Semiparametric Bayesian Difference-in-Differences*

Christoph Breunig† Ruixuan Liu‡ Zhengfei Yu§

June 14, 2025

Abstract

This paper studies semiparametric Bayesian inference for the average treatment effect on the treated (ATT) within the difference-in-differences (DiD) research design. We propose two new Bayesian methods with frequentist validity. The first one places a standard Gaussian process prior on the conditional mean function of the control group. The second method is a double robust Bayesian procedure that adjusts the prior distribution of the conditional mean function and subsequently corrects the posterior distribution of the resulting ATT. We prove new semiparametric Bernstein-von Mises (BvM) theorems for both proposals. Monte Carlo simulations and an empirical application demonstrate that the proposed Bayesian DiD methods exhibit strong finite-sample performance compared to existing frequentist methods. We also present extensions of the canonical DiD approach, incorporating both the staggered design and the repeated cross-sectional design.

KEYWORDS: Difference-in-differences, conditional parallel trends, semiparametric Bayesian inference, Bernstein–von Mises theorem, double robustness.

*We thank Joachim Freyberger, Björn Höppner, Toru Kitagawa, Soonwoo Kwon, Andriy Norets, Aureo de Paula, Jonathan Roth, Liangjun Su 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

The Difference-in-Differences (DiD) method is widely used in causal inference. It is particularly effective for evaluating policy interventions while accounting for unobserved time-invariant heterogeneity. The primary parameter of interest in this context is the average treatment effect on the treated (ATT). One of its key identifying conditions is the (conditional) parallel trends assumption, i.e., treated and control groups would exhibit similar trends absent treatment after adjusting for covariates (Abadie, 2005; Sant'Anna and Zhao, 2020). While the related literature is largely frequentist, this paper introduces a Bayesian framework under conditional parallel trends, avoiding parametric assumptions on model primitives. Our approach yields point estimates and credible sets in a unified manner.

We propose two novel Bayesian methods for inference on the ATT in the DiD framework. First, we propose the *Bayesian procedure using Gaussian process priors*. This method places the Gaussian process prior on the conditional mean function for the control group and a Dirichlet process prior on the remaining part of the distribution in the likelihood. This avoids the need to impose Bayesian modeling on either the conditional mean of the treated group or the propensity score. Our method can be viewed as the Bayesian counterpart of Heckman, Ichimura, and Todd (1997), with the added advantage of enabling automatic uncertainty quantification through the posterior distribution. We show that this Bayesian method satisfies the Bernstein-von Mises (BvM) theorem under regularity conditions and is therefore asymptotically equivalent to semiparametric efficient frequentist estimators. While its asymptotic BvM property does not hold under double robust smoothness conditions, the Bayesian method performs well empirically when the number of continuous covariates is moderate and in scenarios where the overlap assumption is nearly violated. This robustness stems, in part, from the Gaussian process prior being specified solely on the conditional mean function for the control group.

We also provide an extension of our Bayesian procedure, which incorporates robustification via estimated propensity scores and is particularly suited for more complex models, either due to a larger number of continuous covariates or when the underlying conditional mean functions are not smooth. Our *Double Robust Bayesian procedure* adjusts the prior and posterior distributions by incorporating an efficient influence function. By doing so, we leverage the rich frequentist literature on double-robust estimation, specifically Sant'Anna and Zhao (2020) in the DiD framework, without sacrificing many of the desirable properties of the Bayesian approach. Under double-robust smoothness conditions, our robust Bayesian procedure satisfies the semiparametric Bernstein–von Mises (BvM)

theorem, albeit with a "bias term" in the posterior. Specifically, the resulting posterior distribution depends on the unknown true conditional mean and propensity score functions. Our double-robust Bayesian approach addresses this "bias term" by incorporating an explicit posterior correction. Both the prior adjustment and the posterior correction are derived from functional forms closely associated with the efficient influence function of estimating the ATT.

In our Monte Carlo simulations, we find that our methods result in improved empirical coverage probabilities while maintaining competitive confidence interval lengths compared to existing frequentist methods. This finite sample advantage is also observed in low dimensional cases for our Bayesian method that does not involve prior or posterior corrections.[1] This can be explained by the construction of the Bayesian procedure for the ATT, which involves only a prior specification for the conditional mean function in the control arm. In particular, we note that our approach leads to more accurate uncertainty quantification and is less sensitive to estimated propensity scores that are close to boundary values. Our Bayesian methodology requires a prior specification through a likelihood function for the control arm, for which we impose an exponential family structure. In the Gaussian case, for instance, this leads to a computationally efficient procedure with the posterior being multivariate Gaussian, which avoids computational demanding methods like MCMC. We stress that the misspecification of this structure does not have serious consequences for estimation of the ATT. First, as shown in (Kleijn and van der Vaart, 2006, Section 4), nonparametric Bayesian methods possess the same robustness to misspecification as the frequentist M-estimation using least squares. Second, we provide finite sample evidence through simulations, where we find that our Bayesian procedures are not sensitive to misspecifications of the likelihood functions.

In the related literature, nonparametric Bayesian causal inference has recently received considerable interest; see, for example, the numerous applications in Daniels, Linero, and Roy (2024). Ray and van der Vaart (2020) develop the comprehensive theory for establishing the BvM theorem in the missing data framework, employing Gaussian process priors for the conditional mean function. Extending their methodology to the ATT would require nonparametric Bayesian modeling of both the propensity score and the conditional mean function for the treated group, as discussed in Remark 3.1. A key innovation of our proposal is to circumvent this route by building our Bayesian procedure on a

[1] In contrast, a Bayesian method without prior correction performs poorly for the average treatment effect (ATE) as shown by Breunig, Liu, and Yu (2025a).

reparametrization that is particularly convenient for the analysis of the ATT.

Ray and van der Vaart (2020) also propose the novel prior adjustment to the conditional mean, which makes use of the estimated propensity score. Building on this prior adjustment, Breunig, Liu, and Yu (2025a) introduce a debiasing step to further correct the posterior and establish the BvM theorem for the average treatment effect (ATE) under double robustness. Although they outline the extension to general semiparametric models where the parameter of interest can be written as the linear functional of conditional means, this approach does not cover the case of the ATT, because of its ratio form. To address the random denominator that estimates the proportion of treated individuals, we apply a new conditional Slutsky lemma introduced by Yiu, Fong, Holmes, and Rousseau (2023) in the Bayesian context. Also, the correction steps in our Bayesian method share the same motivation as in Breunig, Liu, and Yu (2025a), but differ in their functional form. This is in line with the well known subtle differences between the cases for ATE and ATT; see Hahn (1998). Additionally, we find that the bias term in the BvM for ATT is substantially simpler than that for ATE, which also explains the favorable finite-sample behavior of Bayesian ATT estimators even without posterior corrections. In contrast, Yiu, Fong, Holmes, and Rousseau (2023) suggest a different type of posterior correction that assumes stronger regularity conditions in the context of the ATT. To the best of our knowledge, our proposed double robust BvM theorem for the ATT is the first to relax the Donsker property assumption for the conditional mean (see also Remark 5.2 for a detailed comparison).

Our paper is also connected to the broader literature on robustifying standard Bayesian procedures in econometrics. Regarding Bayesian inference methods for partially or weakly identified models, we refer readers to Chen, Christensen, and Tamer (2018); Giacomini and Kitagawa (2021); Andrews and Mikusheva (2022). Under local misspecification of parametric models, Müller and Norets (2024) establish a novel BvM result utilizing the efficient influence function. There are also scattered results exploring Bayesian methodology to the study of ATT or DiD. In an earlier paper, Chib and Hamilton (2002) developed a semiparametric Bayesian model for the ATT in both cross-sectional and panel data settings. Their semiparametric model differs from our setup in that covariates enter the outcome equation linearly, while the error term is modeled using flexible Dirichlet process mixtures. Recently, in the context of assessing sensitivity to the parallel trends assumption, Kwon and Roth (2024) proposed a Bayesian approach.

The remainder of this paper is organized as follows. Section 2 presents the setup and

introduces the Bayesian framework in the DiD setup. Section 3 outlines our Bayesian methods. In Section 4, we establish inference via semiparametric BvM theorems for our first method. In Section 5, we derive a doubly robust, semiparametric BvM theorem for our second method. Section 6 provides BvM results under primitive conditions when using squared expontential process priors. Section 7 presents finite sample results via simulations and an empirical illustration. Towards the end, we outline two extensions of our methodology: Section 8 provides an extension to the staggered intervention with multiple time periods and repeated cross-sectional data. Proofs of main theoretical results are collected in Appendix A. Supplementary Appendices B–E provide additional technical results and further simulation evidence.

2 Setup and Implementation

This section provides the main setup of the average treatment effect on the treated (ATT) in the difference-in-differences (DiD) design. We first provide standard conditions for the identification of the ATT and introduce additional notations under the Bayesian formulation of the problem.

2.1 Setup

We focus on the canonical DiD design case, where there are two treatment periods and two treatment groups. Let Y_{it} be the outcome of interest for unit i at time t. We assume that researchers have access to outcome data in a pre-treatment period $t = 1$ and in a post-treatment period $t = 2$. Let $D_{it} = 1$ if unit i is treated before time t and $D_{it} = 0$ otherwise. Note that $D_{i1} = 0$ for every i and thus we may write $D_i = D_{i2}$. Using the potential outcome notation, $Y_{it}(0)$ or $Y_{it}(1)$ denotes the outcome of unit i at time t if it does not receive or receives treatment by time t, respectively. Thus, the realized outcome for unit i at time $t = 1$ is $Y_{i1} = Y_{i1}(0)$, and at time $t = 2$ it is $Y_{i2} = D_i Y_{i2}(1) + (1 - D_i) Y_{i2}(0)$. Below, P_0 denotes the frequentist distribution generating the observed data.

A vector of p-dimensional pre-treatment covariates X_i is also available, with cumulative distribution function denoted by F_X.[2] Let $\pi_0(x) = P_0(D_i = 1 \mid X_i = x)$ denote the propensity score, $\pi_0 = P_0(D_i = 1)$ the proportion, and $m_0(x) = \mathbb{E}_0[\Delta Y_i \mid D_i = 0, X_i = x]$ the conditional mean of the differenced outcome across two periods, where $\Delta Y_i := Y_{i2} - Y_{i1}$ and where $\mathbb{E}_0[\cdot]$ denotes the expectation under P_0. The researcher observes an independent

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

and identically distributed (*i.i.d.*) observations of $(Y_{i1}, Y_{i2}, D_i, X_i^\top)^\top$, $i = 1, \ldots, n$. In addition to this canonical panel data setup, we discuss how our results translate to repeated cross-sections data in Section 8.2, while an extension to staggered intervention is provided in Section 8.1. In addition to this canonical panel data setup, we provide an extension to staggered interventions in Section 8.1 and discuss how our results translate to repeated cross-sectional data in Section 8.2. For notational simplicity, we will henceforth suppress the unit index i.

Regarding the causal effect in the canonical DiD setup, the related literature primarily focuses on the average treatment effect on the treated (ATT) given by

$$\tau_0 = \mathbb{E}_0[Y_2(1) - Y_2(0)|D = 1].$$

For its identification, we impose the no anticipation assumption, conditional parallel trends (PTA) given covariates X, and the weak overlap conditions as follows.

Assumption 1. For all x in the support of F_X we have:
(i) $\mathbb{E}_0\left[Y_1(0) \mid D = 1, X = x\right] = \mathbb{E}_0\left[Y_1(1) \mid D = 1, X = x\right]$ (No Anticipation),
(ii) $\mathbb{E}_0\left[Y_2(0) - Y_1(0) \mid D = 1, X = x\right] = \mathbb{E}_0\left[Y_2(0) - Y_1(0) \mid D = 0, X = x\right]$ (PTA),
(iii) $P_0(D = 1) > \varepsilon$ and $P_0(D = 1 \mid X = x) \leqslant 1 - \varepsilon$ for some $\varepsilon > 0$ (Overlap).

Under Assumption 1 the ATT is identified by

$$\tau_0 = \mathbb{E}_0\left[\Delta Y - m_0(X) \mid D = 1\right] = \frac{\mathbb{E}_0\left[D\left(\Delta Y - m_0(X)\right)\right]}{\mathbb{E}_0[D]}. \tag{2.1}$$

One can construct an estimator that replaces the conditional mean function m_0 with an estimator, known as the outcome regression approach, as described in Heckman, Ichimura, and Todd (1997). As noted by (Abadie, 2005, p.6), plug-in estimators based on standard nonparametric estimators of the conditional mean function m_0 can face significant challenges due to the curse of dimensionality. This is where we can capitalize the strength of Bayesian estimation, which allows us to incorporate rich covariate information in the prior distribution. It has also been noted in the recent literature that, in the presence of heterogeneous treatment effects in X, i.e., when $\mathbb{E}_0[Y_2(1) - Y_2(0) \mid X = x, D = 1]$ varies with x, the two-way fixed effect estimator (TWFE) is in general not consistent for the ATT. See also Remark 1 of Sant'Anna and Zhao (2020) for an explicit discussion.

