

# Robust Semiparametric Inference for Bayesian Additive Regression Trees

Christoph Breunig\*      Ruixuan Liu<sup>†</sup>      Zhengfei Yu<sup>‡</sup>

September 21, 2026

## Abstract

Bayesian additive regression trees (BART) provide a class of flexible non-parametric regressions in machine learning. Their rich tree-based structure can generate non-Donsker classes of regression functions. We study posterior inference for the population mean under missing at random. We derive a new Bernstein–von Mises (BvM) theorem showing that even a one-step posterior based on BART and Bayesian-bootstrap reweighting retains a non-negligible bias term in the non-Donsker regime. To remove this term, we introduce RoBART, a posterior correction that adjusts each draw using pilot estimators of the outcome regression and propensity score. We prove a BvM theorem for the corrected posterior and develop a cross-fitted version using fold-specific posteriors. We provide primitive assumptions tailored to BART. We show that the posterior mean coincides exactly with the conventional augmented inverse-probability-weighted (AIPW) estimator and, under cross-fitting, with the corresponding double machine learning (DML) estimator. RoBART therefore provides a posterior distribution with asymptotically valid uncertainty quantification around this well-established frequentist center. Simulations and an empirical illustration demonstrate competitive finite-sample performance relative to existing methods.

*Keywords:* Causal inference, posterior correction, nonparametric Bayesian inference, Bernstein–von Mises theorem, BART, fold-specific posterior.

---

\*University of Bonn. Email: [cbreunig@uni-bonn.de](mailto:cbreunig@uni-bonn.de)

<sup>†</sup>Chinese University of Hong Kong. Email: [ruixuanliu@cuhk.edu.hk](mailto:ruixuanliu@cuhk.edu.hk)

<sup>‡</sup>University of Osaka. Email: [yu.zhengfei@osaka-u.ac.jp](mailto:yu.zhengfei@osaka-u.ac.jp)

# 1 Introduction

Bayesian additive regression trees (BART) is one of the most powerful statistical methods (Chipman et al., 2010). Its empirical success in predictive modeling has been demonstrated in numerous studies. In the context of nonparametric function estimation, BART achieves adaptive, near rate-minimax posterior contraction, as shown by Linero and Yang (2018); Ročková and van der Pas (2020); Ročková (2020); Jeong and Ročková (2023). Beyond prediction, BART has also shown strong performance in semiparametric inference for causal parameters of interest (Dorie et al., 2019; Hahn et al., 2020). However, a theoretical justification for semiparametric inference based on BART under reasonable regularity conditions is still developing.

The difficulty arises partly from the structure of tree-based methods. BART represents regression functions through ensembles of piecewise-constant trees and is therefore intrinsically nonsmooth. For Hölder-type function classes, standard Donsker conditions typically require smoothness exceeding one-half of the covariate dimension. Such conditions become increasingly restrictive in moderately high-dimensional settings and are particularly difficult to reconcile with piecewise-constant tree priors. At the same time, flexible nuisance estimation is essential in missing-data and causal-inference problems (Chernozhukov et al., 2017, 2018; Athey and Wager, 2021), where misspecification of the outcome regression may lead to invalid inference. Our objective is to establish valid posterior inference for the finite-dimensional parameter of interest when the nuisance regression is modeled by BART in a non-Donsker regime where existing theory does not apply. While augmented inverse-probability-weighted (AIPW) estimation, combined with cross-fitting and flexible nuisance estimation as in double machine learning (DML), already provides robust frequentist point estimation in this setting, our focus is on valid posterior uncertainty quantification around the same center.

We consider estimation of the population mean under missing at random and subsequently extend the construction to average treatment effects (ATE). Starting from the AIPW representation, we first consider a baseline one-step posterior that combines a BART posterior for the conditional mean, a pilot estimator of the propensity score, and Bayesian-bootstrap weights. This construction is a specialization of the one-step posterior of Yiu et al. (2025), obtained by replacing their posterior for the propensity score with a point mass at a pilot estimate. We use this one-step posterior as a benchmark for studying posterior inference with BART. We derive a semiparametric Bernstein–von Mises theorem for the resulting one-step posterior. The Bernstein–von Mises limit of the one-step BART posterior is generally shifted by an additional first-order bias term that need not vanish outside the Donsker regime. The same bias term arises in Breunig et al. (2025a) from a different posterior construction based on prior adjustment. Thus, two distinct posterior constructions encounter the same first-order obstruction. We characterize this obstruction for the one-step BART posterior and remove it by a draw-specific correction.

To remove this term, we introduce robust BART, or RoBART, which applies an ex-

explicit correction to each posterior draw using pilot estimators of the conditional mean and propensity score. The correction is analytically simple and preserves the posterior variation generated by the BART posterior and the Bayesian bootstrap. We establish a Bernstein–von Mises theorem for the corrected posterior, yielding  $\sqrt{n}$ -rate inference and the semiparametric efficiency bound under product-rate conditions that permit weaker regularity of the conditional mean function to be offset by stronger regularity of the propensity score, and vice versa.

An important justification for the correction comes from its posterior mean. Conditional on the observed data, the posterior mean of the corrected procedure coincides exactly with the AIPW estimator based on the pilot nuisance estimates. We also develop a cross-fitted version in which BART posteriors are constructed separately within each evaluation fold and the corresponding pilot estimators are trained on the complementary observations. The average of the fold-specific posterior means then coincides exactly with the cross-fitted AIPW estimator (Robins et al., 1994), which, when combined with machine-learning nuisance estimators and cross-fitting, is the DML estimator (Chernozhukov et al., 2018). Thus, RoBART does not introduce a different point estimator from DML. Rather, it provides a corrected posterior distribution for uncertainty quantification around the same semiparametrically robust center.

Our theoretical results cover both the construction of pilot estimators using an auxiliary sample and a cross-fitted procedure that uses the full sample. In contrast to standard empirical-process arguments, the analysis does not require the nuisance-function class to be Donsker. This feature is particularly relevant for BART because its piecewise-constant structure limits the smoothness that can be exploited directly. The conditions instead allow the regularity of the outcome regression and propensity score to compensate for one another through the familiar product-type rate restriction.

Our RoBART procedure is best viewed as a calibrated Bayesian construction in the sense of Rubin (1984). It begins with a prior and a posterior for the nuisance function, but augments the resulting posterior draws with a correction designed to deliver valid frequentist inference. This perspective is also related to the notion of pragmatic Bayesian procedures discussed by Ritov et al. (2014). The resulting credible intervals retain a posterior interpretation conditional on the observed data, while the Bernstein–von Mises theorem establishes their asymptotic frequentist validity.

Admittedly, the use of pilot estimation and cross-fitting deviates from the Bayesian principle in the strict sense. The construction is closely related to the active DML literature (Chernozhukov et al., 2017, 2018). DML combines debiased moment conditions with cross-fitting to permit flexible nuisance estimation without relying on restrictive Donsker conditions. RoBART shares this robustness principle but uses the posterior distribution generated by a nonparametric Bayesian learner to conduct inference. The exact equivalence of the posterior mean and the corresponding DML estimator clarifies the relationship between the two approaches: they share the same point-estimation center, whereas RoBART uses the corrected BART posterior and Bayesian-bootstrap variation to quantify

uncertainty around that center.

Monte Carlo experiments illustrate the more robust performance of RoBART relative to other Bayesian and frequentist methods. In increasingly complex designs, standard BART and the one-step BART can exhibit substantial undercoverage, whereas RoBART maintains coverage closer to the nominal level. Because the posterior mean of RoBART coincides with the corresponding DML estimator, the comparison between these two focuses on uncertainty quantification. Using the same folds and nuisance estimates, RoBART yields better coverage performance than DML, especially for more complex designs. We also apply RoBART in an empirical illustration.

*Related literature.* Ray and van der Vaart (2020) use adjusted priors to weaken regularity requirements in semiparametric Bayesian inference. Breunig et al. (2025a) combine a related data-dependent prior adjustment with an explicit posterior correction and obtain doubly robust Bayesian inference for the ATE. The present paper addresses a different problem: inference when the outcome regression is given by a BART posterior whose function class need not satisfy the Donsker conditions used in existing semiparametric Bayesian asymptotic theory. In addition, the prior adjustment is difficult to incorporate into BART-based modeling of the conditional outcome. Starting from the one-step posterior of Yiu et al. (2025), we characterize the non-negligible bias that appears in this regime. We show that it coincides with the bias generated by the distinct prior adjustment construction of Breunig et al. (2025a), and develop a draw-wise correction device to remove it. For the cross-fitted procedure, we establish a strong equivalence result that shows the posterior mean after our correction coincides exactly with the corresponding AIPW/DML estimator, which is new to the literature. The technical contribution is a BvM theorem that justifies the uncertainty quantification from the corrected posterior distribution under primitive conditions tailored to BART.

As a general nonparametric regression method, BART achieves adaptive, near rate-minimax posterior contraction, a major breakthrough due to Linero and Yang (2018); Ročková and van der Pas (2020); Ročková (2020); Jeong and Ročková (2023). Compared with the extensive results on the posterior convergence for BART, the corresponding inferential theory is relatively scarce. Ročková (2020) establish a semiparametric BvM theorem using BART-type priors for a particular type of linear functional of the regression function under the fixed design. That result requires the Donsker property of the conditional mean. For Bayesian CART (a single tree rather than a forest), Castillo and Ročková (2021) also establish a general semiparametric BvM theorem within the Donsker regime.

*Organization.* 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. We highlight the strong equivalence between our point estimator and the AIPW (or DML) estimator. In Section 3, we derive the BvM theorem under high-level assumptions using pilot estimators from auxiliary data. Section 4 establishes corresponding results for the cross-fitted version under more primitive conditions 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, and the lemmas and auxiliary results used therein are collected in Appendix B. Appendix C extends the framework to inference on average treatment effects. Appendix D contains the posterior contraction and empirical process results for regression trees that are used in the proof of Theorem 8, and Appendix E provides implementation details of BART.

*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  $P[h] = \int h(z) dP(z)$ ,  $\mathbb{P}_n[h] = n^{-1} \sum_{i=1}^n h(Z_i)$ , and  $\mathbb{G}_n[h] = \sqrt{n}(\mathbb{P}_n - P)[h]$ . We further write  $\mathbb{P}_n^*[h] = \sum_{i=1}^n W_{ni}h(Z_i)$  for the bootstrap empirical measure induced by the Bayesian-bootstrap weights  $W^{(n)} = (W_{n1}, \dots, W_{nn})^\top$ . For two sequences  $\{a_n\}$  and  $\{b_n\}$  of positive numbers, we write  $a_n \lesssim b_n$  if  $\limsup_{n \rightarrow \infty} (a_n/b_n) < \infty$ , 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 connections to and differences from 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  $\{(R_i Y_i, R_i, X_i^\top)^\top\}_{i=1}^n$  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 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

$$m_\eta(x) = \int y f_\eta(y \mid x) dy.$$

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

$$p_\eta(z) = \frac{d\mu_\eta(x)}{dx} \pi_\eta^r(x) (1 - \pi_\eta(x))^{1-r} f_\eta^r(y \mid x). \quad (2.1)$$

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

$$\chi_0 = \int m_0(x) d\mu_0(x),$$

where  $\mu_0$  denotes the marginal distribution of  $X_i$  under  $P_0$ . For our approach, it is useful to consider the following augmented representation of  $\chi_0$ :

$$\chi_0 = \int (m_0(x) + \gamma_0(r, x)(y - m_0(x))) dP_0(y, r, x), \quad (2.2)$$

where the Riesz representer is given by

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

with  $\pi_0(x) := \mathbb{P}_0(R_i = 1 \mid X_i = x)$ . In particular,  $\mathbb{E}_0[m_0(X_i)\gamma_0(R_i, X_i)] = \chi_0$ ; see Chernozhukov et al. (2018).

Let  $W^{(n)} := (W_{n1}, \dots, W_{nn})$  denote the Bayesian-bootstrap weights of Rubin (1981), defined by  $W_{ni} = e_i / \sum_{j=1}^n e_j$  for  $e_i \stackrel{iid}{\sim} \text{Exp}(1)$ ,  $i = 1, \dots, n$ . Motivated by the augmented representation in (2.2), we define the Bayesian-bootstrap corrected draw

$$\chi_{\eta, \hat{\gamma}} := \sum_{i=1}^n W_{ni} (m_{\eta}(X_i) + \hat{\gamma}(R_i, X_i)(Y_i - m_{\eta}(X_i))), \quad (2.4)$$

where  $m_{\eta}(\cdot)$  is a stochastic function endowed with a prior and

$$\hat{\gamma}(r, x) := \frac{r}{\hat{\pi}(x)}$$

is the estimated Riesz representer based on a pilot estimator  $\hat{\pi}$  of  $\pi_0$  computed using an auxiliary data set. The conditional distribution of  $\chi_{\eta, \hat{\gamma}}$  given the observed data  $Z^{(n)} := (Z_1, \dots, Z_n)$  is jointly induced by the posterior law of  $m_{\eta}$ , denoted by  $\Pi_m(\cdot \mid Z^{(n)})$ , and the distribution of the Bayesian-bootstrap weights  $W^{(n)}$ . The latter distribution is known and straightforward to simulate. Following the notation for bootstrap laws used by van der Vaart and Wellner (1996), we denote it by  $\Pi_W(\cdot \mid Z^{(n)})$ . We refer to the resulting conditional law of  $\chi_{\eta, \hat{\gamma}}$  as the baseline one-step posterior for  $\chi_0$ . While our general results apply to a broad class of nonparametric priors for  $m_{\eta}$ , we particularly analyze BART priors under more primitive conditions.<sup>1</sup>

---

<sup>1</sup>Posterior draws of the vector  $(m_{\eta}(X_1), \dots, m_{\eta}(X_n))$  can be obtained using the R package **BART**; see Sparapani et al. (2021).

