Double Robust Bayesian Inference on Average Treatment Effects*

Christoph Breunig† Ruixuan Liu‡ Zhengfei Yu§

October 5, 2024

Abstract

We propose a double robust Bayesian inference procedure on the average treatment effect (ATE) under unconfoundedness. For our new Bayesian approach, we first adjust the prior distributions of the conditional mean functions, and then correct the posterior distribution of the resulting ATE. Both adjustments make use of pilot estimators motivated by the semiparametric influence function for ATE estimation. We prove asymptotic equivalence of our Bayesian procedure and efficient frequentist ATE estimators by establishing a new semiparametric Bernstein-von Mises theorem under double robustness; i.e., the lack of smoothness of conditional mean functions can be compensated by high regularity of the propensity score and vice versa. Consequently, the resulting Bayesian credible sets form confidence intervals with asymptotically exact coverage probability. In simulations, our method provides precise point estimates of the ATE through the posterior mean and credible intervals that closely align with the nominal coverage probability. Furthermore, our approach achieves a shorter interval length in comparison to existing methods. We illustrate our method in an application to the National Supported Work Demonstration following LaLonde [1986] and Dehejia and Wahba [1999].

KEYWORDS: Average treatment effects, unconfoundedness, double robustness, nonparametric Bayesian inference, Bernstein–von Mises theorem, Gaussian processes.

*We thank the anonymous reviewers, as well as Xiaohong Chen, Yanqin Fan, Essie Maasoumi, Yichong Zhang, and numerous seminar and conference participants for helpful comments and illuminating discussions. 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 21K01419.

†Department of Economics, University of Bonn. Email: cbreunig@uni-bonn.de

‡CUHK Business School, Chinese University of Hong Kong. Email: ruixuanliu@cuhk.edu.hk

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

1 Introduction

This paper proposes a double robust Bayesian approach for estimating the average treatment effect (ATE) under unconfoundedness, given a set of pretreatment covariates. Our new Bayesian procedure involves both prior and posterior adjustments. First, following Ray and van der Vaart [2020], we adjust the prior distributions of the conditional mean function using an estimator of the propensity scores. Second, we use this propensity score estimator together with a pilot estimator of the conditional mean to correct the posterior distribution of the ATE. The adjustments in both steps are closely related to the functional form of the semiparametric influence function for ATE estimation under unconfoundedness. They do not only shift the center but also change the shape of the posterior distribution. For our robust Bayesian procedure, we derive a new Bernstein–von Mises (BvM) theorem, which means that this posterior distribution, when centered at any efficient estimator, is asymptotically normal with the efficient variance in the semiparametric sense. The key innovation of our paper is that this result holds under double robust smoothness assumptions within the Bayesian framework.

Despite the recent success of Bayesian methods, the literature on ATE estimation is predominantly frequentist-based. For the missing data problem specifically, it was shown that conventional Bayesian approaches (i.e., using uncorrected priors) can produce inconsistent estimates, unless some unnecessarily strong smoothness conditions on the underlying functions were imposed; see the results and discussion in Robins and Ritov [1997] or Ritov et al. [2014]. Once the prior distribution was adjusted using some pre-estimated propensity score, Ray and van der Vaart [2020] recently established a novel semiparametric BvM theorem under weaker smoothness requirement for the propensity score function.[1] However, a minimum differentiability of order $p/2$ is still required for the conditional mean function in the outcome equation, where p denotes the dimensionality of covariates. In this paper, we are interested in Bayesian inference under double robustness that allows for a trade-off between the required levels of smoothness in the propensity score and the conditional mean functions.

Under double robust smoothness conditions, we show that Bayesian methods, which use propensity score adjusted priors as in Ray and van der Vaart [2020], satisfy the BvM Theorem only up to a "bias term" depending on the unknown true conditional mean and propensity score functions. In this paper, our robust Bayesian approach accounts

[1] Strictly speaking, the main objective in Ray and van der Vaart [2020] concerns the mean response in a missing data model, which is equivalent to observing one arm (either the treatment or control) of the causal setup.

for this bias term in the BvM Theorem by considering an explicit posterior correction. Both the prior adjustment and the posterior correction are based on functional forms that are closely related to the efficient influence function for the ATE, see Hahn [1998]. We show that the corrected posterior satisfies the BvM Theorem under double robust smoothness assumptions. Our novel procedure combines the advantages of Bayesian methodology with the robustness features that are the strengths of frequentist procedures. Our credible intervals are Bayesianly justifiable, as the uncertainty quantification is made conditional on the observed data [Rubin, 1984] and can be also interpreted as frequentist confidence intervals with asymptotically exact coverage probability. Our procedure is inspired by the double machine learning (DML), as well as the bias-corrected matching approach from Abadie and Imbens [2011], as our robustification of an initial procedure removes some non-negligible bias and remains asymptotically valid under weaker regularity conditions. While the main part of our theoretical analysis focuses on the ATE of binary outcomes, also considered by Ray and van der Vaart [2020], we outline extensions of our methodology to continuous and multinomial cases, as well as to other causal parameters.

In both simulations and an empirical illustration using the National Supported Work Demonstration data, we provide evidence that our procedure performs well compared to existing Bayesian and frequentist approaches. In our Monte Carlo simulations, we find that our method results in improved empirical coverage probabilities, while maintaining very competitive lengths for confidence intervals. This finite sample advantage is also observed over Bayesian methods that rely solely on prior corrections. In particular, we note that our approach leads to more accurate uncertainty quantification and is less sensitive to estimated propensity scores being close to boundary values.

While the BvM theorem for parametric Bayesian models is well-established [van der Vaart, 1998], the semiparametric version is still being studied very actively when nonparametric priors are used [Castillo, 2012, Castillo and Rousseau, 2015, Ray and van der Vaart, 2020]. To the best of our knowledge, our new semiparametric BvM theorem is the first one that possesses the double robustness property. Our paper is also connected to another active research area concerning Bayesian inference for parameters in econometric models, which is robust to partial or weak identification [Chen et al., 2018, Giacomini and Kitagawa, 2021, Andrews and Mikusheva, 2022]. The framework and the approach we take is different. Nonetheless, they share the same scope of tailoring the Bayesian inference procedure to new challenges in contemporary econometrics.

2 Setup and Implementation

This section provides the main setup of the average treatment effect (ATE). We motivate the new Bayesian methodology and detail the practical implementation.

2.1 Setup

We consider a family of probability distributions $\{P_\eta : \eta \in \mathcal{H}\}$ for some parameter space $\mathcal{H}$, where the (possibly infinite dimensional) parameter η characterizes the probability model. Let η_0 be the true value of the parameter and denote $P_0 = P_{\eta_0}$, which corresponds to the frequentist distribution generating the observed data.

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 now focus on the binary outcome case where both $Y_i(1)$ and $Y_i(0)$ take values in $\{1, 0\}$. An extension to multinomial or continuous outcomes is provided in Section 6. 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.[2] Let $\pi_0(x) = P_0(D_i = 1|X_i = x)$ denote the propensity score and $m_0(d, x) = P_0(Y_i = 1|D_i = d, X_i = x)$ the conditional mean. Suppose that the researcher observe an independent and identically distributed (i.i.d.) observations of $Z_i = (Y_i, D_i, X_i^\top)^\top$ for $i = 1, \ldots, n$. The joint density of Z_i is given by p_{π_0, m_0, f_0} where

$$p_{\pi,m,f}(z) = \pi(x)^d(1 - \pi(x))^{1-d} m(d, x)^y (1 - m(d, x))^{(1-y)} f(x). \tag{2.1}$$

The parameter of interest is the ATE given by $\tau_0 = \mathbb{E}_0[Y_i(1) - Y_i(0)]$, where $\mathbb{E}_0[\cdot]$ denotes the expectation under P_0. For its identification, we impose the following standard assumption of unconfoundedness and overlap [Rosenbaum and Rubin, 1984, Imbens, 2004, Imbens and Rubin, 2015].

Assumption 1. (i) $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid X_i$ and (ii) there exists $\bar{\pi} > 0$ such that $\bar{\pi} < \pi_0(x) < 1 - \bar{\pi}$ for all x in the support of F_0.

We introduce additional notations from the Bayesian perspective, following the similar setup from Ray and van der Vaart [2020]. For the purpose of assigning prior distributions to (π, m) in the Bayesian procedure, it is convenient to transform them by a link function. We make use of the Logistic function $\Psi(t) = 1/(1 + e^{-t})$ here. Specifically, we consider the

[2] If X_i does not have a density we can simply consider the conditional density of (Y_i, D_i) given $X_i = x$ instead of the joint density of (Y_i, D_i, X_i).

reparametrization of (π, m, f) given by $\eta = (\eta^\pi, \eta^m, \eta^f)$. We index the probability model as P_η, in line with the notation introduced at the first paragraph of this section, where

$$\eta^\pi = \Psi^{-1}(\pi), \quad \eta^m = \Psi^{-1}(m), \quad \eta^f = \log f. \tag{2.2}$$

Below, we write $m_\eta = \Psi(\eta^m)$, $\pi_\eta = \Psi(\eta^\pi)$, and $f_\eta = \exp(\eta^f)$ to make the dependence on η explicit. Given any prior on the triplet $(\eta^\pi, \eta^m, \eta^f)$, the Bayesian solution to the estimation and inference of the ATE is to obtain the posterior distribution of

$$\tau_\eta = \mathbb{E}_\eta \left[m_\eta(1, X) - m_\eta(0, X) \right], \tag{2.3}$$

where $\mathbb{E}_\eta[\cdot]$ denotes the expectation denotes the expectation under P_η. Our aim is to examine large-sample behavior of the posterior of τ_η and compare Bayesian methods with frequentist estimators based on the true probability distribution P_0. In the same vein, the true parameter of interest becomes $\tau_0 = \tau_{\eta_0}$.

The construction of our double robust Bayesian procedure in Section 2.2 has fundamental connection to the efficient influence function. For any generic component η, the efficient influence function (see Hahn [1998], Hirano et al. [2003]) is given by

$$\widetilde{\tau}_\eta(z) = m_\eta(1, x) - m_\eta(0, x) + \gamma_\eta(d, x)(y - m_\eta(d, x)) - \tau_\eta \tag{2.4}$$

for the Riesz representor γ_η, which is given by

$$\gamma_\eta(d, x) = \frac{d}{\pi_\eta(x)} - \frac{1 - d}{1 - \pi_\eta(x)}. \tag{2.5}$$

We write $\widetilde{\tau}_0 = \widetilde{\tau}_{\eta_0}$ and $\gamma_0 = \gamma_{\eta_0}$. Both the prior adjustment and posterior correction of our approach require a pilot estimator for γ_0. Under Assumption 1, the true Riesz representor γ_0 is well defined.

2.2 Double Robust Bayesian Point Estimators and Credible Sets

We build upon the ATE expression in (2.3) to develop our doubly robust inference procedure. Our approach is based on nonparametric prior processes for η^m and η^f. For the latter, we consider the Dirichlet process, which is a default prior on spaces of probability measures. This choice is also convenient for posterior computation via the Bayesian bootstrap; see Remark 2.1. For the former, we make use of Gaussian process priors, along with an adjustment that involves a preliminary estimator of γ_0. Gaussian process priors

are also closely related to spline smoothing (Wahba [1990]). Their posterior contraction properties (see Ghosal and Van der Vaart [2017]), together with excellent finite sample behavior (see [Rassmusen and Williams, 2006]), make Gaussian process priors popular in the related literature. Since τ_η does not depend on η^π, the specification of a prior on the propensity score is not required.

We consider pilot estimators $\widehat{\pi}$ of the propensity score π_0 and $\widehat{m}$ of the conditional mean function m_0, which both are based on an auxiliary sample. We consider a plug-in estimator for the Riesz representor γ_0 given by

$$\widehat{\gamma}(d,x) = \frac{d}{\widehat{\pi}(x)} - \frac{1-d}{1-\widehat{\pi}(x)}. \tag{2.6}$$

Below, we also make use of the notation $\widehat{\Gamma} = n^{-1}\sum_{i=1}^n |\widehat{\gamma}(D_i, X_i)|$, which we use for scale normalization in our prior adjustment below (see also Section 4.2 for more details). The use of an auxiliary data for pilot estimators simplifies the technical analysis related to the propensity score adjusted priors; see Ray and van der Vaart [2020]. Also, it provides an effective way to control some negligible higher-order terms, see our Lemma C.2 in the online supplement; cf. related discussion on the sample splitting in the DML type methods on Page C6 of Chernozhukov et al. [2018]. In practice, we use the full data twice and do not split the sample, as we have not observed any over-fitting or loss of coverage thereby. Algorithm 1 describes our double robust Bayesian inference procedure.

Algorithm 1 Double Robust Bayesian Procedure

Input: Data $Z_i = (Y_i, D_i, X_i^\top)^\top$ for $i = 1, \ldots, n$, number of posterior draws B, initial estimators $\widehat{\gamma}$ and $\widehat{m}$, and $\lambda \sim N(0, \sigma_n^2)$ where $\sigma_n = (\log n)/(\sqrt{n}\,\widehat{\Gamma})$.

Prior Specification:

(a) Select a Gaussian process prior W^m.

(b) Set prior for $m_\eta(d, X_i) = \Psi\left(\eta^m(d, X_i)\right)$, where $\eta^m(d, X_i) = W^m(d, X_i) + \lambda\,\widehat{\gamma}(d, X_i)$.

Posterior Computation:

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

(a) Obtain the posterior draw of $(m_\eta^s(d, X_i))_{i=1}^n$ using the adjusted prior.