2.2 A Bayesian Framework

We now provide the formal Bayesian setup to the ATT in the DiD context. We consider a family of probability distributions $\{P_\eta : \eta \in \mathcal{H}\}$ for some parameter space $\mathcal{H}$. 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. Under P_η where $\eta = (\pi, f_X, f_{\Delta Y|D,X})$, the joint density function of $Z = (\Delta Y, D, X^\top)^\top$ can thus be written as

$$p_\eta(y, d, x) = \underbrace{f_X(x)\pi^d(x)(1-\pi(x))^{1-d} f^d_{\Delta Y|D,X}(y \mid 1, x)}_{=:f(y,d,x)} f^{1-d}_{\Delta Y|D,X}(y \mid 0, x), \tag{2.2}$$

where $f_{\Delta Y|D,X}(y \mid d, x)$ for $d = 1$ is the unrestricted conditional density of ΔY given (D, X), while for $d = 0$, we impose the exponential family condition as in (2.4). Here, f denotes the joint density of $(D\Delta Y, D, X^\top)^\top$ under P_η, and the corresponding cumulative distribution function is denoted by F. Importantly, specifying only the prior distribution on density function f and the conditional mean function m is sufficient for identifying the ATT parameter of interest. Specifically, there is no need to additionally parameterize the propensity score π, the marginal density of X, or the conditional density for the control group, $f_{\Delta Y|D,X}(y \mid 0, x)$, which is a deterministic function of $m(x)$ due to the exponential family assumption imposed in (2.4) below.

We consider the following reparametrization of (m, f) given by $\eta = (\eta^m, \eta^f)$. A central insight of this paper is to show that a nonparametric process prior specification on the conditional mean function m and the density f is sufficient for the Bayesian inference on the ATT parameter. We index the probability model by P_η, where

$$m_\eta = q^{-1}(\eta^m) \quad \text{and} \quad f_\eta = \exp(\eta^f)$$

for some known, invertible function $q(\cdot)$, which we specify below. We can write the ATT depending on a hyperparameter η as

$$\tau_\eta := \frac{\mathbb{E}_\eta[D\Delta Y - D m_\eta(X)]}{\mathbb{E}_\eta[D]}, \tag{2.3}$$

where $\mathbb{E}_\eta$ denotes the expectation under P_η and in this case, is the integral with respect to the density f_η.

As we saw above, we only need to impute the conditional mean of the outcome in the control group, making it unnecessary to impose a model for the treated group. We assume

that the distribution of ΔY, conditional on $D = 0$ and X, belongs to the "single-parameter" exponential family, where the unknown parameter is the nonparametric conditional mean function $m(x) = \mathbb{E}[\Delta Y \mid D = 0, X = x]$. Specifically, we assume that the conditional density function is given by

$$f_{\Delta Y|D,X}(y \mid 0, x) = c(y) \exp\left[q(m(x))ay - A(m(x))\right], \tag{2.4}$$

where $A(m) = \log \int c(y) \exp\left[q(m)ay\right] dy$, some constant $a > 0$, and the function $q(\cdot)$ links the conditional mean to the "natural parameter" of the exponential family. We also restrict the sufficient statistic to be linear in y. The exponential family assumption implies the conditional mean equation $\mathbb{E}[\Delta Y \mid D = 0, X = x] = A'(m(x))/(a\, q'(m(x)))$, which corresponds to generalized regression models. Interestingly, in the exponential family examples provided below, the previous equation implies $\mathbb{E}[\Delta Y \mid D = 0, X = x] = m(x)$ and hence the exponential family assumption (2.4) does not impose functional form assumptions on the conditional mean function m in these cases and, in particular, does not restrict the ATT parameter.

The family (2.4) 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 $\mathrm{Var}(\Delta Y|D = 0, X) = \sigma^2$ 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 $a = 1/\sigma$. For the normal case, we treat σ as a hyperparameter and estimate it together with hyperparameters in Gaussian process prior by maximizing marginal likelihood. We note that while a generalization to multinomial outcomes, as in Breunig, Liu, and Yu (2025a), is possible, we do not consider this case explicitly in this paper.

A high-level assumption in our BvM theorem requires the posterior contraction of m_η to the true conditional mean m_0. There are cases that this holds even if the exponential family is misspecified. Generally, the posterior of m_η will contract on near the point (pseudo-true value) in the support of the prior that minimize the Kullback–Leibler (KL) divergence with respect to the true data generating probability. Related posterior contraction results and cases where the pseudo-truth concides with m_0 can be found in Kleijn and van der Vaart (2006). This aligns with our finite sample results, which are not sensitive to deviations from exponential family distributions, as shown in Appendix E. Beyond the exponential family, one can also consider the flexiable nonparametric Bayesian approach in Norets and Pelenis (2022) to model the conditional distribution of the outcome for the control group.

Remark 2.1 (ATT in Cross-Sectional Setting)**.** *The results of our paper contribute to the literature of ATT using cross-sectional data, i.e., where an i.i.d. sample of* $(Y_i, D_i, X_i^\top)^\top$ *for* $i = 1, \ldots, n$ *is available. In this case, a specific example captured by the single-parameter exponential family is when the outcome variable is binary, where* $q(m) = \log(m/(1-m))$*,* $A(m) = -\log(1-m)$*, and* $c(y) = a = 1$*. This binary outcome case does not require any distributional assumptions. Interestingly, the sample ATT, given by* $(\sum_{i=1}^n D_i)^{-1} \sum_{i=1}^n D_i(m_0(1, X_i) - m_0(0, X_i))$*, requires only a prior on the conditional mean functions, without the need for to specify a Dirichlet process prior. On the other hand, a prior for the conditional mean function of the treatment group is also necessary in this case. We do not address a Bayesian approach for the sample ATT in this paper.*

3 Bayesian Point Estimators and Credible Sets

We now present two Bayesian procedures that build on flexible prior processes, enabling semiparametric inference on the parameter of interest. The first corresponds to a nonparametric Bayesian approach based on standard Gaussian process priors. The second involves Bayesian methods with frequentist modifications, incorporating an adjustment to the prior along with a posterior correction.

3.1 Semiparametric Bayesian Inference

We first consider nonparametric Bayesian inference, which builds on a standard Gaussian process prior for the conditional mean function combined with an independent Dirichlet process prior for the conditional expectation in (2.3). The proposed method does not include a propensity score adjustment, which prevents it from achieving double robustness, as we see in the next section. In our simulation results, however, we find that the proposed method is robust even in cases of near overlap failure.

The use of Gaussian process priors for the conditional mean has the following motivation. The mode of a posterior stemming from Gaussian process priors can be derived by a minimization problem involving the corresponding norm of a so-called reproducing kernel Hilbert space (RKHS). Gaussian process (GP) priors share close ties with spline estimation (Wahba, 1990), a connection that—along with their strong finite-sample performance—has fueled their popularity in machine learning (Rassmusen and Williams, 2006; Murphy, 2023). For other notable applications in econometrics, see Kasy (2018), Chib, Shin, and Simoni (2018) and Florens and Simoni (2021).

The Dirichlet process is default prior on spaces of probability measures. By the definition of the ATT τ_η, we assign a Dirichlet process prior to model the distribution F_η, which induces the so-called Bayesian bootstrap when the base measure of the Dirichlet process is taken to be zero; see (Rubin, 1981) and also Chamberlain and Imbens (2003).

Algorithm 1 Bayesian Procedure using Standard Gaussian Process Priors

Input: Data $Z_i = (\Delta Y_i, D_i, X_i^\top)^\top$ for $1 \leqslant i \leqslant n$ and number of posterior draws B.

Prior Specification: Select a Gaussian process prior W^m and set the prior for

$$m_\eta(x) = q^{-1}\left(\eta^m(x)\right) \qquad \text{and} \qquad \eta^m(x) = W^m(x). \tag{3.1}$$

Posterior Computation:

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

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

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

(c) Calculate a posterior draw for the ATT:

$$\tau_\eta^s = \frac{\sum_{i=1}^n M_{ni}^s D_i \big(\Delta Y_i - m_\eta^s(X_i)\big)}{\sum_{i=1}^n M_{ni}^s D_i}. \tag{3.2}$$

end for

Output: $\{\tau_\eta^s : s = 1, \ldots, B\}$

The Bayesian Algorithm 1 allows for simultaneous point estimation and uncertainty quantification. Our $100 \cdot (1-\alpha)\%$ credible set $\mathcal{C}_n(\alpha)$ for the ATT parameter τ_0 is computed by

$$\mathcal{C}_n(\alpha) = \big\{\tau : q(\alpha/2) \leqslant \tau \leqslant q(1-\alpha/2)\big\}, \tag{3.3}$$

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

For the choice of the prior process W^m, we use a Gaussian process with mean μ and the squared exponential (SE) covariance function $K(\cdot, \cdot)$ (Rassmusen and Williams, 2006, p.83) given by

$$K(x, x') := \nu^2 \exp\left(-\sum_{l=1}^p a_{ln}^2 (x_l - x_l')^2/2\right), \tag{3.4}$$

where the hyperparameter ν^2 is the kernel variance and $a_{1n}, \ldots, a_{pn}$ are rescaling

parameters that reflect the relevance of each covariate in predicting η^m. In practice, the hyperparameters μ, ν, and $a_{1n}, \ldots, a_{pn}$ can be chosen by maximizing the marginal likelihood. When the exponential family specification in (2.4) takes the Gaussian form, Step (a) of posterior computation in Algorithm 1 is analytically tractable and computationally very efficient, see Supplementary Appendix D for details. For non-Gaussian cases, one can use Laplace approximation or Monte Carlo sampling for Step (a).

3.2 A Double Robust Version

Our Bayesian approach relies on prior correction via inverse propensity score weighting (IPW) in the least favorable direction, as specified by the efficient influence function. In contrast to Abadie (2005), we do not incorporate IPW directly. Importantly, we make use of IPW for the prior and posterior correction of our Bayesian procedure. This resembles Sant'Anna and Zhao (2020) who combine OR and IPW to achieve doubly robust estimation in the frequentist setting. Following Hahn (1998); Hirano, Imbens, and Ridder (2003) or, in the DiD setup Sant'Anna and Zhao (2020), the efficient influence function for the ATT is given by

$$\widetilde{\tau}_\eta(\Delta Y, D, X) = \gamma_\eta(D, X)(\Delta Y - m_\eta(X)) - \frac{D}{\pi_\eta}\tau_\eta, \tag{3.5}$$

with its Riesz representer γ_η given by

$$\gamma_\eta(d, x) = \frac{d}{\pi_\eta} - \frac{1-d}{\pi_\eta}\frac{\pi_\eta(x)}{1-\pi_\eta(x)}. \tag{3.6}$$

We show in the Supplemental Appendix B that the Riesz representer γ_η determines the *least favorable direction* associated with the Bayesian submodel with the largest variance. Our prior adjustment using this Riesz representer provides exact invariance under shifts in nonparametric components along this direction. This extends the work of Ray and van der Vaart (2020) on unconditional average treatment effects to the ATT case, where the least favorable function, given by the efficient influence function, takes a different functional form. In a similar vein to Breunig, Liu, and Yu (2025a), we use the Riesz representer to correct for posterior bias under double robust smoothness conditions.

Our prior and posterior adjustments depend on a preliminary estimator of γ_0. A pilot estimator for the propensity score $\pi_0(\cdot)$ is denoted by $\widehat{\pi}(\cdot)$, based on an auxiliary sample, as is the estimator of the treated proportion π_0, which is taken to be the sample mean of the treatment indicators. We also make use of a pilot estimator $\widehat{m}$ for the conditional mean

function m_0. We consider a plug-in estimator for the Riesz representer γ_0 given by

$$\widehat{\gamma}(d,x) = \frac{d}{\widehat{\pi}} - \frac{1-d}{\widehat{\pi}} \frac{\widehat{\pi}(x)}{1-\widehat{\pi}(x)} = \frac{d-\widehat{\pi}(x)}{(1-\widehat{\pi}(x))\widehat{\pi}}. \tag{3.7}$$

One could also consider an estimation proportion based on the sample information of D_i alone, yet this would complicate the theoretical analysis without improving the finite sample performance of our procedure. The use of an auxiliary data for the estimation of unknown functional parameters simplifies the technical analysis and is common in the related Bayesian literature; see Ray and van der Vaart (2020) for propensity score adjusted priors in the case of missing data. 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 2 describes our double robust Bayesian procedure that approximates the posterior distribution of τ_η given in equation (2.3). Let n_c denote the number of observations in the control arm. Based on simulations, we recommend the following choices of the pilot estimators. The initial estimator $\widehat{\gamma}$, given in (3.7), is implemented based on logistic regression for the propensity scores $\pi(x)$ and the sample average of the treated proportion π. The pilot estimator for $m(x)$ is implemented by $\widehat{m}(x) = \sum_{s=1}^{B} m_\eta^s(x)/B$, where m_η^s is obtained in Step (a) of the posterior computation in Algorithm 1, and B denotes the number of posterior draws.