To motivate this construction from a more conventional Bayesian perspective, suppose that  $\eta$  indexes a probability model  $P_\eta$  with corresponding conditional mean function  $m_\eta$  and propensity score  $\pi_\eta$ . The associated model-based functional is

$$\chi_\eta = \int \left( m_\eta(x) + \frac{r}{\pi_\eta(x)} (y - m_\eta(x)) \right) dP_\eta(y, r, x). \quad (2.5)$$

The conditional distribution of (2.4) can be viewed as a modular approximation to the posterior distribution of (2.5). In particular, replacing  $\pi_\eta$  by the pilot estimator  $\hat{\pi}$  amounts to placing a degenerate distribution at  $\hat{\pi}$ , while uncertainty about  $m_\eta$  is retained through its nonparametric posterior. Related uses of preliminary propensity-score estimators in semiparametric Bayesian inference are considered by Ray and van der Vaart (2020). Alternatively, one could place an independent nondegenerate prior on the propensity score and propagate the resulting posterior uncertainty through the corrected posterior.

For the integrating probability measure, we use the Bayesian bootstrap. More precisely, the Bayesian bootstrap can be viewed as the noninformative limiting posterior arising from a sequence of Dirichlet process models as the prior concentration parameter tends to zero. Equation (2.4) can be viewed as a specialization of the one-step posterior of Yiu et al. (2025): uncertainty about  $m_\eta$  is retained through its nonparametric posterior, the posterior for the propensity score is replaced by a point mass at a pilot estimator, and the integrating distribution is represented by the Bayesian bootstrap. We use this construction as the baseline one-step posterior. Section 2.2 shows why, when the conditional mean is modeled by BART, this baseline generally requires an additional draw-wise correction outside the Donsker regime.

## 2.2 Motivation of Robust Bayesian Procedure

Our construction is guided by the efficient influence function for the mean response under missing at random. For a generic nuisance parameter  $\eta = (m_\eta, \pi_\eta)$ , consider the influence-function map

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

where the corresponding Riesz representer is

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

Evaluated at the true nuisance parameter  $\eta_0 = (m_0, \pi_0)$ , (2.6) is the efficient influence function for the target parameter. Let  $\hat{\pi}$  denote a pilot estimator of the propensity score based on the auxiliary sample, and define  $\hat{\gamma}(r, x) = r/\hat{\pi}(x)$ . The posterior correction also uses a pilot estimator  $\hat{m}$  of the conditional mean function  $m_0$ . In our implementation,  $\hat{m}$  is taken to be the posterior mean of a BART model fitted to the auxiliary sample.

Estimating the nuisance functions on independent auxiliary data simplifies the analysis by rendering certain empirical-process remainder terms asymptotically negligible. Sample splitting has a long history in semiparametric estimation (Schick, 1986; van der Vaart, 1998) and is central to the recent DML literature (Chernozhukov et al., 2018). Related auxiliary-sample constructions have also been used for propensity-score-adjusted Bayesian priors in missing-data problems (Ray and van der Vaart, 2020). The auxiliary-sample formulation allows us to present the main argument transparently, while the cross-fitted extension avoids the need for a separate sample.

We show that the one-step Bayesian procedure  $\chi_{\eta, \hat{\gamma}}$  defined in (2.4), which uses a pilot estimator of the propensity score, generally admits a Bernstein–von Mises approximation (see Theorem 2) only after accounting for the additional centering bias

$$\begin{aligned} b_{0, \eta} &= \mathbb{P}_n [m_\eta(X) + \gamma_0(R, X)(Y - m_\eta(X))] - \mathbb{P}_n [m_0(X) + \gamma_0(R, X)(Y - m_0(X))] \\ &= \mathbb{P}_n [(1 - \gamma_0(R, X))(m_\eta(X) - m_0(X))]. \end{aligned}$$

This term need not be asymptotically negligible when the posterior for the outcome regression concentrates too slowly for standard Donsker arguments to apply.

The purpose of the posterior correction is to remove the draw-specific centering bias induced by nonparametric Bayesian modeling of the outcome regression, while preserving the posterior variation generated by the BART posterior and the Bayesian-bootstrap weights. The correction also has a useful point-estimation interpretation: it recenters the posterior at the conventional doubly robust estimator based on the pilot nuisance estimates.

Our proposal considers an estimator of the bias term  $b_{0, \eta}$  given by

$$\hat{b}_\eta = \mathbb{P}_n [(1 - \hat{\gamma}(R, X))(m_\eta(X) - \hat{m}(X))],$$

where  $\hat{m}$  and  $\hat{\gamma}$  are pilot estimators computed from an auxiliary sample. The resulting corrected posterior draw is  $\check{\chi}_\eta = \chi_{\eta, \hat{\gamma}} - \hat{b}_\eta$ . An important feature of this correction is that the conditional posterior mean of  $\check{\chi}_\eta$  coincides exactly with the conventional AIPW estimator based on the pilot estimators.

To see this, let  $\mathbb{E}_\Pi$  denote expectation with respect to the joint posterior distribution of  $m_\eta$  and the Bayesian-bootstrap weights, conditional on  $Z^{(n)}$ , where the pilot estimators  $\hat{m}$  and  $\hat{\gamma}$  are treated as fixed. Since the Bayesian-bootstrap weights are independent of  $m_\eta$  and satisfy  $\mathbb{E}_\Pi[W_{ni} | Z^{(n)}] = 1/n$ , we obtain

$$\mathbb{E}_\Pi [\chi_{\eta, \hat{\gamma}} | Z^{(n)}] = \mathbb{P}_n [\mathbb{E}_\Pi [m_\eta(X) | Z^{(n)}] + \hat{\gamma}(R, X) (Y - \mathbb{E}_\Pi [m_\eta(X) | Z^{(n)}])].$$

Moreover, since  $\hat{b}_\eta$  is affine in  $m_\eta$ , we have

$$\begin{aligned} \mathbb{E}_\Pi [\hat{b}_\eta | Z^{(n)}] &= \mathbb{P}_n [\mathbb{E}_\Pi [m_\eta(X) | Z^{(n)}] + \hat{\gamma}(R, X) (Y - \mathbb{E}_\Pi [m_\eta(X) | Z^{(n)}])] \\ &\quad - \mathbb{P}_n [\hat{m}(X) + \hat{\gamma}(R, X)(Y - \hat{m}(X))]. \end{aligned}$$

It follows that the posterior mean of the corrected draw  $\check{\chi}_\eta = \chi_{\eta, \hat{\gamma}} - \hat{b}_\eta$  is

$$\mathbb{E}_\Pi [\check{\chi}_\eta \mid Z^{(n)}] = \mathbb{P}_n [\hat{m}(X) + \hat{\gamma}(R, X)(Y - \hat{m}(X))]. \quad (2.8)$$

Thus, the conditional posterior mean of the corrected Bayesian procedure coincides exactly with the conventional AIPW estimator based on  $(\hat{m}, \hat{\pi})$ . The BART posterior draws and Bayesian-bootstrap weights determine the posterior distribution around this center, but not the center itself. In the cross-fitted construction, averaging the fold-specific posterior means analogously yields the conventional cross-fitted AIPW, or DML, estimator.

## 2.3 Robust Bayesian Procedure with Auxiliary Data

We first present the implementation of our robust inference procedure under the assumption that an auxiliary sample is available for computing the pilot estimators. This auxiliary-sample formulation provides a convenient starting point for presenting the main idea. In the next subsection, we extend the approach to a  $K$ -fold cross-fitted version, which makes efficient use of the full data.

The inputs to the algorithm are the observable data  $\{(R_i Y_i, R_i, X_i^\top)^\top\}_{i=1}^n$ , the pilot estimators  $\hat{\gamma}$  and  $\hat{m}$  as discussed above, and the number of posterior draws  $S$ . The posterior distribution of  $(m_\eta(X_i))_{i=1}^n := (m_\eta(X_1), \dots, m_\eta(X_n))$  is generated by applying BART to estimate  $m_\eta(\cdot)$  using the data  $\{(Y_i, X_i^\top) : R_i = 1\}_{i=1}^n$ .

---

### Algorithm 1: Posterior Computation

---

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

- (a) Obtain one draw from the posterior of  $(m_\eta(X_i))_{i=1}^n$ , and denote it by  $(m_\eta^s(X_i))_{i=1}^n$ .
- (b) Generate  $W_{ni}^s = e_i^s / \sum_{j=1}^n e_j^s$  with  $e_i^s \stackrel{iid}{\sim} \text{Exp}(1)$  for all  $1 \leq i \leq n$ .
- (c) Calculate  $\check{\chi}_\eta^s = \chi_{\eta, \hat{\gamma}}^s - \hat{b}_\eta^s$ , where

$$\chi_{\eta, \hat{\gamma}}^s = \sum_{i=1}^n W_{ni}^s (m_\eta^s(X_i) + \hat{\gamma}(R_i, X_i)(Y_i - m_\eta^s(X_i))), \quad (2.9)$$

and the debiasing term is

$$\hat{b}_\eta^s = \frac{1}{n} \sum_{i=1}^n (\hat{\gamma}(R_i, X_i) - 1) (\hat{m}(X_i) - m_\eta^s(X_i)). \quad (2.10)$$

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

---

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

$$\mathcal{C}_n(\alpha) = \{\chi : q(\alpha/2) \leq \chi \leq q(1 - \alpha/2)\}, \quad (2.11)$$

where  $q(a)$  denotes the  $a$ -quantile of  $\{\check{\chi}_\eta^s : s = 1, \dots, S\}$ . We take the Bayesian point

estimator (the posterior mean) by averaging the simulation draws:  $\bar{\chi}_\eta = S^{-1} \sum_{s=1}^S \tilde{\chi}_\eta^s$ .

**Remark 1** (Estimation of the Riesz Representer). *Our proposal builds on a plug-in approach that employs pilot estimators of the Riesz representer. Our approach is motivated by the following considerations. First, by adopting sample splitting for the pilot estimators, our proposal makes it possible to relax the Donsker condition. Sample splitting has long been used in semiparametric models (Schick, 1986; van der Vaart, 1998) and has regained popularity in the DML literature (Chernozhukov et al., 2017, 2018). 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 Chernozhukov et al. (2022a). 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 prior for the propensity score (Ray and van der Vaart, 2020).*

## 2.4 Cross-fitted Robust Bayesian Inference Procedure

This subsection describes the implementation of RoBART using  $K$ -fold cross-fitting, inspired by the double machine learning literature (Chernozhukov et al., 2018), in particular by the fold-specific bootstrap method in Bach et al. (2024). In the context of difference-in-differences designs, Breunig et al. (2024) develop a similar cross-fitted approach involving both prior and posterior adjustments.

Algorithm 2 is the  $K$ -fold cross-fitting version of Algorithm 1. We randomly partition the observation indices  $\{1, \dots, n\}$  into  $K$  folds  $(I_k)_{k=1}^K$ , each of size  $n/K$ , assuming for notational simplicity that  $n$  is divisible by  $K$ . For each fold  $k$ , we make draws of the fold-specific posterior for the conditional mean function and Bayesian-bootstrap weights. The fold-specific posterior of the conditional mean is obtained via Bayes' rule given the BART prior and fold-specific likelihood function. This differs from cross-fitting in the frequentist double machine learning literature, where the conditional mean function is estimated using observations outside the evaluation fold and evaluated on observations within that fold. We make independent draws over different folds. Let  $\hat{\pi}^{(-k)}(\cdot)$  and  $\hat{m}^{(-k)}(\cdot)$  denote estimators of the propensity score  $\pi_0(\cdot)$  and the conditional mean function  $m_0(\cdot)$ ,

respectively, constructed using all observations not belonging to fold  $k$ .

---

**Algorithm 2:** RoBART for  $\mathbb{E}_0[Y_i]$  with  $K$ -fold cross-fitting

---

(a) Take a  $K$ -fold random partition  $(I_k)_{k=1}^K$  of the observation indices. The size of each fold  $I_k$  is  $n/K$ .

**for**  $k = 1, \dots, K$  **do**

(b) Use the subsample  $\{(R_i, X_i^\top)\}_{i \notin I_k}$  to construct the estimator  $\hat{\pi}^{(-k)}(x)$ . Set  $\hat{\gamma}^{(-k)}(r, x) = r / \hat{\pi}^{(-k)}(x)$ .

(c) Use the subsample  $\{(Y_i, X_i^\top) : R_i = 1\}_{i \notin I_k}$  to construct the estimator  $\hat{m}^{(-k)}(x)$ .

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

(d) Make a draw from the posterior of  $(m_\eta(X_i))_{i \in I_k}$  using the subsample  $\{(Y_i, X_i^\top) : R_i = 1\}_{i \in I_k}$ , and denote it by  $(m_\eta^{k,s}(X_i))_{i \in I_k}$ .

(e) Generate  $W_i^{k,s} = e_i^{k,s} / \sum_{j \in I_k} e_j^{k,s}$  with  $e_i^{k,s} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$  for all  $i \in I_k$ .