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

(c) Calculate the corrected posterior draw for the ATE:

$$\check{\tau}_\eta^s = \tau_\eta^s - \widehat{b}_\eta^s, \tag{2.7}$$

$$\tau_\eta^s = \sum_{i=1}^n M_{ni}^s \big(m_\eta^s(1, X_i) - m_\eta^s(0, X_i)\big) \quad \text{and} \quad \widehat{b}_\eta^s = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}[m_\eta^s - \widehat{m}](Z_i), \tag{2.8}$$

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

end for

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

Given the draws from the corrected posterior calculated in Algorithm 1, we obtain the point estimate and credible set as follows. The Bayesian point estimator is $\overline{\tau}_\eta = \frac{1}{B}\sum_{s=1}^B \check{\tau}_\eta^s$. The $100 \cdot (1-\alpha)\%$ credible set for the ATE parameter τ_0 is given by

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

where $q_n(a)$ denotes the a-th quantile of $\{\check{\tau}_\eta^s : s = 1, \ldots, B\}$.

For the implementation of our pilot estimator $\widehat{\gamma}$ given in (2.6), we recommend using propensity scores estimated by the Logistic Lasso. For the implementation of the pilot estimator $\widehat{m}$, we adopt the posterior mean of the uncorrected Gaussian process priors (without adjustment), as in Ghosal and Roy [2006]. Section 4.2 provides details about our default prior choice and the construction of our adjusted prior. To approximate the posterior distribution, we make use of the Laplace approximation, but one can also resort to the Markov Chain Monte Carlo (MCMC) algorithms. The parameter σ_n controls the relative weight placed on the prior adjustment relative to the standard unadjusted prior on η^m (e.g., a Gaussian prior with a squared exponential covariance function). Regarding

the tuning parameter σ_n, we emphasize that our finite sample results are not sensitive to its choice, as we show in the online supplementary Appendix H.

Remark 2.1 (Bayesian bootstrap)**.** *Under unconfoundedness and the reparametrization in* (2.2)*, the ATE can be written as* $\tau_\eta = \int[\Psi(\eta^m(1,x)) - \Psi(\eta^m(0,x))]dF_\eta(x)$*. With independent priors on* η^m *and* F_η*, their posteriors also become independent. It is thus sufficient to consider the posterior for* η^m *and* F_η *separately. We place a Dirichlet process prior for* F_η *with the base measure to be zero. Consequently, the posterior law of* F_η *coincides with the Bayesian bootstrap [Rubin, 1981]; also see Chamberlain and Imbens [2003]. One key advantage of the Bayesian bootstrap is that it allows us to incorporate a broad class of data generating processes, whose posterior can be easily sampled. Replacing* F_η *by the standard empirical cumulative distribution function does not provide sufficient randomization of* F_η*, as it yields an underestimation of the asymptotic variance; see [Ray and van der Vaart, 2020, p. 3008]. In principle, one could consider other types of bootstrap weights; however, these generally do not correspond to the posterior of any given prior distribution.*

3 Main Theoretical Results

In this section, we derive the Bernstein-von Mises (BvM) theorem which establishes the asymptotic equivalance between our Bayesian procedure and the frequentist-type semiparametric efficient one for the ATE. We consider an asymptotically efficient estimator $\widehat{\tau}$ with the following linear representation:

$$\widehat{\tau} = \tau_0 + \frac{1}{n}\sum_{i=1}^{n} \widetilde{\tau}_0(Z_i) + o_{P_0}(n^{-1/2}), \tag{3.1}$$

where $\widetilde{\tau}_0 = \widetilde{\tau}_{\eta_0}$ is the efficient influence function in accordance with (2.4). Below, we denote $Z^{(n)} = (Z_1, \ldots, Z_n)$. By virtue of the BvM Theorem, two conditional distributions $\sqrt{n}(\tau_\eta - \widehat{\tau})|Z^{(n)}$ and $\sqrt{n}(\widehat{\tau} - \tau_\eta)|\eta = \eta_0$ are asymptotically equivalent under the underlying sampling distribution. Another important consequence of the BvM theorem is about the asymptotic normality and efficiency of the Bayesian point estimator. That is, $\sqrt{n}(\overline{\tau}_\eta - \tau_0)$ is asymptotically normal with mean zero and variance $\mathrm{v}_0 = \mathbb{E}_0[\widetilde{\tau}_0^2(Z_i)]$. Thus, $\overline{\tau}_\eta$ achieves the semiparametric efficiency bound of Hahn [1998].

3.1 Least Favorable Direction

Our prior correction through the Riesz representor γ_0 is motivated by the least favorable direction of Bayesian submodels. We first provide such least favorable calculations, which are closely linked to the semiparametric efficiency. Consider the one-dimensional submodel $t \mapsto \eta_t$ defined by the path

$$\pi_t(x) = \Psi(\eta^\pi + t\mathfrak{p})(x), \quad m_t(d,x) = \Psi(\eta^m + t\mathfrak{m})(d,x), \quad f_t(x) = \frac{f(x)e^{t\mathfrak{f}(x)}}{\int e^{t\mathfrak{f}(x)} f(x)dx}, \tag{3.2}$$

for a given direction $(\mathfrak{p}, \mathfrak{m}, \mathfrak{f})$ with $\int \mathfrak{f}(x)f(x)dx = 0$. The difficulty of estimating the parameter τ_{η_t} for the submodels depends on the direction $(\mathfrak{p}, \mathfrak{m}, \mathfrak{f})$. Among them, let $\xi_\eta = (\xi_\eta^\pi, \xi_\eta^m, \xi_\eta^f)$ be the *least favorable direction* that is associated with the most difficult submodel, i.e., the one that gives rise to the largest asymptotic optimal variance for estimating τ_{η_t}. Let p_{η_t} denote the joint density of Z depending on $\eta_t := (\pi_t, m_t, f_t)$. Taking derivative of the logarithmic density $\log p_{\eta_t}(z)$ with respect to t and evaluating at $t = 0$ gives the score operator:

$$B_\eta(\mathfrak{p}, \mathfrak{m}, \mathfrak{f})(z) = B_\eta^\pi \mathfrak{p}(z) + B_\eta^m \mathfrak{m}(z) + B_\eta^f \mathfrak{f}(z), \tag{3.3}$$

where $B_\eta^\pi \mathfrak{p}(z) = (d - \pi_\eta(x))\mathfrak{p}(x)$, $B_\eta^m \mathfrak{m}(z) = (y - m_\eta(d,x))\mathfrak{m}(d,x)$ and $B_\eta^f \mathfrak{f}(z) = \mathfrak{f}(x)$. The least favorable direction is defined as the solution ξ_η which solves the equation $B_\eta \xi_\eta = \widetilde{\tau}_\eta$, see Ghosal and Van der Vaart [2017, p.370]. We immediately obtain the following.

Lemma 3.1. *Consider the submodel* (3.2)*. Let Assumption 1 hold for P_η with any η under consideration, then the least favorable direction for estimating the ATE parameter in* (2.3) *is:*

$$\xi_\eta(d,x) = (0, \gamma_\eta(d,x), m_\eta(1,x) - m_\eta(0,x) - \tau_\eta), \tag{3.4}$$

where the Riesz representer γ_η is given in (2.5)*.*

Lemma 3.1 motivates the adjustment of the prior distribution as considered in our Bayesian procedure in Section 2.2. Our prior correction, which takes the form of the (estimated) least favorable direction, provides an exact invariance under a shift of nonparametric components in this direction. It provides additional robustness against posterior inaccuracy in the "most difficult direction", i.e., the one inducing the largest bias in the average treatment effects. We also note that Lemma 3.1 extends the result in Section 2.1 in Ray and van der Vaart [2020] for the missing data problem, which is equivalent as observing only one arm (either the treatment or control arm), to the context

of ATE estimation that involves both arms.

3.2 Assumptions for Inference

We now provide additional notations and assumptions. The posterior distribution plays an important role in the following analysis and is given by

$$\Pi\left((\pi,m)\in A, F\in B|Z^{(n)}\right)=\int_B \frac{\int_A\prod_{i=1}^n p_{\pi,m}(Y_i,D_i|X_i)d\Pi(\pi,m)}{\int\prod_{i=1}^n p_{\pi,m}(Y_i,D_i|X_i)d\Pi(\pi,m)}d\Pi(F|X^{(n)})$$

where $p_{\pi,m}$ denotes the conditional density of (Y_i,D_i) given X_i, given by (2.1) divided by the marginal density of X_i. We write $\mathcal{L}_\Pi(\sqrt{n}(\tau_\eta-\widehat{\tau})|Z^{(n)})$ for the marginal posterior distribution of $\sqrt{n}(\tau_\eta-\widehat{\tau})$. We focus on the case that η^π has a prior that is independent of the prior for (η^m,F). Because the factorization of the likelihood function (2.1) into (η^m,η^π,F) separately, so the posterior of η^π is also independent of the posterior for (η^m,F). Due to the fact that τ_η does not depend on η^π, it is unnecessary to further discuss a prior or posterior distribution on η^π.

We first introduce high-level assumptions and discuss primitive conditions for those in the next section. Below, we consider some measurable sets $\mathcal{H}_n^m$ of functions η^m such that $\Pi(\eta^m\in\mathcal{H}_n^m|Z^{(n)})\to_{P_0}1$. To abuse the notation for convenience, we also denote $\mathcal{H}_n=\{\eta:\eta^m\in\mathcal{H}_n^m\}$ when we index the conditional mean function m_η by its subscript η. We introduce the notation $\|\phi\|_{2,F_0}:=\sqrt{\int\phi^2(x)dF_0(x)}$ for all $\phi\in L^2(F_0):=\{\phi:\|\phi\|_{2,F_0}<\infty\}$, as well as the supremum norm $\|\cdot\|_\infty$.

Assumption 2. [Rates of Convergence] The estimators $\widehat{\pi}$ and $\widehat{m}$, which are based on an auxiliary sample independent of $Z^{(n)}$, satisfy $\|\widehat{\pi}-\pi_0\|_{2,F_0}=O_{P_0}(r_n)$ and for $d\in\{0,1\}$:

$$\|\widehat{m}(d,\cdot)-m_0(d,\cdot)\|_{2,F_0}=O_{P_0}(\varepsilon_n)\quad\text{and}\quad\sup_{\eta\in\mathcal{H}_n}\|m_\eta(d,\cdot)-m_0(d,\cdot)\|_{2,F_0}\leqslant\varepsilon_n,$$

where $\max\{\varepsilon_n,r_n\}\to0$ and $\sqrt{n}\,\varepsilon_n r_n\to0$. Further, $\|\widehat{\gamma}\|_\infty=O_{P_0}(1)$.

We adopt the standard empirical process notations as follows. For a function h of a random vector Z_i that follows distribution P_0, we let $P_0[h]=\int h(z)dP(z),\mathbb{P}_n[h]=n^{-1}\sum_{i=1}^n h(Z_i)$, and $\mathbb{G}_n[h]=\sqrt{n}\,(\mathbb{P}_n-P_0)\,[h]$. Below, we make use of the notations $\bar{m}_\eta(\cdot)=m_\eta(1,\cdot)-m_\eta(0,\cdot)$ and $\bar{m}_0(\cdot)=m_0(1,\cdot)-m_0(0,\cdot)$.

Assumption 3. [Complexity] For $\mathcal{G}_n=\{\bar{m}_\eta(\cdot):\eta\in\mathcal{H}_n\}$ it holds $\sup_{\bar{m}_\eta\in\mathcal{G}_n}|(\mathbb{P}_n-P_0)\bar{m}_\eta|=$

$o_{P_0}(1)$ and

$$\sup_{\eta \in \mathcal{H}_n} \left| \mathbb{G}_n \left[(\widehat{\gamma} - \gamma_0) (m_\eta - m_0) \right] \right| = o_{P_0}(1). \tag{3.5}$$

Recall the propensity score-dependent prior on m given by $m_\eta(\cdot) = \Psi(\eta^m(\cdot))$ where $\eta^m(\cdot) = W^m(\cdot) + \lambda \widehat{\gamma}(\cdot)$. The restriction on λ is made through its hyperparameter $\sigma_n > 0$. For two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \lesssim b_n$ if $\limsup_{n\to\infty}(a_n/b_n) < \infty$, and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Assumption 4. [Prior Stability] For $d \in \{0, 1\}$, $W^m(d, \cdot)$ is a continuous stochastic process independent of the normal random variable $\lambda \sim N(0, \sigma_n^2)$, where $\sigma_n \lesssim 1$, $n\sigma_n^2 \to \infty$ and that satisfies: (i) $\Pi\left(\lambda : |\lambda| \leqslant u_n \sigma_n^2 \sqrt{n} \mid Z^{(n)}\right) \to_{P_0} 1$, for some deterministic sequence $u_n \to 0$ and (ii) $\Pi\left((w, \lambda) : w + (\lambda + tn^{-1/2})\widehat{\gamma} \in \mathcal{H}_n^m \mid Z^{(n)}\right) \to_{P_0} 1$ for any $t \in \mathbb{R}$.

Discussion of Assumptions: Assumption 2 imposes sufficiently fast convergence rates for the pilot estimators for the conditional mean function m_0 and the propensity score π_0. In practice, one can explore the recent proposals from Chernozhukov et al. [2020, 2022]. Note that one can also use Bayesian point estimators such as the posterior mean of the Gaussian process for $\widehat{m}$ and $\widehat{\pi}$. The posterior convergence rate for the conditional mean m_η can be derived in the same spirit of Ray and van der Vaart [2020]. The rate restriction is more likely to be satisfied if one function is easier to estimate, which resembles Theorem 1 conditions (i) and (ii) of Farrell [2015]. Remark 4.1 illustrates that under classical smoothness assumptions, this condition is less restrictive than the method of Ray and van der Vaart [2020] or other approaches for semiparametric estimation of ATEs as found in Chen et al. [2008] or Farrell et al. [2021]. Assumption 4 incorporates Conditions (3.9) and (3.10) from Theorem 2 in Ray and van der Vaart [2020], and it is imposed to check the invariance property of the adjusted prior distribution. These restrictions are mild and extend beyond the Gaussian processes considered in Section 4 for concreteness.

