup_{m \in \mathcal{M}_n(\varepsilon_n)} |\mathbb{G}_n(m_\eta - m_0)| \lesssim \sqrt{n}\varepsilon_n, \tag{B.8}$$

*where $\mathcal{M}_n(\varepsilon_n) := \{m_\eta - m_0 : \|m_\eta - m_0\|_{2,\mu_0} \lesssim \varepsilon_n\}$.*

*Proof.* The functional class $\mathcal{M}_n(\varepsilon_n)$ recenters $\mathcal{M}_n$ by subtracting $m_0$ and it also restricts to functions around the truth with a radius of order $\varepsilon_n$. We resort to the following inequality as in Equation (2.6) of [21]:

$$\mathbb{E}_0 \sup_{m_\eta \in \mathcal{M}_n(\sigma)} |\mathbb{G}_n(m_\eta)| \lesssim \inf_{0 \leqslant \iota \leqslant \sigma/2} \left\{ \sqrt{n}\iota + \int_\iota^\sigma \sqrt{\log N_{[]}(\varepsilon, \mathcal{M}_n, L^2(\mu_0))}\, \mathrm{d}\varepsilon \right\}$$

The entropy integral term $\int_\iota^\sigma \sqrt{\log N_{[]}(\varepsilon, \mathcal{M}_n, L^2(\mu_0))}\, \mathrm{d}\varepsilon$ comes from $L^2$ chaining with bracketing in the Gaussian regime using Bernstein's inequality, starting from $\sigma$ to $\iota$. The residual term $\sqrt{n}\iota$ comes from the bound $\|m\|_{1,\mu_0} \leqslant \|m\|_{2,\mu_0}$ towards the end level $\iota$ of the $L^2$ chaining, where $\|\cdot\|_{1,\mu_0}$ denotes the $L^1(\mu_0)$–norm.

For the bracketing entropy bound, we first bound the entropy number in $L_\infty$ norm. Let $m_1, \ldots, m_N$ be the center of $\|\cdot\|_\infty$–balls of radius $\varepsilon$ that covers the functional class $\mathcal{M}_n$, one could obtain the pair of brackets $[m_j - \varepsilon, m_j + \varepsilon]$. Each bracket has $L^2(\mu_0)$-size of at most $2\varepsilon$; see the proof of Corollary 2.7.2 of [38]. Given the entropy number in $L_\infty$-norm established in Lemma B.5, we also have the upper bound for the bracketing entropy number as $\log N_{[]}(\varepsilon_n, \mathcal{M}_n, L^2(\mu_0)) \lesssim n\varepsilon_n^2$. By taking $\sigma = \varepsilon_n$ and $\iota = \sigma/2$, we then get the desired upper bound $\mathbb{E}_0 \sup_{m \in \mathcal{M}(\varepsilon_n)} |\mathbb{G}_n(m_\eta)| \lesssim \sqrt{n}\varepsilon_n$. □

# C Potential Outcomes and Treatment Effects

As in [28] and [4], the proposed robust method can be applied to causal inference, which is fundamentally related to the missing data literature [24]. In this section, we outline the extension of our methodology to this topic. In accordance with the theory for the missing data problem, we note that the additional bias correction terms take the same forms in as [4, 3]. For individual $i$, consider a treatment indicator $D_i \in \{0, 1\}$. The observed outcome $Y_i$ is determined by $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ where $(Y_i(1), Y_i(0))$ are the potential outcomes of individual $i$ associated with $D_i = 1$ or 0. The covariates for individual $i$ are denoted by $X_i$, a vector of dimension $p$, with the distribution $F_0$ and the density $f_0$. The key functional component is the conditional mean function

$$m_0(d, x) := \mathbb{E}_0[Y_i | D_i = d, X_i = x].$$

One of the most commonly used causal parameter is the average treatment effect (ATE):

$$\tau_0 := \mathbb{E}_0[m_0(1, X) - m_0(0, X)].$$

We can also express it in the one-step updated form incorporating the correction term:

$$\tau_0 = \mathbb{E}_0 \left[ (m_0(1, X) - m_0(0, X)) + \gamma_0^{ATE}(D, X)(Y - m_0(D, X)) \right],$$