(f) Calculate  $\check{\chi}_\eta^{k,s} = \chi_{\eta, \hat{\gamma}}^{k,s} - \hat{b}_\eta^{k,s}$ , where

$$\chi_{\eta, \hat{\gamma}}^{k,s} = \sum_{i \in I_k} W_i^{k,s} (m_\eta^{k,s}(X_i) + \hat{\gamma}^{(-k)}(R_i, X_i)(Y_i - m_\eta^{k,s}(X_i))), \quad (2.12)$$

and the debiasing term is

$$\hat{b}_\eta^{k,s} = \frac{K}{n} \sum_{i \in I_k} (\hat{\gamma}^{(-k)}(R_i, X_i) - 1) (\hat{m}^{(-k)}(X_i) - m_\eta^{k,s}(X_i)). \quad (2.13)$$

**Output:**  $\{\check{\chi}_\eta^s = \frac{1}{K} \sum_{k=1}^K \check{\chi}_\eta^{k,s} : s = 1, \dots, S\}$

---

### 3 BvM Theorems under High-level Assumptions

Beyond the equivalence of point estimation, we aim to develop valid inference via showing new Bernstein–von Mises (BvM) theorems. Although our primary interest lies in establishing the BvM theorem for BART, we first develop the theory under high-level conditions before specializing the analysis. The robustness achieved through our additional debiasing step is therefore applicable to other types of priors beyond the BART framework. We begin with a version that relies on pilot estimators computed from an external sample in order to illustrate the main idea. This reduces the notational burden and highlights the core issue. In the next section, we prove the BvM theorem for the cross-fitted version. The proof essentially reduces to establishing the BvM theorem for the fold-specific posterior, which follows from the results developed in this section.

The one-step posterior of  $\chi_{\eta, \hat{\gamma}}$  is determined jointly by the posterior of the conditional mean and the Bayesian-bootstrap law. Recall that the conditional density associated with  $m_\eta$  is denoted by  $f_\eta$ . 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 Breunig et al. (2025a). Accordingly, we denote it by  $f_{m_\eta}(\cdot | x)$  hereafter. The posterior distribution is formally given

by

$$\Pi(m_\eta \in A, W^{(n)} \in B \mid Z^{(n)}) = \int_B \frac{\int_A \prod_{i=1}^n f_{m_\eta}^{R_i}(Y_i \mid X_i) d\Pi_m(m_\eta)}{\int \prod_{i=1}^n f_{m_\eta}^{R_i}(Y_i \mid X_i) d\Pi_m(m_\eta)} d\Pi_W(W^{(n)} \mid Z^{(n)}).$$

Regarding the centering point in the asymptotic normal approximation, we can consider any asymptotically efficient estimator  $\hat{\chi}$  with the following linear representation:

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

where  $\tilde{\chi}_0 = \tilde{\chi}_{\eta_0}$  is the efficient influence function given in (2.6) under  $\eta = \eta_0$ . Denote its variance by  $v_0 = \mathbb{E}_0[\tilde{\chi}_0^2(Z)]$ . We write  $\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta - \hat{\chi}) \mid Z^{(n)})$  for the posterior distribution of  $\sqrt{n}(\chi_\eta - \hat{\chi})$ .

Below, we consider some measurable sets  $\mathcal{H}_n^m$  of functions  $m_\eta$  such that  $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \rightarrow_{P_0} 1$ . With a slight abuse of notation, 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  $\|\phi\|_{2,\mu_0} := \sqrt{\int \phi^2(x) d\mu_0(x)}$  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  $\{(R_i Y_i, R_i, X_i^\top)^\top\}_{i=1}^n$ , 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  $\inf_{x \in \mathcal{X}} \pi_0(x) \geq c$  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  $m_0(x)$ .

**Assumption 2** (Rates of Convergence). The pilot estimators  $\hat{\pi}$  and  $\hat{m}$ , which are based on an auxiliary sample independent of  $Z^{(n)}$ , satisfy  $\|\hat{\pi} - \pi_0\|_{2,\mu_0} = O_{P_0}(r_n)$  and  $\|1/\hat{\pi}\|_\infty = O_{P_0}(1)$ . In addition,

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

for some measurable sets  $\mathcal{H}_n = \{\eta : m_\eta \in \mathcal{H}_n^m\}$  such that  $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \rightarrow_{P_0} 1$ , where  $\max\{\varepsilon_n, r_n\} \rightarrow 0$  and  $\sqrt{n} \varepsilon_n r_n \rightarrow 0$ .

Assumption 2 imposes sufficiently fast convergence rates of the estimators of the conditional mean function  $m_0$  and the propensity score  $\pi_0$ . In practice, one can explore the recent proposals from Chernozhukov et al. (2022b) and Hirshberg and Wager (2021). The posterior contraction rate for the conditional mean can be derived by verifying high-level assumptions of Ghosal et al. (2000).

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

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

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

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

(ii) Furthermore, we assume that

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

Assumption 3 (i) restricts the function class  $\mathcal{G}_n$  to form a  $P_0$ -Glivenko–Cantelli class; see Section 2.4 of van der Vaart and Wellner (1996). The moment conditions on the envelope functions follow Yiu et al. (2025) and 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  $\hat{\gamma}$  and  $m_\eta$ , which sets us apart from the existing literature. In contrast to this assumption, Ray and van der Vaart (2020) and Yiu et al. (2025) 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 Breunig et al. (2025a), this is weaker than directly imposing the Donsker property on  $(m_\eta - m_0)$ . For the Hölder smooth class, Breunig et al. (2025a) show that the low-order differentiability of the conditional mean can be compensated by exploiting 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 Sudakov’s inequality (van der Vaart and Wellner, 1996).

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*  $d_{BL}$  on probability distributions on  $\mathbb{R}$ ; see Chapter 11 of Dudley (2002).

**Theorem 2.** *Let Assumptions 1–3 hold. Then, for the baseline one-step posterior, we have*

$$d_{BL}(\mathcal{L}_\Pi(\sqrt{n}(\chi_{\eta, \hat{\gamma}} - \hat{\chi} - b_{0, \eta}) \mid Z^{(n)}), N(0, \mathbf{v}_0)) \rightarrow_{P_0} 0,$$

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

The two parts of Assumption 3 play distinct roles in the proof. Part (i) delivers the conditional convergence of the Laplace transform of the Bayesian-bootstrap process through the  $P_0$ -Glivenko–Cantelli property of  $\mathcal{G}_n$ , while part (ii) controls the remainder term  $\mathbb{P}_n[(\gamma_0 - \hat{\gamma})(m_\eta - m_0)]$  arising from the pilot estimation of the propensity score.

---

<sup>2</sup>Because a nonparametric prior is also assigned to the propensity score, Yiu et al. (2025) further require the propensity score function to belong to a Donsker class.

The previous result shows that, under our double-robust smoothness conditions, the baseline one-step posterior satisfies the BvM theorem only up to a bias term. Specifically, the bias term is given by

$$\begin{aligned} b_{0,\eta} &= \mathbb{P}_n[m_\eta(X) + \gamma_0(Y - m_\eta)] - \mathbb{P}_n[m_0 + \gamma_0(Y - m_0)] \\ &= \mathbb{P}_n \left[ m_\eta(X) + \frac{R}{\pi_0(X)}(Y - m_\eta(X)) \right] - \mathbb{P}_n \left[ m_0(X) + \frac{R}{\pi_0(X)}(Y - m_0(X)) \right], \end{aligned}$$

which measures the empirical influence-function discrepancy between  $m_\eta$  and  $m_0$ . The difference from frequentist doubly robust estimation arises because  $m_\eta$  is a random function under the nonparametric Bayesian posterior. As a consequence, the resulting bias term need not be asymptotically negligible. In general, it vanishes asymptotically only under Donsker conditions on the conditional mean function; see Remark 9.

**Remark 3** (Bias Equivalence for Prior Correction). *Under the product-rate condition imposed in Assumption 2, Theorem 3.1 of Breunig et al. (2025a) establishes a semiparametric BvM theorem for the prior-adjustment approach of Ray and van der Vaart (2020), denoted by  $\chi_\eta^{PA}$ , with bias term  $b_{0,\eta}$ . Combined with Theorem 2 above, this implies that the prior-adjusted posterior and the one-step posterior considered here exhibit the same first-order bias. In particular, by the triangle inequality for the bounded Lipschitz distance, we have the asymptotic equivalence*

$$d_{BL}(\mathcal{L}_\Pi(\sqrt{n}(\chi_\eta^{PA} - \hat{\chi} - b_{0,\eta}) \mid Z^{(n)}), \mathcal{L}_\Pi(\sqrt{n}(\chi_{\eta,\hat{\gamma}} - \hat{\chi} - b_{0,\eta}) \mid Z^{(n)})) \rightarrow_{P_0} 0.$$

*In other words, the one-step posterior exhibits exactly the same bias term as the plug-in version using the adjusted prior of Ray and van der Vaart (2020).*

We next relax the Donsker condition by allowing for pilot estimators of the unknown functional parameters in  $b_{0,\eta}$ , while retaining the weaker product-rate condition in (3.4). We explicitly correct the posterior distribution, following our proposed methodology in Algorithm 1. Recall the definition of the bias correction term  $\hat{b}_\eta$  given in Algorithm 1:

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

Under our conditions,  $\hat{b}_\eta$  consistently estimates the posterior-specific bias  $b_{0,\eta}$  uniformly over the relevant posterior set, so that subtracting  $\hat{b}_\eta$  removes the first-order centering error. In the present one-step construction, this correction also yields the exact AIPW representation discussed in Section 2.

**Theorem 4.** *Let Assumptions 1–3 hold. Then, we have*

$$d_{BL}(\mathcal{L}_\Pi(\sqrt{n}(\check{\chi}_\eta - \hat{\chi}) \mid Z^{(n)}), N(0, \mathbf{v}_0)) \rightarrow_{P_0} 0.$$

An important consequence of the BvM results is the frequentist validity of the Bayesian

credible set built from the corrected posterior. For any  $\alpha \in (0, 1)$ , the Bayesian credible set  $\mathcal{C}_n(\alpha)$  for the corrected draw  $\check{\chi}_\eta = \chi_{\eta, \hat{\gamma}} - \hat{b}_\eta$  satisfies  $P_0(\chi_0 \in \mathcal{C}_n(\alpha)) \rightarrow 1 - \alpha$ .

Below, let the AIPW estimator by Robins et al. (1994) be  $\hat{\chi}_{\text{AIPW}} := \mathbb{P}_n[\hat{m}(X) + \hat{\gamma}(R, X)(Y - \hat{m}(X))]$ . Under our assumptions, it is straightforward to see that  $\hat{\chi}_{\text{AIPW}}$  satisfies the asymptotic linear representation in (3.1) and hence achieves the semiparametric efficiency bound. Together with the discussion in Section 2.2, this yields the following result.

**Corollary 5.** *Let Assumptions 1–3 hold. Then, we have*

$$\check{\chi}_\eta = \hat{\chi}_{\text{AIPW}} + (\mathbb{P}_n^* - \mathbb{P}_n)[m_\eta(X) + \hat{\gamma}(R, X)(Y - m_\eta(X))] \quad (3.6)$$

and

$$d_{BL}(\mathcal{L}_\Pi(\sqrt{n}(\check{\chi}_\eta - \hat{\chi}_{\text{AIPW}}) \mid Z^{(n)}), N(0, V_0)) \rightarrow_{P_0} 0.$$

This corollary provides a direct link between the corrected posterior and conventional semiparametric estimation. Its finite-sample center is exactly the AIPW estimator based on the pilot nuisance estimates, while its asymptotic distribution is Gaussian with the semiparametric efficiency bound. Thus, the correction does not alter the first-order efficient point estimator, but instead constructs a posterior distribution around this center whose uncertainty is generated jointly by the posterior variation in the conditional mean function and the Bayesian bootstrap. This distinction will also be useful for interpreting the finite-sample comparison with DML in Section 5.

For intuition, note that the difference  $\check{\chi}_\eta - \hat{\chi}_{\text{AIPW}}$  in (3.6) resembles a centered bootstrap version of the AIPW estimator. Hence, conditional weak convergence to the desired normal limiting distribution is to be expected. That said, the technical challenge we address concerns uncertainty arising from the nonparametric Bayesian modeling of  $m_\eta$ , which differs substantially from the semiparametric bootstrap theory underlying existing frequentist constructions.

If the Donsker property holds, our procedure remains valid even when the pilot estimators are computed over the same sample. This can be easily seen as the debiasing term becomes asymptotically negligible. In other related causal inference problems, we observe that the incorporation of such a debiasing term can improve the finite-sample performance even without any sample splitting (Breunig et al., 2025a, 2024).

## 4 A BvM Theorem for the Cross-fitted Posterior under Primitive Conditions

Now we present the theoretical justification for the  $K$ -fold cross-fitted posterior under primitive conditions tailored to BART priors. The key ingredient of our construction involves the fold-specific posterior for the conditional mean function.

## 4.1 Tree Constructions

We now provide a detailed specification of the BART-type priors. Because the construction applies to any fold-specific posterior, we drop the additional subscript or superscript on  $k$  to lower the notational burden.

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 of the form  $\{x_j \leq \tau\}$  versus  $\{x_j > \tau\}$  where  $x_j$  signifies that the  $j$ -th coordinate 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 of the split-net from Ročková and van der Pas (2020). Given an increasing integer sequence  $b_n$ , a split-net  $\mathcal{X} = \{x_i = (x_{i1}, \dots, x_{ip})^\top \in [0, 1]^p, i = 1, \dots, b_n\}$  is a discrete collection of  $b_n$  split-candidates  $x_i$  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  $[\Omega]_j = \{x_j \in \mathbb{R} : (x_1, \dots, x_p)^\top \in \Omega\}$ .