Assumption 3 restricts the functional class $\mathcal{G}_n$ to form a P_0-Glivenko-Cantelli class; see Section 2.4 of van der Vaart and Wellner [1996]. This imposes a new stochastic equicontinuity condition as in (3.5) on the product structure involving $\widehat{\gamma}$ and m_η, which further relaxes the corresponding one, namely $\sup_{\eta \in \mathcal{H}_n^m} \mathbb{G}_n [m_\eta - m_0] = o_{P_0}(1)$, from Ray and van der Vaart [2020]. In the next section, we demonstrate that our formulation allows for double robustness under Hölder classes (see Remark 4.1). Hence, the complexity of the functional class $(m_\eta - m_0)$ can be compensated by sufficient regularity of the corresponding Riesz representor and vice versa. In essence, a condition similar to our Assumption 3 is also used in the frequentist literature; see Section 2 of Benkeser et al. [2017]. Nonetheless,

the technical argument differs substantially from the frequentist's study, because we mainly need the condition (3.5) to control changes in the likelihood under perturbations along the estimated and true least favorable directions. This is unique to Bayesian analysis with nonparametric priors.

3.3 A Double Robust Bernstein-von Mises Theorem

We now establish a new Bernstein–von Mises theorem, which establishes the asymptotic normality of the posterior distribution, modulo a "bias term". In a next step, we show that posterior correction, as proposed in our procedure, eliminates this "bias term". This asymptotic equivalence result is established using the bounded Lipschitz distance. For two probability measures P, Q defined on a metric space $\mathcal{X}$, we define the bounded Lipschitz distance as

$$d_{BL}(P,Q) = \sup_{f \in BL(1)} \left| \int_{\mathcal{Z}} f(dP - dQ) \right|, \tag{3.6}$$

where

$$BL(1) = \left\{ f : \mathcal{Z} \mapsto \mathbb{R}, \sup_{z \in \mathcal{Z}} |f(z)| + \sup_{z \neq z'} \frac{|f(z) - f(z')|}{\|z - z'\|_{\ell_2}} \leqslant 1 \right\}.$$

Here, $\|\cdot\|_{\ell_2}$ denotes the vector ℓ_2 norm.

Below is our main statement about the asymptotic behavior of the posterior distribution of τ_η. As in the modern Bayesian paradigm, the exact posterior is rarely of closed-form, and one needs to rely on certain Monte Carlo simulations, such as the implementation procedure in Section 2.2, to approximate this posterior distribution, as well as the resulting point estimator and credible set.

Theorem 3.1. *Let Assumptions 1–4 hold. Then we have*

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

where $b_{0,\eta} := \mathbb{P}_n[\gamma_0(m_0 - m_\eta) - (\bar{m}_0 - \bar{m}_\eta)]$.

We emphasize that the above BvM theorem is not feasible for applications, because it depends on the "bias term" $b_{0,\eta}$, which depends on the unknown conditional mean m_0. Nonetheless, it provides an important theoretical benchmark. One can follow the existing literature on semiparametric BvM theorems to impose the so-called "no-bias" condition, but this generally leads to strong smoothness restrictions and may not be satisfied when the dimensionality of covariates is large relative to the smoothness properties of the underlying functions; see the discussion on page 395 of van der Vaart [1998].

This "bias term" in our context consists of two key components, with the first involving unknown true functions and the second depending on the posterior of m_η. We consider pilot estimators for the unknown functional parameters in $b_{0,\eta}$. The correction term $\widehat{b}_\eta$, as introduced in (2.8), results in a feasible Bayesian procedure that satisfies the BvM theorem under double robustness, as demonstrated below.

Theorem 3.2. *Let Assumptions 1–4 hold. Then we have*

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

We now show how Theorem 3.2 can provide frequentist justification of Bayesian methods to construct the point estimator and the confidence sets. Recall that $\overline{\tau}_\eta$ represents the posterior mean. Introduce a Bayesian credible set $\mathcal{C}_n(\alpha)$ for τ_η, which satisfies $\Pi(\tau_\eta \in \mathcal{C}_n(\alpha)|Z^{(n)}) = 1 - \alpha$ for a given nominal level $\alpha \in (0, 1)$. The next result shows that $\mathcal{C}_n(\alpha)$ also forms a confidence interval in the frequentist sense for the ATE parameter whose coverage probability under P_0 converges to $1 - \alpha$.

Corollary 3.1. *Let Assumptions 1–4 hold. Then under P_0, we have*

$$\sqrt{n}\,(\overline{\tau}_\eta - \tau_0) \Rightarrow N(0, \mathrm{v}_0). \tag{3.7}$$

Also, for any $\alpha \in (0, 1)$ we have $P_0\big(\tau_0 \in \mathcal{C}_n(\alpha)\big) \to 1 - \alpha$.

To the best of our knowledge, this is the first BvM theorem that entails the double robustness. We discuss the distinction with Theorem 2 in Ray and van der Vaart [2020]. Their work laid the theoretical foundation that supports the usefulness of propensity score in Bayesian analysis of the ATE. They showed that propensity score adjustment via priors can allow for weak regularity conditions on the propensity score function, coining the corresponding property as the single robustness. Our analysis differs from Ray and van der Vaart [2020] in two crucial ways. First, we improve on their Lemma 3 by showing that it is possible to verify the prior stability condition for propensity score-adjusted priors under the product structure in Assumption 3, modulo the "bias term" $b_{0,\eta}$. This separation is essential to identify the source of the restrictive condition, such as the Donsker property on m_η, which is mainly used to eliminate $b_{0,\eta}$. Second, our proposal introduces an explicit debiasing step, borrowing key insights from recent developments in the DML literature.

Remark 3.1 (Connection with frequentist robust estimation)**.** *In our BvM theorem, we do not restrict the centering estimator $\widehat{\tau}$, as long as it admits the linear representation as*

in (3.1)*. A popular frequentist estimator for the ATE that achieves double robustness is*

$$\widehat{\tau} = n^{-1}\sum_{i=1}^{n}\big(\widehat{m}(1,X_i)-\widehat{m}(0,X_i)\big) + n^{-1}\sum_{i=1}^{n}\widehat{\gamma}(D_i,X_i)\big(Y_i-\widehat{m}(D_i,X_i)\big) \tag{3.8}$$

based on frequentist-type pilot estimators $\widehat{m}$ of the conditional mean function m_0 and $\widehat{\gamma}$ of the Riesz representer γ_0; see Robins and Rotnitzky [1995] and more recently Chernozhukov et al. [2020, 2022]. The double robust or double machine learning estimator (3.8) *recenters the plug-in type functional by an explicit correction factor that depends on the Riesz representor.*[3] *Our main result establishes the asymptotic equivalence of our estimator and* (3.8)*.This not only offers frequentist validity to our Bayesian procedure but also provides doubly robust frequentist methods with a Bayesian interpretation.*

Remark 3.2 (Parametric Bayesian Methods)**.** *A couple of recent papers propose doubly robust Bayesian recipes for ATE inference, under parametric model restrictions. Saarela et al. [2016] considered a Bayesian procedure based on an analog of the double robust frequentist estimator given in Equation* (3.8)*, replacing the empirical measure with the Bayesian bootstrap measure. However, there was no formal BvM theorem presented therein. Another recent paper by Yiu et al. [2020] explored Bayesian exponentially tilted empirical likelihood with a set of moment constraints that are of a double-robust type. They proved a BvM theorem for the posterior constructed from the resulting exponentially tilted empirical likelihood under parametric specifications. Luo et al. [2023] provided Bayesian results for ATE estimation in a partial linear model, which implies homogeneous treatment effects. They also assign parametric priors to the propensity score. Their BvM Theorem allows for misspecification only in a parametric nonlinear component of the outcome equation. It is not clear how to extend their analysis to incorporate flexible nonparametric modeling strategies.*

4 Illustration with Gaussian Process Priors

We illustrate the general methodology by placing the Gaussian process prior on $\eta^m(d,\cdot)$ in relation to the conditional mean functions for $d \in \{0,1\}$. The Gaussian process regression has been extensively used among the machine learning community [Rassmusen and Williams, 2006, Murphy, 2023], and started to gain popularity among economists [Kasy, 2018]. Our study further strengthened the appealing features of this modern Bayesian

[3]Another popular method in the statistics literature is the targeted learning approach [Van der Laan and Rose, 2011, Benkeser et al., 2017].

toolkit. We provide primitive conditions used in our main results in the previous section. In addition, we provide details on the implementation using Gaussian process priors and discuss the data-driven choices of tuning parameters.

4.1 Inference Based on Gaussian Process Priors

Let $(W(t) : t \in \mathbb{R}^p)$ be a centered, homogeneous Gaussian random field with covariance function of the following form $\mathbb{E}[W(s)W(t)] = \phi(s-t)$, for a given continuous function $\phi : \mathbb{R}^p \mapsto \mathbb{R}$. We consider $W(t)$ as a Borel measurable map in the space of continuous functions on $[0,1]^p$, equipped with the supremum norm $\|\cdot\|_\infty$. The Gaussian process is completely determined by the covariance function. For example, the covariance function of the squared exponential process is given by $\mathbb{E}[W(s)W(t)] = \exp(-\|s-t\|_{\ell_2}^2)$, as its name suggests. In this section, we focus on the squared exponential process prior, which is one of the most commonly used priors in applications; see Rassmusen and Williams [2006] and Murphy [2023]. We also consider a rescaled Gaussian process $\big(W(a_n t) : t \in [0,1]^p\big)$. Intuitively speaking, a_n^{-1} can be thought as a bandwidth parameter. For a large a_n (or equivalently a small bandwidth), the prior sample path $t \mapsto W(a_n t)$ is obtained by shrinking the long sample path $t \mapsto W(t)$. Thus, it employs more randomness and becomes suitable as a prior model for less regular functions, see van der Vaart and van Zanten [2008, 2009].

Below, $\mathcal{C}^{s_m}([0,1]^p)$ denotes a Hölder space with the smoothness index s_m. Specifically, we illustrate our theory with the case where $m_0(d,\cdot) \in \mathcal{C}^{s_m}([0,1]^p)$ for $d \in \{0,1\}$. Given such a Hölder-type smoothness condition, we choose

$$a_n \sim n^{1/(2s_m+p)} (\log n)^{-(1+p)/(2s_m+p)}, \tag{4.1}$$

which coincides (up to some logarithm factor) with the minimax posterior contraction rate for the conditional mean function $m_\eta(d,\cdot)$ given by $\varepsilon_n = n^{-s_m/(2s_m+p)} (\log n)^{s_m(1+p)/(2s_m+p)}$; see Section 11.5 of Ghosal and Van der Vaart [2017]. The particular choice of a_n mimics the corresponding kernel bandwidth based on any kernel smoothing method. Other choices of a_n will generally make the convergence rate slower. Nonetheless, as long as the propensity score is estimated with a sufficiently fast rate, our BvM theorem still holds.

Proposition 4.1 (Squared Exponential Process Priors)**.** *The estimator $\widehat{\gamma}$ satisfies $\|\widehat{\gamma}\|_\infty = O_{P_0}(1)$ and $\|\widehat{\gamma} - \gamma_0\|_\infty = O_{P_0}\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$ for some $s_\pi > 0$. Suppose $m_0(d,\cdot) \in \mathcal{C}^{s_m}([0,1]^p)$ for $d \in \{0,1\}$ and some $s_m > 0$ with $\sqrt{s_\pi\, s_m} > p/2$. Also, $\|\widehat{m}(d,\cdot) - m_0(d,\cdot)\|_{2,F_0} = O_{P_0}\big((n/\log n)^{-s_m/(2s_m+p)}\big)$. Consider the propensity score-dependent prior on m given by $m(d,x) = \Psi\left(W^m(d,x) + \lambda\,\widehat{\gamma}(d,x)\right)$, where $W^m(d,\cdot)$ is the rescaled squared*

exponential process for $d \in \{0,1\}$, with its rescaling parameter a_n of the order in (4.1) *and* $(n/\log n)^{-s_m/(2s_m+p)} \lesssim u_n\sigma_n$ *for some deterministic sequence* $u_n \to 0$, *and* $\sigma_n \lesssim 1$. *Then, the corrected posterior distribution for the ATE satisfies Theorem 3.1.*

Remark 4.1 (Double Robust Hölder Smoothness)**.** *Proposition 4.1 requires* $\sqrt{s_\pi\, s_m} > p/2$, *which represents a trade-off between the smoothness requirement for* m_0 *and* π_0. *This encapsulate the double robustness; i.e., a lack of smoothness of the conditional mean function* m_0 *can be mitigated by exploiting the regularity of the propensity score and vice versa. Referring to the Hölder class* $\mathcal{C}^{s_m}([0,1]^p)$, *its complexity measured by the bracketing entropy of size* ε *is of order* $\varepsilon^{-2\upsilon}$ *for* $\upsilon = d/(2s_m)$. *One can show that the key stochastic equicontinuity assumption in Ray and van der Vaart [2020], i.e., their condition (3.5), is violated by exploring the Sudkov lower bound [Han, 2021] when* $\upsilon > 1$ *or equivalently when* $s_m < p/2$. *In contrast, our framework accommodates this non-Donsker regime as long as* $\sqrt{s_\pi\, s_m} > p/2$, *which enables us to exploit the product structure and a fast convergence rate for estimating the propensity score. Our methodology is not restricted to the case where propensity score belongs to a Hölder class per se. For instance, under a parametric restriction (such as in logistic regression) or an additive model with unknown link function, the possible range of the posterior contraction rate* ε_n *for the conditional mean function can be substantially enlarged. In the case* $s_m > p/2$, *the bias term becomes asymptotically negligible, i.e.,* $b_{0,\eta} = o_{P_0}(n^{-1/2})$. *This allows for smoothness robustness only with respect to the propensity score and is also known as single robustness. In this case, no posterior correction is required, see Ray and van der Vaart [2020].*