Algorithm 2 also leads to simultaneous point estimation and uncertainty quantification. The $100 \cdot (1-\alpha)\%$ credible set $\mathcal{C}_n(\alpha)^{DR}$ for the ATT parameter τ_0 is as in (3.3), but here $q(a)$ denotes the a quantile of $\{\check{\tau}_\eta^s : s = 1, \ldots, B\}$. The Bayesian point estimator by $\overline{\tau}_\eta^{DR} = B^{-1} \sum_{s=1}^{B} \check{\tau}_\eta^s$.

Remark 3.1 (Distinction with ATE)**.** *With cross-sectional i.i.d. data on (Y_i, D_i, X_i), Breunig, Liu, and Yu (2025a) study the Bayesian inference for the ATE. The posterior of the ATE builds on $\int [m_\eta(1,x) - m_\eta(0,x)] \mathrm{d}F_{X,\eta}(x)$, where one assigns Gaussian process priors on the conditional means $(m_\eta(1,\cdot), m_\eta(0,\cdot))$ and places a Dirichlet process prior on $F_{X,\eta}(\cdot)$. An adoption of their framework for our analysis of the ATT would lead to an alternative Bayesian method based on*

$$\tau_\eta = \frac{\int \pi_\eta(x)[m_\eta(1,x) - m_\eta(0,x)] \mathrm{d}F_{X,\eta}(x)}{\int \pi_\eta(x) \mathrm{d}F_{X,\eta}(x)},$$

which requires prior specification for each component of $(m_\eta(0,\cdot), m_\eta(1,\cdot), \pi_\eta(\cdot), F_{X,\eta}(\cdot))$. Fortunately, this is not necessary. The key observation of our current approach is that the last three components are all contained in $F_\eta(\cdot)$. As a result, we do not need to specify them

Algorithm 2 Double Robust Bayesian Procedure

Input: Data $Z_i = (\Delta Y_i, D_i, X_i^\top)^\top$ for $i = 1, \ldots, n$, number of posterior draws B, initial estimators $\widehat{\gamma}$ and $\widehat{m}$.

Prior Specification: Set the adjusted prior:

$$m_\eta(x) = q^{-1}\left(\eta^m(x)\right), \text{where} \qquad \eta^m(x) = W^m(x) + \lambda\,\widehat{\gamma}(0,x), \tag{3.8}$$

where W^m is the Gaussian process in Algorithm 1 independent of $\lambda \sim N(0, \varsigma_n^2)$, where $\varsigma_n = \nu \log n_c/(\sqrt{n_c}\,\Gamma_n)$, ν is the hyperparameter in (3.4), and $\Gamma_n = \sum_{i=1}^n |\widehat{\gamma}(0,X_i)(1 - D_i)|/n_c$.

Posterior Computation:

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

(a) Generate the s-th draw of the posterior of $(m_\eta(X_i))_{i=1}^n$ using the adjusted prior in (3.8) and the data from the control arm; denote it as $(m_\eta^s(X_i))_{i=1}^n$.

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

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

$$\check{\tau}_\eta^s = \tau_\eta^s - \widehat{b}_\eta^s, \tag{3.9}$$

where τ_η^s is given in (3.2) but using the propensity score adjusted prior from (3.8) and the posterior correction $\widehat{b}_\eta^s$ is given by

$$\widehat{b}_\eta^s = \frac{1}{n}\sum_{i=1}^n \widehat{\gamma}(D_i, X_i)(\widehat{m} - m_\eta^s)(X_i). \tag{3.10}$$

end for

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

separately when analyzing the ATT.

Remark 3.2 (Posterior Recentering)**.** *A posterior debiasing step for the posterior is required by Theorem 5.1 and, as shown in Theorem 5.2, our posterior correction term in* (3.10) *indeed allows for a derivation of the BvM result under double robust smoothness assumptions. On the other hand, the posterior correction is not required if one is willing to impose Donsker type smoothness conditions on the conditional mean function m_η, i.e., if the smoothness of m_η exceeds* $\dim(X_i)/2$*, which is an implication of Corollary 5.1.*

Posterior corrections were also proposed by Breunig, Liu, and Yu (2025a) in the context of average treatment effects (ATEs) using cross-sectional data. In their case, the bias correction term is given by $\widehat{b}_\eta^{ATE,s} = n^{-1}\sum_{i=1}^n \boldsymbol{\tau}\left[m_\eta^s - \widehat{m}\right](Z_i)$, where $\boldsymbol{\tau}[m](z) := m(1,x) - m(0,x) + \widehat{\gamma}^{ATE}(d,x)(y - m(d,x))$ is an estimator of the efficient influence function for the ATE. See also Remark 5.1 for the a more explicit comparison of the biases in both cases. We observe that double-robust Bayesian inference for the ATT, as given in (3.10)*, involves a simpler form of posterior correction compared to that for the ATE.*

Remark 3.3 (Comparison with Frequentist Estimators)**.** *Our approach is also inspired by existing frequentist methods to conduct inference on the ATT. Heckman, Ichimura, and Todd (1997) propose the following outcome imputed estimator for the ATT:*

$$\widehat{\tau}_n = \frac{\sum_{i=1}^n D_i\big(\Delta Y_i - \widehat{m}(X_i)\big)}{\sum_{i=1}^n D_i},$$

where $\widehat{m}(\cdot)$ stands for the kernel smoothing estimator of the conditional mean in the control group. The double robust version from Sant'Anna and Zhao (2020) is

$$\widehat{\tau}_n^{DR} = \frac{1}{n}\sum_{i=1}^n \widehat{\gamma}_n(D_i, X_i)\big(\Delta Y_i - \widehat{m}(X_i)\big),$$

for some pilot estimators of the propensity score and the conditional mean of the control group. In contrast, our Bayesian estimator does not directly shift the parameter via the estimated Riesz representer $\widehat{\gamma}$; rather, it enters indirectly via prior and posterior adjustments.

4 Bayesian Inference with Gaussian Process Priors

In this section, we establish a Bernstein-von Mises Theorem using standard Gaussian process priors as considered in our Bayesian procedure in Algorithm 1.

4.1 High-level Assumptions

We now provide additional notations used for the derivation of our semiparametric Bernstein-von Mises Theorem. Recall that we restrict the joint density for the control arm only, imposing the exponential family restriction as in (2.4). We denote the observed data corresponding to the treated part as $Z^{(n)}_{\text{Treat}} := (X_i, D_i, D_i\Delta Y_i)$. We express the posterior as follows:

$$\Pi\left(m \in A, F \in B \mid Z^{(n)}\right) = \int_B \frac{\int_A \prod_{i=1}^n f^{1-D_i}_{\Delta Y|D,X}\left(\Delta Y_i \mid 0, X_i\right) \mathrm{d}\Pi(m)}{\int \prod_{i=1}^n f^{1-D_i}_{\Delta Y|D,X}\left(\Delta Y_i \mid 0, X_i\right) \mathrm{d}\Pi(m)} \mathrm{d}\Pi(F \mid Z^{(n)}_{\text{Treat}}),$$

where the conditional density $f_{\Delta Y|D,X}$ is a function of the conditional mean m by the exponential family restriction given in (2.4). Here, we used the fact that independent priors are placed on the conditional mean m and the distribution function F.

We first introduce assumptions, which are high-level, and discuss primitive conditions for those in the next section. Below, we consider some measurable sets $\mathcal{H}^m_n$ of functions η^m is understood only for the control arm such that $\Pi(\eta^m \in \mathcal{H}^m_n \mid Z^{(n)}) \to_{P_0} 1$. To abuse the notation for convenience, we also denote $\mathcal{H}_n = \{\eta : \eta^m \in \mathcal{H}^m_n\}$ when we index the conditional mean function m_η by its subscript η. We write the expression $\|\phi\|_{2,F_0} := \sqrt{\int \phi^2(z) \mathrm{d}F_0(z)}$ for all $\phi \in L^2(F_0) := \{\phi : \|\phi\|_{2,F_0} < \infty\}$. When we consider the conditional moment function m below, the integral simplifies to one that depends only on the marginal distribution of X under P_0.

Assumption 2 (Rates of Convergence)**.** For some $\varepsilon_n \to 0$, $\sup_{\eta \in \mathcal{H}_n} \|m_\eta - m_0\|_{2,F_0} \leqslant \varepsilon_n$.

The posterior contraction rate for the conditional mean can be derived by modifying the classical results of Ghosal, Ghosh, and van der Vaart (2000). In the related literature, the requirement $\varepsilon_n = o(n^{-1/4})$ is stated explicitly in order to eliminate second-order remainder terms; see Condition (C) in Castillo (2012). This also aligns with the usual cut-off rate of the nonparametric components in frequentist semiparametric models (Newey, 1994). Note that the ATT τ_η is linear in m_η, so that we do not need to deal with these second-order terms. Nevertheless, the posterior contraction rate also plays a crucial role in the next two assumptions related to the stochastic equicontinuity and prior stability. For the concrete example involving the Hölder class for the conditional mean function, we need to impose sufficient smoothness so that this contraction rate indeed satisfies $\sqrt{n}\varepsilon_n^2 = o(1)$. If the exponential family structure (2.4) is misspecified, the posterior contracts to the point in the support of the prior that is closest to the true distribution (as measured by the Kullback-Leibler divergence). Specifically, if one starts with the Gaussian model, the

contraction result required by Assumption 2 can be established utilizing Theorem 4.1 of Kleijn and van der Vaart (2006), provided that the true conditional mean lies within the prior's support.

We adopt the standard empirical process notation as follows. For a function h of a random vector $Z = (Y, D, X^\top)^\top$ that follows distribution P, we let $P[h] = \int h(z)\mathrm{d}P(z)$, $\mathbb{P}_n[h] = n^{-1}\sum_{i=1}^n h(Z_i)$, and $\mathbb{G}_n[h] = \sqrt{n}\left(\mathbb{P}_n - P\right)[h]$. The next set of assumptions restrict the complexity of the conditional mean functions. The first part requires the class $\{m_\eta : \eta \in \mathcal{H}_n\}$ to be Glivenko-Cantelli, plus some mild moment conditions on its envelope function. The second part imposes the stochastic equicontinuity, which is holds when the conditional mean function belong to a Donsker class. For the Hölder class considered in Section 6, this enforces the sufficient smoothness of those functions relative to the dimensionality of covariates.

Assumption 3 (Complexity)**.** (i) $\sup_{\eta\in\mathcal{H}_n}|(\mathbb{P}_n - P_0)m_\eta| = o_{P_0}(1)$ and $\{m_\eta : \eta \in \mathcal{H}_n\}$ has an envelope function $M(\cdot)$ with $P_0 M^{2+\delta} < \infty$ for some constant $\delta > 0$ and (ii) $\sup_{\eta\in\mathcal{H}_n}|\mathbb{G}_n\left[m_\eta - m_0\right]| = o_{P_0}(1)$.

The next assumption concerns the prior stability condition, which is common to semiparametric Bayesian inference Ghosal and Van der Vaart (2017). This facilitates the technical proof for which we need to consider the perturbation along the least favorable direction. For standard parametric models, the absolute continuity of the prior density suffices. However, for nonparametric priors, the very notion of a Radon-Nikodym density is non-trivial, and one needs to apply the Cameron-Martin theorem; see Proposition I.20 in Ghosal and Van der Vaart (2017). For that purpose, we introduce some necessary terminologies related to the general Gaussian process. Such a process determines a so-called reproducing kernel Hilbert space (RKHS) $(\mathbb{H}^m, \|\cdot\|_{\mathbb{H}^m})$.