The construction of each individual tree starts with some root node in  $[0, 1]^p$  at depth  $\delta = 0$ , 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 (i) the probability that a node at depth  $\delta$  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  $\delta \in \{0, 1, 2, \dots\}$  is split (hence nonterminal) with some prior probability depending on the depth. When it comes to point (ii), we follow Linero (2018) and Linero and Yang (2018) in using a sparse Dirichlet prior. This approach chooses a splitting covariate  $j$  from a proportion vector  $\vartheta = (\vartheta_1, \dots, \vartheta_p)^\top$  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  $[\mathcal{X}]_j \cap \text{int}([\Omega]_j)$  for a given split-net  $\mathcal{X}$ . The procedure continues until all nodes are 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  $j = 1, \dots, p$ .

We pick a fixed number of trees  $T$  and assume an independent product prior for the tree ensemble,  $\Pi(\mathcal{E}) = \prod_{t=1}^T \Pi(\mathcal{T}^t)$ . Chipman et al. (2010) recommend taking  $T = 200$  based on extensive numerical evidence. For a given split-net  $\mathcal{X}$  and for each  $1 \leq t \leq T$ , we denote by  $\mathcal{T}^t = (\Omega_1^t, \dots, \Omega_{L^t}^t)$  an  $\mathcal{X}$ -tree partition of size  $L^t$ , with step heights  $\beta^t = (\beta_1^t, \dots, \beta_{L^t}^t) \in \mathbb{R}^{L^t}$ . An additive tree-based learner is fully described by a tree ensemble  $\mathcal{E} = \{\mathcal{T}^1, \dots, \mathcal{T}^T\}$  and terminal node parameters  $\mathcal{B} = (\beta^1, \dots, \beta^T)^\top \in \mathbb{R}^{\sum_{t=1}^T L^t}$  as follows

$$m_{\mathcal{E}, \mathcal{B}}(x) = \sum_{t=1}^T \sum_{l=1}^{L^t} \beta_l^t \mathbb{1}\{x \in \Omega_l^t\}. \quad (4.1)$$

Given an ensemble  $\mathcal{E}$  of trees, we denote  $\mathcal{M}_{\mathcal{E}} := \{m_{\mathcal{E}, \mathcal{B}}(x) : \mathcal{B} \in \mathbb{R}^{\sum_{t=1}^T L^t}\}$ . If  $\mathcal{E}$  consists of

a single tree  $\mathcal{T}$ , we denote  $\mathcal{M}_{\mathcal{E}}$  by  $\mathcal{M}_{\mathcal{T}}$ .

## 4.2 Primitive Conditions

Now we introduce primitive conditions about the conditional mean function. Binary decision trees or forests offer good approximation properties for the 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 or priors. 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  $\phi : [0, 1]^p \rightarrow \mathbb{R}$ , with some  $\alpha > 0$ , where  $\|\cdot\|_2$  denotes the Euclidean norm. The true conditional mean function is assumed to belong to the following class of functions.

**Definition 6.** *We denote the space of uniformly  $\alpha$ -Hölder continuous functions that only depend on some subset of covariates  $\mathcal{S}$  as follows,*

$$\mathcal{F}_p(\alpha, \mathcal{S}) := \left\{ m : [0, 1]^p \rightarrow \mathbb{R} : \|m\|_{\mathcal{H}^\alpha} < \infty \text{ and } m \text{ is constant in the directions } \{1, \dots, p\} \setminus \mathcal{S} \right\}$$

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

BART can also exploit sparsity in the relevant regressors and perform variable selection. 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

$$\mathcal{F}_p^{add}(\boldsymbol{\alpha}, \mathcal{S}_0) := \left\{ m_0(x) = \sum_{t=1}^{T_0} m_0^t(x), \text{ such that } m_0^t \in \mathcal{F}_p(\alpha^t, \mathcal{S}_0^t) \right\},$$

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

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

related to the rate of posterior contraction for  $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$ . When considering the additive case  $m_0 \in \mathcal{F}_p^{add}(\boldsymbol{\alpha}, \mathcal{S}_0)$ , 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 fails to deliver the stochastic equicontinuity. The same issue occurs for the additive class whenever  $\alpha^t < q_0^t/2$  for any  $1 \leq t \leq T_0$ .

Next, we list assumptions on the outcome model, the split-net, and the prior, followed by a remark.

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

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

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

**Assumption 5** (Tree-based Partition). The split-net  $\mathcal{X}$  satisfies the following conditions:

(i)  $\max_{1 \leq j \leq p} \log b_j(\mathcal{X}) \lesssim \log n$ . (ii) One can construct an  $\mathcal{X}$ -tree partition  $\hat{\mathcal{T}}$  such that there exists  $m_{0,\hat{\mathcal{T}},\hat{\beta}}$  with its step heights  $\hat{\beta}$ , satisfying  $\|m_0 - m_{0,\hat{\mathcal{T}},\hat{\beta}}\|_\infty \lesssim \varepsilon_n$  if  $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$ . For  $m_0 \in \mathcal{F}_p^{\text{add}}(\alpha, \mathcal{S}_0)$ , we assume this tree-based approximation exists for each individual component in the additive function, that is,  $\|m_0^t - m_{0,\hat{\mathcal{T}}^t,\hat{\beta}^t}\|_\infty \lesssim \varepsilon_n^t$  for  $t = 1, \dots, T_0$ .

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

$$\text{d}\Pi(\mathcal{B} | L^1, \dots, L^T) = \prod_{t=1}^T \prod_{l=1}^{L^t} \phi_T(\beta_l^t) \text{d}\beta_l^t,$$

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

**Remark 7** (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 Xie and Xu (2020) and Jeong and Ročková (2023). 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 such 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 Jeong and Ročková (2023).*

*Assumption 6 lists the specification of priors in 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 Chipman et al. (2010) does not account for sparsity of the true regression model. Linero (2018) suggests the sparse Dirichlet prior on splitting coordinates to perform variable selection. Following Ročková and Saha (2020), the splitting probabilities decay exponentially with respect to the depth  $\delta$ , 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 Ghosal et al. (2000); see Xie and Xu (2020) and Jeong and Ročková (2023).

### 4.3 A New BvM Theorem for BART

In our Algorithm 2, when we sample from the posterior distributions of  $\chi_{\eta, \hat{\gamma}}^k$  for  $k = 1, \dots, K$ , the draws from the posterior distribution of the conditional mean function  $m_{\eta}^k$  and the Bayesian-bootstrap weights  $W^{(n_k)}$  are generated independently across folds. As such, the *posterior* law under consideration in our algorithm can be written as the convolution of fold-specific posterior distributions. For a pair of independent random variables  $U_1$  and  $U_2$  with marginal laws  $\mathcal{U}_1$  and  $\mathcal{U}_2$ , respectively, we denote their convolution by  $\mathcal{L}_{\mathcal{U}_1}(U_1) * \mathcal{L}_{\mathcal{U}_2}(U_2)$ , which represents the law of their sum.

Recall from Section 2 that  $K$  divides  $n$ , so that every fold is of the same size, i.e.,  $|I_1| = \dots = |I_K| = n/K$ . We define the DML estimator by  $\hat{\chi}_{\text{DML}} = K^{-1} \sum_{k=1}^K \hat{\chi}^k$  where in this section:

$$\hat{\chi}^k := \frac{K}{n} \sum_{i \in I_k} (\hat{m}^{(-k)}(X_i) + \hat{\gamma}^{(-k)}(R_i, X_i) (Y_i - \hat{m}^{(-k)}(X_i))).$$

For each fold  $k$ , our corrected posterior estimator is given by  $\check{\chi}_{\eta}^k = \chi_{\eta, \hat{\gamma}}^k - \hat{b}_{\eta}^k$ . As a point estimator, we use the average of the posterior means of the fold-specific corrected estimators. Formally,

$$\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{\Pi} [\check{\chi}_{\eta}^k \mid Z^{(n_k)}] = \frac{1}{K} \sum_{k=1}^K \hat{\chi}^k, \quad (4.2)$$

where the equality follows from the same argument as in (2.8), since the pilot estimators used for fold  $k$  are computed using observations outside  $I_k$ .

The numerical equivalence of the posterior mean and the DML estimator is agnostic to the type of nonparametric priors for the conditional mean, as well as the pilot estimators  $(\hat{m}, \hat{\pi})$ . Nonetheless, our simulation results favor BART-type modeling and reusing its posterior mean as the pilot estimator  $\hat{m}^{(-k)}$ . This is also close to the findings of Dorie et al. (2019), where targeted MLE (TMLE) with BART point estimators is among the top performers in their large-scale data competition.

Our main theoretical contribution is that, besides the point estimator, the credible sets constructed from the fold-specific posteriors are valid. Next, we present a formal statement of the validity of the  $K$ -fold posterior cross-fitting procedure under primitive conditions for the BART priors. Regarding the convolution in the stated theorem, we do not claim conditional independence of the fold-specific posteriors under the ground-truth probability law  $P_0$ , conditional on the entire observed data  $Z^{(n)}$ . Rather, we deliberately generate posterior draws of the conditional mean function and the bootstrap weights separately across different folds. After applying Bayes' rule to those components, the posterior laws  $\mathcal{L}_{\Pi}(\cdot \mid Z^{(n_k)})$ , which encode the randomness of the conditional mean function and the bootstrap weights, are made independent over different folds.

**Theorem 8.** *Let Assumptions 1 and 4–6 hold. In addition, we assume that  $\|1/\hat{\pi}^{(-k)}\|_\infty = O_{P_0}(1)$  for  $1 \leq k \leq K$ . If  $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$ ,  $\|\hat{m}^{(-k)} - m_0\|_{2, \mu_0} = O_{P_0}(\varepsilon_n)$ , and  $\sqrt{n} \varepsilon_n \|\hat{\pi}^{(-k)} - \pi_0\|_\infty \rightarrow_{P_0} 0$ , then*

$$d_{BL} \left( \mathcal{L}_\Pi \left( \frac{\sqrt{n}}{K} (\check{\chi}_\eta^1 - \hat{\chi}^1) \mid Z^{(n_1)} \right) * \dots * \mathcal{L}_\Pi \left( \frac{\sqrt{n}}{K} (\check{\chi}_\eta^K - \hat{\chi}^K) \mid Z^{(n_K)} \right), N(0, \mathbf{v}_0) \right) \rightarrow_{P_0} 0.$$

*The same conclusion holds if  $m_0 \in \mathcal{F}_p^{add}(\alpha, \mathcal{S}_0)$ ,  $\|\hat{m}^{(-k)} - m_0\|_{2, \mu_0} = O_{P_0}(\varepsilon_n^{add})$  and  $\sqrt{n} \varepsilon_n^{add} \|\hat{\pi}^{(-k)} - \pi_0\|_\infty \rightarrow_{P_0} 0$ .*

Theorem 8 establishes a BvM result for BART without requiring the Donsker property for either the propensity score or the conditional mean function. We first argue that it suffices to consider each fold-specific posterior separately. Then, the desired result follows from verifying the high-level assumptions from the previous section.

**Remark 9** (Donsker Class). *One can establish the BvM theorem 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  $m_0 \in \mathcal{F}_p^{add}(\alpha, \mathcal{S}_0)$  if  $\alpha^t > q_0^t/2$  for all  $1 \leq t \leq T_0$ . For semiparametric Bayesian inference, such Donsker properties are imposed in Assumption 2(c) in Yiu et al. (2025) or Condition (3.12) in Ray and van der Vaart (2020). Our results exploit the regularity of the propensity score, expressed in terms of a new stochastic equicontinuity condition, to restore the frequentist validity of our robust procedure.*

**Remark 10** (Smoothed BART). *Given the nonsmooth feature of BART, it is natural to consider the smoothed BART (see Linero and Yang (2018)), because it exploits the smoothness of the conditional mean. The same trade-off between the smoothness of the conditional mean and propensity score continues to hold when the additional bias correction is applied. Consider functions in the Hölder smooth class, i.e., the space of functions on  $[0, 1]^p$  with bounded partial derivatives up to order  $\lfloor \alpha \rfloor$ , where  $\lfloor \alpha \rfloor$  is the largest integer strictly less than  $\alpha$  and such that the partial derivatives of order  $\lfloor \alpha \rfloor$  are Hölder continuous of order  $\alpha - \lfloor \alpha \rfloor$ . If the true function only depends on at most  $q_0$  covariates, Theorem 2 in Linero and Yang (2018) states that the posterior rate of contraction is of the order  $O(n^{-\alpha/(2\alpha+q_0)} \log(n))$ . If the conditional mean function and propensity score function are in such Hölder smooth classes with smoothness indices  $(\alpha_m, \alpha_\pi)$  and they only depend on  $q_0$  regressors, our robust approach is asymptotically normal when  $\sqrt{\alpha_m \alpha_\pi} > q_0/2$ . In comparison, the Donsker class restriction will force  $\min\{\alpha_m, \alpha_\pi\} > q_0/2$ .*

## 5 Numerical Studies

In this section, we first examine the finite-sample performance of the proposed robust BART procedure for the mean response  $\chi_0 = \mathbb{E}_0[Y_i]$  in missing-data models. Then, building on Appendix C, which extends the robust BART to the average treatment effect (ATE) framework, we apply the method to the well-known National Health and Nutrition Examination

Survey data and 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  $i = 1, \dots, n$ , we generate the observed variables  $(R_i Y_i, R_i, X_i^\top)$  by  $X_i = (X_{i1}, \dots, X_{i5})^\top$  where  $X_{i1}, X_{i2}, X_{i3} \sim N(0, 1)$ ,  $X_{i4} \sim \text{Bernoulli}(0.5)$ , and  $X_{i5}$  is a categorical variable taking values  $\{1, 2, 3\}$  with equal probability. The distributions of  $R_i$  and  $Y_i$  are given by

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

where  $\Phi$  denotes the cumulative distribution function of the standard normal distribution. We analyze the finite-sample effects of varying the complexity of the conditional mean function  $m$  and the propensity score for sample sizes  $n \in \{1000, 2000\}$ . Throughout our simulations, the number of Monte Carlo replications is set to 1,000.