4.2 Implementation of Gaussian Process Priors

We provide details on the Gaussian process prior placed on $\eta^m(d,x)$ and its posterior computation. Algorithm 1 sets the adjusted prior as $\eta^m(d,x) = W^m(d,x) + \lambda\,\widehat{\gamma}(d,x)$: the first component $W^m(d,x)$ is a zero-mean Gaussian process with the commonly used squared exponential (SE) covariance function [Rassmusen and Williams, 2006, p.83]. That is, $K\left((d,x),(d',x')\right) := \nu^2 \exp\left(-a_{0n}^2(d-d')^2/2 - \sum_{l=1}^p a_{ln}^2(x_l - x_l')^2/2\right)$ where the hyperparameter ν^2 is the kernel variance and $a_{0n},\ldots,a_{pn}$ are rescaling parameters that reflect the relevance of treatment and each covariate in predicting η^m. They are selected by maximizing the marginal likelihood. Conditional on the data used to obtain the propensity score estimator $\widehat{\pi}$, the prior for η^m has zero mean and the covariance kernel K^c including an additional term based on the estimated Riesz representer $\widehat{\gamma}$ is given by $K^c\left((d,x),(d',x')\right) = K\left((d,x),(d',x')\right) + \sigma_n^2\widehat{\gamma}(d,x)\,\widehat{\gamma}(d',x')$, cf. related constructions from

Ray and Szabó [2019] and Ray and van der Vaart [2020]. The parameter σ_n, representing the standard deviation of λ, controls the weight of the prior adjustment relative to the standard Gaussian process. The choice $\sigma_n = (\log n)/(\sqrt{n}\,\widehat{\Gamma})$ in Algorithm 1 satisfies the rate condition in Assumption 4 with probability approaching one. It is similar to the choice suggested by Ray and Szabó [2019, page 6], which is proportional to $1/(\sqrt{n}\,\widehat{\Gamma})$. The factor $\widehat{\Gamma}$ normalizes the second term (adjustment term) of K^c to have the same scale as the unadjusted covariance K. Supplementary Appendix H shows that the finite sample performance of the double robust Bayesian approaches remains stable across different choices of σ_n.

Utilizing Gaussian process priors with zero mean and covariance function K^c, and incorporating the available data, we generate posterior draws of the vector $[\eta^m(d, X_1), \cdots, \eta^m(d, X_n)]^\top$ for $d \in \{0, 1\}$. This can be achieved through the Laplace approximation method detailed in online Appendix G.

For the implementation of the pilot estimator $\widehat{\gamma}$ given in (2.6), we recommend Logistic Lasso for the propensity scores, with the penalty parameter chosen by cross-validation [Friedman et al., 2010]. As a pilot estimator $\widehat{m}$ in Algorithm 1 for posterior correction, we use the uncorrected posterior mean $\sum_{s=1}^{B} \tau_\eta^s / B$, where τ_η^s is calculated following the first expression in (2.8), but using Gaussian process priors without adjustment. When the rescaling parameter a_n is as stated in Proposition 4.1, the convergence rate of $\widehat{m}$ is $O_{P_0}\big((n/\log n)^{-s_m/(2s_m+p)}\big)$. This can be shown by combining Theorems 11.22, 11.55 and 8.8 from Ghosal and Van der Vaart [2017].

5 Numerical Results

In this section, we apply our method to one version of the Lalonde–Dehejia–Wahba data that contains a treated sample of 185 men from the National Supported Work (NSW) experiment and a control sample of 2490 men from the Panel Study of Income Dynamics (PSID). The data has been used by LaLonde [1986], Dehejia and Wahba [1999], Abadie and Imbens [2011], and Armstrong and Kolesár [2021], among others. We refer readers to LaLonde [1986], and Dehejia and Wahba [1999] for reviews of the data.[4]

[4]The data is available on Dehejia's website: http://users.nber.org//∼ rdehejia/nswdata2.html.

5.1 Simulations

In this section, we consider a simulation study where the observations are randomly drawn from a large sample generated by the Wasserstein Generative Adversarial Networks (WGAN) method from the the job-training real data, see ath [2024]. We view their simulated data as the population and repeatedly draw our simulation samples (each consisting of 185 treated observations and 2490 control observations) for each of the 1000 Monte Carlo replications. We slightly depart from previous studies by focusing on a binary outcome Y: the employment indicator for the year 1978, which is defined as an indicator for positive earnings. The treatment D is the participation in the NSW program. We are interested in the average treatment effect of the NSW program on the employment status. For the set of covariates, we follow Abadie and Imbens [2011] and include nine variables: age, education, black, Hispanic, married, earnings in 1974, earnings in 1975, unemployed in 1974, and unemployed in 1975. We implement our double robust Bayesian method (DR Bayes) following Algorithm 1, using $B = 5000$ posterior draws and the pilot estimator $\widehat{\gamma}$ and $\widehat{m}$, as detailed at the end of Section 4.2. We compare DR Bayes to two other Bayesian procedures: First, we consider the prior adjusted Bayesian method (PA Bayes) proposed by Ray and van der Vaart [2020], which constructs the point estimate and credible interval based on τ_η^s in (2.8). Second, we examine an unadjusted Bayesian method (Bayes) which is also based on τ_η^s but is generated using Gaussian process priors without the adjustment term $\lambda\widehat{\gamma}$.

We also compare our method to frequentist estimators. Match/Match BC corresponds to the nearest neighbor matching estimator and its bias-corrected version, which adjusts for differences in covariate values through regression by Abadie and Imbens [2011]. DR TMLE corresponds to the doubly robust targeted maximum likelihood estimator by Benkeser et al. [2017]. DML refers to the double/debiased machine learning estimator from Chernozhukov et al. [2017], where the nuisance functions π_0 and m_0 are estimated using random forests (which outperformed DML combined with other nuisance function estimators, such as Lasso, in our simulation setup). Since the job-training data contains a sizable proportion of units with propensity score estimates very close to 0 and 1, we follow Crump et al. [2009] and discard observations with the estimated propensity score outside the range $[t, 1-t]$, with the trimming threshold $t \in \{0.10, 0.05, 0.01\}$.[5]

[5]Crump et al. [2009] suggested a simple rule of thumb with a threshold of $t = 0.10$, while ath [2024] used $t = 0.05$. Applying the optimal trimming rule proposed by Crump et al. [2009] to our simulated samples yields an average optimal trimming threshold 0.073.

Table 1: Simulation results using WGAN-generated data. Trimming is based on $\widehat{\pi} \in [t, 1-t]$ and $\bar{n}$ = the average sample size after trimming. CP = coverage probability of 95% credible/confidence interval, CIL = average length of the 95% credible/confidence interval.

Methods	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
	$t = 0.10(\bar{n} = 240)$			$t = 0.05(\bar{n} = 363)$			$t = 0.01(\bar{n} = 664)$		
Bayes	-0.040	0.683	0.147	-0.010	0.841	0.149	-0.006	0.911	0.120
PA Bayes	-0.008	0.981	0.260	0.033	0.949	0.254	0.047	0.897	0.308
DR Bayes	-0.024	0.983	0.223	0.014	0.970	0.221	0.023	0.952	0.258
Match	0.027	0.933	0.334	0.048	0.908	0.323	0.033	0.965	0.323
Match BC	0.040	0.880	0.347	0.065	0.816	0.334	0.083	0.804	0.339
DR TMLE	0.015	0.832	0.300	0.039	0.746	0.282	0.039	0.668	0.242
DML	0.045	0.927	0.524	0.052	0.870	0.393	0.054	0.918	0.522

Table 1 presents the finite sample performance of the Bayesian and frequentist methods mentioned above. We use the full data twice in computing the prior/posterior adjustments and the posterior distribution of the conditional mean function. Online Appendix H reports the performance of DR Bayes using sample-splitting, which results in similar coverage but a larger credible interval length due to the halved sample size.

Concerning the Bayesian methods for estimating the ATE, Table 1 reveals that unadjusted Bayes yields to highly inaccurate coverage except for the case with trimming constant $t = 0.01$. If the prior is corrected using the propensity score adjustment, the results improve significantly. Nevertheless, our DR Bayes method demonstrates two further improvements: First, DR Bayes leads to smaller average confidence lengths in each case while simultaneously improving the coverage probability. This can be attributed to a reduction in bias and/or more accurate uncertainty quantification via our posterior correction. Second, when the trimming threshold is small (i.e., $t = 0.01$), propensity score estimators can be less accurate, leading to reduced coverage probabilities of PA Bayes. Our double robust Bayesian method, on the other hand, is still able to provide accurate coverage probabilities. In other words, DR Bayes exhibits more stable performance than PA Bayes with respect to the trimming threshold.[6]

Our DR Bayes also exhibits encouraging performances when compared to frequentist methods. It provides a more accurate coverage than bias-corrected matching, DR TMLE

[6]In additional simulations without trimming ($t = 0$), we find that all double robust methods, including DR Bayes, substantially under-cover and/or inflate the length of their confidence intervals. This is consistent with Crump et al. [2009], who point out that propensity score estimates close to the boundaries tend to induce substantial bias and large variances in estimating the ATE. We also note that unadjusted Bayes severely undercovers in this case.

and DML. Compared with the matching estimator that exhibits a similarly good coverage performance, DR Bayes yields considerably shorter credible intervals.

5.2 An Empirical Illustration

We apply the Bayesian and frequentist methods considered above to the real job-training data. Similar to the simulation exercise, we consider a varying choice of the threshold constant $t \in \{0.10, 0.05, 0.01\}$.[7] The ATE point estimates and confidence intervals are presented in Table 2. As a benchmark, the experimental data that uses both treated and control groups in NSW ($n = 445$) yields an ATE estimate (treated-control mean difference) of 0.111 with a 95% confidence interval $[0.026, 0.196]$. As we see from Table 2, the unadjusted Bayesian method yields larger estimates. The adjusted Bayesian methods (PA and DR Bayes), on the other hand, produce estimates comparable to the experimental estimate. PA Bayes finds that the job training program enhanced the employment by 9.0% to 17.0% across different trimming thresholds, and DR Bayes estimates the effect from 12.1% to 18.4%. Among frequentist estimators, the matching estimator and its bias-corrected version produce similar estimates as PA and DR Bayes, but with wider confidence intervals. DR TMLE produces negative estimates for $t = 0.10$ when all other estimates are positive. For $t = 0.10$ and 0.05, DML yields similar point estimates as PA and DR Bayes, but with less estimation precision. In the case $t = 0.01$ where the overlapping condition is closer to violation, however, its point estimate and confidence interval length become considerably larger than other methods.

Table 2: Estimates of ATE for the job-training data: trimming based on $\widehat{\pi} \in [t, 1-t]$, $\bar{n}$ = sample size after trimming. ATE = point estimate, 95% CI = 95% credible/confidence interval, CIL = 95% credible/confidence interval length.

Methods	$t = 0.10(\bar{n} = 245)$			$t = 0.05(\bar{n} = 398)$			$t = 0.01(\bar{n} = 740)$		
	ATE	95% CI	CIL	ATE	95% CI	CIL	ATE	95% CI	CIL
Bayes	0.213	[0.120, 0.301]	0.181	0.214	[0.132, 0.292]	0.161	0.198	[0.140, 0.251]	0.112
PA Bayes	0.158	[0.019, 0.288]	0.270	0.170	[0.045, 0.281]	0.236	0.090	[-0.078, 0.233]	0.311
DR Bayes	0.178	[0.061, 0.293]	0.231	0.184	[0.064, 0.294]	0.230	0.121	[-0.031, 0.250]	0.281
Match	0.188	[0.022, 0.355]	0.333	0.140	[-0.029, 0.309]	0.338	0.079	[-0.111, 0.269]	0.380
Match BC	0.157	[-0.006, 0.321]	0.327	0.145	[-0.021, 0.310]	0.331	0.180	[-0.004, 0.365]	0.369
DR TMLE	-0.023	[-0.171, 0.125]	0.296	0.073	[-0.074, 0.220]	0.294	0.071	[-0.146, 0.289]	0.435
DML	0.172	[0.018, 0.327]	0.308	0.150	[-0.010, 0.310]	0.320	0.258	[-0.183, 0.699]	0.882

[7] Applying the optimal trimming rule proposed by Crump et al. [2009] yields an optimal threshold of 0.064.

6 Extensions

This section extends the binary variable Y to encompass general cases, including continuous, counting, and multinomial outcomes. First, we examine the class of single-parameter exponential families, where the conditional density function is solely determined by the nonparmatric conditional mean function. This covers continuous outcomes and counting variables. Second, we consider the "vector" case of exponential families for multinomial outcomes. For both classes, we derive the novel correction to the Bayesian procedure and delegate more technical discussions to the online Appendices D and F. Additionally, we outline extensions to other causal parameters of interest.

6.1 A Single-parameter Exponential Family

In this part, we assume that the distribution of Y_i conditional on D_i and X_i belongs to the "single-parameter" exponential family, where the unknown parameter is the nonparametric conditional mean function $m(d,x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$. The conditional density function is given by

$$f_{Y|D,X}(y \mid d,x) = c(y)\exp\left[q(m(d,x))ay - A(m(d,x))\right], \tag{6.1}$$

where $A(m) = \log\int c(y)\exp\left[q(m)ay\right]\mathrm{d}y$, and the function $q(\cdot)$ links the mean to the "natural parameter" of the exponential family. We also restrict the sufficient statistic to be linear in y.