Our Bayesian method based on standard Gaussian process priors in Algorithm 1 does not include a correction involving the Riesz representer γ_0 as defined in (3.6). Yet to establish prior stability, an approximation condition for γ_0 is imposed, requiring sufficient regularity of the propensity score $\pi_0(\cdot)$. We introduce the ball in $\mathbb{H}^m$ centered at the true Riesz representer γ_0 given by

$$\mathbb{H}^m(r_n) := \left\{h \in \mathbb{H}^m : \|h - \gamma_0\|_\infty \leqslant r_n \text{ and } \|h\|_{\mathbb{H}^m} \leqslant \sqrt{n}r_n\right\}$$

for some rate r_n, where $\|\cdot\|_\infty$ denotes the supremum norm.

Assumption 4 (Prior Stability)**.** There exists $\overline{\gamma}_n \in \mathbb{H}^m(\zeta_n)$ for a sequence $\zeta_n = o(1)$

with $\sqrt{n}\,\varepsilon_n\zeta_n = o(1)$ where ε_n is the posterior contraction rate in Assumption 2. Further, $\Pi(\eta^m \in \mathcal{H}_n^m - t\overline{\gamma}_n n^{-1/2}|Z^{(n)}) \to_{P_0} 1$ for every $t \in \mathbb{R}$.

Assumption 4 imposes an approximation condition to the Riesz representer γ_0 via the restriction $\overline{\gamma}_n \in \mathbb{H}^m(\zeta_n)$. Based on this assumption, we provide the proof of this prior stability in Supplementary Appendix C.3. In comparison, the prior correction weakens the requirement with the help of a pilot estimator of the propensity score, pioneered by Ray and van der Vaart (2020). Under propensity score adjusted priors analyzed in the next section, Breunig, Liu, and Yu (2025a) the approximation condition even holds under double robustness.

4.2 A BvM Theorem

We now establish a Bernstein-von Mises Theorem for our nonparametric Bayesian method based on standard Gaussian process priors. When it comes to the centering point of the posterior, we consider an asymptotically efficient estimator $\widehat{\tau}$ with the following linear representation:

$$\widehat{\tau} = \tau_0 + \frac{1}{n}\sum_{i=1}^{n} \widetilde{\tau}_0(Z_i) + o_{P_0}(n^{-1/2}), \tag{4.1}$$

where $\widetilde{\tau}_0 = \widetilde{\tau}_{\eta_0}$ is the efficient influence function given in (3.5). Below, we write $\mathcal{L}_\Pi(\sqrt{n}(\tau_\eta - \widehat{\tau})|Z^{(n)})$ for the marginal posterior law of $\sqrt{n}(\tau_\eta - \widehat{\tau})$.

This asymptotic equivalence result is established using the so called *bounded Lipschitz distance*. For two probability measures P, Q defined on a metric space $\mathcal{Z}$, we define the bounded Lipschitz distance as

$$d_{BL}(P,Q) = \sup_{f \in BL(1)} \left| \int_{\mathcal{Z}} f(\mathrm{d}P - \mathrm{d}Q) \right|, \tag{4.2}$$

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, that is derived from the Bayes rule given the prior specification and the observed data $Z^{(n)}$. 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 3, to approximate this posterior distribution, as well as the resulting point estimator and credible set.

Theorem 4.1. *Let Assumptions 1–4 hold. Then, using standard Gaussian process priors*

(3.1) *on* η^m *and an independent Dirichlet process prior on* F*, we have*

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

As a result, the posterior mean $\overline{\tau}_\eta$ *given in Section 3 satisfies* $\sqrt{n}\,(\overline{\tau}_\eta - \tau_0) \Rightarrow N(0, \mathrm{v}_0)$ *under* P_0*. Furthermore, for any* $\alpha \in (0,1)$*, the Bayesian credible set* $\mathcal{C}_n(\alpha)$ *given in Section 3 satisfies* $P_0\big(\tau_0 \in \mathcal{C}_n(\alpha)\big) \to 1 - \alpha$*.*

Theorem 4.1 establishes the BvM result for our Bayesian procedure using standard Gaussian process priors. The entropy condition uniformly over $\eta \in \mathcal{H}_n$ is satisfied if m_η is sufficiently smooth, that is, if m_η belong to a fixed F_0-Donsker class and, in particular, rules out double robustness. On the other hand, note that the asymptotic equivalence is obtained without any adjustment of prior or correction to posterior distributions, so the full Bayesian flavor is preserved.

5 Bayesian Inference under Double Robustness

In this section, we show that the Bayesian procedure in Algorithm 2, which employs prior and posterior adjustments, satisfies the Bernstein-von Mises Theorem under double robust smoothness conditions. Herein, we clarify the notion of double robustness. In the earlier development, the focus is typically on developing working parametric models for either the propensity score or the conditional mean function, and the double robust estimation hedges against the risk of model misspecification. However, implausible parametric assumptions on the data generating process are of limited applicability to complex phenomena in economics. Recent advances in the double machine learning literature have led to a number of important developments in causal inference, utilizing flexible nonparametric or machine learning algorithms. In this context, double robustness means the possibility to trade off the estimation accuracy between nuisance functions.

5.1 High-level Assumptions

Below, we present the assumptions that enable double-robust inference through our propensity score adjustments to the prior and posterior distributions.

Assumption 5 (DR Rates of Convergence)**.** The estimators $\widehat{\pi}$ and $\widehat{m}$, which are based on

an auxiliary sample independent of $Z^{(n)}$, satisfy $\|\widehat{\pi} - \pi_0\|_{2,F_0} = O_{P_0}(r_n)$,

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

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

Assumption 5 imposes sufficiently fast convergence rates for the estimators for the conditional mean function m_0 and the propensity score π_0. The posterior convergence rate for the conditional mean can be derived by modifying the classical results of Ghosal, Ghosh, and van der Vaart (2000) by accommodating the propensity score-adjusted prior, in the same spirit of Ray and van der Vaart (2020). We refer to Breunig, Liu, and Yu (2025a) who showed that this assumption allows for double robustness under Hölder type smoothness assumptions.

Assumption 6 (DR Stochastic Equicontinuity)**.** $\sup_{\eta \in \mathcal{H}_n} |\mathbb{G}_n [(\gamma_0 - \widehat{\gamma})(m_\eta - m_0)]| = o_{P_0}(1)$.

Assumption 6 restricts the functional class $\mathcal{H}_n$ to form a P_0-Glivenko-Cantelli class; see Section 2.4 of van der Vaart and Wellner (2023) and imposes a stochastic equicontinuity condition on a product structure involving $\widehat{\gamma}$ and m_η. Hence, the complexity of the functional class $(m_\eta - m_0)$ can be compensated by certain high regularity of the corresponding Riesz representer and vice versa. This condition adapts the complexity requirement of Breunig, Liu, and Yu (2025a) by only restricting the control arm.

Recall the propensity score-dependent prior on m given in (3.8), i.e., $m(\cdot) = q^{-1}(W^m(\cdot) + \lambda\widehat{\gamma}(\cdot))$. Below, we restrict the behavior for λ through its hyperparameter $\varsigma_n > 0$. 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 7 (DR Prior Stability)**.** W^m is a continuous stochastic process independent of the normal random variable $\lambda \sim N(0, \varsigma_n^2)$, where $\varsigma_n \lesssim 1$, $n\varsigma_n^2 \to \infty$ and that satisfies (i) $\Pi\left(\lambda : |\lambda| \leqslant u_n\varsigma_n^2\sqrt{n} \mid Z^{(n)}\right) \to_{P_0} 1$, 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 $t \in \mathbb{R}$.

Assumption 7 incorporates Conditions (3.9) and (3.10) from Theorem 2 in Ray and van der Vaart (2020), and it is imposed to establish the stability property of the adjusted prior distribution. We will provide sufficient conditions for Assumption 7 in Section 6.

5.2 Double Robust BvM Theorems

We now establish a semiparametric Bernstein–von Mises theorem for our double robust Bayesian procedure given in Algorithm 2.

Theorem 5.1. *Let Assumptions 1, 3(i), 5, 6, and 7 hold. Consider the propensity score adjusted prior* (3.8) *on* η^m *and an independent Dirichlet process prior on* F*. Then we have*

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

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

Theorem 5.1 shows that, under double-robust smoothness conditions, the BvM theorem holds only up to a "bias term" $b_{0,\eta}$, which depends on the unknown conditional mean m_0. This biased posterior makes the BvM not feasible in practice. We also emphasize that the derivation of this result is different to the BvM results in Breunig, Liu, and Yu (2025a), as we need to control the denominator in the asymptotic expansions.

Remark 5.1 (Comparison of Bias in ATE/ATT Posteriors)**.** *Breunig, Liu, and Yu (2025a) showed that, for inference on the ATE in the cross-sectional case, the BvM holds only for a biased posterior under double robust smoothness conditions, see also Remark 3.2. This "bias term" is closely related to the influence function of the ATE, which takes the following form*

$$b_{0,\eta}^{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\underbrace{\frac{D_i}{\pi_0(X_i)} - \frac{1-D_i}{1-\pi_0(X_i)}}_{=:\gamma_0^{ATE}(D_i,X_i)}\right)(m_0(D_i,X_i) - m_\eta(D_i,X_i)) - (\bar{m}_0(X_i) - \bar{m}_\eta(X_i))\right\},$$

where $\bar{m}_0(\cdot) = m_0(1,\cdot) - m_0(0,\cdot)$, $\bar{m}_\eta(\cdot) = m_\eta(1,\cdot) - m_\eta(0,\cdot)$*, and the Riesz representer* γ_0^{ATE} *as given in the ATE case, see Breunig, Liu, and Yu (2025a). Referring to the influence function of the ATT, we can also express it in terms of the conditional mean* $m_0(D,X)$ *involving both treated and control groups, cf. Equation (8.5) in Van der Laan and Rose (2011). Therefore, we have the following expression for the bias term in the ATT case:*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\underbrace{\frac{D_i}{\pi_0} - \frac{1-D_i}{\pi_0}\frac{\pi_0(X_i)}{1-\pi_0(X_i)}}_{=\gamma_0(D_i,X_i)}\right)(m_0(D_i,X_i) - m_\eta(D_i,X_i)) - \frac{D_i}{\pi_0}(\bar{m}_0(X_i) - \bar{m}_\eta(X_i))\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\gamma_0(D_i,X_i)\,(m_0(0,X_i) - m_\eta(0,X_i)) = b_{0,\eta}^{ATT},$$

where the simplification occurs because the term $(D_i/\pi_0)(m_0(1, X_i) - m_\eta(1, X_i))$ *cancels out in the difference.*

The resulting simplification of the bias term aligns with our simulation results, which show that standard Gaussian process priors also provide accurate coverage for the ATT in many cases.

The next result is an immediate implication of Theorem 5.1. Specifically, it provides a Bernstein-von Mises Theorem for Bayesian procedures that do not rely on posterior correction. This can be achieved if the bias term is asymptotically negligible uniformly over the class of hyperparameters, which requires more restrictive smoothness conditions on the conditional mean function m_0.

Corollary 5.1. *Let Assumptions 1, 3(i), 5, 6, and 7 hold. Consider the propensity score adjusted prior* (3.8) *on* η^m *and an independent Dirichlet process prior on* F*. If, in addition,* $b_{0,\eta} = o_{P_0}(n^{-1/2})$ *uniformly for* $\eta \in \mathcal{H}_n$*, then we have*

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

While Corollary 5.1 allows for arbitrarily low regularity of propensity scores, it requires the conditional mean function to be sufficiently smooth; specifically, the smoothness of m must be greater than or equal to $\dim(X_i)/2$ (also referred to as the Donsker property). This condition is also called *single robustness* by Ray and van der Vaart (2020), and indeed, this corollary extends their findings to the inference on the ATT. Also, as they point out, propensity score adjusted priors (3.8) relax the uniformity condition $\sup_{\eta \in \mathcal{H}_n} |\mathbb{G}_n [m_\eta - m_0]| = o_{P_0}(1)$ used in Theorem 4.1 under standard Gaussian process priors.

Under double robust assumptions, however, the Bayesian procedure that achieves the BvM equivalence in Theorem 5.1 is not feasible, because it depends on the term $b_{0,\eta}$, which is a function of the unknown conditional mean m_0. Our objective is to maintain double robust conditions, while considering pilot estimators for the unknown functional parameters in $b_{0,\eta}$. The correction term $\widehat{b}_\eta$, as introduced in (3.10), results in a feasible Bayesian procedure that satisfies the BvM theorem, as demonstrated below.