where its Riesz reprenster is

$$\gamma_0^{ATE} = \frac{D}{\pi_0(X)} - \frac{1 - D}{1 - \pi_0(X)}.$$

One can simply apply our current approach to the population mean for the treated and control outcomes separately and study their difference, as demonstrated by [4] based on

the Gaussian process priors with prior and posterior adjustment. Our current work extends [4] to a flexiable modeling scheme that allows for the BART type priors. For concreteness, we outline the algorithm for making posterior draws of the ATE as follows. Let $\widehat{m}$ and $\widehat{\pi}$ be pilot estimators for the conditional mean and propensity score computed over some external sample. We denote the estimated Riesz reprenster as

$$\widehat{\gamma}^{ATE} = \frac{D}{\widehat{\pi}(X)} - \frac{1-D}{1-\widehat{\pi}(X)}.$$

**Algorithm 2** Posterior Computation of ATE

**for** $s = 1, \ldots, S$ **do**

(a) Generate the $s$-th draw using the BART prior and the data; denote it as $(m_\eta^s(D_i, X_i))_{i=1}^n$.

(b) Draw Bayesian bootstrap weights $W_{ni}^s = e_i^s / \sum_{j=1}^n e_j^s$ where $e_i^s \overset{iid}{\sim} \text{Exp}(1)$, $1 \leqslant i \leqslant n$.

(c) Calculate the recentered posterior draw for the ATT $\check{\tau}_\eta^s := \tau_\eta^s - \widehat{b}_\eta^s$ where

$$\tau_\eta^s = \sum_{i=1}^n W_{ni}^s \left[ (m_\eta^s(D_i, X_i) - m_\eta^s(D_i, X_i)) + \widehat{\gamma}^{ATE}(D_i, X_i)\big(Y_i - m_\eta^s(0, X_i)\big) \right].$$

and

$$\widehat{b}_\eta^s = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\tau}[m_\eta^s(\cdot,\cdot) - \widehat{m}(\cdot,\cdot)](Z_i), \tag{C.1}$$

where $\boldsymbol{\tau}[m](z) := m(1,x) - m(0,x) + \widehat{\gamma}^{ATE}(d,x)(y - m(d,x))$.

**end for**

**Output:** $\{\check{\tau}_\eta^s : s = 1, \ldots, S\}$.

Besides the ATE, the average treatment effect on the treated (ATT) is of policy interest, especially to economists [22]. In order to facilitate the comparison with [40] who also studied ATT, we outline our strategy to model the ATT, which is defined as follows:

$$\theta_0 = \frac{\mathbb{E}_0[D_i(Y_i - m_0(0, X_i))]}{\mathbb{E}_0[D_i]}.$$

Its one-step updated version is

$$\theta_0 = \frac{\mathbb{E}_0[D_i(Y_i - m_0(0, X_i))]}{\mathbb{E}_0[D_i]} + \mathbb{E}_0\left[\gamma_0^{ATT}(D, X)(Y - m_0(D, X))\right],$$

where the Riesz representer is given by

$$\gamma_0^{ATT}(D,X) = \frac{D}{\pi_0} - \frac{1-D}{\pi_0}\frac{\pi_0(X)}{1-\pi_0(X)},$$

with $\pi_0 = \mathbb{E}_0[D_i] = \mathbb{E}_0[\pi_0(X)]$. In this case, the estimated Riesz representer is

$$\widehat{\gamma}^{ATT}(D,X) = \frac{D}{\widehat{\pi}} - \frac{1-D}{\widehat{\pi}}\frac{\widehat{\pi}(X)}{1-\widehat{\pi}(X)},$$

where $\widehat{\pi} = \frac{1}{n}\sum_{i=1}^n D_i$. We only need to obtain the pilot estimator and the posterior draws for the conditional mean in the control arm.

**Algorithm 3** Posterior Computation of ATT

**for** $s = 1, \ldots, S$ **do**

(a) Generate the $s$-th draw using the BART prior and the data from the control arm; denote it as $(m_\eta^s(0, X_i))_{i=1}^n$.

(b) Draw Bayesian bootstrap weights $W_{ni}^s = e_i^s / \sum_{j=1}^n e_j^s$ where $e_i^s \overset{iid}{\sim} \text{Exp}(1)$, $1 \leqslant i \leqslant n$.