The family (6.1) not only encompasses the Bernoulli distribution (with $q(m) = \log(m/(1-m))$, $A(m) = -\log(1-m)$, and $c(y) = a = 1$), as considered in the previous sections, but also allows for counting and continuous outcomes. For instance, when $a = 1$, the Poisson distribution corresponds to the choices $c(y) = 1/(y!)$, $q(m) = \log m$, and $A(m) = m$, while the exponential distribution is represented by $c(y) = 1$, $q(m) = -1/m$, and $A(m) = \log m$. Furthermore, the normal distribution with $\mathrm{Var}(Y|D,X) = \sigma^2$ for some $\sigma > 0$, is captured by $c(y) = \exp(-y^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$, $q(m) = m/\sigma$, $A(m) = m^2/(2\sigma^2)$, and $a = 1/\sigma$. We emphasize that model (6.1) does not impose functional form assumptions on the conditional mean function m. The joint density of (Y_i, D_i, X_i) can be written as

$$p_{\pi,m,f}(y,d,x) = \pi(x)^d(1-\pi(x))^{1-d}c(y)\exp\left[q(m(d,x))ay - A(m(d,x))\right]f(x). \tag{6.2}$$

We consider the same reparametrization of (π, m, f) as in (2.2) except that now the second component of η uses the general link function q satisfying $\eta^m = q(m)$. We now state the

least favorable direction for the exponential family case, which serves as motivation for the prior adjustment.

Lemma 6.1. *For the joint distribution* (6.2) *and the submodel* $t \mapsto \eta_t$ *defined by the path* $m_t(d,x) = q^{-1}(\eta^m + t\mathfrak{m})(d,x)$ *with* (π_t, f_t) *as defined in* (3.2)*, the least favorable direction for estimating the ATE parameter in* (2.3) *is:*

$$\xi_\eta(d,x) = \left(0, \frac{1}{a}\gamma_\eta(d,x), m_\eta(1,x) - m_\eta(0,x) - \tau_\eta\right), \tag{6.3}$$

where the Riesz representer γ_η *is given in* (2.5)*.*

For the outcome family with $a = 1$, which includes Bernoulli, Poisson and exponential distributions, the least favorable direction for ATE estimation coincides with the one as given in Lemma 3.1. To implement the double robust Bayesian procedure for general outcomes, one can still follow Algorithm 1, with the logistic function Ψ replaced by the inverse link function q^{-1}. For the normal (homoscedastic) outcome where prior adjustment $\lambda\widehat{\gamma}(d,x)$ in Algorithm 1 becomes $\lambda\widehat{\gamma}(d,x)/a$, the hyperparameter a can be determined together with other parameters of the Gaussian process by optimizing the marginal likelihood as in Ray and Szabó [2019]. In Proposition F.1, in the online supplementary appendix, we provide primitive conditions for the BvM Theorem to hold under double robust smoothness conditions.

6.2 Multinomial Outcomes

We now assume that the dependent variable Y_i takes values in a finite set, specifically $Y_i \in \{0, 1, \ldots, J\}$. The ATE can then be written as $\tau_\eta = \sum_{j=0}^{J} j\, \mathbb{E}_\eta\left[m_{\eta,j}(1,X) - m_{\eta,j}(0,X)\right]$, where the choice probabilities are given by $m_{\eta,j}(d,x) = \Psi_j\left(\eta^{m_1}, \cdots, \eta^{m_J}\right)$ with the multinomial logit specification:

$$\Psi_0\left(\eta^{m_1}, \cdots, \eta^{m_J}\right) = \frac{1}{1 + \sum_{l=1}^{J} \exp(\eta^{m_l})} \quad \text{and} \quad \Psi_j\left(\eta^{m_1}, \cdots, \eta^{m_J}\right) = \frac{\exp(\eta^{m_j})}{1 + \sum_{l=1}^{J} \exp(\eta^{m_l})},$$

for $j = 1, \ldots, J$. The multinomial logit specification implies $m_{\eta,0}(d,x) = 1 - \sum_{j=1}^{J} m_{\eta,j}(d,x)$. We now provide the least favorable direction for multinomial outcomes in the presence of multinomial outcomes and discuss its consequences for prior adjustment below.

Lemma 6.2. *Consider the submodel* $t \mapsto \eta_t$ *defined by the path* $m_{t,j}(d,x) = \Psi(\eta^{m_j} + t\mathfrak{m}_j)(d,x)$*,* $1 \leqslant j \leqslant J$*, with* (π_t, f_t) *as defined in* (3.2)*. Under Assumption 1, the least*

favorable direction for estimating the ATE parameter is:

$$\xi_\eta(d,x) = (0, \gamma_\eta(d,x), 2\gamma_\eta(d,x), \ldots, J\gamma_\eta(d,x), m_\eta(1,x) - m_\eta(0,x) - \tau_\eta)\,,$$

where the Riesz representer γ_η is given in (2.5)*.*

We emphasize that the least favorable direction calculation is not a trivial extension of Hahn [1998] or Ray and van der Vaart [2020]. This is because there are J nonparametric components involved in the conditional probability function of the multinomial outcomes given covariates, and we need to consider the perturbation of those J components together. Nonetheless, we show that the efficient influence function is of the same generic form as derived in Hahn [1998]. In the proof of 6.2, we compute the derivative of the parameter mapping along the path considered herein. We derive inner products involving the least favorable direction for each nonparametric component consisting of the conditional choice probabilities. The extension to the multinomial case had not been considered in the literature to our knowledge, and it offers a result of independent interest.

Lemma 6.2 motivates the following modification of our double robust Bayesian estimator based on the propensity score-dependent prior on $m_{\eta,j}$ for $1 \leqslant j \leqslant J$:

$$m_{\eta,j}(d,x) = \Psi_j\left(\eta^{m_1}, \cdots, \eta^{m_J}\right) \qquad \text{and} \qquad \eta^{m_j}(d,x) = W^{m_j}(d,x) + \lambda\, j\widehat{\gamma}(d,x),$$

where $W^{m_j}(d,\cdot)$ is a continuous stochastic process independent $\lambda \sim N(0,\sigma_n^2)$ for $\sigma_n > 0$. We may then follow the implementation as described in Section 2.2 using $m_\eta(d,x) = \sum_{j=0}^{J} j\, m_{\eta,j}(d,x)$.

6.3 Other Causal Parameters

We now extend our procedure to general linear functionals of the conditional mean function. We do so only for binary outcomes, as the modification to other types of outcomes follows as above. Recall that the observable data consists of *i.i.d.* observations of $Z = (Y, D, X^\top)^\top$. The causal parameter of interest is $\tau_0 = \mathbb{E}_0[\psi(Z, m_0)]$, where the function ψ is linear with respect to the conditional mean function m_0. We introduce the Riesz representor $\gamma_0(d,x)$ satisfying $\mathbb{E}_0[\psi(Z,m)] = \mathbb{E}_0[\gamma_0(D,X)m(D,X)]$. Let $\widehat{m}$ and $\widehat{\gamma}$ be pilot estimators for the conditional mean and Riesz representor, respectively, computed over an external sample. Our double robust Bayesian procedure can be extended by considering the corrected posterior distribution for τ_η as follows: $\check{\tau}_\eta^s = \sum_{i=1}^n M_{ni}^s \psi(Z_i, m_\eta^s) - n^{-1}\sum_{i=1}^n \boldsymbol{\tau}[m_\eta^s - \widehat{m}](Z_i)$, $s = 1, \ldots, B$, where here $\boldsymbol{\tau}[m](z) := \psi(z,m) + \widehat{\gamma}(d,x)(y - m(d,x))$. The derivations of the

least favorable directions in the following two examples are provided in online Appendix E.

Example 6.1 (Average Policy Effects)**.** The policy effect from changing the distribution of X is $\tau_\eta^P = \int m_\eta(x)d(G_1(x) - G_0(x))$, where the known distribution functions G_1 and G_0 have their supports contained in the support of the marginal covariate distribution F_η. Following the general setup, $\psi(z, m_\eta) = \psi(m_\eta) := \int m_\eta(x)d(G_1(x) - G_0(x))$ with its Riesz representor $\gamma_\eta^P(x) = (g_1(x) - g_0(x))/f_\eta(x)$, where g_1 and g_0 stand for the density function of G_1 and G_0, respectively.

Example 6.2 (Average Derivative)**.** For a continuous scalar (treatment) variable D, the average derivative is given by $\tau_\eta^{AD} = \mathbb{E}_\eta[\partial_d m_\eta(D, X)]$, where $\partial_d m$ denotes the partial derivatives of m with respect to the continuous treatment D. Thus, we have $\psi(Z, m_\eta) = \partial_d m_\eta(D, X)$ with its Riesz representor given by $\gamma_\eta^{AD}(D, X) = \partial_d \pi_\eta(D, X)/\pi_\eta(D, X)$, where here π_η denotes the conditional density function of D given X.

A Proofs of Main Results

In the Appendix, $C > 0$ denotes a generic constant, whose value might change line by line. We introduce additional subscripts when there are multiple constant terms in the same display. For two sequences a_n, b_n, we write $a_n \lesssim b_n$, if $a_n \leqslant Cb_n$. In the following, we denote the log-likelihood based on $Z^{(n)} = (Z_i)_{i=1}^n$ as

$$\ell_n(\eta) = \sum_{i=1}^n \log p_\eta(Z_i) = \ell_n^\pi(\eta^\pi) + \ell_n^m(\eta^m) + \ell_n^f(\eta^f),$$

where each term is the logarithm of the factors involving only π or m or f. Recall the definition of the measurable sets $\mathcal{H}_n^m$ of functions η^m such that $\Pi(\eta^m \in \mathcal{H}_n^m \mid Z^{(n)}) \to_{P_0} 1$. We introduce the conditional prior $\Pi_n(\cdot) := \Pi(\cdot \cap \mathcal{H}_n^m)/\Pi(\mathcal{H}_n^m)$. The following posterior Laplace transform of $\sqrt{n}(\tau_\eta - \widehat{\tau} - b_{0,\eta})$ given by

$$I_n(t) = \mathbb{E}^{\Pi_n}\left[e^{t\sqrt{n}(\tau_\eta - \widehat{\tau} - b_{0,\eta})} \mid Z^{(n)}\right], \quad \forall t \in \mathbb{R} \tag{A.1}$$

plays a crucial role in establishing the BvM theorem [Castillo, 2012, Castillo and Rousseau, 2015, Ray and van der Vaart, 2020]. To abuse the notation slightly, we define a perturbation of $\eta = (\eta^\pi, \eta^m)$ along the least favorable direction, restricted to the

components corresponding to π and m:

$$\eta_t(\eta) := \left(\eta^\pi, \eta^m - \frac{t}{\sqrt{n}}\xi_0^m\right). \tag{A.2}$$

We explicitly write the perturbation of η^m by $\eta_t^m := \eta_t(\eta^m) = \eta^m - t\xi_0^m/\sqrt{n}$. Recall that ξ_0^m coincides with the Riesz representer γ_0 by Lemma 3.1.

Proof of Theorem 3.1. Since the estimated least favorable direction $\widehat{\gamma}$ is based on observations that are independent of $Z^{(n)}$, we may apply Lemma 2 of Ray and van der Vaart [2020]. It suffices to handle the ordinary posterior distribution with $\widehat{\gamma}$ set equal to a deterministic function γ_n. By Lemma 1 of Castillo and Rousseau [2015], it is sufficient to show that the Laplace transform $I_n(t)$ given in (A.1) satisfies

$$I_n(t) \to_{P_0} \exp\left(t^2 \mathrm{v}_0/2\right), \tag{A.3}$$

for every t in a neighborhood of 0, where the limit at the right hand side of (A.3) is the Laplace transform of a $N(0, \mathrm{v}_0)$ distribution. Note that we can write $\tau_\eta = \int \bar{m}_\eta \, \mathrm{d}F_\eta$. Further, let $\widehat{\tau} = \int \bar{m}_0 \, \mathrm{d}F_0 + \mathbb{P}_n[\widetilde{\tau}_0]$, which satisfies (3.1).

The Laplace transform $I_n(t)$ can thus be written as

$$\int\int_{\mathcal{H}_n^m} \frac{\exp\left(t\sqrt{n}(\int \bar{m}_\eta dF_\eta - \bar{m}_0 dF_0 - b_{0,\eta}) - t\mathbb{G}_n[\widetilde{\tau}_0] + \ell_n^m(\eta^m) - \ell_n^m(\eta_t^m)\right) \exp\left(\ell_n^m(\eta_t^m)\right)}{\int_{\mathcal{H}_n^m} \exp\left(\ell_n^m(\eta^{m\prime})\right) \mathrm{d}\Pi(\eta^{m\prime})} \mathrm{d}\Pi(\eta^m) \mathrm{d}\Pi(F_\eta | Z^{(n)}).$$

The expansion in Lemma B.1 gives the following identity for all t in a sufficiently small neighborhood around zero and uniformly for $\eta^m \in \mathcal{H}_n^m$:

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_0 \rho^{m_0}] + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] + t\sqrt{n}\int (\bar{m}_0 - \bar{m}_\eta) \mathrm{d}F_0 + \frac{t^2}{2} P_0 (B_0^m \xi_0^m)^2 + o_{P_0}(1),$$

where we make use of the notation $\rho^m(y, d, x) = y - m(d, x)$ and the score operator $B_0^m = B_{\eta_0}^m$ defined through (3.3).