Theorem 5.2. *Let Assumptions 1, 3(i), 5, 6, and 7 hold. Consider the propensity score adjusted prior* (3.8) *on* η^m *and an independent Dirichlet process prior on* F*. Then we have*

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

where $\widehat{b}_\eta = \mathbb{P}_n[(\widehat{m}-m_\eta)\widehat{\gamma}]$. *As a result, the posterior mean* $\overline{\tau}_\eta^{DR}$ *given in Section 3.2 satisfies* $\sqrt{n}\left(\overline{\tau}_\eta^{DR} - \tau_0\right) \Rightarrow N(0, \mathrm{v}_0)$ *under* P_0. *Furthermore, for any* $\alpha \in (0,1)$, *the Bayesian credible set* $\mathcal{C}_n^{DR}(\alpha)$ *given in Section 3.2 satisfies* $P_0\big(\tau_0 \in \mathcal{C}_n^{DR}(\alpha)\big) \to 1-\alpha$.

Theorem 5.2 shows that the Bayesian method proposed in Algorithm 2, $\check{\tau}_\eta = \tau_\eta - \widehat{b}_\eta$, achieves the BvM result under double robust smoothness conditions. The following remark clarifies the relationship when considering posterior correction alone, in which case BvM results are available only under more restrictive smoothness assumptions on the propensity score and the conditional mean function.

Remark 5.2. *Building on the idea of a one-step update in frequentist semiparametric estimation, Yiu, Fong, Holmes, and Rousseau (2023) propose a different method of posterior correction (without prior adjustment) that involves the efficient influence function. When applying their methodology to the ATT, it is evident that both the conditional mean function and the propensity score must satisfy the Donsker property, cf. Assumption 4(c) therein. In contrast, the relaxation of the Donsker property is one of the key technical innovation of our double robust Bayesian inference.*

6 Illustration under Low-level Conditions

In this section, we provide primitive conditions for the assumptions used to derive the BvM Theorems. To do so, we focus on squared exponential process priors as an example of Gaussian process priors. Moreover, we consider specific smoothness classes to derive the explicit regularity conditions implied by our high-level assumptions.

A Gaussian process (GP) is completely characterized by its mean and covariance functions (Rassmusen and Williams, 2006). Below we consider a GP prior, which has mean zero and the covariance function specified by $\mathbb{E}[W(s)W(t)] = \exp(-\|s-t\|_{\ell_2}^2)$. This so-called squared exponential process prior, which is one of the most commonly used priors in applications; see Rassmusen and Williams (2006) and Murphy (2023). Following (Breunig, Liu, and Yu, 2025a), we consider a rescaled Gaussian process $\big(W(a_n t) : t \in [0,1]^p\big)$. Intuitively, a_n^{-1} can be thought as a bandwidth parameter. For a large a_n, the prior sample path $t \mapsto W(a_n t)$ is obtained by shrinking the long sample path $t \mapsto W(t)$. Thus, it incorporates 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 $s_m > 0$. Specifically, we illustrate our theory with the case where $m_0 \in \mathcal{C}^{s_m}([0,1]^p)$. Given such

a Hölder-type smoothness condition, we choose

$$a_n \sim n^{1/(2s_m+p)}(\log n)^{-(1+p)/(2s_m+p)}. \tag{6.1}$$

The particular choice of a_n mimics the corresponding kernel bandwidth based on any kernel smoothing method. Note that the minimax posterior contraction rate for the conditional mean function m_η 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).

Proposition 6.1 (Unadjusted Squared Exponential Process Priors)**.** *Suppose $m_0 \in \mathcal{C}^{s_m}([0,1]^p)$ and $\pi_0 \in \mathcal{C}^{s_\pi}([0,1]^p)$ under the smoothness conditions $\min(s_\pi, s_m) > p/2$. Consider the prior on m given by $m(x) = q^{-1}\left(W^m(x)\right)$, where W^m is the rescaled squared exponential process, with its rescaling parameter a_n of the order in* (6.1)*, combined with an independent Dirichlet process prior on F. Then, under Assumption 1, the posterior distribution for the ATT satisfies Theorem 4.1.*

Proposition 6.1 makes explicit the smoothness requirements for the BvM Theorem to hold when standard Gaussian process priors are placed on the conditional mean function m. This result shows that the smoothness of both the conditional mean function and the propensity score function must exceed $\dim(X)/2$. Conversely, in situations where one is confident that these regularity conditions are met, no additional modifications to the Bayesian procedures are necessary to achieve the BvM result.

Proposition 6.2 (Adjusted 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 $m_0 \in \mathcal{C}^{s_m}([0,1]^p)$ and some $s_m > 0$ with $\sqrt{s_\pi\, s_m} > p/2$. Also, $\|\widehat{m} - m_0\|_{2,F_0} = O_{P_0}\big((n/\log n)^{-s_m/(2s_m+p)}\big)$. Consider a Dirichlet process prior on F combined with the independent prior on m given by $m(x) = q^{-1}\left(W^m(x) + \lambda\,\widehat{\gamma}(0,x)\right)$, where W^m is the rescaled squared exponential process, with rescaling parameter a_n satisfying* (6.1) *and $(n/\log n)^{-s_m/(2s_m+p)} \lesssim u_n\varsigma_n$ for some deterministic sequence $u_n \to 0$, and $\varsigma_n \lesssim 1$. Then, under Assumption 1, the corrected posterior distribution for the ATT satisfies Theorem 5.1.*

Proposition 6.2 requires $\sqrt{s_\pi\, s_m} > p/2$, which represents a trade-off between the smoothness requirement for m_0 and π_0. This corresponds to the *double robustness* property; i.e., a lack of smoothness of the conditional mean function m_0 can be mitigated by exploiting the regularity of the propensity scoreπ_0, and vice versa.

7 Finite Sample Results

This section investigates the finite sample performance of the proposed Bayesian estimation/inference approaches and then apply them to the well-known DiD study of Card and Krueger (1994).

7.1 Simulation Evidence

We now present Monte Carlo simulation results to compare our proposed semiparametric Bayesian methods with existing frequentist approaches. Consider the following data generating process (DGP) for observed variables $(Y_{i1}, Y_{i2}, D_i, X_i^\top)^\top$ given by

$$X_i \sim \mathcal{N}\left((1,-1,1,-1,\ldots,(-1)^{p-1})^\top, \Sigma\right) \quad \text{and} \quad D_i \mid X_i \sim \text{Bernoulli}\left(\Psi\left[g(X_i)\right]\right),$$

where the covariance matrix $\Sigma = (\Sigma_{jk})_{1\leqslant j,k,\leqslant p}$ is determined by $\Sigma_{jk} = 0.5^{|k-j|}$. We generate outcomes in two periods:

$$\begin{aligned} Y_{i1} &= h(X_i) + D_i\mu(X_i) + \alpha_i + \epsilon_{i1}, \\ Y_{i2}(d) &= 2 + 2\mu(X_i) + D_i\mu(X_i) + \alpha_i + \epsilon_{i2}(d), \text{ for } d = 0, 1. \end{aligned}$$

We consider the following four different designs based on different specifications of the functions g and h:

Design I: $g(x) = 0.5\sum_{j=1}^p x_j/j$, $h(x) = \sum_{j=1}^p x_j/j$,

Design II: $g(x) = 0.5\sum_{j=1}^p x_j/j$, $h(x) = 0.8\sum_{j=1}^p x_j/j + 0.2\sum_{j=1}^p x_j^2/j$,

Design III: $g(x) = \left(0.5\sum_{j=1}^p x_j/j + 0.5\sum_{j=1}^p x_j^2/j\right)/4$, $h(x) = \sum_{j=1}^p x_j/j$,

Design IV: $g(x) = \left(0.5\sum_{j=1}^p x_j/j + 0.5\sum_{j=1}^p x_j^2/j\right)/4$, $h(x) = 0.8\sum_{j=1}^p x_j/j + 0.2\sum_{j=1}^p x_j^2/j$.

The fixed effect α_i and the error terms are standard normal, with $(\alpha_i, \epsilon_{i1}, \epsilon_{i2}(0), \epsilon_{i2}(1))^\top \sim \mathcal{N}(0, I_4)$, where I_4 denotes the four-dimensional identity matrix.[3] The true ATT is zero for all cases. Our DGPs follow the structure of DGP1 in Sant'Anna and Zhao (2020) but allow for more covariates and a possibly nonlinear function $h(\cdot)$. In our simulations, we analyze the effect of varying the dimension of covariates $p \in \{5, 10, 20\}$ and varying sample

[3]Tables A4 and A5 in the Supplementary Appendix E present additional simulation results for cases where the error terms follow chi-squared distributions or include heteroskedasticity. The finite-sample performance in these cases is similar to that observed in Table 1 for standard normal errors.

sizes $n \in \{500, 1000\}$. Throughout our simulations, the number of Monte Carlo replication is set to 1000.

Our nonparametric Bayesian (hereinafter **Bayes**) and the double robust Bayesian (**DR Bayes**) methods are implemented following Algorithms 1 and 2 in Section 3, using the MATLAB package `GPML` to draw posteriors. Both Bayesian methods are implemented based on $B = 5000$ posterior draws. Here, we apply the exponential family specification in (2.4) to the Gaussian case. The resulting posterior distribution of the conditional mean function is available in closed form (see Supplementary Appendix D for details), eliminating the need for computationally costly Monte Carlo samplers like MCMC.[4] The tuning parameter ς_n for DR Bayes, which corresponds to the standard deviation of the adjusted prior, is set according to the prior specification step in Algorithm 2. In Supplementary Appendix E, Table A1 demonstrates that the performance of DR Bayes is stable with respect to the value of ς_n. DR Bayes in Table 1 uses the full sample twice in computing the prior/posterior adjustments and the posteriors of the conditional mean function. As shown in Table A2 in Supplementary Appendix E, results from sample splitting are comparable to those in Table 1.

We also compare the Bayesian methods to several frequentist DiD estimators. **DR** corresponds to the improved doubly robust DiD estimator proposed by Sant'Anna and Zhao (2020). **OR**, the outcome regression approach, refers to the sample analog of (2.1) where the conditional mean m_0 is estimated by a linear regression of ΔY_i on X_i using the sample of the control arm. Two types of inverse propensity score weighted (IPW) estimators are considered: $\mathbf{IPW^{HT}}$ refers to the IPW estimator in Abadie (2005), which is of the Horvitz and Thompson (1952) type. $\mathbf{IPW^{Hájek}}$ refers to the Hájek (1971) type IPW estimator that normalizes the weights to sum up to one.[5] **TWFE** corresponds to the standard two-way-fixed effect model that regresses Y_{it} on D_i, t, the interaction $D_i \times t$ and X_i. **DML** corresponds to the double/debiased machine learning ATT estimator of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017) or Chang (2020), where the nuisance function π_0 is estimated by logistic LASSO and m_0 estimated by random forests.[6] Table 1 presents the finite sample (mean) bias of the point estimator, coverage

[4] If the conditional density function in (2.4) belongs to another distribution in the exponential family, the posterior of the conditional mean can also be approximated using an analytical approximation, such as the Laplace method, see Riihimäki and Vehtari (2014).

[5] The expression for the Hájek–type IPW DiD estimator is given in equation (4.1) of Sant'Anna and Zhao (2020).

[6] We apply random forest to estimate m_0 to cope with the nonlinear function forms in Designs II and IV, and to match DML with our Bayesian procedures that estimate m_0 nonparametrically. Frequentist DiD estimators, except for DML, are implemented using the R package `DRDID`, while DML is implemented using the R package `DoubleML`.

probability (CP) and the average length (CIL) of the 95% credible/confidence interval for the Bayesian and frequentist methods mentioned above.

Table 1: Simulation results for Designs I and II, correctly-specified propensity score.