We consider four designs for the function  $e(\cdot)$  that determines the propensity score by  $\pi(x) = \Phi(e(x))$  and for the conditional mean function  $m(\cdot)$ . In all four designs, neither  $e(\cdot)$  nor  $m(\cdot)$  is additively separable in  $x$ . The complexity of the nuisance functions increases as more interaction terms are added.

$$\text{Design I: } e(x) = -0.2x_1 + 0.4x_1x_3, \quad m(x) = 1 + x_1x_2 + x_1x_3 + x_2 + x_4 + h(x_5),$$

$$\text{Design II: } e(x) = -0.2x_1 + 0.4x_1x_3, \quad m(x) = 1 + x_1x_3 + x_2x_3 + x_2x_4 + h(x_5),$$

$$\text{Design III: } e(x) = -0.2x_1 + 0.4x_1x_3 + 0.4x_2x_3, \quad m(x) = 1 + x_1x_3 + x_2x_3 + x_2x_4 + h(x_5),$$

$$\text{Design IV: } e(x) = -0.2x_1 + 0.4x_1x_3 + 0.4x_2x_3, \quad m(x) = 1 + x_1x_3 + x_2x_3 + x_2x_4 + \tilde{h}(x),$$

where  $h(x_5) = (5\mathbb{1}\{x_5 = 1\} - \mathbb{1}\{x_5 = 2\} - 1)/2$  and  $\tilde{h}(x) = ((4x_1 + 1)\mathbb{1}\{x_5 = 1\} - \mathbb{1}\{x_5 = 2\} - 1)/2 + \sin(x_4x_5)$ , with  $\mathbb{1}\{\cdot\}$  denoting the indicator function.

We evaluate six 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** (Sparapani et al., 2021): the posterior of  $m_\eta(\cdot)$  is computed using the function `gbart` with argument `type = wbart`. We draw 2,000 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 E for details. **One-step BART** applies the posterior correction of Yiu et al. (2025), which uses posteriors of both  $m(\cdot)$  and  $\pi(\cdot)$  obtained via BART or its probit variant.<sup>3</sup> **One-step BART <sub>$\hat{\gamma}$</sub>**  as given in (2.4) is a variant of the one-step BART that uses a pilot estimate for the propensity score  $\pi(\cdot)$ .<sup>4</sup> **RoBART**, the robust BART method described in Algorithm 2,

<sup>3</sup>We implement probit BART by calling `gbart` with `type = pbart`.

<sup>4</sup>It can be viewed as assigning a degenerate prior to the propensity score  $\pi(\cdot)$ . Its implementation follows Algorithm 2 but without the debiasing term  $\hat{b}_\eta^k$ .

combines the BART posterior of  $m_\eta(\cdot)$ , a plug-in estimator for the propensity score  $\pi_0(\cdot)$ , and the debiasing term  $\widehat{b}_\eta$  in (2.13). RoBART applies cross-fitting with  $K = 5$  folds. The propensity score is estimated using a probit LASSO that includes all pairwise interactions of  $X$ 's as regressors. The debiasing term also involves an estimator for  $m_0(\cdot)$ , which is constructed by the posterior mean of  $m_\eta(\cdot)$  produced by BART. We also consider two variants of a frequentist double machine learning estimator (Chernozhukov et al., 2017, 2018): **DML–BART** estimates  $\pi(\cdot)$  using a probit LASSO and  $m(\cdot)$  using the BART posterior mean; **DML–RF** applies the same estimator for  $\pi(\cdot)$  but a random forest estimator for  $m(\cdot)$ . Both DML variants use cross-fitting with  $K = 5$  folds and report the usual Wald confidence interval based on the sample variance of the estimated orthogonal score. DML–BART is deliberately constructed from exactly the same partition  $(I_k)_{k=1}^K$  and the same pilot estimators  $(\widehat{m}^{(-k)}, \widehat{\pi}^{(-k)})$  as RoBART, so that the equivalence of Section 2 applies and the two procedures differ only in how they quantify uncertainty.

Table 1 reports the Monte Carlo bias, the coverage probability (CP) of the nominal 95% interval estimates, and the average interval length (CIL) for each method. Standard BART is biased and undercovers in every design, and its coverage remains far below the nominal level when the sample size doubles. One-step BART reduces the bias by 16% to 28% and raises coverage by 6 to 20 percentage points, but the improvement is not sufficient, especially when the design complexity increases. The variant One-step BART $_{\widehat{\gamma}}$  also reduces the bias and improves the coverage at the sample size  $n = 1000$ , but stops doing so at  $n = 2000$ . RoBART, on the other hand, produces substantially smaller bias and close-to-nominal empirical coverage probabilities throughout all designs. These results align with our theoretical discussion below Theorem 8. As every design contains additive components depending on more than one covariate, the sieve of tree ensembles that carries the posterior mass may fail the stochastic equicontinuity condition  $\sup_{\eta \in \mathcal{H}_n^m} |\mathbb{G}_n[m_\eta - m_0]| = o_{P_0}(1)$ , and the bias term  $b_{0,\eta}$  of Theorem 2 need not be asymptotically negligible. By explicitly removing  $b_{0,\eta}$ , our RoBART demonstrates robust performance throughout all designs with its advantage over standard and one-step BART increasing with the complexity of nuisance functions. The cross-fitted nature of RoBART is likely to contribute to its coverage as well through the reduced dependence between the pilot estimators entering the debiasing term and the subsample that generates the posterior draw. The comparison of the RoBART and DML–BART rows is a direct numerical check of the equivalence established in Section 2. Because the two procedures use the same folds and the same pilot estimators, the average of the fold-specific posterior means of RoBART is algebraically identical to the cross-fitted DML. The table confirms this: the two rows report the same bias to three decimals in six of the eight design and sample size cells and differ by 0.001 in the remaining two, the residual discrepancy being simulation error from approximating the posterior mean by the average of  $S$  finite posterior draws. In terms of uncertainty quantification, RoBART exhibits better coverage performance than DML–BART in all eight cells, and the gap widens with the complexity of the design, from 0.7 and 1.4 percentage points in Designs I and II to 3.0 and 4.3 points in Designs III and IV at  $n = 1000$ , and to 3.9 points in

Table 1: Finite-sample performance of BART-based and DML inference methods for the mean response with missing data. Bias, coverage probability (CP) of the nominal 95% interval estimates and average interval length (CIL), computed over 1,000 Monte Carlo replications.

| $n$  | Method                          | Design I |       |       | Design II |       |       | Design III |       |       | Design IV |       |       |
|------|---------------------------------|----------|-------|-------|-----------|-------|-------|------------|-------|-------|-----------|-------|-------|
|      |                                 | Bias     | CP    | CIL   | Bias      | CP    | CIL   | Bias       | CP    | CIL   | Bias      | CP    | CIL   |
| 1000 | BART                            | 0.088    | 0.813 | 0.331 | 0.087     | 0.797 | 0.315 | 0.166      | 0.448 | 0.315 | 0.174     | 0.426 | 0.315 |
|      | One-step BART                   | 0.069    | 0.879 | 0.343 | 0.067     | 0.868 | 0.326 | 0.134      | 0.646 | 0.328 | 0.146     | 0.594 | 0.328 |
|      | One-step BART $_{\hat{\gamma}}$ | 0.077    | 0.882 | 0.433 | 0.069     | 0.882 | 0.418 | 0.132      | 0.716 | 0.482 | 0.139     | 0.710 | 0.496 |
|      | RoBART                          | 0.010    | 0.948 | 0.421 | 0.007     | 0.947 | 0.407 | 0.029      | 0.920 | 0.471 | 0.041     | 0.918 | 0.480 |
|      | DML-BART                        | 0.011    | 0.941 | 0.413 | 0.007     | 0.933 | 0.395 | 0.029      | 0.890 | 0.440 | 0.040     | 0.875 | 0.441 |
|      | DML-RF                          | 0.039    | 0.936 | 0.444 | 0.035     | 0.933 | 0.420 | 0.076      | 0.853 | 0.472 | 0.071     | 0.870 | 0.477 |
| 2000 | BART                            | 0.059    | 0.816 | 0.236 | 0.057     | 0.812 | 0.224 | 0.106      | 0.547 | 0.224 | 0.116     | 0.473 | 0.224 |
|      | One-step BART                   | 0.043    | 0.875 | 0.241 | 0.041     | 0.886 | 0.229 | 0.081      | 0.708 | 0.231 | 0.093     | 0.635 | 0.231 |
|      | One-step BART $_{\hat{\gamma}}$ | 0.081    | 0.778 | 0.275 | 0.075     | 0.784 | 0.270 | 0.132      | 0.539 | 0.342 | 0.152     | 0.452 | 0.332 |
|      | RoBART                          | 0.014    | 0.932 | 0.266 | 0.011     | 0.941 | 0.261 | 0.020      | 0.941 | 0.334 | 0.031     | 0.927 | 0.320 |
|      | DML-BART                        | 0.014    | 0.928 | 0.263 | 0.011     | 0.933 | 0.251 | 0.020      | 0.902 | 0.294 | 0.031     | 0.888 | 0.292 |
|      | DML-RF                          | 0.037    | 0.901 | 0.312 | 0.033     | 0.922 | 0.296 | 0.064      | 0.818 | 0.361 | 0.061     | 0.837 | 0.361 |

both Designs III and IV at  $n = 2000$ . The price is a moderate increase in the interval length. The superior coverage performance of RoBART likely reflects its propagation of the BART posterior uncertainty about  $m_\eta$  together with the Bayesian-bootstrap variation, compared with DML’s fixed pilot  $\hat{m}^{(-k)}$ . DML-RF, which replaces the BART pilot for the conditional mean by a random forest, produces larger bias in every cell and worse coverage than DML-BART, except in Design II at  $n = 1000$  where the two coincide. Its intervals are longer than those of DML-BART throughout, and comparable to or longer than those of RoBART except in Design IV at  $n = 1000$ .

## 5.2 Empirical Application

We apply the BART-based methods to a subsample of data from the National Health and Nutrition Examination Survey (NHANES) 2007–2008, previously analyzed by Chan et al. (2016).<sup>5</sup> 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

<sup>5</sup>The data file `nhanes_bmi.csv` is publicly available from Ding (2023), *Replication Data for: A First Course in Causal Inference*, Harvard Dataverse, V4, <https://doi.org/10.7910/DVN/ZX3VEV>.

in the family). The sample size is 2330.

We adapt BART and DML methods from the simulation study to the ATE setting (see Appendix C for RoBART for the ATE), and present the results in Table 2. As in the simulation study, RoBART applies cross-fitting with  $K = 5$  folds, and DML-BART is computed from the same partition and the same pilot estimators as RoBART. In addition to the methods considered in the simulation study, we replicate two frequentist estimates from Chan et al. (2016): IPW refers to the inverse propensity score weighting estimator (Hájek-type) and CAL refers to the calibration estimator recommended by Chan et al. (2016).<sup>6</sup>

Table 2 shows that all estimated effects of school meal programs on BMI are small relative to the sample mean BMI of 20.11 (standard deviation 5.42). In absolute value, BART, DML-RF and IPW yield comparatively larger point estimates, while RoBART, DML-BART, CAL and One-step BART lead to estimates closer to zero. Every 95% interval estimate contains zero, so the sign of the average effect cannot be determined once sampling uncertainty is taken into account.

The application also reproduces, on real data, the numerical equivalence established in Section 2 and documented in the simulations. RoBART and DML-BART report the same point estimate of 0.07. The two procedures differ only in how they quantify uncertainty, and here the RoBART credible interval  $[-0.36, 0.50]$  is marginally shorter than the corresponding Wald interval  $[-0.37, 0.51]$ . Replacing the BART pilot for the outcome regression by a random forest, as in DML-RF, moves the point estimate to 0.15. One-step BART produces the most dispersed posterior and hence the widest interval among the methods considered. The variant One-step BART $_{\hat{\gamma}}$  leads to a narrower credible interval than One-step BART, although it is still wider than those of the other BART-based methods. Overall, our RoBART, DML-BART and the calibration estimator of Chan et al. (2016) combine point estimates close to zero with the shortest intervals, leading to comparatively precise and stable inference.

## 6 Conclusion

This paper studies semiparametric posterior inference with BART outside the Donsker regime. We show that the natural one-step BART posterior can retain a centering bias and develop RoBART to remove this bias while preserving the posterior variation generated by BART and the Bayesian bootstrap. For the cross-fitted procedure, we establish a Bernstein-von Mises theorem under primitive conditions tailored to BART. Our point estimator, the average of the fold-specific posterior means in the cross-fitted version, coincides exactly with the corresponding DML estimator. This strong equivalence further justifies our correction procedure. Our theoretical and numerical results demonstrate that RoBART

---

<sup>6</sup>We take the calibration estimator with exponential tilting as a representative; other versions of calibration estimates reported in Chan et al. (2016) are similar in terms of the magnitude and confidence interval.