Next, we plug this into the exponential part in the definition of $I_n(t)$, which then gives

$$\begin{aligned}
&\int\int_{\mathcal{H}_n^m} \frac{\exp\left(t\sqrt{n}\left(\int(\bar{m}_\eta dF_\eta - \bar{m}_0 dF_0) + \int(\bar{m}_0 - \bar{m}_\eta)dF_0 - b_{0,\eta}\right) + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] + \ell_n^m(\eta_t^m)\right)}{\int_{\mathcal{H}_n}\exp\left(\ell_n^m(\eta^{m\prime})\right)\mathrm{d}\Pi(\eta^{m\prime})}\mathrm{d}\Pi(\eta^m)\mathrm{d}\Pi(F_\eta|Z^{(n)}) \\
&\qquad\times\exp\left(-t\mathbb{G}_n[\widetilde{\tau}_0] + t\mathbb{G}_n[\gamma_0\rho^{m_0}] + \frac{t^2}{2}P_0(B_0^m\xi_0^m)^2 + o_{P_0}(1)\right) \\
&= \int\int_{\mathcal{H}_n^m} \frac{\exp\left(t\sqrt{n}\left(\int\bar{m}_\eta \mathrm{d}(F_\eta - F_0) - b_{0,\eta}\right) + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)]\right)\exp\left(\ell_n^m(\eta_t^m)\right)}{\int_{\mathcal{H}_n}\exp\left(\ell_n^m(\eta^{m\prime})\right)\mathrm{d}\Pi(\eta^{m\prime})}\mathrm{d}\Pi(\eta^m)\mathrm{d}\Pi(F_\eta|Z^{(n)}) \\
&\qquad\times\exp\left(-t\mathbb{G}_n[\widetilde{\tau}_0] + t\mathbb{G}_n[\gamma_0\rho^{m_0}] + \frac{t^2}{2}P_0(B_0^m\xi_0^m)^2 + o_{P_0}(1)\right).
\end{aligned}$$

Because all variables have been integrated out in the integral in the denominator, it is a constant relative to either m_η or F_η. By Fubini's Theorem, the double integral without this normalizing constant is

$$\int_{\mathcal{H}_n^m}\exp\left(t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] - t\sqrt{n}b_{0,\eta} + \ell_n^m(\eta_t^m)\right)\int\exp\left(t\sqrt{n}\int\bar{m}_\eta\mathrm{d}(F_\eta - F_0)\right)\mathrm{d}\Pi(F_\eta|Z^{(n)})\mathrm{d}\Pi(\eta^m).$$

By the assumed P_0-Glivenko-Cantelli property for $\mathcal{G}_n = \{\bar{m}_\eta : \eta \in \mathcal{H}_n\}$ in Assumption 3, i.e., $\sup_{\bar{m}_\eta\in\mathcal{G}_n}|(\mathbb{P}_n - P_0)\bar{m}_\eta| = o_{P_0}(1)$, and the boundedness of $\bar{m}_\eta$, we apply Lemma C.4. Further, we may apply the convergence of m_η imposed in Assumption 2, so that the above display becomes

$$\begin{aligned}
&e^{o_{P_0}(1)}\int_{\mathcal{H}_n^m}\exp\left(t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] - t\sqrt{n}b_{0,\eta} + \ell_n^m(\eta_t^m)\right)\exp\left(t\sqrt{n}\int\bar{m}_\eta\mathrm{d}(\mathbb{F}_n - F_0) + \frac{t^2}{2}\|\bar{m}_0 - F_0\bar{m}_0\|_{2,F_0}^2\right)\mathrm{d}\Pi(\eta^m) \\
&\qquad= e^{o_{P_0}(1)}\exp\left(t\sqrt{n}\int\bar{m}_0\mathrm{d}(\mathbb{F}_n - F_0) + \frac{t^2}{2}\|\bar{m}_0 - F_0\bar{m}_0\|_{2,F_0}^2\right) \\
&\qquad\qquad\times\int_{\mathcal{H}_n^m}\exp\Big(\underbrace{t\mathbb{G}_n[\gamma_0(m_0 - m_\eta) - (\bar{m}_0 - \bar{m}_\eta)] - t\sqrt{n}b_{0,\eta}}_{=0} + \ell_n^m(\eta_t^m)\Big)\mathrm{d}\Pi(\eta^m),
\end{aligned}$$

where $F_0\bar{m}_0 \equiv \int\bar{m}_0(x)\mathrm{d}F_0(x)$ and $F_n\bar{m}_0 \equiv 1/n\sum_{i=1}^n\bar{m}_0(X_i)$. We take a closer examination about the empirical process term in the integral. Note that $dm(d,x) = dm(1,x)$ and $(1-d)m(d,x) = (1-d)m(0,x)$ for any $m(\cdot,\cdot)$ and x. Thus, we get

$$\begin{aligned}
\mathbb{G}_n[\gamma_0(m_0 - m_\eta) - (\bar{m}_0 - \bar{m}_\eta)] &= \mathbb{G}_n\left[\left(\frac{d(m_0(1,x) - m_\eta(1,x))}{\pi_0(x)} - \frac{(1-d)(m_0(0,x) - m_\eta(0,x))}{1-\pi_0(x)}\right)\right] \\
&\qquad - \mathbb{G}_n\left[(m_0(1,x) - m_0(0,x)) - (m_\eta(1,x) - m_\eta(0,x))\right] \\
&= \mathbb{G}_n\left[\left(\frac{(d - \pi_0(x))(m_0(1,x) - m_\eta(1,x))}{\pi_0(x)} - \frac{(\pi_0(x) - d)(m_0(0,x) - m_\eta(0,x))}{1-\pi_0(x)}\right)\right]. \quad \text{(A.4)}
\end{aligned}$$

Note that both term are centered, so that one can replace the operator $\mathbb{G}_n$ with $\sqrt{n}\mathbb{P}_n$ therein. Therefore, it cancels this bias term $b_{0,\eta}$ exactly.

Further, observe that $\mathbb{G}_n[\gamma_0\rho^{m_0}] - \mathbb{G}_n[\tilde{\tau}_0] = -\mathbb{G}_n[\bar{m}_0]$ and $\mathbb{G}_n[\bar{m}_0] = \sqrt{n}\int \bar{m}_0 d(\mathbb{F}_n - F_0)$ by the definition of the efficient influence function given in (2.4). As we insert these in the previous expression for $I_n(t)$, we obtain for all t in a sufficiently small neighborhood around zero and uniformly for $\eta \in \mathcal{H}_n$:

$$
\begin{aligned}
I_n(t) = \exp\Bigg(& \underbrace{-t\mathbb{G}_n[\bar{m}_0] + t\sqrt{n}\int \bar{m}_0 \mathrm{d}(\mathbb{F}_n - F_0)}_{=0} + \frac{t^2}{2}\Big(\underbrace{P_0(B_0^m\xi_0^m)^2 + \overbrace{\|\bar{m}_0 - F_0\bar{m}_0\|_{2,F_0}^2}^{=P_0(B_0^f\xi_0^f)^2}}_{=P_0(B_0\xi_0)^2} \Big) + o_{P_0}(1) \Bigg) \\
& \times \frac{\int_{\mathcal{H}_n^m} \exp\big(\ell_n^m(\eta_t^m)\big)\mathrm{d}\Pi(\eta^m)}{\int_{\mathcal{H}_n^m} \exp\big(\ell_n^m(\eta^{m\prime})\big)\mathrm{d}\Pi(\eta^{m\prime})} \\
= \exp & \left(\frac{t^2}{2} P_0(B_0\xi_0)^2 \right) + o_{P_0}(1),
\end{aligned}
$$

where the last line follows from the prior invariance condition established in Lemma B.2. This implies (A.3) using that $P_0(B_0\xi_0)^2 = P_0\widetilde{\tau}_0^2 = \mathrm{v}_0$ by the Lemma 3.1. □

Proof of Theorem 3.2. It is sufficient to show that $\sup_{\eta\in\mathcal{H}_n} \left| b_{0,\eta} - \widehat{b}_\eta \right| = o_{P_0}(n^{-1/2})$, where $b_{0,\eta} = \mathbb{P}_n[\gamma_0(m_0 - m_\eta) + \bar{m}_\eta - \bar{m}_0]$ and $\widehat{b}_\eta = \mathbb{P}_n[\widehat{\gamma}(\widehat{m} - m_\eta) + \bar{m}_\eta - \widehat{\bar{m}}]$. We make use of the decomposition

$$
b_{0,\eta} - \widehat{b}_\eta = \mathbb{P}_n[\gamma_0(m_0 - m_\eta) - \widehat{\gamma}\rho^{m_\eta}] - \mathbb{P}_n[\bar{m}_0 - \widehat{\bar{m}} - \widehat{\gamma}\rho^{\widehat{m}}]. \tag{A.5}
$$

Consider the first summand on the right hand side of the previous equation. We have uniformly for $\eta \in \mathcal{H}_n$:

$$
\begin{aligned}
\mathbb{P}_n[\gamma_0(m_0 - m_\eta) - \widehat{\gamma}\rho^{m_\eta}] &= -\mathbb{P}_n[\widehat{\gamma}\rho^{m_0}] + \mathbb{P}_n[(\gamma_0 - \widehat{\gamma})(m_0 - m_\eta)] \\
&= -\mathbb{P}_n[\widehat{\gamma}\rho^{m_0}] + o_{P_0}(n^{-1/2}),
\end{aligned}
$$

where the last equation follows from the following derivation:

$$
\begin{aligned}
\sqrt{n}\sup_{\eta\in\mathcal{H}_n} |\mathbb{P}_n[(\gamma_0 - \widehat{\gamma})(m_0 - m_\eta)]| &\leqslant \sup_{\eta\in\mathcal{H}_n} |\mathbb{G}_n[(\gamma_0 - \widehat{\gamma})(m_0 - m_\eta)]| \\
&\quad + \sqrt{n}\sup_{\eta\in\mathcal{H}_n} |P_0[(\gamma_0 - \widehat{\gamma})(m_0 - m_\eta)]| \\
&\leqslant o_{P_0}(1) + O_{P_0}(1) \times \sqrt{n}\|\pi_0 - \widehat{\pi}\|_{2,F_0} \sup_{\eta\in\mathcal{H}_n} \|m_\eta - m_0\|_{2,F_0} = o_{P_0}(1),
\end{aligned}
$$

using the Cauchy-Schwarz inequality, Assumption 2, and Assumption 3. Consider the

second summand on the right hand side of (A.5). From Lemma C.8 we infer

$$\mathbb{P}_n[\widehat{\bar{m}} + \widehat{\gamma}\rho^{\widehat{m}} - \bar{m}_0] = \mathbb{P}_n[\gamma_0\rho^{m_0}] + o_{P_0}(n^{-1/2}).$$

Consequently, decomposition (A.5) together with the asymptotic expansion of each summand yields

$$\sup_{\eta\in\mathcal{H}_n}\left|b_{0,\eta} - \widehat{b}_\eta\right| \leqslant |\mathbb{P}_n[(\gamma_0 - \widehat{\gamma})\rho^{m_0}]| + o_{P_0}(n^{-1/2}) = o_{P_0}(n^{-1/2}),$$

where the last equation is due to the equation (C.6). □

Proof of Corollary 3.1. The weak convergence of the Bayesian point estimator directly follows from our asymptotic characterization of the posterior and the argmax theorem; see the proof of Theorem 10.8 in van der Vaart [1998]. The corrected Bayesian credible set $\mathcal{C}_n(\alpha)$ satisfies $\Pi(\check{\tau}_\eta \in \mathcal{C}_n(\alpha) \mid Z^{(n)}) = 1-\alpha$ for any $\alpha \in (0,1)$. In particular, we have

$$\Pi\left(\sqrt{n/\mathrm{v}_0}(\tau_\eta - \widehat{\tau} - \widehat{b}_\eta) \in \sqrt{n/\mathrm{v}_0}(\mathcal{C}_n(\alpha) - \widehat{\tau}) \mid Z^{(n)}\right) = 1-\alpha.$$

Now the definition of the estimator $\widehat{\tau}$ given in (3.1) yields $\sqrt{n}\widehat{\tau} = \sqrt{n}\big(\tau_0 + \mathbb{P}_n\widetilde{\tau}_0\big) + o_{P_0}(1)$. For any set A, we write $\mathbb{N}(A) := \int_A e^{-u^2/2}/\sqrt{2\pi}\,\mathrm{d}u$. Theorem 3.1 implies

$$\mathbb{N}\left(\sqrt{n/\mathrm{v}_0}(\mathcal{C}_n(\alpha) - \tau_0 - \mathbb{P}_n\widetilde{\tau}_0)\right) \to_{P_0} 1-\alpha.$$

We may thus write $\mathcal{C}_n(\alpha) = \sqrt{\mathrm{v}_0/n}\,\mathcal{A}_n(\alpha) + \tau_0 + \mathbb{P}_n\widetilde{\tau}_0 + o_{P_0}(1)$ for some set $\mathcal{A}_n(\alpha)$ satisfying $\mathbb{N}(\mathcal{A}_n(\alpha)) \to_{P_0} 1-\alpha$. Therefore, the frequentist coverage of the Bayesian credible set is

$$P_0\left(\tau_0 \in \mathcal{C}_n(\alpha)\right) = P_0\left(\tau_0 \in \sqrt{\mathrm{v}_0/n}\,\mathcal{A}_n(\alpha) + \tau_0 + \mathbb{P}_n\widetilde{\tau}_0\right) = P_0\left(-\frac{\mathbb{G}_n\widetilde{\tau}_0}{\sqrt{\mathrm{v}_0}} \in \mathcal{A}_n(\alpha)\right) \to 1-\alpha,$$

noting that $\mathbb{G}_n\widetilde{\tau}_0$ is asymptotically normal with mean zero and variance v_0 under P_0. □

Proof of Proposition 4.1. Note that $\widehat{\gamma}$ is based on an auxiliary sample and hence we can treat $\widehat{\gamma}$ below as a deterministic function denoted by γ_n satisfying the rate restrictions $\|\gamma_n\|_\infty = O(1)$ and $\|\gamma_n - \gamma_0\|_\infty = O\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$. Regarding the conditional mean functions, we consider the set $\mathcal{H}^m_{n,d} := \{w_d + \lambda\gamma_n : (w_d,\lambda) \in \mathcal{W}_{n,d}\}$, where for $d \in \{1,0\}$ and some constant $C > 0$:

$$\mathcal{W}_{n,d} := \left\{(w_d,\lambda) : w_d \in \mathcal{B}^m_n, |\lambda| \leqslant C\sigma_n\sqrt{n}\varepsilon_n\right\} \cap \left\{(w_d,\lambda) : \|\Psi(w_d(\cdot) + \lambda\gamma_n) - m_0(d,\cdot)\|_{2,F_0} \leqslant \varepsilon_n\right\}, \tag{A.6}$$

where $\mathcal{B}_n^m$ in the first restriction for the Gaussian process $W(d,\cdot)$ is a regularity class of functions defined in the equation (C.7) in the online Supplementary Appendix C. We write $\mathcal{H}_n^m = \mathcal{H}_{n,1}^m \times \mathcal{H}_{n,0}^m$.