Design		Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
I	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.033	0.962	0.615	0.048	0.927	0.628	0.076	0.909	0.650
	DR Bayes	0.010	0.960	0.628	0.012	0.934	0.657	0.028	0.926	0.691
	DR	-0.012	0.943	0.734	-0.004	0.908	0.664	0.001	0.903	0.620
	OR	0.001	0.956	0.585	0.004	0.937	0.592	0.007	0.941	0.609
	IPW^{HT}	0.025	0.939	1.717	-0.019	0.933	1.987	-0.013	0.923	2.311
	$\text{IPW}^{\text{Hájek}}$	0.026	0.938	1.075	0.004	0.909	1.202	0.011	0.909	1.313
	TWFE	2.306	0.000	1.114	2.575	0.000	1.141	2.690	0.000	1.154
	DML	0.012	0.964	0.883	0.076	0.938	1.061	0.213	0.915	1.278
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.018	0.943	0.428	0.018	0.949	0.434	0.035	0.938	0.445
	DR Bayes	0.004	0.948	0.438	-0.001	0.950	0.448	0.006	0.948	0.464
	DR	-0.002	0.936	0.524	-0.006	0.934	0.479	-0.002	0.928	0.451
	OR	0.004	0.948	0.413	-0.004	0.945	0.416	0.002	0.951	0.423
	IPW^{HT}	-0.002	0.944	1.236	0.001	0.949	1.345	0.005	0.939	1.506
	$\text{IPW}^{\text{Hájek}}$	0.004	0.936	0.788	0.003	0.941	0.850	0.010	0.933	0.922
	TWFE	2.310	0.000	0.790	2.573	0.000	0.808	2.694	0.000	0.820
	DML	0.003	0.963	0.594	0.039	0.961	0.710	0.144	0.929	0.875
II	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.086	0.908	0.635	0.111	0.872	0.649	0.139	0.857	0.672
	DR Bayes	0.034	0.933	0.637	0.043	0.904	0.665	0.063	0.908	0.697
	DR	0.013	0.941	0.774	0.025	0.897	0.723	0.048	0.904	0.685
	OR	0.263	0.695	0.713	0.274	0.684	0.725	0.280	0.695	0.740
	IPW^{HT}	0.038	0.926	2.144	-0.023	0.915	2.505	-0.015	0.920	2.934
	$\text{IPW}^{\text{Hájek}}$	0.039	0.926	1.392	0.006	0.898	1.499	0.017	0.898	1.574
	TWFE	2.165	0.000	1.263	2.359	0.000	1.282	2.450	0.000	1.290
	DML	0.014	0.951	1.010	0.070	0.917	1.191	0.204	0.903	1.414
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.047	0.929	0.440	0.058	0.914	0.449	0.078	0.894	0.460
	DR Bayes	0.015	0.942	0.442	0.014	0.937	0.453	0.026	0.931	0.467
	DR	0.008	0.939	0.558	0.011	0.927	0.525	0.028	0.915	0.503
	OR	0.257	0.505	0.507	0.264	0.457	0.512	0.274	0.457	0.520
	IPW^{HT}	-0.001	0.942	1.563	0.004	0.938	1.708	0.012	0.934	1.935
	$\text{IPW}^{\text{Hájek}}$	0.006	0.926	1.035	0.006	0.932	1.073	0.018	0.923	1.135
	TWFE	2.162	0.000	0.896	2.356	0.000	0.909	2.457	0.000	0.918
	DML	0.000	0.940	0.666	0.037	0.948	0.780	0.140	0.914	0.970

Concerning the Bayesian DiD for estimating the ATT, Table 1 shows that the nonparameteric Bayes performs well in Design I, but undercovers by 9% to 14% in Design II when $p = 10$ and 20. DR Bayes improves the coverage probability of nonparameteric Bayesian inference in these cases and performs well across both designs, different dimensions p, and sample sizes n. The point estimator produced by DR Bayes also leads to a smaller bias than the nonparametric Bayes. On the other hand, the nonparametric Bayes yields shorter confidence intervals. We also see that for large values of p, nonparametric Bayes tends to undercover.

In Table 1, the frequentist DiD estimators DR, two types of IPW, and DML – each of which is double robust or at least robust to misspecification in the conditional mean function – exhibit good coverage performance in both designs. Among them, DR produces slightly longer CIs than our DR Bayes in most cases; IPW estimators yield longer CIs than most of other methods, including both Bayesian procedures; and DML yields slightly shorter CIs than DR Bayes in Design I but noticeably longer CIs in Design II. Unsurprisingly, OR suffers from severe undercoverage in Design II. TWFE performs poorly in both designs, where the time trend is linearly correlated with the covariates X, as also documented in Sant'Anna and Zhao (2020).

Table 2 presents the finite sample performance of aforementioned DiD estimators when the function $g(\cdot)$, used to specification of the propensity score, is nonlinear. DR Bayes, DR, and IPW estimate the propensity score using logistic regression, while DML uses logistic LASSO. Table 2 illustrates the impact of misspecifying the propensity score on these methods. DR Bayes maintains reasonably good performance in Design III. Although its performance deteriorates in Design IV, particularly as dimensionality of X increases, it still outperforms frequentist estimators, including double-robust methods like DR and DML. Nonparametric Bayes, using standard Gaussian process priors, avoids estimating the propensity score under misspecification and hence performs well in Design III while outperforming all other methods in Design IV.

Table 2: Simulation results for Designs III and IV, misspecified propensity score.

Design		Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
III	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.029	0.937	0.593	0.030	0.941	0.607	0.060	0.900	0.659
	DR Bayes	0.019	0.925	0.559	0.020	0.924	0.577	0.047	0.871	0.610
	DR	-0.026	0.951	0.691	-0.012	0.930	0.631	-0.009	0.911	0.614
	OR	0.000	0.955	0.561	0.004	0.942	0.569	0.003	0.926	0.604
	IPW^{HT}	0.301	0.760	0.931	0.243	0.835	0.959	0.250	0.871	1.244
	$\text{IPW}^{\text{Hájek}}$	0.220	0.783	0.745	0.185	0.838	0.785	0.190	0.864	0.944
	TWFE	1.258	0.021	1.178	1.047	0.120	1.258	1.083	0.123	1.321
	DML	0.099	0.919	0.700	0.171	0.875	0.779	0.313	0.792	0.943
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.015	0.944	0.411	0.012	0.938	0.417	0.018	0.927	0.439
	DR Bayes	0.010	0.938	0.390	0.007	0.926	0.398	0.010	0.921	0.419
	DR	-0.014	0.938	0.488	-0.008	0.944	0.446	-0.010	0.931	0.438
	OR	0.002	0.944	0.395	-0.002	0.950	0.398	-0.002	0.936	0.418
	IPW^{HT}	0.294	0.569	0.644	0.238	0.686	0.620	0.240	0.744	0.737
	$\text{IPW}^{\text{Hájek}}$	0.213	0.638	0.524	0.180	0.740	0.525	0.182	0.759	0.600
	TWFE	1.268	0.000	0.835	1.047	0.010	0.892	1.072	0.014	0.943
	DML	0.074	0.895	0.467	0.123	0.852	0.516	0.228	0.726	0.621
IV	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.088	0.920	0.614	0.095	0.883	0.626	0.152	0.816	0.683
	DR Bayes	0.075	0.901	0.567	0.082	0.862	0.585	0.135	0.796	0.614
	DR	0.151	0.856	0.729	0.179	0.797	0.687	0.197	0.760	0.678
	OR	0.253	0.659	0.645	0.248	0.681	0.653	0.257	0.686	0.691
	IPW^{HT}	0.493	0.579	1.129	0.446	0.656	1.169	0.467	0.733	1.531
	$\text{IPW}^{\text{Hájek}}$	0.393	0.596	0.936	0.368	0.637	0.945	0.379	0.670	1.070
	TWFE	1.401	0.004	1.287	1.235	0.057	1.343	1.264	0.046	1.398
	DML	0.165	0.850	0.736	0.273	0.735	0.802	0.457	0.538	0.969
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.052	0.914	0.423	0.057	0.897	0.429	0.076	0.881	0.452
	DR Bayes	0.045	0.891	0.393	0.049	0.882	0.402	0.064	0.859	0.423
	DR	0.153	0.780	0.516	0.176	0.695	0.486	0.186	0.660	0.485
	OR	0.247	0.444	0.456	0.235	0.479	0.458	0.245	0.482	0.480
	IPW^{HT}	0.477	0.330	0.790	0.437	0.378	0.764	0.456	0.469	0.912
	$\text{IPW}^{\text{Hájek}}$	0.377	0.403	0.667	0.358	0.403	0.647	0.372	0.431	0.709
	TWFE	1.403	0.000	0.912	1.230	0.000	0.952	1.257	0.001	0.998
	DML	0.122	0.816	0.485	0.208	0.664	0.525	0.361	0.376	0.643

7.2 Empirical application: Minimum Wage

We apply Bayesian DiD methods to the well-known minimum wage study of Card and Krueger (1994).[7] The outcome variables Y_{1i} and Y_{2i} are full time equivalent (FTE) employment of fast-food stores in New Jersey and Pennsylvania before and after New Jersey's raise of minimum wage. The treatment variable takes one for fast-food stores in New Jersey and zero otherwise. The set of covariates X includes the twelve store characteristics surveyed before the minimum wage change: indicator for company ownership, three chain type dummies, numbers of managers, cash registers and hours open per weekday, time to the first wage raise, indicator for offering recruitment bonus, item prices of medium soda, small french fries and a main course. We would like to see whether the findings of Card and Krueger (1994) which considers as store characteristics the company ownership indicator and chain type dummies in their regression-adjusted model would change if more covariates are included and a flexible functional form of $m_0(x)$ is allowed.

The sample size is 307. Since the data contains a non-negligible proportion of units with propensity score estimates very close to 1, we follow Crump, Hotz, Imbens, and Mitnik (2009) and discard observations with the estimated propensity score outside the range $(0, 1-t]$, with the trimming threshold $t \in \{0.05, 0.01\}$.

Table 3: Estimates of ATT for the minimum wage increase: sample trimmed based on estimated propensity score within $(0, 1-t]$, n_t and n_c are the number of treated and control units after trimming.

	$t = 0.05 (n_t = 116, n_c = 56)$			$t = 0.01 (n_t = 177, n_c = 57)$		
	ATT	95% CI	CIL	ATT	95% CI	CIL
Bayes	1.907	[-1.427, 5.256]	6.683	1.990	[-0.853, 4.813]	5.666
DR Bayes	2.024	[-0.958, 4.959]	5.917	2.006	[-0.724, 4.790]	5.514
DR	2.894	[-0.749, 6.538]	7.287	3.664	[-0.325, 7.652]	7.976
OR	3.633	[-0.781, 8.047]	8.828	4.432	[-0.334, 9.197]	9.531
IPW^{HT}	1.783	[-1.661, 5.226]	6.887	1.417	[-1.994, 4.827]	6.821
$\text{IPW}^{\text{Hájek}}$	1.468	[-1.685, 4.620]	6.305	1.119	[-1.771, 4.008]	5.779
TWFE	2.561	[-0.841, 5.964]	6.805	2.691	[-0.604, 5.986]	6.590
DML	2.291	[-4.047, 8.629]	12.677	3.027	[-2.574, 8.628]	11.202

Table 3 presents the ATT estiamtes for Bayesian and frequentist methods. As we see, all methods produce positive but insignificant ATT estimates for the impact of minimum wage,

[7] The data is available on https://davidcard.berkeley.edu/data_sets.

which is in line with the findings of Card and Krueger (1994). For example, nonparametric Bayes and DR Bayes yield ATT point estimates ranging from 1.907 to 2.024 and confidence intervals covering 0 with the length from 5.514 to 6.683. Bayesian methods also provide shorter confidence intervals than most of the frequentist methods including the widely-used TWFE estimator, except that the credible interval produced by semiparametric Bayes is slightly longer than the confidence interval of Hájek–type IPW when $t = 0.05$.

If we do not trim the propensity score, the failure of the overlap condition prevents us from using estimators that involve the inverse propensity score. Among estimators that do not use the propensity score, nonparametric Bayes gives an ATT estimate of 1.935, with a 95% confidence interval of $[-0.460, 4.341]$, for the full sample without any trimming ($t = 0, \bar{n}_t = 249, \bar{n}_c = 58$). OR yields an estimated ATT of 3.351, with a 95% CI of $[-1.233, 7.936]$. TWFE provides an estimated ATT of 2.635, with a 95% CI of $[-0.622, 5.891]$. It turns out that our semiparametric Bayesian method continue to yield stable results when the overlap condition is nearly violated. In sum, our Bayesian methods, which allows a flexible form of the conditional mean function $m_0(x)$ as well as a rich set of covariate, generate comparable ATT estimate with the original findings in Card and Krueger (1994).[8] Therefore, our Bayesian DiD methods confirms the robustness of findings in the classic literature against model specifications.

8 Extensions

We now provide extensions to the canonical DiD panel data setup and show that our Bayesian DiD methods, described in Section 3, can be conveniently extended to cases such as multiple periods with staggered entry and repeated cross sections.

8.1 Extension to Multiple Periods and Staggered Entry

The Bayesian DiD methods described in Section 3 can be conveniently extended to the cases with multiple periods and staggered intervention (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak, Jaravel, and Spiess, 2024). The related literature focuses on the identification of disaggregated causal parameters and some proper aggregation of these parameters. This section extends our

[8]When covariates are not included in the model, Card and Krueger (1994) report the difference-in-difference estimate of 2.76 (standard error 1.36), and the regression adjusted model with controls for chain and ownership dummies yields an estimate of 2.30 (standard error 1.20).

Bayesian method for inference on the disaggregated ATT, specifically the group-time ATT proposed by Callaway and Sant'Anna (2021).