Table 2: Bayesian and frequentist ATE estimates of the effect of school meal programs on children’s and youths’ BMI,  $n = 2330$ .

| Method                          | ATE   | 95% CI          | CIL  |
|---------------------------------|-------|-----------------|------|
| BART                            | 0.15  | $[-0.30, 0.60]$ | 0.90 |
| One-step BART                   | -0.02 | $[-0.79, 0.66]$ | 1.45 |
| One-step BART $_{\hat{\gamma}}$ | 0.04  | $[-0.44, 0.51]$ | 0.95 |
| RoBART                          | 0.07  | $[-0.36, 0.50]$ | 0.86 |
| DML-BART                        | 0.07  | $[-0.37, 0.51]$ | 0.88 |
| DML-RF                          | 0.15  | $[-0.30, 0.60]$ | 0.90 |
| IPW                             | -0.14 | $[-0.62, 0.34]$ | 0.96 |
| CAL                             | -0.04 | $[-0.48, 0.40]$ | 0.88 |

provides a theoretically sound posterior distribution for uncertainty quantification around an established robust and efficient frequentist center.

RoBART and DML share the same point-estimation center when constructed using the same folds and nuisance estimators, so their comparison isolates differences in uncertainty quantification. Although the two procedures have the same first-order asymptotic distribution, our simulations show that RoBART achieves coverage closer to the nominal level in finite samples, with the improvement becoming more pronounced as the nuisance functions become more complex. This pattern is consistent with the different ways in which uncertainty is quantified: RoBART propagates posterior uncertainty about the conditional mean function together with Bayesian-bootstrap variation, whereas the conventional DML Wald interval is based on the estimated orthogonal score and treats the estimated nuisance functions as fixed. These differences vanish asymptotically under our conditions but can matter substantially in finite samples.

## Acknowledgments

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.

## A Proofs of Main Results

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

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

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  $\Pi(m_\eta \in \mathcal{H}_n^m | Z^{(n)}) \rightarrow_{P_0} 1$ . We introduce the conditional prior  $\Pi_n(\cdot) := \Pi(\cdot \cap \mathcal{H}_n^m) / \Pi(\mathcal{H}_n^m)$ . Our BvM theorem makes use of a frequentist estimator  $\hat{\chi}$  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. For simplicity of notation, we introduce

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

The frequentist estimator  $\hat{\chi}$  is required to satisfy the linear representation with the efficient influence function given in (2.6) under  $P_0$ . We set  $\hat{\chi} = \mathbb{P}_n[m_0 + \hat{\gamma}\rho^{m_0}]$  and indeed due to Lemma C.2 of Breunig et al. (2025b) this choice is justified since

$$\hat{\chi} = \mathbb{P}_n[m_0 + \hat{\gamma}\rho^{m_0}] = \mathbb{P}_n[m_0 + \gamma_0\rho^{m_0}] + o_{P_0}(n^{-1/2}) = \chi_0 + \mathbb{P}_n\tilde{\chi}_0 + o_{P_0}(n^{-1/2}),$$

which is used in the proofs presented below.

*Proof of Theorem 2.* As in Theorem 2 in Ray and van der Vaart (2020), we assume that the propensity score estimator  $\hat{\pi}$  is constructed from observations that are independent of  $Z^{(n)}$ , as imposed in Assumption 2. Accordingly, we treat  $\hat{\pi}$  as a deterministic sequence  $\pi_n$  in the sequel and define  $\gamma_n(r, x) = r/\pi_n(x)$ . Consequently, we can write  $\chi_{\eta, \gamma_n} = \mathbb{P}_n^*[m_\eta + \gamma_n\rho^{m_\eta}]$ .

By Theorem 1.13.1 in van der Vaart and Wellner (1996), we can show this by establishing the convergence of the conditional Laplace transform

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

We proceed with the following decomposition:

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

by the definition of the bias term  $b_{0, \eta}$ . For the remainder term  $r_{n, \eta}$ , we further express it as

$$\begin{aligned}
r_{n, \eta} &= (\mathbb{P}_n - P_0) [(\gamma_0 - \gamma_n) (m_\eta - m_0)] + P_0 [(\gamma_0 - \gamma_n) (m_\eta - m_0)], \\
|r_{n, \eta}| &\leq |(\mathbb{P}_n - P_0) [(\gamma_0 - \gamma_n) (m_\eta - m_0)]| + \frac{1}{c} \|1/\pi_n\|_\infty \|\pi_n - \pi_0\|_{2, \mu_0} \|m_\eta - m_0\|_{2, \mu_0}, \quad (\text{A.1})
\end{aligned}$$

by employing the Cauchy–Schwarz inequality and the overlap condition in Assumption 1. Hence, we obtain  $\sup_{\eta \in \mathcal{H}_n} \sqrt{n} |r_{n, \eta}| = o_{P_0}(1)$ , by invoking the stochastic equicontinuity condition (3.4) imposed in Assumption 3 for the first term, and the convergence rate conditions in Assumption 2 for the second term on the right-hand side of the inequality (A.1).

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

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

Thus, the conditional Laplace transform becomes

$$I_n(t) = \frac{\iint_{\eta \in \mathcal{H}_n} e^{(t\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)(\tilde{\chi}_0 + \vartheta_{n, \eta}))} d\Pi_W(W^{(n)} \mid Z^{(n)}) e^{\ell_n(m_\eta)} d\Pi(m_\eta)}{\int_{\eta \in \mathcal{H}_n} e^{\ell_n(m_\eta)} d\Pi(m_\eta)} \times e^{o_{P_0}(1)}.$$

Under Assumption 3 (i), we apply Lemma 12 to the function class  $\{\tilde{\chi}_0 + \vartheta_{n, \eta} : \eta \in \mathcal{H}_n\}$ . 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 12 gives us

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

Under Assumption 2 we show in Lemma 14 that

$$\sup_{\eta \in \mathcal{H}_n} |\text{Var}_0(\tilde{\chi}_0(Z) + \vartheta_{n, \eta}(Z)) - \text{Var}_0(\tilde{\chi}_0(Z))| \rightarrow 0. \quad (\text{A.2})$$

We emphasize that the above uniform convergence only requires the influence function

$\tilde{\chi}_0 + \vartheta_{n,\eta}$  to be in the  $P_0$ -Glivenko–Cantelli class, not necessarily the  $P_0$ -Donsker class. Therefore, we have

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

which concludes the proof.  $\square$

The original setup in Breunig et al. (2025a) concerns estimation of the average treatment effect. Our missing-data example can be viewed as applying the proposal of Breunig et al. (2025a) to the treated (or control) arm only. For concreteness, we adapt the proof of their Corollary 3.1 to show that the feasible estimator  $\hat{b}_\eta$  approximates  $b_{0,\eta}$  uniformly well in the missing-data problem.

*Proof of Theorem 4.* It is sufficient to show that  $\sup_{\eta \in \mathcal{H}_n} |b_{0,\eta} - \hat{b}_\eta| = o_{P_0}(n^{-1/2})$ , where  $b_{0,\eta} = \mathbb{P}_n[(\gamma_0 - 1)(m_0 - m_\eta)]$  and  $\hat{b}_\eta = \mathbb{P}_n[(\hat{\gamma} - 1)(\hat{m} - m_\eta)]$ . We make use of the decomposition

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

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

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

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

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

Under Assumption 2, we apply Lemma 13 to get  $\mathbb{P}_n \left[ \frac{R - \pi_0(X)}{\pi_0(X)} (m_0(X) - \hat{m}(X)) \right] = o_{P_0}(n^{-1/2})$ . Hence, we obtain  $\mathbb{P}_n[(\gamma_0 - 1)(m_0 - \hat{m})] = o_{P_0}(n^{-1/2})$ , which completes the proof.  $\square$

We present a proof of the validity of the  $K$ -fold posterior cross-fitting procedure. Recall that we partition the sample of size  $n$  into  $K$  folds of sizes  $n_k \asymp n/K$ ,  $k = 1, \dots, K$ . For each fold  $k$ , let  $\hat{\chi}^k$  denote the frequentist DML estimator computed on fold  $k$ , serving as the centering point in the Bernstein–von Mises theorem. Denote by  $\ell_{n_k}(m_\eta^k)$  the likelihood

function in fold  $k$ , and let  $Z^{(n_k)}$  and  $W^{(n_k)}$  denote the observed data and Bayesian-bootstrap weights in fold  $k$ , respectively. Let  $a \vee b$  denote the maximum of  $a$  and  $b$ .

*Proof of Theorem 8.* First, we argue that it suffices to check the conditional convergence for each fold-specific posterior in order to get the desired result. For each fold  $k$ , define the posterior Laplace transform

$$\begin{aligned} I_{n_k}(t) &= \mathbb{E}_{\Pi} \left[ \exp \left( t \frac{\sqrt{n}}{K} (\tilde{\chi}_{\eta}^k - \hat{\chi}^k) \right) \mid Z^{(n_k)} \right] \\ &= \frac{\int \exp \left( t \frac{\sqrt{n}}{K} (\chi_{\eta, \hat{\gamma}}^k - \hat{\chi}^k - \hat{b}_{\eta}^k) \right) d\Pi_W(W^{(n_k)} \mid Z^{(n_k)}) e^{\ell_{n_k}(m_{\eta}^k)} d\Pi(m_{\eta}^k)}{\int e^{\ell_{n_k}(m_{\eta}^k)} d\Pi(m_{\eta}^k)}. \end{aligned}$$

Note that  $I_{n_k}(t)$  depends on the entire dataset  $(Z^{(n_1)}, \dots, Z^{(n_K)})$  through pilot estimation, but the sampling from each posterior distribution is made independently across the  $K$  folds. Once we verify the high-level assumptions that lead to Theorem 2 and Theorem 4, for each fixed  $t$  in a neighborhood of zero, we get  $I_{n_k}(t) \rightarrow_{P_0} \exp(v_0 t^2 / (2K))$ . Since posterior draws are generated independently across folds, the joint Laplace transform of

$$\mathcal{L}_{\Pi} \left( \frac{\sqrt{n}}{K} (\chi_{\eta, \hat{\gamma}}^1 - \hat{\chi}^1 - \hat{b}_{\eta}^1) \mid Z^{(n_1)} \right) * \dots * \mathcal{L}_{\Pi} \left( \frac{\sqrt{n}}{K} (\chi_{\eta, \hat{\gamma}}^K - \hat{\chi}^K - \hat{b}_{\eta}^K) \mid Z^{(n_K)} \right) \quad (\text{A.4})$$

factorizes into

$$\prod_{k=1}^K I_{n_k}(t) \rightarrow_{P_0} \prod_{k=1}^K \exp \left( \frac{v_0 t^2}{2K} \right) = \exp \left( \frac{v_0 t^2}{2} \right).$$

This limit is precisely the Laplace transform of a centered Gaussian distribution with variance  $v_0$ . By Theorem 1.13.1 of van der Vaart and Wellner (1996), this conditional convergence of Laplace transforms implies the distributional convergence in bounded Lipschitz distance. Thus, the posterior law in (A.4) converges conditionally to the correct Gaussian limit.

Now, we check the high-level assumptions that lead to our BvM theorem for the fold-specific posterior. In the remaining proof, we suppress the subscript or superscript related to the  $k$ -th fold to lower the notational burden. The statement is understood to be with respect to any given fold, for  $k = 1, \dots, K$ .

First, we provide the construction of measurable sets  $\mathcal{H}_n$  such that  $\Pi(\eta \in \mathcal{H}_n \mid Z^{(n)}) \rightarrow_{P_0} 1$  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  $\Pi(m_{\eta} \in \mathcal{H}_n^m \mid Z^{(n)}) \rightarrow_{P_0} 1$ . Referring to the sieve set defined by equation (D.1), we work with

$$\mathcal{H}_n^m := \{m_{\eta} : m_{\eta} \in \mathcal{M}_n, \|m_{\eta} - m_0\|_{2, \mu_0} \lesssim \varepsilon_n\}. \quad (\text{A.5})$$

We prove the following posterior contraction in Lemma D.1:

$$\Pi(m \in \mathcal{M}_n : \|m - m_0\|_{2,\mu_0} > C_n \varepsilon_n \mid Z^{(n)}) \rightarrow_{P_0} 0, \quad (\text{A.6})$$

with  $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)} \sqrt{\log n}$  for any  $C_n \rightarrow \infty$  as  $n, p \rightarrow \infty$ . In Lemma D.1, we have shown that  $\Pi(m_\eta \notin \mathcal{M}_n) \rightarrow_{P_0} 0$ . Combined with Lemma 1 in Ghosal and van der Vaart (2007), this yields  $\Pi(m_\eta \notin \mathcal{M}_n \mid Z^{(n)}) \rightarrow_{P_0} 0$ . Taken together, we have shown that  $\Pi(m_\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \rightarrow_{P_0} 1$ .

Under the uniform boundedness assumption of the conditional mean, the requirement for the envelope functions in (3.3) is satisfied. Regarding the  $P_0$ -Glivenko–Cantelli property of  $\{m_\eta - m_0\}$  for  $m_\eta \in \mathcal{G}_n$ , it is sufficient to verify this for the sieve set  $\mathcal{M}_n$ . Let  $\|\cdot\|_n$  denote the empirical  $L^2(\mathbb{P}_n)$ -norm. For  $\mathcal{M}_n$  with  $\bar{L}_n \sim n \varepsilon_n^2 / \log n$  and  $\bar{s}_n \sim n \varepsilon_n^2 / \log(p \vee n)$ , the proof of Lemma D.1 shows that  $\log N(\varepsilon, \mathcal{M}_n, \|\cdot\|_n) \lesssim n^{q_0/(2\alpha+q_0)} \log n = o(n)$  for any given  $\varepsilon > 0$ . This satisfies the requirement about the growth of the entropy number in Theorem 2.4.3 of van der Vaart and Wellner (1996), which verifies the  $P_0$ -Glivenko–Cantelli property.