We first verify Assumption 2 with $\varepsilon_n = (n/\log n)^{-s_m/(2s_m+p)}$. The posterior contraction rate is shown in our Lemma C.3. Referring to the product rate condition, i.e., $\sqrt{n}\varepsilon_n r_n = o(1)$ for $r_n \sim (n/\log n)^{-s_\pi/(2s_\pi+p)}$. This is satisfied if $2s_m/(2s_m+p) + 2s_\pi/(2s_\pi+p) > 1$, which can be rewritten as $\sqrt{s_\pi\, s_m} > p/2$.

We now verify Assumption 3. It is sufficient to deal with the resulting empirical process $\mathbb{G}_n$. Note that the Cauchy-Schwartz inequality implies

$$\begin{aligned}
|P_0(m_\eta - m_0)| &= |\mathbb{E}_0[D(m_\eta(1,X) - m_0(1,X))] + \mathbb{E}_0[(1-D)(m_\eta(0,X) - m_0(0,X))]| \\
&\leqslant \sqrt{\mathbb{E}_0[(m_\eta(1,X) - m_0(1,X))^2]} + \sqrt{\mathbb{E}_0[(m_\eta(0,X) - m_0(0,X))^2]} \\
&= \|m_\eta(1,\cdot) - m_0(1,\cdot)\|_{2,F_0} + \|m_\eta(0,\cdot) - m_0(0,\cdot)\|_{2,F_0}.
\end{aligned}$$

Consequently, from Lemma C.5 we infer

$$\begin{aligned}
\mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n[(\gamma_n - \gamma_0)(m_\eta - m_0)]| &\leqslant 4\|\gamma_n - \gamma_0\|_\infty \mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n[m_\eta - m_0]| \\
&\quad + \|\gamma_n - \gamma_0\|_{2,F_0} \sup_{\eta\in\mathcal{H}_n} \Big(\|m_\eta(1,\cdot) - m_0(1,\cdot)\|_{2,F_0} + \|m_\eta(0,\cdot) - m_0(0,\cdot)\|_{2,F_0} \Big) \\
&\lesssim (n/\log n)^{-s_\pi/(2s_\pi+p)} \mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n[m_\eta - m_0]| + (n/\log n)^{-s_\pi/(2s_\pi+p)} (n/\log n)^{-s_m/(2s_m+p)} \\
&= (n/\log n)^{-s_\pi/(2s_\pi+p)} \mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n[m_\eta - m_0]| + o(1).
\end{aligned}$$

Note that if $s_m > p/2$, from Lemma C.9 we infer $\mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} \mathbb{G}_n[m_\eta - m_0] = o(1)$. Thus it remains to consider the case $s_m \leqslant p/2$. By the entropy bound presented in the proof of Lemma C.3, we have $\log N(\varepsilon_n, \mathcal{H}_n^m, L^2(F_0)) \lesssim \varepsilon_n^{-2\upsilon}$, with $\upsilon = p/(2s_m)$ modulo some $\log n$ term on the right hand of the bound. Because $\Psi(\cdot)$ is monotone and Lipschitz, a set of ε-covers in $L^2(F_0)$ for $\eta^m \in \mathcal{H}_n^m$ translates into a set of ε-covers for m_η. In this case, the empirical process bound of [Han, 2021, p.2644] yields

$$\mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n[m_\eta - m_0]| \lesssim L_n n^{(\upsilon-1)/(2\upsilon)} = O(L_n n^{1/2 - s_m/p}),$$

where L_n represents a term that diverges at certain polynomial order of $\log n$. Consequently,

we obtain

$$(n/\log n)^{-s_\pi/(2s_\pi+p)}\mathbb{E}_0 \sup_{\eta\in\mathcal{H}_n^m} |\mathbb{G}_n[m_\eta - m_0]| = o(1),$$

which is satisfied under the smoothness restriction $-s_\pi/(2s_\pi + p) + 1/2 - s_m/p < 0$ or equivalently $4s_\pi s_m + 2ps_m > p^2$. This condition automatically holds given $\sqrt{s_\pi\, s_m} > p/2$.

Finally, it remains to verify Assumption 4. By the univariate Gaussian tail bound, the prior mass of the set $\Lambda_n := \{\lambda : |\lambda| > u_n\sigma_n^2\sqrt{n}\}$ satisfies $\Pi(\lambda \in \Lambda_n) \leqslant 2\exp(-u_n^2\sigma_n^2 n/2)$. Also, the Kullback-Leibler neighborhood around η_0^m has prior probability at least $e^{-n\varepsilon_n^2}$. We may thus apply Lemma 4 of Ray and van der Vaart [2020], which yields $\Pi(\lambda \in \Lambda_n \mid Z^{(n)}) \to_{P_0} 0$, as imposed in Assumption 4(i).

Regarding Assumption 4(ii), we need to show the posterior probability of the shifted version of $\mathcal{H}_n^m$ is tending to one. Considering $\mathcal{H}_n^m$ itself, the first set in the intersection of (A.6) that defines $\mathcal{W}_{n,d}$ is seen to have posterior probability tending to one by the result in (II) of Lemma C.3, combined with the univariate Gaussian tail probability bound

$$\Pi(|\lambda| \geqslant C\sigma_n\sqrt{n}\varepsilon_n) \leqslant 2\exp(-Cn\varepsilon_n^2/2).$$

The second set in the intersection of (A.6) has posterior probability tending to one by Lemma 17 of Ray and van der Vaart [2020]. Hence, $\mathcal{H}_n^m$ has posterior probability going to one. Next, we consider $\mathcal{H}_n^m + t\gamma_n/\sqrt{n}$, for any $t \in \mathbb{R}$. To slightly abuse the notation, we write $\eta_d^m = w_d + \lambda\gamma_n$ for $d \in \{0, 1\}$ in the sequel. By the Lipschitz continuity of the Logistic link function, we have $\|\Psi(\eta_d^m) - \Psi(\eta_d^m + t\gamma_n/\sqrt{n})\|_{2,F_0} \leqslant |t|\|\gamma_n\|_\infty/\sqrt{n}$ for $d \in \{0, 1\}$. Therefore, we get $\mathcal{H}_{n,d}^m + t\gamma_n/\sqrt{n} \supset \Xi_{n,d,t}$ with probability P_0 approaching one, where

$$\Xi_{n,d,t} := \left\{\eta_d^m : \|\eta_d^m\|_{\mathbb{H}} \leqslant C\sqrt{n}\varepsilon_n - |t|\|\gamma_n\|_\infty, \|\Psi(\eta_d^m) - m_0(d,\cdot)\|_{2,F_0} \leqslant \varepsilon_n - |t|\|\gamma_n\|_\infty/\sqrt{n}\right\}$$

and $\|\cdot\|_{\mathbb{H}}$ denotes the norm of the Reproducing Kernel Hilbert Space associated with the squared exponential process; see online supplementary Appendix C for a formal definition. Because $\sqrt{n}\varepsilon_n \to \infty$ and $\|\gamma_n\|_\infty = O(1)$, the posterior probability of $\Xi_{n,d,t}$ tends to one following similar arguments concerning the set $\mathcal{H}_n^m$, after replacing ε_n with a multiple of itself for $d \in \{1, 0\}$. Hence, the posterior probability of $\mathcal{H}_n^m + t\gamma_n/\sqrt{n}$ is seen to tend to one, which completes the proof. □

B Key Lemmas

We now present key lemmas used in the derivation of our BvM Theorem. We introduce $\eta_u := (\eta^\pi, \eta_u^m)$ where

$$\eta_u^m = \eta^m - tu\xi_0^m/\sqrt{n}, \quad \text{for} \quad u \in [0,1]. \tag{B.1}$$

This defines a path from $\eta_{u=0} = (\eta^\pi, \eta^m)$ to $\eta_{u=1} = (\eta^\pi, \eta_t^m)$. We also write $g(u) := \log p_{\eta_u^m}$, for $u \in [0,1]$, so that $\log p_{\eta^m} - \log p_{\eta_t^m} = g(0) - g(1)$, cf. the proof of Theorem 1 in Ray and van der Vaart [2020].

Lemma B.1. *Let Assumptions 1 and 2 hold. Then, we have uniformly for $\eta \in \mathcal{H}_n$:*

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_0\rho^{m_0}] + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] + t\sqrt{n}\int(\bar{m}_0 - \bar{m}_\eta)\mathrm{d}F_0 + \frac{t^2}{2}P_0(B_0^m\xi_0^m)^2 + o_{P_0}(1).$$

Proof. We start with the following decomposition:

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_0\rho^{m_0}] + \underbrace{\sqrt{n}\mathbb{G}_n[\log p_{\eta^m} - \log p_{\eta_t^m} - \frac{t}{\sqrt{n}}\gamma_0\rho^{m_0}]}_{\text{Stochastic Equicontinuity}} + \underbrace{nP_0[\log p_{\eta^m} - \log p_{\eta_t^m}]}_{\text{Taylor Expansion}}.$$

From the calculation in the proof of Lemma C.1, we have $g'(0) = -\frac{t}{\sqrt{n}}\gamma_0\rho^{m_0} + \frac{t}{\sqrt{n}}\gamma_0(m_\eta - m_0)$. Then, we infer for the stochastic equicontinuity term that

$$\sqrt{n}\mathbb{G}_n[\log p_{\eta^m} - \log p_{\eta_t^m} - \frac{t}{\sqrt{n}}\gamma_0\rho^{m_0}] + t\mathbb{G}_n[\gamma_0(m_\eta - m_0)] = o_{P_0}(1),$$

uniformly in $\eta^m \in \mathcal{H}_n^m$. We can thus write uniformly in $\eta^m \in \mathcal{H}_n^m$:

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_0\rho^{m_0}] + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] + nP_0[\log p_{\eta^m} - \log p_{\eta_t^m}] + o_{P_0}(1).$$

The rest of the proof involves a standard Taylor expansion for the third term on the right hand side of the above equation. By the equation (C.4) in the proof of Lemma C.1, we get

$$-nP_0g'(0) = t\sqrt{n}P_0[\gamma_0\rho^{m_0}] + t\sqrt{n}P_0[\gamma_0(m_0 - m_\eta)] = t\sqrt{n}\int(\bar{m}_0 - \bar{m}_\eta)\mathrm{d}F_0,$$

by the fact that $P_0[\gamma_0\rho^{m_0}] = 0$ and the definition of the Riesz representor γ_0 in (2.5). Regarding the second-order term in the Taylor expansion in the equation (C.5) of the proof

of Lemma C.1, we get

$$g^{(2)}(0) = -\frac{t^2}{n}\gamma_0^2 m_0(1-m_0) - \frac{t^2}{n}\gamma_0^2(m_\eta(1-m_\eta) - m_0(1-m_0)).$$

Considering the score operator $B_0^m = B_{\eta_0}^m$ defined in (3.3), we have

$$\begin{aligned} P_0(B_0^m\xi_0^m)^2 &= \mathbb{E}_0\left[\gamma_0^2(D,X)(Y-m_0(D,X))^2\right] \\ &= \mathbb{E}_0\left[\frac{D}{\pi_0^2(X)}(Y(1)-m_0(1,X))^2\right] + \mathbb{E}_0\left[\frac{1-D}{(1-\pi_0(X))^2}(Y(0)-m_0(0,X))^2\right]. \end{aligned}$$

Consequently, by the unconfoundedness imposed in Assumption 1(i) and the binary nature of Y, we have $\mathbb{E}_0[Y(d)^2|D=d,X=x] = \mathbb{E}_0[Y(d)|D=d,X=x] = m_0(d,x)$. We thus obtain

$$\begin{aligned} P_0(B_0^m\xi_0^m)^2 &= \mathbb{E}_0\left[\frac{D}{\pi_0^2(X)}m_0(1,X)(1-m_0(1,X))\right] + \mathbb{E}_0\left[\frac{1-D}{(1-\pi_0(X))^2}m_0(0,X)(1-m_0(0,X))\right] \\ &= P_0[\gamma_0^2 m_0(1-m_0)]. \end{aligned}$$

Then, by employing Assumption 1(ii), i.e., $\bar{\pi} < \pi_0(x) < 1-\bar{\pi}$ for all x, it yields uniformly for $\eta \in \mathcal{H}_n$:

$$\begin{aligned} &-nP_0 g^{(2)}(0) - t^2 P_0(B_0^m\xi_0^m)^2 = t^2 P_0[\gamma_0^2(m_\eta(1-m_\eta) - m_0(1-m_0))] \\ &\quad = t^2 P_0[\gamma_0^2(m_\eta - m_0)(1-m_0)] + t^2 P_0[\gamma_0^2 m_\eta(m_0 - m_\eta)] \\ &\quad \leqslant 2t^2\,\mathbb{E}_0\left[\frac{D}{\pi_0^2(X)}\big|m_\eta(1,X) - m_0(1,X)\big|\right] + 2t^2\,\mathbb{E}_0\left[\frac{1-D}{(1-\pi_0(X))^2}\big|m_\eta(0,X) - m_0(0,X)\big|\right] \\ &\quad \leqslant \frac{2t^2}{\bar{\pi}^2}\Big(\|m_\eta(1,\cdot) - m_0(1,\cdot)\|_{2,F_0} + \|m_\eta(0,\cdot) - m_0(0,\cdot)\|_{2,F_0}\Big) = o_{P_0}(1), \end{aligned}$$

where the last equation is due to the posterior contraction rate of the conditional mean function $m(d,\cdot)$ imposed in Assumption 2. Consequently, we obtain, uniformly for $\eta \in \mathcal{H}_n$,

$$\begin{aligned} nP_0[\log p_{\eta^m} - \log p_{\eta_t^m}] &= -n(P_0 g'(0) + P_0 g^{(2)}(0)) + o_{P_0}(1) \\ &= t^2 P_0(B_0^m\xi_0^m)^2 + t\sqrt{n}\int(\bar{m}_0 - \bar{m}_\eta)\,\mathrm{d}F_0 + o_{P_0}(1), \end{aligned}$$

which leads to the desired result. □

The next lemma verifies the prior stability condition under our double robust smoothness conditions.