Suppose the available panel data consists of T periods indexed by $t = 1, \ldots, T$ and the earliest treatment intervention occurs at period S. We assume that the treatment intervention remains once a unit get treated. As a result, the entire path of treatment assignment for each unit can be summarized by his/her first treated period (cohort), denoted by the cohort variable $G_i \in \{S, \ldots, T, \infty\}$, where $G_i = \infty$ means the unit i never gets treated. Let the cohort indicators D_{ig} denote whether unit i first receives treatment in period $g \in \{S, \ldots, T, \infty\}$, where $D_{i\infty} = 1$ indicates that unit i never receives treatment.

We assume that never-treated units exist. The potential outcomes depend on cohorts and thus are denoted as $Y_{it}(g)$ for $g \in \{S, \ldots, T\}$ and $Y_{it}(0)$ for $G_i = \infty$. Obviously, $\sum_{g=S}^{T} D_{ig} + D_{i\infty} = 1$. The realized outcome for unit i at time t is $Y_{it} = Y_{it}(0) + \sum_{g=S}^{T} D_{ig} \left(Y_{it}(g) - Y_{it}(0)\right)$.

We focus on the analysis of treatment effect heterogeneity by allowing the ATT to vary with the cohort g ($g \neq \infty$) and the time period $t \geqslant g$:

$$\tau_0^{g,t} = \mathbb{E}_0[Y_t(g) - Y_t(0) | D_g = 1], \text{ for } g = S, \ldots, T \text{ and } t = g, \ldots, T.$$

Suppose a vector of pre-treatment covariates X_i is also available, a vector of dimension p, with the distribution F_0 and the density f_0. The researcher observes independent and identically distributed observations of $(Y_{i1}, \ldots, Y_{iT}, D_{iS}, \ldots, D_{iT}, D_{i\infty}, X_i)$, $i = 1, \ldots, n$.

Applying the identification strategy in Callaway and Sant'Anna (2021), the ATT parameters $\tau_0^{g,t}$ for $g \in \mathcal{G} := \{S, \ldots, T\}$ and $t = g, \ldots, T$ can be identified under Assumption 8 below.[9] For the identification of the ATT, we follow the setup by Callaway and Sant'Anna (2021) and impose the following conditions, which correspond to their Assumptions 3, 4, and 6 (in their $\delta = 0$ case).

Assumption 8. For all x in the support of F_X and $g \in \mathcal{G}$ we have:
(i) $\mathbb{E}_0 \left[Y_t(g) \mid D_g = 1, X = x\right] = \mathbb{E}_0 \left[Y_t(0) \mid D_g = 1, X = x\right]$ for all $t \in \{1, \ldots, g-1\}$,
(ii) $\mathbb{E}_0 \left[Y_t(0) - Y_1(0) \mid D_g = 1, X = x\right] = \mathbb{E}_0 \left[Y_t(0) - Y_1(0) \mid D_\infty = 1, X = x\right]$ for all $t \in \{g, \ldots, T\}$,
(iii) $P_0 \left(D_g = 1\right) > \varepsilon$ and $P_0 \left(D_g = 1 \mid D_g + D_\infty = 1, X = x\right) \leqslant 1 - \varepsilon$ for some $\varepsilon > 0$.

Assumption 8(i) is a "no anticipation" assumption, Assumption 8(ii) is a conditional parallel trend assumption based on the never-treated cohort, and Assumption 8(iii) is an

[9] Callaway and Sant'Anna (2021) propose two identification strategies, depending on the whether the parallel trend assumption is imposed on the never-treated cohort or "Not-Yet-Treated" cohorts. Here we consider the former version.

overlap restriction. Under Assumption 8, Callaway and Sant'Anna (2021) show that the ATT in the staggered entry case is identified by

$$\tau_0^{g,t} = \mathbb{E}_0\left[\Delta_g Y_t - m_0^{g,t}(X) \mid D_g = 1\right], \text{ for } g \in \mathcal{G} \text{ and } t = g, \ldots, T, \infty, \tag{8.1}$$

where the difference operator Δ_g is defined by $\Delta_g Y_t := Y_t - Y_{g-1}$ and the conditional mean function $m_0^{g,t}(x) := \mathbb{E}_0\left[\Delta_g Y_t \mid D_\infty = 1, X = x\right]$.

The identification result in (8.1) uses the cohort g (i.e., $D_g = 1$) as the treated group and the "never treated" cohort ($D_\infty = 1$) as the control group. Using the transformed cross-sectional data $(\Delta_g Y_{it}, D_{ig}, X_i)$ for $i = 1, \ldots, n$ and following the notation in Section 2.2, we can write ATT for a given pair (g, t) under a family of probability distributions $\{P_\eta : \eta \in \mathcal{H}\}$ as

$$\tau_\eta^{g,t} := \frac{\mathbb{E}_\eta[D_g \Delta_g Y_t - D_g m_\eta^{g,t}(X)]}{\mathbb{E}_\eta[D_g]}, \tag{8.2}$$

where $\mathbb{E}_\eta$ denotes the expectation with respect to the distribution of $(\Delta_g Y_t, D_g, X)$. The Bayesian DiD procedures in Section 3 can be applied in the staggered DiD case to obtain the posterior draws $\left\{(\tau_\eta^{g,t})^s : s = 1, \ldots, B\right\}$. Specifically, this can be achieved by replacing ΔY_i, D_i, $m_\eta(\cdot)$, π_η, and $\pi_\eta(\cdot)$ in Algorithm 1 or 2 by ΔY_{it}, D_{ig}, $m_\eta^{g,t}(\cdot)$, $\pi_\eta^g := \mathbb{E}_\eta[D_g]$ and $\pi_\eta^g(\cdot) := P_\eta\left(D_g = 1 \mid D_g + D_\infty = 1, X = \cdot\right)$, respectively, as defined in this section.

The first resulting Bayesian estimator is denoted by $\tau_\eta^{g,t}$, while the second, double-robust Bayesian method is denoted by $\check{\tau}_\eta^{g,t}$ for a cohort $g \in \mathcal{G}$. The next result is an immediate implication of Theorem 4.1 and Theorem 5.2, and its proof is thus omitted.

Corollary 8.1. *Let Assumption 8 hold, and suppose that for any $g \in \mathcal{G}$:*

(i) Assumptions 2–4 hold under the g–specific components, i.e., $(\Delta Y_i, D_i, m_\eta(\cdot), \pi_\eta, \pi_\eta(\cdot))$ are replaced by $(\Delta Y_{it}, D_{ig}, m_\eta^{g,t}(\cdot), \pi_\eta^g, \pi_\eta^g(\cdot))$. Then, the Bayesian method $\tau_\eta^{g,t}$ satisfies the BvM result in Theorem 4.1.

(ii) If Assumptions 3(i), 5, 6, and 7 hold under the g–specific components. Then, the double robust Bayesian method $\check{\tau}_\eta^{g,t}$ satisfies the BvM result in Theorem 5.2.

Corollary 8.1 pertains to inference on cohort-specific ATTs and establishes BvM results for our two Bayesian methods, employing either standard Gaussian process priors or prior/posterior adjustments via the cohort-specific propensity score. We note that extending this framework to aggregate ATTs is highly non-trivial, as it requires a joint modeling of outcome variables across different cohorts and time periods. Hence, distinct prior and likelihood specifications in the Bayesian methodology, as well as prior/posterior

adjustments of the double robust version, are needed. A thorough investigation is therefore left for future research.

8.2 Repeated Cross Sections

Our method also allows for repeated cross-sections following Abadie (2005), as also considered by Sant'Anna and Zhao (2020). In this case, we consider a dummy variable T_i that takes the value two if observation i is only observed in the post-treatment period, and one if observation i is only observed in the pre-treatment period. Define $Y_i = (T_i - 1)Y_{i2} + (2 - T_i)\, Y_{i1}$. The available data is $\{Y_i, D_i, T_i, X_i\}_{i=1}^n$. Let n_2 and n_1 be the sample sizes for the post-treatment and pre-treatment periods, respectively, such that $n = n_2 + n_1$; let $\mathbb{P}(T = 2) \in (0, 1)$. The following assumption restates Assumption 3.3 of Abadie (2005).

Assumption 9. Conditional on $T_i = 1$, (Y_i, D_i, X_i) are i.i.d. from the distribution of (Y_1, D, X); conditional on $T_i = 2$, (Y_i, D_i, X_i) are i.i.d. from the distribution of (Y_2, D, X).

Under Assumptions 1, we can write

$$\begin{aligned}\tau_0 \;=\; & \mathbb{E}_0[Y_2 \mid D = 1] - \mathbb{E}_0[Y_1 \mid D = 1] \\ & - \mathbb{E}_0\left[\mathbb{E}_0[Y_2 \mid D = 0, X = x] - \mathbb{E}_0[Y_1 \mid D = 0, X = x] \mid D = 1\right].\end{aligned}$$

Then using Assumption 9, we can identify ATT as

$$\tau_0 = \mathbb{E}_0[Y \mid D = 1, T = 2] - \mathbb{E}_0[Y \mid D = 1, T = 1] - \mathbb{E}_0\left[m_0(X, 2) - m_0(X, 1) \mid D = 1\right],$$

where $m_0(x, t) \equiv \mathbb{E}_0[Y \mid D = 0, X = x, T = t]$ for $t = 1, 2$.

With an analogous reparametrization as in the panel data case, we obtain

$$\tau_\eta = \frac{\mathbb{E}_\eta[DY\mathbb{1}_{\{T=2\}}]}{\mathbb{E}_\eta[D\mathbb{1}_{\{T=2\}}]} - \frac{\mathbb{E}_\eta[DY\mathbb{1}_{\{T=1\}}]}{\mathbb{E}_\eta[D\mathbb{1}_{\{T=1\}}]} - \frac{\mathbb{E}_\eta\left[D(m_\eta(X, 2) - m_\eta(X, 1))\right]}{\mathbb{E}_\eta\left[D\right]}.$$

Interestingly, the analysis of the last conditional expectation involves a difference of conditional moment function as for the average treatment effect and can be analyzed similarly to Breunig, Liu, and Yu (2025a) in absence of prior corrections. For our double robust method in repeated cross-sections, we emphasize that the efficient influence function takes a different functional form (see Sant'Anna and Zhao (2020)). This translates to a modified prior and posterior adjustments of our double robust Bayesian procedure in

Algorithm 2. While this procedure would be analogous to our double robust method, a full derivation of its asymptotic properties lies beyond the scope of this paper.

9 Conclusion

This paper introduces new semiparametric Bayesian procedures that satisfy the Bernstein-von Mises results in the DiD setup. Our first proposal, based on standard Gaussian process priors, provides a Bayesian analog to the outcome regression in Heckman, Ichimura, and Todd (1997). Through simulations, we show that it performs well in models that are not overly complex and, since no propensity score specification is required, it is not sensitive to the overlap issues. Our second, double robust proposal incorporates prior/posterior corrections based on estimated propensity scores. In simulations it works well for complex models, i.e., when the number of covariates is large. Overall, our Bayesian methods exhibit remarkable finite sample performance, while adapting to the functional form of the conditional mean function. Although our focus is primarily on the DiD panel data case, we also discuss extensions to the repeated cross-section case and staggered interventions.

References

Abadie, A. (2005): "Semiparametric difference-in-differences estimators," *The Review of Economic Studies*, 72(1), 1–19.

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

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting event-study designs: robust and efficient estimation," *The Review of Economic Studies*, p. rdae007.

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

——— (2025b): "Supplement to 'Double Robust Bayesian Inference on Average Treatment Effects'," *Econometrica Supplemental Material*, 93.

Callaway, B., and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of econometrics*, 225(2), 200–230.

Card, D., and A. B. Krueger (1994): "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *The American Economic Review*, 84(4), 772.

Castillo, I. (2012): "A semiparametric Bernstein–von Mises theorem for Gaussian process priors," *Probability Theory and Related Fields*, 152, 53–99.

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

Chamberlain, G., and G. Imbens (2003): "Nonparametric applications of Bayesian inference," *Journal of Business and Economic Statistics*, 21, 12–18.

Chang, N.-C. (2020): "Double/debiased machine learning for difference-in-differences models," *The Econometrics Journal*, 23(2), 177–191.

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

Cheng, G., and J. Huang (2010): "Bootstrap consistency for general semiparametric M-estimate," *The Annals of Statistics*, 38, 2884–2915.

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

Chib, S., and B. Hamilton (2002): "Semiparametric Bayes analysis of longitudinal data treatment models," *Journal of Econometrics*, 110, 67–89.

Chib, S., M. Shin, and A. Simoni (2018): "Bayesian estimation and comparison of moment condition models," *Journal of the American Statistical Association*, 113, 1656–1668.