For the stochastic equicontinuity condition, we apply the multiplier inequality as stated in Lemma 11:

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

In the last inequality, we have applied the upper bound  $\mathbb{E}_0 \sup_{m_\eta \in \mathcal{M}_n(\varepsilon_n)} |\mathbb{G}_n(m_\eta - m_0)| \lesssim \sqrt{n} \varepsilon_n$ , as given in Lemma D.2, where  $\mathcal{M}_n(\varepsilon_n)$  is defined.

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

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

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

## B Useful Lemmas and Auxiliary Results

### B.1 Useful Lemmas

The following lemma is in the same spirit as Lemma 9 in Ray and van der Vaart (2020) with one important difference: 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 Ledoux and Talagrand (1991) instead of Proposition A.1.10 of van der Vaart and Wellner (1996). This allows us to relax the positive range restriction in Ray and van der Vaart (2020).

**Lemma 11.** *Consider a set  $\mathcal{H}$  of measurable functions  $h : \mathcal{Z} \rightarrow \mathbb{R}$  and a bounded measurable function  $\varphi$ . We have*

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

We now state the following generalization of Theorem 1 from Ray and van der Vaart (2021), where the function 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 Yiu et al. (2025). As discussed by Ray and van der Vaart (2021) on page 2225 therein, this uniformity refers to the marginal posterior distributions of the process  $\tilde{\chi}_0 + \vartheta_{n,\eta}$  for  $\eta \in \mathcal{H}_n$ . It is not about the distributional convergence of this process as a random element in  $\ell^\infty(\mathcal{H}_n)$ .

**Lemma 12.** *Suppose  $\mathcal{G}_n$  is a sequence of separable classes of measurable functions with envelope functions  $G_n$ , such that*

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

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

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

### B.2 Smaller Order Terms

The next lemma verifies that the term arising in the debiasing step is asymptotically negligible. Its proof is in the same spirit as bounding the  $R_{n,2}$  term in Lemma C.8 of Breunig et al. (2025b).

**Lemma 13.** *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.,  $\|\hat{m} - m_0\|_{2,\mu_0} \rightarrow_{P_0} 0$ .*

If, in addition, the true propensity score is uniformly bounded away from zero, then we have

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

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

$$\begin{aligned} & \mathbb{E}_0 \left[ \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i - \pi_0(X_i)}{\pi_0(X_i)} (\hat{m}(X_i) - m_0(X_i)) \right)^2 \middle| X_1, \dots, X_n, \hat{m} \right] \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{m}(X_i) - m_0(X_i))^2 \frac{\text{Var}_0(R_i | X_i)}{\pi_0^2(X_i)} \end{aligned}$$

using that  $\text{Var}_0(R_i | X_i) = \pi_0(X_i)(1 - \pi_0(X_i))$ . Consequently, by the overlap condition  $1 \lesssim \pi_0(X_i)$  for all  $1 \leq i \leq n$ , we obtain

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

where the last equality is due to the  $L^2(\mu_0)$ -convergence for the pilot estimator  $\hat{m}$ . Consequently, the result follows from Markov's inequality.  $\square$

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 2.

**Lemma 14.** *Let  $\pi_n(\cdot)$  be a sequence of functions uniformly bounded away from zero and  $\|\pi_n - \pi_0\|_{2, \mu_0} \rightarrow 0$ . Uniformly in  $\eta \in \mathcal{H}_n$ , assume  $\|m_\eta - m_0\|_{2, \mu_0} \rightarrow 0$ . Then we have*

$$\sup_{\eta \in \mathcal{H}_n} |\text{Var}_0(\tilde{\chi}_0 + \vartheta_{n, \eta}) - \text{Var}_0(\tilde{\chi}_0)| \rightarrow 0.$$

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

$$\begin{aligned} |\text{Var}_0(\tilde{\chi}_0(Z) + \vartheta_{n, \eta}(Z)) - \text{Var}_0(\tilde{\chi}_0(Z))| &\leq |\mathbb{E}_0(\tilde{\chi}_0(Z) + \vartheta_{n, \eta}(Z))^2 - \mathbb{E}_0 \tilde{\chi}_0^2(Z)| + (\mathbb{E}_0 \vartheta_{n, \eta}^2(Z))^2 \\ &\leq 2\sqrt{\mathbb{E}_0 \tilde{\chi}_0^2(Z)} \sqrt{\mathbb{E}_0 \vartheta_{n, \eta}^2(Z)} + 2\mathbb{E}_0 \vartheta_{n, \eta}^2(Z) \\ &= 2(\sqrt{\text{Var}_0} \|\vartheta_{n, \eta}\|_{2, P_0} + \|\vartheta_{n, \eta}\|_{2, P_0}^2). \end{aligned}$$

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 assumption that the conditional variance of  $R_i(Y_i - m_0(X_i))$  is in  $L^2(\mu_0)$ , i.e.,  $\int \sigma_0^2(x) d\mu_0(x) < \infty$ . Thus, we obtain

$$\|\vartheta_{n,\eta}\|_{2,P_0} \lesssim \|m_\eta - m_0\|_{2,\mu_0} + \|\pi_n - \pi_0\|_{2,\mu_0} + \|m_\eta - m_0\|_{2,\mu_0} \|\pi_n - \pi_0\|_{2,\mu_0},$$

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

## C Potential Outcomes and Treatment Effects

As in Ray and van der Vaart (2020) and Breunig et al. (2025a), our robust procedure extends naturally to causal inference. Building on the connection between causal inference and missing-data problems (Imbens, 2004), the additional bias correction terms take the same form as in Breunig et al. (2025a, 2024). 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. We assume unconfoundedness, i.e., that  $(Y_i(1), Y_i(0))$  is independent of  $D_i$  given  $X_i$ , as well as the usual overlap condition. 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 parameters 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:

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

where its Riesz representer is

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

The procedure can be applied separately to the mean potential outcomes under treatment and control and then combined to obtain the ATE, as in Breunig et al. (2025a) for Gaussian process priors with prior and posterior adjustment. Here, we extend this con-

struction to BART-type priors. The corresponding algorithm for posterior draws of the ATE is given below. 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 representer as

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

Throughout this appendix, let  $\boldsymbol{\tau}[\cdot]$  denote the linear ATE score operator, defined for a candidate function  $m(\cdot, \cdot)$  and an observation  $z = (y, d, x^\top)^\top$  by

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

---

**Algorithm 3:** Posterior Computation of the ATE

---

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

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

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

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

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

and

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

with the ATE score operator  $\boldsymbol{\tau}[\cdot]$  defined in (C.1).

**Output:**  $\{\tilde{\tau}_\eta^s : s = 1, \dots, S\}$

---

Besides the ATE, the average treatment effect on the treated (ATT) is of policy interest, especially to economists (Heckman et al., 1997). In order to facilitate the comparison with Yiu et al. (2025), who also study the ATT, we outline our strategy to model the ATT, which is defined as follows:  $\theta_0 = \mathbb{E}_0[D_i(Y_i - m_0(0, X_i))]/\mathbb{E}_0[D_i]$ . Its one-step updated version is

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

where the Riesz representer is given by

$$\gamma_0^{\text{ATT}}(D, X) = \frac{D}{\bar{\pi}_0} - \frac{1-D}{\bar{\pi}_0} \frac{\pi_0(X)}{1-\pi_0(X)},$$

where  $\bar{\pi}_0 := \mathbb{E}_0[D_i] = \mathbb{E}_0[\pi_0(X)]$  denotes the scalar treatment proportion, to be distinguished from the propensity score function  $\pi_0(\cdot)$ . In this case, the estimated Riesz repre-

sender is

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

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

Algorithm 4 describes the  $K$ -fold cross-fitted version of RoBART for the ATE.

---

**Algorithm 4:** RoBART for the ATE with  $K$ -fold cross-fitting

---

(a) Take a  $K$ -fold random partition  $(I_k)_{k=1}^K$  of observation indices. The size of each fold  $I_k$  is  $n/K$ .

**for**  $k = 1, \dots, K$  **do**

(b) Use the subsample  $\{(D_i, X_i^\top)^\top\}_{i \notin I_k}$  to construct the estimator  $\hat{\pi}^{(-k)}(x)$ . Set  $\hat{\gamma}_{\text{ATE}}^{(-k)}(d, x) = d/\hat{\pi}^{(-k)}(x) - (1-d)/(1 - \hat{\pi}^{(-k)}(x))$ .

(c) Use the subsample  $\{(Y_i, D_i, X_i^\top)\}_{i \notin I_k}$  to construct the estimator  $\hat{m}^{(-k)}(d, x)$ .

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

(d) Generate the  $s$ -th draw of the posterior of  $(m_\eta(D_i, X_i))_{i \in I_k}$  using the BART prior and the subsample  $\{(Y_i, D_i, X_i^\top)\}_{i \in I_k}$ ; denote it as  $(m_\eta^{k,s}(D_i, X_i))_{i \in I_k}$ .

(e) Draw Bayesian-bootstrap weights  $W_i^{k,s} = e_i^{k,s} / \sum_{j \in I_k} e_j^{k,s}$  with  $e_i^{k,s} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ ,  $i \in I_k$ .

(f) Calculate the recentered posterior draw for the ATE:  $\check{\tau}_\eta^{k,s} := \tau_\eta^{k,s} - \hat{b}_\eta^{k,s}$ , where

$$\tau_\eta^{k,s} = \sum_{i \in I_k} W_i^{k,s} \left[ (m_\eta^{k,s}(1, X_i) - m_\eta^{k,s}(0, X_i)) + \hat{\gamma}_{\text{ATE}}^{(-k)}(D_i, X_i)(Y_i - m_\eta^{k,s}(D_i, X_i)) \right],$$

and

$$\hat{b}_\eta^{k,s} = \frac{K}{n} \sum_{i \in I_k} \boldsymbol{\tau}^{(-k)}[m_\eta^{k,s}(\cdot, \cdot) - \hat{m}^{(-k)}(\cdot, \cdot)](Z_i), \quad (\text{C.3})$$

where  $\boldsymbol{\tau}^{(-k)}[\cdot]$  is the ATE score operator (C.1) with  $\hat{\gamma}^{\text{ATE}}$  replaced by the fold-specific  $\hat{\gamma}_{\text{ATE}}^{(-k)}$ .

**Output:**  $\{\check{\tau}_\eta^s = \frac{1}{K} \sum_{k=1}^K \check{\tau}_\eta^{k,s} : s = 1, \dots, S\}$ .

---

## D Technical Results for 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 (Chipman et al., 2010). 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  $[\Omega]_j = \{x_j \in \mathbb{R} : (x_1, \dots, x_p)^\top \in \Omega\}$ . For a given split-net  $\mathcal{X}$ , we call each point  $x_i = (x_{i1}, \dots, x_{ip})^\top$  a split-candidate. For a given splitting coordinate  $j$  and a split-net  $\mathcal{X}$ , a split-point will be chosen from  $[\mathcal{X}]_j \cap \text{int}([\Omega]_j)$  to

dichotomize a box  $\Omega$ . Note that  $[\mathcal{X}]_j = \{x_{ij} \in [0, 1], i = 1, \dots, b_n\}$  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  $(x_{1j}, \dots, x_{b_n j})$ .

In this appendix, we prove the posterior rate of contraction for BART. The first rigorous analysis covering the random-design case appears in Jeong and Ročková (2023) for a more general piecewise heterogeneous anisotropic Hölder class. Our setup can be seen as a special case of Jeong and Ročková (2023) 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 Ghosal and van der Vaart (2017). 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.

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

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

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

$$\mathcal{M}_n := \mathcal{M}_{\bar{s}_n, \bar{L}_n, \bar{M}} := \bigcup_{\mathcal{E}: |S| \leq \bar{s}_n, L^t \leq \bar{L}_n, 1 \leq t \leq T} \mathcal{M}_{\mathcal{E}, \bar{M}}, \quad (\text{D.1})$$

where  $S \subseteq \{1, \dots, p\}$  is the set of splitting variables used by  $\mathcal{E}$  (with cardinality  $s = |S|$ ),  $\bar{L}_n \sim n \varepsilon_n^2 / \log n$ ,  $\bar{s}_n \sim n \varepsilon_n^2 / \log(p \vee n)$ , 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 Jeong and Ročková (2023).

**Lemma D.1.** *Under the assumptions stated in Theorem 8 for  $m_0 \in \mathcal{F}_p(\alpha, \mathcal{S}_0)$ , we have*

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

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

*Proof.* We prove the result for  $m_0$  in the Hölder continuous class and outline the 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.1 in the supplementary material of Xie and Xu (2020) 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  $B_n := \{m_\eta : \max\{K(f_{m_0}, f_{m_\eta}), V(f_{m_0}, f_{m_\eta})\} \leq \varepsilon_n^2\}$ . To check the posterior rate of contraction, we verify the high-level assumptions from Theorem 2.1 in Ghosal et al. (2000):

$$\Pi(m_\eta \in B_n) \geq c_1 \exp(-c_2 n \varepsilon_n^2), \quad (\text{D.2})$$

$$\log N(\varepsilon_n, \mathcal{M}_n, \rho_H) \leq c_3 n \varepsilon_n^2, \quad (\text{D.3})$$

$$\Pi(m_\eta \notin \mathcal{M}_n) \leq \exp(-c_4 n \varepsilon_n^2), \quad (\text{D.4})$$

for positive constants  $c_1, \dots, c_4$ .