Lemma B.2. *Let Assumptions 1–4 hold. Then we have*

$$\frac{\int_{\mathcal{H}_n^m} \exp\left(\ell_n^m(\eta_t^m)\right) \mathrm{d}\Pi(\eta^m)}{\int_{\mathcal{H}_n^m} \exp\left(\ell_n^m(\eta^{m\prime})\right) \mathrm{d}\Pi(\eta^{m\prime})} \to_{P_0} 1, \tag{B.2}$$

for a sequence of measurable sets $\mathcal{H}_n^m$ such that $\Pi(\eta^m \in \mathcal{H}_n^m | Z^{(n)}) \to_{P_0} 1$.

Proof. Since $\widehat{\gamma}$ is based on an auxiliary sample, it is sufficient to consider deterministic functions γ_n with the same rates of convergence as $\widehat{\gamma}$. Denote the corresponding propensity score by π_n. By Assumption 4, we have $\lambda \sim N(0, \sigma_n^2)$ and

$$\frac{\int_{\mathcal{H}_n^m} \exp\left(\ell_n^m(\eta_t^m)\right) \mathrm{d}\Pi(\eta^m)}{\int_{\mathcal{H}_n^m} \exp\left(\ell_n^m(\eta^{m\prime})\right) \mathrm{d}\Pi(\eta^{m\prime})} = \frac{\int_{\Theta_n} e^{\ell_n^m(w+\lambda\gamma_n - t\gamma_0/\sqrt{n})} \phi_{\sigma_n}(\lambda)\, \mathrm{d}\lambda \mathrm{d}\Pi(w)}{\int_{\Theta_n} e^{\ell_n^m(w+\lambda\gamma_n)} \phi_{\sigma_n}(\lambda)\, \mathrm{d}\lambda \mathrm{d}\Pi(w)} + o_{P_0}(1), \tag{B.3}$$

where ϕ_{σ_n} denotes the probability density function of a $N(0, \sigma_n^2)$ random variable and the set Θ_n is defined by $\Theta_n = \{(w, \lambda) : w + \lambda\gamma_n \in \mathcal{H}_n^m, |\lambda| \leqslant 2u_n\sigma_n^2\sqrt{n}\}$ where $u_n \to 0$ imposed in Assumption 4 and $u_n n\sigma_n^2 \to \infty$.

Considering the log likelihood ratio of two normal densities together with the constraint $|\lambda| \leqslant 2u_n\sigma_n^2\sqrt{n}$, it is shown on page 3015 of Ray and van der Vaart [2020] that

$$\left|\log \frac{\phi_{\sigma_n}(\lambda)}{\phi_{\sigma_n}(\lambda - t/\sqrt{n})}\right| \leqslant \frac{|t\lambda|}{\sqrt{n}\sigma_n^2} + \frac{t^2}{2n\sigma_n^2} \to 0.$$

We show at the end of the proof that $|\ell_n^m(w + \lambda\gamma_n - t\gamma_0/\sqrt{n}) - \ell_n^m(w + \lambda\gamma_n - t\gamma_n/\sqrt{n})| = o_{P_0}(1)$, uniformly for $(w, \lambda) \in \Theta_n$. Consequently, the numerator of this leading term in (B.3) becomes

$$\int_{B_n} e^{\ell_n^m(w+\lambda\gamma_n - t\gamma_0/\sqrt{n})} \phi_{\sigma_n}(\lambda)\, \mathrm{d}\lambda \mathrm{d}\Pi(w) = e^{o_{P_0}(1)} \int_{B_n} e^{\ell_n^m(w+\gamma_n(\lambda - t/\sqrt{n}))} \phi_{\sigma_n}(\lambda - t/\sqrt{n})\, \mathrm{d}\lambda \mathrm{d}\Pi(w).$$

By the change of variables $\lambda - t/\sqrt{n} \mapsto \lambda'$ on the numerator and using the notation $\Theta_{n,t} = \{(w, \lambda) : (w, \lambda + t/\sqrt{n}) \in \Theta_n\}$, the prior invariance property becomes

$$e^{o_{P_0}(1)} \frac{\int_{\Theta_{n,t}} e^{\ell_n^m(w+\lambda'\gamma_n)} \phi_{\sigma_n}(\lambda')\, \mathrm{d}\lambda' \mathrm{d}\Pi(w)}{\int_{\Theta_n} e^{\ell_n^m(w+\lambda\gamma_n)} \phi_{\sigma_n}(\lambda)\, \mathrm{d}\lambda \mathrm{d}\Pi(w)} = e^{o_{P_0}(1)} \frac{\Pi(\Theta_{n,t} | X^{(n)})}{\Pi(\Theta_n | X^{(n)})}.$$

The desired result would follow from $\Pi(\Theta_n | X^{(n)}) = 1 - o_{P_0}(1)$ and $\Pi(\Theta_{n,t} | X^{(n)}) = 1 - o_{P_0}(1)$. The first convergence directly follows from Assumption 4. The set $\Theta_{n,t}$ is the intersection of these two conditions in Assumption 4, except that the restriction on λ

in $\Theta_{n,t}$ is $|\lambda + t/\sqrt{n}| \leqslant 2u_n\sqrt{n}\sigma_n^2$ instead of $|\lambda| \leqslant u_n\sqrt{n}\sigma_n^2$. By construction, we have $t/\sqrt{n} = o(u_n\sqrt{n}\sigma_n^2)$, so that $\Pi(\Theta_{n,t}|X^{(n)}) = 1 - o_{P_0}(1)$.

We finish the proof by establishing the following result:

$$\sup_{\eta^m \in \mathcal{H}_n^m} \left|\ell_n^m(\eta^m - t\gamma_n/\sqrt{n}) - \ell_n^m(\eta^m - t\gamma_0/\sqrt{n})\right| = o_{P_0}(1). \tag{B.4}$$

We denote $\eta_{n,t}^m = \eta^m - t\gamma_n/\sqrt{n}$ and $\eta_t^m = \eta^m - t\gamma_0/\sqrt{n}$. Consider the following decomposition of the log-likelihood:

$$\begin{aligned}\ell_n^m(\eta_{n,t}^m) - \ell_n^m(\eta_t^m) &= \ell_n^m(\eta_{n,t}^m) - \ell_n^m(\eta^m) + \ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) \\ &= n\mathbb{P}_n[\log p_{\eta_{n,t}^m} - \log p_{\eta^m}] + n\mathbb{P}_n[\log p_{\eta^m} - \log p_{\eta_t^m}].\end{aligned}$$

Next, we apply third-order Taylor expansions in Lemma C.1 separately to the two terms in the brackets of the above display making use of the notation $\rho^m(y, d, x) = y - m(d, x)$:

$$\begin{aligned}n\mathbb{P}_n[\log p_{\eta_{n,t}^m} - \log p_{\eta^m}] &= -t\sqrt{n}\mathbb{P}_n\left[\gamma_n\rho^{m_\eta}\right] - \frac{t^2}{2}\mathbb{P}_n\left[\gamma_n^2 m_\eta\left(1 - m_\eta\right)\right] - \frac{t^3}{\sqrt{n}}\mathbb{P}_n\left[\gamma_n^3\Psi^{(2)}(\eta_{u^*}^m)\right], \\ n\mathbb{P}_n[\log p_{\eta^m} - \log p_{\eta_t^m}] &= t\sqrt{n}\mathbb{P}_n\left[\gamma_0\rho^{m_\eta}\right] + \frac{t^2}{2}\mathbb{P}_n\left[\gamma_0^2 m_\eta\left(1 - m_\eta\right)\right] + \frac{t^3}{\sqrt{n}}\mathbb{P}_n\left[\gamma_0^3\Psi^{(2)}(\eta_{u^{**}}^m)\right],\end{aligned}$$

for some intermediate points $u^*, u^{**} \in (0, 1)$, cf. the equation (B.1). Combining the previous calculation yields

$$\begin{aligned}\ell_n^m(\eta_{n,t}) - \ell_n^m(\eta_t) = t\sqrt{n}\mathbb{P}_n[(\gamma_0 - \gamma_n)\rho^{m_\eta}] - \frac{t^2}{2}\mathbb{P}_n[dm_\eta(1 - m_\eta)(\gamma_n^2 - \gamma_0^2)] \\ + \frac{t^3}{\sqrt{n}}\mathbb{P}_n\left[(\gamma_0^3 - \gamma_n^3)\left(\Psi^{(2)}(\eta_{u^{**}}^m) - \Psi^{(2)}(\eta_{u^*}^m)\right)\right] =: T_1 + T_2 + T_3.\end{aligned}$$

In order to control T_1, we evaluate

$$T_1 = t\mathbb{G}_n[(\gamma_0 - \gamma_n)\rho^{m_0}] + t\mathbb{G}_n[(\gamma_0 - \gamma_n)(m_0 - m_\eta)] + t\sqrt{n}P_0[(\gamma_0 - \gamma_n)\rho^{m_\eta}].$$

Note that the first term is centered, so it becomes $t\sqrt{n}\mathbb{P}_n[(\gamma_0 - \gamma_n)\rho^{m_0}]$. We apply Lemma C.2 to conclude that it is of smaller order. The middle term is negligible by our Assumption

3. Referring to the last term, the Cauchy–Schwarz inequality yields

$$\begin{aligned}&\sup_{\eta\in\mathcal{H}_n}\left|\sqrt{n}P_0[(\gamma_n-\gamma_0)(m_\eta-m_0)]\right|\\&\lesssim\sqrt{2\,n}\,\|\pi_n-\pi_0\|_{2,F_0}\sup_{\eta\in\mathcal{H}_n}\Big(\|m_\eta(1,\cdot)-m_0(1,\cdot)\|_{2,F_0}+\|m_\eta(0,\cdot)-m_0(0,\cdot)\|_{2,F_0}\Big)=o_{P_0}(1),\end{aligned}$$

where the last equality is due to Assumption 2. We thus obtain $T_1 = o_{P_0}(1)$ uniformly in $\eta \in \mathcal{H}_n^m$. Consider T_2. We note that $\|m_\eta(1-m_\eta)\|_\infty \leqslant 1$ uniformly in $\eta \in \mathcal{H}_n^m$. Hence, we obtain

$$P_0|T_2| \leqslant \frac{t^2}{2}P_0|\gamma_n^2-\gamma_0^2| = \frac{t^2}{2}P_0[(\gamma_n-\gamma_0)(\gamma_n+\gamma_0)] \lesssim \frac{t^2}{2}\|\pi_n-\pi_0\|_{2,F_0}\to 0$$

as $\pi_n \to \pi_0$ in $L^2(F_0)$-norm by Assumption 2. Thus, $T_2 = o_{P_0}(1)$ uniformly in $\eta \in \mathcal{H}_n$. Finally, we control T_3 by evaluating $|T_3| \lesssim t^3 n^{-1/2}\mathbb{P}_n(\|\gamma_n\|_\infty^3 + \|\gamma_0\|_\infty^3) = o_{P_0}(1)$ uniformly in $\eta \in \mathcal{H}_n^m$, which shows (B.4). □

References

Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 240(2):105076, 2024.

A. Abadie and G. W. Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

I. Andrews and A. Mikusheva. Optimal decision rules for weak gmm. *Econometrica*, 90(2): 715–748, 2022.

T. B. Armstrong and M. Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177, 2021.

D. Benkeser, M. Carone, M. V. D. Laan, and P. Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

I. Castillo. A semiparametric bernstein–von mises theorem for gaussian process priors. *Probability Theory and Related Fields*, 152:53–99, 2012.

I. Castillo and J. Rousseau. A bernstein–von mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383, 2015.

G. Chamberlain and G. W. Imbens. Nonparametric applications of bayesian inference. *Journal of Business & Economic Statistics*, 21(1):12–18, 2003.

X. Chen, H. Hong, and A. Tarozzi. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

X. Chen, T. M. Christensen, and E. Tamer. Monte carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018, 2018.

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, W. Newey, and R. Singh. De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2020.

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

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94 (448):1053–1062, 1999.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

S. Ghosal and A. Roy. Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429, 2006.

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

R. Giacomini and T. Kitagawa. Robust bayesian inference for set-identified models. *Econometrica*, 89(4):1519–1556, 2021.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

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

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

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

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

M. Kasy. Optimal taxation and insurance using machine learning—sufficient statistics and beyond. *Journal of Public Economics*, 167:205–219, 2018.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

Y. Luo, D. J. Graham, and E. J. McCoy. Semiparametric bayesian doubly robust causal estimation. *Journal of Statistical Planning and Inference*, 225:171–187, 2023.

K. P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press, 2023.

C. Rassmusen and C. Williams. *Gaussian processes for machine learning*. MIT, 2006.

K. Ray and B. Szabó. Debiased bayesian inference for average treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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 and Y. Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

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

D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

O. Saarela, L. Belzile, and D. Stephens. A bayesian view of doubly robust causal inference. *Biometrika*, 103(3):667–681, 2016.

M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer, 2011.

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.

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

A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B):2655 – 2675, 2009.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

A. Yiu, R. J. Goudie, and B. D. Tom. Inference under unequal probability sampling with the bayesian exponentially tilted empirical likelihood. *Biometrika*, 107(4):857–873, 2020.