(c) Calculate the recentered posterior draw for the ATT $\check{\theta}_\eta^s := \theta_\eta^s - \widehat{b}_\eta^s$ where

$$\theta_\eta^s = \sum_{i=1}^n W_{ni}^s \widehat{\gamma}^{ATT}(D_i, X_i)\big(Y_i - m_\eta^s(0, X_i)\big).$$

and

$$\widehat{b}_\eta^s = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}[m_\eta^s(0,\cdot) - \widehat{m}(0,\cdot)](Z_i), \tag{C.2}$$

where $\boldsymbol{\tau}[m](z) := m(0,x) + \widehat{\gamma}^{ATT}(d,x)(y - m(0,x))$.

**end for**

**Output:** $\{\check{\theta}_\eta^s : s = 1, \ldots, S\}$.

# D Implementation of BART

We implement BART using the R package BART [36]. Specifically, the posterior of $m_\eta(X_i)$ is obtained by applying the function gbart with the argument type = wbart for continuous outcomes. We generate 2,000 posterior draws after discarding a burn-in of 500. The number of trees is set to $T = 200$. All BART priors follow the default choices recommended in the package. We briefly summarize these priors below and refer readers to Appendix B of [36] and Section 2 of [13] for further details. While some priors specified below differs from

those in Assumption 6, our simulation results are consistent with Theorem 4.1.

Consider the model $Y \sim N(m_\eta(x), \sigma^2)$ given $X = x$. The conditional mean $m_\eta(x)$ is approximated by an additive tree learner $\sum_{t=1}^{T} \sum_{k=1}^{K^t} \beta_k^t \mathbb{1}\{x \in \Omega_k^t\}$, where $(\Omega_1^t, \ldots, \Omega_{K^t}^t) := \mathcal{T}^t$ denotes the structure (partition) of the $t$-th tree of size $K^t$, and $(\beta_1^t, \ldots, \beta_{K^t}^t) := \boldsymbol{\beta}^t \in \mathbb{R}^{K^t}$ denotes the step heights (leaf values) at $K^t$ terminal nodes of the $t$-th tree. The structure of each binary tree $\mathcal{T}^t$ consists of three components: (i) the number of internal nodes, (ii) the splitting variable at each internal node, and (iii) the splitting value for that variable. The priors are as follows: (i) the probability that a node at depth $l$ is internal is $0.95(1 + l)^{-2}$; (ii) the splitting variable is chosen from $p$ covariates with the probability vector $(s_1, \ldots, s_p) \sim \mathrm{Dir}(\theta/p, \ldots, \theta/p)$, where the hyperparameter $\theta$ is induced from $\theta/(\theta + p) \sim \mathrm{Beta}(0.5, 1)$; and (iii) the splitting value is chosen uniformly among the observed values of the splitting variable. Conditional on $\mathcal{T}^t$, each step-height $\beta_k^t$ follows a normal prior with mean $\sum_{i=1}^{n} Y_i/n$ and standard deviation $(\max_i Y_i - \min_i Y_i)/(4\sqrt{T})$. The noise standard deviation $\sigma$ is endowed with inverse $\chi^2$ prior with $\nu = 3$ degrees of freedom and scale parameter $\lambda$ chosen so that the 0.9 quantile of the prior equals $\hat{\sigma}$, the residual standard deviation from a linear regression of $Y$ on $X$.

Posterior for the additive tree is drawn via a Gibbs sampler. Because the priors on the step-heights $\boldsymbol{\beta}^t$ and $\sigma$ are conjugate, conditional on the tree structure $\mathcal{T}^t$, these parameters can be drawn directly: $\beta_k^t$ from a normal posterior and $\sigma$ from an inverse $\chi^2$ distribution. Conversely, conditional on $\boldsymbol{\beta}^t$ and $\sigma$, the tree structure $\mathcal{T}^t$ is updated via a Metropolis–Hastings step detailed in Appendix C of [36].