Referring to the first inequality (D.2), it follows by the direct calculation of the normal likelihood:

$$B_n \supset \{m_\eta : \|m_\eta - m_0\|_{2, \mu_0} \leq C_1 \varepsilon_n\}, \quad (\text{D.5})$$

under Assumption 6 (iii). Then we construct an approximating ensemble denoted by  $\hat{\mathcal{E}} = \{\hat{\mathcal{T}}^1, \dots, \hat{\mathcal{T}}^T\}$ . By restricting the function space to the one constructed by  $\hat{\mathcal{E}}$ , we have

$$\Pi(m_\eta : \|m_\eta - m_0\|_{2, \mu_0} \leq C_1 \varepsilon_n) \geq \underbrace{\Pi(\hat{\mathcal{E}})}_{\textcircled{1}} \underbrace{\Pi(m_\eta \in \mathcal{M}_{\hat{\mathcal{E}}} : \|m_\eta - m_0\|_{2, \mu_0} \leq C_1 \varepsilon_n)}_{\textcircled{2}}. \quad (\text{D.6})$$

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  $\hat{\mathcal{T}}$  generating an approximating function  $m_{0, \hat{\mathcal{T}}, \hat{\beta}}$  such that  $\|m_0 - m_{0, \hat{\mathcal{T}}, \hat{\beta}}\|_\infty \lesssim \varepsilon_n$ . An approximating ensemble  $\hat{\mathcal{E}}$  can be constructed by setting  $\hat{\mathcal{T}}^1$  to be  $\hat{\mathcal{T}}$  and  $\hat{\mathcal{T}}^t = [0, 1]^p$  for  $t = 2, \dots, T$  so that the rest of the trees are root nodes with no splits. Under Assumption 6 (i)(ii), we have

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

where the last inequality follows from Lemma 4 in Jeong and Ročková (2023).

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

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

A similar calculation as in the proof of Lemma 5 in Jeong and Ročková (2023) leads to the lower bound of the right hand side of the above inequality. By Assumption 6 (iii) where the step-height priors are independent and compactly supported, we can simply

replace the Gaussian small-ball probability bounds with the volume of order  $C\nu_n^{\widehat{L}_*}$  with  $\nu_n = \varepsilon_n \sigma_{\max}^{-1}(A^{-1})/\sqrt{\widehat{L}_*}$  for  $\widehat{L}_* = \widehat{L} + T - 1$ , by lower bounding the volume of  $\{\mathcal{B} \in \mathbb{R}^{\widehat{L}_*} : \|\mathcal{B}\|_2 \leq C\nu_n\}$  for some non-singular matrix  $A$ . Hence, we have

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

The above two inequalities combined with (D.5) and (D.6) give us the desired bound in (D.2).

Next, we bound the entropy number in (D.3). 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 Ghosal and van der Vaart (2017). We proceed to upper bound  $\log N(\varepsilon, \mathcal{M}_n, \|\cdot\|_\infty)$  for the stronger supremum norm. Define  $\mathcal{E}_{S,L^1,\dots,L^T}$  as the collection of all tree ensembles  $\mathcal{E}$  with given  $S, L^1, \dots, L^T$ . By construction, we have the upper bound

$$N(\varepsilon_n, \mathcal{M}_n, \|\cdot\|_\infty) \leq \sum_{S: |S| \leq \bar{s}_n} \sum_{(L^1, \dots, L^T): L^t \leq \bar{L}_n} \sum_{\mathcal{E} \in \mathcal{E}_{S,L^1,\dots,L^T}} N(\varepsilon_n, \mathcal{M}_{\mathcal{E},\bar{M}}, \|\cdot\|_\infty),$$

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

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

Following the proof of Lemma 6 of Jeong and Ročková (2023), the covering number of the sieve set is thus upper bounded by

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

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

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

using  $\varepsilon_n = n^{-\alpha/(2\alpha+q_0)} \sqrt{\log n}$  and the choices of  $\bar{L}_n$  and  $\bar{s}_n$  given below the sieve set (D.1). 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 condition

(2.3) in Ročková and van der Pas (2020). It boils down to checking

$$\sum_{t=1}^T \Pi(L^t > \bar{L}_n) + \Pi(S : s > \bar{s}_n \mid L^t \leq \bar{L}_n, t = 1, \dots, T) \lesssim e^{-n\varepsilon_n^2},$$

with proper choices of  $\bar{L}_n$  and  $\bar{s}_n$  to be made later. Under Assumption 6 (i)(ii), one can apply Corollary 7 of Ročková and Saha (2020) to get

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

for  $\bar{L}_n = \lfloor Cn\varepsilon_n^2/\log n \rfloor$ . 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 Jeong and Ročková (2023) that

$$\Pi(S : s > \bar{s}_n \mid L^t \leq \bar{L}_n, t = 1, \dots, T) \lesssim e^{-n\varepsilon_n^2}.$$

This leads to the bound in (D.4) 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 for the construction of the approximating ensemble  $\hat{\mathcal{E}}$ . For each  $1 \leq t \leq T_0$ , there exists a tree partition  $\hat{\mathcal{T}}^t$  generating an approximating function  $m_{0,\hat{\mathcal{T}}^t,\hat{\beta}^t}$  such that  $\|m_0^t - m_{0,\hat{\mathcal{T}}^t,\hat{\beta}^t}\|_\infty \lesssim \varepsilon_n^t$ . An approximating ensemble  $\hat{\mathcal{E}}$  can be constructed by setting  $\hat{\mathcal{T}}^t$  as aforementioned with  $1 \leq t \leq T_0$ , and taking the rest of the 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 Jeong and Ročková (2023), and hence omitted.  $\square$

The next lemma bounds the local empirical process term without assuming the Donsker property. Compared with the more refined analysis in Han (2021), the following bound is not sharp. Nonetheless, it is sufficient to show the validity of our debiasing step, under the rate condition  $\sqrt{n}\varepsilon_n\|\hat{\pi}_n - \pi_0\|_\infty \rightarrow 0$ .

**Lemma D.2.** *Given the sieve set  $\mathcal{M}_n$  over the tree ensembles in (D.1) and the posterior rate of contraction  $\varepsilon_n$ , we have*

$$\mathbb{E}_0 \sup_{m_\eta \in \mathcal{M}_n(\varepsilon_n)} |\mathbb{G}_n(m_\eta - m_0)| \lesssim \sqrt{n}\varepsilon_n, \quad (\text{D.7})$$

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

*Proof.* The function 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 Han (2021):

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

The entropy integral term  $\int_{\iota}^{\sigma} \sqrt{\log N_{[]}(\varepsilon, \mathcal{M}_n, L^2(\mu_0))} d\varepsilon$  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  $\|m\|_{1,\mu_0} \leq \|m\|_{2,\mu_0}$  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. If  $m_1, \dots, m_N$  are the centers of  $\|\cdot\|_\infty$ -balls of radius  $\varepsilon$  that cover  $\mathcal{M}_n$ , one obtains the 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 van der Vaart and Wellner (1996). Given the entropy number in  $L^\infty$ -norm established in Lemma D.1, 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  $\mathbb{E}_0 \sup_{m_\eta \in \mathcal{M}_n(\varepsilon_n)} |\mathbb{G}_n(m_\eta)| \lesssim \sqrt{n\varepsilon_n}$ .  $\square$

## E Implementation of BART

We implement BART using the R package BART (Sparapani et al., 2021). 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 Sparapani et al. (2021) and Section 2 of Chipman et al. (2010) for further details. Some of the priors specified below differ from those in Assumption 6, and we make the discrepancies explicit at the end of this appendix.

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  $\sum_{t=1}^T \sum_{l=1}^{L^t} \beta_l^t \mathbb{1}\{x \in \Omega_l^t\}$ , where  $\mathcal{T}^t := (\Omega_1^t, \dots, \Omega_{L^t}^t)$  denotes the structure (partition) of the  $t$ -th tree of size  $L^t$ , and  $\beta^t := (\beta_1^t, \dots, \beta_{L^t}^t) \in \mathbb{R}^{L^t}$  denotes the step heights (leaf values) at  $L^t$  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  $\delta$  is internal is  $0.95(1 + \delta)^{-2}$ ; (ii) the splitting variable is chosen from  $p$  covariates with the proportion vector  $\vartheta = (\vartheta_1, \dots, \vartheta_p)^\top \sim \text{Dir}(\theta/p, \dots, \theta/p)$ , where the hyperparameter  $\theta$  is induced from  $\theta/(\theta + p) \sim \text{Beta}(0.5, 1)$ ; and (iii) the splitting value is chosen uniformly among the observed values of the splitting variable. Conditional on  $\mathcal{T}^t$ , each step-height  $\beta_l^t$  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  $X$ .

The posterior for the additive tree is drawn via a Gibbs sampler. Because the priors on the step-heights  $\beta^t$  and  $\sigma$  are conjugate, conditional on the tree structure  $\mathcal{T}^t$ , these parameters can be drawn directly:  $\beta_l^t$  from a normal posterior and  $\sigma$  from an inverse  $\chi^2$  distribution. Conversely, conditional on  $\beta^t$  and  $\sigma$ , the tree structure  $\mathcal{T}^t$  is updated via a

Metropolis–Hastings step detailed in Appendix C of Sparapani et al. (2021).

To implement the one-step corrected posterior method of Yiu et al. (2025) (One-step BART), we need posterior draws of the propensity score function  $\pi_\eta(x)$ . For this, we implement probit BART by calling the function `gbart` with `type = pbart` in R package BART. The algorithm employs the data-augmentation technique of Albert and Chib (1993), introducing a latent variable  $U_i$  with a truncated normal distribution given  $(R_i, X_i)$  and the additive tree model  $(\mathcal{T}^t, \beta^t)_{t=1}^T$ :

$$U_i \mid R_i, X_i, (\mathcal{T}^t, \beta^t)_{t=1}^T \sim \text{TN}_{A_i} \left( \sum_{t=1}^T \sum_{l=1}^{L^t} \beta_l^t \mathbb{1}\{X_i \in \Omega_l^t\}, 1 \right),$$

where  $A_i = (0, \infty)$  (positive real line) if  $R_i = 1$  and  $(-\infty, 0)$  (negative real line) if  $R_i = 0$ , and  $\text{TN}_A$  denotes a normal distribution truncated to  $A$ . Here the binary variable entering the probit step is the response indicator  $R_i$  in the missing data problem, and the treatment indicator  $D_i$  in the treatment effect setting of Appendix C; we write  $U_i$  for the latent variable to avoid a clash with the observation vector  $Z_i = (R_i Y_i, R_i, X_i^\top)^\top$ . The latent  $U_i$  are then treated as the continuous outcome for BART. Section 4 of Sparapani et al. (2021) gives more details on BART for binary outcomes.

## References

- J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.
- S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.
- P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler. DoubleML – An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3), 2024.
- C. Breunig, R. Liu, and Z. Yu. Semiparametric Bayesian difference-in-differences. *arXiv preprint arXiv:2412.04605*, 2024.
- C. Breunig, R. Liu, and Z. Yu. Double robust Bayesian inference on average treatment effects. *Econometrica*, 93(2):539–568, 2025a.
- C. Breunig, R. Liu, and Z. Yu. Supplement to ‘Double robust Bayesian inference on average treatment effects’. *Econometrica Supplemental Material*, 93, 2025b.
- I. Castillo and V. Ročková. Uncertainty quantification for Bayesian CART. *Annals of Statistics*, 49:3482–3509, 2021.
- K. C. G. Chan, S. C. P. Yam, and Z. Zhang. 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, 2016.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.
- V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90:1501–1535, 2022a.
- V. Chernozhukov, W. Newey, and R. Singh. De-biased machine learning of global and local parameters using regularized Riesz representers. *Econometrics Journal*, 25:576–601, 2022b.
- H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
- P. Ding. Replication data for: A first course in causal inference. Harvard Dataverse, 2023. Version V4. DOI: 10.7910/DVN/ZX3VEV.
- V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.
- R. M. Dudley. *Real Analysis and Probability*, volume 74. Cambridge University Press, 2002.
- S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.
- S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.
- S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28:500–531, 2000.
- P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15:965–1056, 2020.
- Q. Han. Set structured global empirical risk minimizers are rate optimal in general dimensions. *The Annals of Statistics*, 49(5):2642–2671, 2021.
- J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation

- estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605–654, 1997.
- D. A. Hirshberg and S. Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227, 2021.
- G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.
- S. Jeong and V. Ročková. The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension. *Journal of Machine Learning Research*, 24(337):1–65, 2023.
- M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 1991.
- A. R. Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113:626–636, 2018.
- A. R. Linero and Y. Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:1087–1110, 2018.
- K. Ray and A. van der Vaart. Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999–3020, 2020.
- K. Ray and A. van der Vaart. On the Bernstein-von Mises theorem for the Dirichlet process. *Electronic Journal of Statistics*, 15(1):2224–2246, 2021.
- Y. Ritov, P. J. Bickel, A. C. Gamst, and B. J. K. Kleijn. The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statistical Science*, 29(4):619–639, 2014.
- J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89: 846–866, 1994.
- V. Ročková. On semi-parametric inference for BART. *International Conference on Machine Learning*, pages 8137–8146, 2020.
- V. Ročková and E. Saha. On theory for BART. In *The 22nd international conference on artificial intelligence and statistics*, pages 2839–2848, 2020.
- V. Ročková and S. van der Pas. Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48:2108–2131, 2020.
- D. B. Rubin. Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.
- D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied

- statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.
- A. Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14:1139–1151, 1986.
- R. Sparapani, C. Spanbauer, and R. McCulloch. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97:1–66, 2021.
- A. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.
- A. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.
- F. Xie and Y. Xu. Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Analysis*, 15:159–186, 2020.
- A. Yiu, E. Fong, C. Holmes, and J. Rousseau. Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1025–1054, 2025.