To implement the one-step corrected posterior method of [40] (One-step BART), we need posterior draws of the propensity score function $\pi_\eta(x)$. For this, we use BART with a binary outcome and logistic link, implemented by calling gbart with type = lbart. The algorithm employs the data-augmentation technique of [1], introducing a latent variable $Z_i$ with a truncated normal distribution given $(Y_i, X_i)$ and the additive tree model $(\mathcal{T}^t, \boldsymbol{\beta}^t)_{t=1}^T$:

$$Z_i \mid Y_i, X_i, (\mathcal{T}^t, \boldsymbol{\beta}^t)_{t=1}^T \sim \mathbb{TN}_{A_i}\left(\sum_{t=1}^{T} \sum_{k=1}^{K^t} \beta_k^t \mathbb{1}\{x \in \Omega_k^t\}, 4V_i^2\right),$$

where $A_i = (0, \infty)$ (positive real line) if $Y_i = 1$ and $(-\infty, 0)$ (negative real line) if $Y_i = 0$, and $\mathbb{TN}_A$ denotes a normal distribution truncated to $A$. The auxiliary variable $V_i$ is drawn from a Kolmogorov distribution.[7] The latent $Z_i$ are then treated as the continuous outcome for BART. Section 4 of [36] gives more details on BART for binary outcomes.

[7] [2] show that logistic random variables can be generated as a mixture of normal and Kolmogorov components.

# References


[1] 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.

[2] Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B*, 36(1):99–102.

[3] Breunig, C., Liu, R., and Yu, Z. (2024). Semiparametric bayesian difference-in-differences. *arXiv preprint arXiv:2412.04605*.

[4] Breunig, C., Liu, R., and Yu, Z. (2025a). Double robust bayesian inference on average treatment effects. *Econometrica*, 93(2):539–568.

[5] Breunig, C., Liu, R., and Yu, Z. (2025b). Supplement to 'double robust bayesian inference on average treatment effects'. *Econometrica Supplemental Material*, 93.

[6] Castillo, I. and Rockova, V. (2021). Uncertainty quantification for bayesian cart. *Annals of Statistics*, 49:3482–3509.

[7] Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):673–700.

[8] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265.

[9] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

[10] Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica*, 90:1501–1535.

[11] Chernozhukov, V., Newey, W., and Singh, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

[12] Chernozhukov, V., Newey, W., and Singh, R. (2022c). De-biased machine learning of global and local parameters using regularized riesz representers. *Econometrics Journal*, 25:576–601.

[13] 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.

[14] Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 2019:43–68.

[15] Dudley, R. M. (2002). *Real Analysis and Probability*, volume 74. Cambridge University Press.

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

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

[18] Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

[19] Hahn, P., Murray, J., and Carvalho, C. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113:626–636.

[20] Hahn, P., Murray, J., and Carvalho, C. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15:965–1056.

[21] Han, Q. (2021). Set structured global empirical risk minimizers are rate optimal in general dimensions. *The Annals of Statistics*, 49(5):2642–2671.

[22] Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605–654.

[23] Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227.

[24] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

[25] Jeong, S. and Rocková, V. (2023). The art of bart: Minimax optimality over nonhomogeneous smoothness in high dimension. *Journal of Machine Learning Research*, 24(337):1–65.

[26] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

[27] 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)*, 80:1087–1110.

[28] Ray, K. and van der Vaart, A. (2020). Semiparametric bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020.

[29] Ray, K. and van der Vaart, A. (2021). On the bernstein-von mises theorem for the dirichlet process. *Electronic Journal of Statistics*, 15(1):2224–2246.

[30] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89:846–866.

[31] Rocková, V. (2020). On semi-parametric inference for bart. *International Conference on Machine Learning*, pages 8137–8146.

[32] Rocková, V. and Saha, E. (2020). On theory for bart. *In The 22nd international conference on artificial intelligence and statistics*, pages 2839–2848.

[33] Rocková, V. and van der Pas, S. (2020). Posterior concentration for bayesian regression trees and forests. *The Annals of Statistics*, 48:2108–2131.

[34] Rubin, D. (1981). Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

[35] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14:1139–1151.

[36] Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: The bart r package. *Journal of Statistical Software*, 97:1–66.

[37] van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

[38] van der Vaart, A. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

[39] Xie, F. and Xu, Y. (2020). Adaptive bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Analysis*, 15:159–186.

[40] Yiu, A., Fong, E., Holmes, C., and Rousseau, J. (2025). Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1025–1054.