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

Daniels, M., A. Linero, and J. Roy (2024): *Bayesian nonparametrics for causal inference and missing data*. CRC Press.

De Chaisemartin, C., and X. d'Haultfoeuille (2020): "Two-way fixed effects estimators with heterogeneous treatment effects," *American economic review*, 110(9), 2964–2996.

Florens, J.-P., and A. Simoni (2021): "Gaussian processes and Bayesian moment estimation," *Journal of Business and Economic Statistics*, 39, 482–492.

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

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

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

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

Hájek, J. (1971): "Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu," *Foundations of statistical inference*, 326.

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

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

Horvitz, D. G., and D. J. Thompson (1952): "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, 47(260), 663–685.

Jonathan, L. (2019): "Tutorial: deriving the efficient influence curve for large models," *arxiv preprint*, arXiv:1903.01706v3.

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

Kleijn, B. J. K., and A. W. van der Vaart (2006): "Misspecification in infinite-dimensional Bayesian statistics," *The Annals of Statistics*, 34, 837–877.

Kwon, S., and J. Roth (2024): "(Empirical) Bayes Approaches to Parallel Trends," in *AEA Papers and Proceedings*, vol. 114, pp. 606–609.

Müller, U. K., and A. Norets (2024): "Locally Robust Efficient Bayesian Inference," Discussion paper.

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

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 62, 1349–1382.

Norets, A., and J. Pelenis (2022): "Adaptive Bayesian Estimation of Discrete-Continuous Distributions Under Smoothness and Sparsity," *Econometrica*, 90(3), 1355–1377.

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

Ray, K., and B. Szabó (2019): "Debiased Bayesian inference for average treatment effects," *Advances in Neural Information Processing Systems*, 32.

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

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

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

Sant'Anna, P. H., and J. Zhao (2020): "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, 219(1), 101–122.

Sun, L., and S. Abraham (2021): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of econometrics*, 225(2), 175–199.

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

van der Vaart, A. (1998): *Asymptotic statistics*. Cambridge University Press.

van der Vaart, A., and J. A. Wellner (2023): *Weak convergence and empirical processes, 2nd Edition*. Springer.

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

——— (2009): "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth," *The Annals of Statistics*, 37, 2655–2675.

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

Yiu, A., E. Fong, C. Holmes, and J. Rousseau (2023): "Semiparametric posterior corrections," *arXiv preprint*, arXiv:2306.06059.

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. We also show in the Supplementary Appendix B that γ_η determines the least favorable direction of Bayesian submodels. For simplicity of notation we write $\sum_i$ instead of $\sum_{i=1}^n$ below.

Our theoretical results relies of key decomposition of the frequentist estimator $\widehat{\tau}$ implied by asymptotic efficiency. The minimal asymptotic variance for estimating the ATT can be written in terms of the information norm as

$$P_0(\gamma_0)^2 = P_0\widetilde{\tau}_0^2 = \mathrm{v}_0, \tag{A.1}$$

which is used in the results below. In the following, we denote the log-likelihood based on $Z^{(n)}$ as

$$\ell_n(\eta) = \sum_i \log p_\eta(Z_i) = \ell_n^m(\eta^m) + \ell_n^f(\eta^f),$$

where each term is the logarithm of the factors involving only m or f. Note that we only put a prior distribution on η^m and η^f, and thus the consideration of the likelihood above is sufficient, as shown in the following proofs.

Define the set $\mathcal{H}_n$ that contains m_η with posterior probability going to 1. Recall the definition of the measurable sets $\mathcal{H}_n^m$ of functions η such that $\Pi(\eta \in \mathcal{H}_n^m \mid Z^{(n)}) \to_{P_0} 1$. We introduce the conditional prior $\Pi_n(\cdot) := \Pi(\cdot \cap \mathcal{H}_n^m)/\Pi(\mathcal{H}_n^m)$. Below, we make use of the notation $\upsilon_\eta := \pi_\eta/\pi_0$.

As we show the conditional weak convergence via examining the convergence of the conditional Laplace transform, the following posterior Laplace transform of $\sqrt{n}\upsilon_\eta(\tau_\eta - \widehat{\tau}) - b_{0,\eta}$ given for all $t \in \mathbb{R}$ by

$$I_n(t) = \mathbb{E}^{\Pi_n}\left[e^{t\sqrt{n}[\upsilon_\eta(\tau_\eta - \widehat{\tau}) - b_{0,\eta}]} \mid Z^{(n)}\right], \tag{A.2}$$

plays a crucial role in establishing the BvM theorem (Castillo, 2012; Castillo and Rousseau, 2015; Ray and van der Vaart, 2020). See also Lemma C.1 in the Supplementary Appendix C. Recall the "bias term" given in Theorem 5.1 is

$$\begin{aligned} b_{0,\eta} :&= \frac{1}{n}\sum_i \gamma_0(D_i, X_i)[m_0(X_i) - m_\eta(X_i)] \\ &= \frac{1}{n}\sum_i \left(\frac{D_i}{\pi_0} - \frac{1 - D_i}{\pi_0}\frac{\pi_0(X_i)}{1 - \pi_0(X_i)}\right)[m_0(X_i) - m_\eta(X_i)]. \end{aligned}$$

The "de-biasing term" of our posterior correction is given by

$$\begin{aligned} \widehat{b}_\eta :&= \frac{1}{n}\sum_i \widehat{\gamma}(D_i, X_i)[\widehat{m}(X_i) - m_\eta(X_i)] \\ &= \frac{1}{n}\sum_i \left(\frac{D_i}{\widehat{\pi}} - \frac{1 - D_i}{\widehat{\pi}}\frac{\widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)}\right)[\widehat{m}(X_i) - m_\eta(X_i)]. \end{aligned}$$

Because the expansions in the proof of both BvM theorems largely coincide, we decide to keep the bias term explicit even in proving the non-double robust version where the bias term is asymptotically negligible.

Proof of Theorem 4.1. We begin with a useful decomposition of $\tau_\eta - \widehat{\tau}$. We may assume $\widehat{\tau} = \tau_0 + \mathbb{P}_n[\widetilde{\tau}_0]$, which satisfies (4.1). Consequently, from the definition of the efficient influence function in (3.5) we infer

$$\widehat{\tau} - \tau_0 = \frac{1}{\pi_0 n}\sum_i \left(\frac{D_i - \pi_0(X_i)}{1 - \pi_0(X_i)}(\Delta Y_i - m_0(X_i)) - D_i\tau_0\right)$$

In addition, the definition of the Bayesian procedure in (3.2) implies

$$\tau_\eta - \tau_0 = \frac{\mathbb{E}_\eta\left[D\left(\Delta Y - m_\eta(X)\right) - D\tau_0\right]}{\mathbb{E}_\eta[D]}.$$

Thus, using the notation $\upsilon_\eta = \pi_\eta/\pi_0$, we obtain the decomposition

$$\begin{aligned}
\upsilon_\eta\sqrt{n}(\tau_\eta - \widehat{\tau}) &= \frac{\pi_\eta}{\pi_0}\sqrt{n}(\tau_\eta - \tau_0 - (\widehat{\tau} - \tau_0)) \\
&= \frac{1}{\pi_0}\sqrt{n}\,\mathbb{E}_\eta\left[D\left(\Delta Y - m_\eta(X)\right) - D\tau_0\right] \\
&\quad - \frac{1}{\pi_0\sqrt{n}}\sum_i\left(\frac{D_i - \pi_0(X_i)}{1-\pi_0(X_i)}\left(\Delta Y_i - m_0(X_i)\right) - D_i\tau_0\right) \\
&\quad + \underbrace{\frac{1}{\sqrt{n}}\sum_i\left(\frac{D_i - \pi_0(X_i)}{1-\pi_0(X_i)}\left(\Delta Y_i - m_0(X_i)\right) - D_i\tau_0\right)(1-\upsilon_\eta)}_{=:R_{n,\eta}}.
\end{aligned}$$

Considering the second summand on the right hand side, we make use of the relation

$$\begin{aligned}
&\frac{1}{\pi_0 n}\sum_i\left(\frac{D_i - \pi_0(X_i)}{1-\pi_0(X_i)}\left(\Delta Y_i - m_0(X_i)\right) - D_i\tau_0\right) \\
&= \frac{1}{\pi_0 n}\sum_i\Big(D_i(\Delta Y_i - m_0(X_i) - \tau_0) - (1-D_i)\frac{\pi_0(X_i)}{1-\pi_0(X_i)}\left(\Delta Y_i - m_0(X_i)\right)\Big)
\end{aligned}$$

and hence arrive at the following decomposition

$$\begin{aligned}
\upsilon_\eta t\sqrt{n}(\tau_\eta - \widehat{\tau}) - tR_{n,\eta} &= \underbrace{\frac{t}{\sqrt{n}}\sum_i\left(\mathbb{E}_\eta\left[\frac{D}{\pi_0}(\Delta Y - m_\eta(X) - \tau_0)\right] - \frac{D_i}{\pi_0}(\Delta Y_i - m_\eta(X_i) + \tau_0)\right)}_{=t\sqrt{n}(P_\eta - \mathbb{P}_n)\left[\frac{d}{\pi_0}(\Delta y - m_\eta(x) - \tau_0)\right]} \\
&\quad + \underbrace{\frac{t}{\sqrt{n}}\sum_i\frac{D_i}{\pi_0}(m_0(X_i) - m_\eta(X_i))}_{\text{ⓐ}} \\
&\quad + \underbrace{\frac{t}{\sqrt{n}}\sum_i\frac{1-D_i}{\pi_0}\frac{\pi_0(X_i)}{1-\pi_0(X_i)}\left(\Delta Y_i - m_0(X_i)\right)}_{\text{ⓑ}}.
\end{aligned}$$

In order to show the conditional (on the observed data) convergence of the posterior distribution in the bounded Lipschitz distance, it is sufficient to show the pointwise convergence of the posterior Laplace transform for every t in a neighborhood of 0, by Theorem 1.13.1 of van der Vaart and Wellner (2023). The Laplace transform given in

(A.2) can be written for all $t \in \mathbb{R}$ as

$$I_n(t) = \int \frac{\int_{\mathcal{H}_n} \left[e^{t\sqrt{n}v_\eta(\tau_\eta - \widehat{\tau}) - tR_{n,\eta} - t\sqrt{n}b_{0,\eta}}\right] e^{\ell_n^m(\eta^m)} \mathrm{d}\Pi(\eta^m)}{\int_{\mathcal{H}_n} e^{\ell_n^m(\eta^{m\prime})} \mathrm{d}\Pi(\eta^{m\prime})} \mathrm{d}\Pi(F_\eta \mid Z^{(n)}_{\text{Treated}}).$$

We consider the perturbation via the least favorable direction for the loglikelihood part that depends on m_η. Specifically, we introduce

$$\eta_t^m := \eta_t\left(\eta^m\right) := \eta^m - \frac{t}{\sqrt{n}}\gamma_{c,0} \quad \text{where} \quad \gamma_{c,0} := -\frac{(1-d)}{\pi_\eta}\frac{\pi_\eta(x)}{1-\pi_\eta(x)},$$

which defines a perturbation of η^m along the control arm of the least favorable direction $\gamma_{c,0}$.

We further evaluate for the Laplace transform for all $t \in \mathbb{R}$:

$$I_n(t) = \int \frac{\int_{\mathcal{H}_n} \left[e^{t\sqrt{n}v_\eta(\tau_\eta - \widehat{\tau}) - tR_{n,\eta} - t\sqrt{n}b_{0,\eta}}\right] e^{\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m)} e^{\ell_n^m(\eta_t^m)} d\Pi(\eta^m)}{\int_{\mathcal{H}_n} e^{\ell_n^m(\eta^{m\prime})} \mathrm{d}\Pi(\eta^{m\prime})} \mathrm{d}\Pi(F_\eta \mid Z^{(n)}_{\text{Treated}}).$$

By Lemma C.6 we further obtain for the likelihood functions uniformly for $\eta \in \mathcal{H}_n$:

$$\begin{aligned}
\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = &\underbrace{- \frac{t}{\sqrt{n}} \sum_{i=1}^n \frac{1-D_i}{\pi_0} \frac{\pi_0(X_i)}{1-\pi_0(X_i)} \left(\Delta Y_i - m_0(X_i)\right)}_{\text{ⓒ}} \\
&\underbrace{+ \frac{t^2}{2} \mathbb{E}_0 \left[\frac{1-D}{\pi_0^2} \frac{\pi_0^2(X)}{(1-\pi_0(X))^2} (\Delta Y - m_0(X))^2\right]}_{\text{ⓓ}} \\
&\underbrace{- \frac{t}{\sqrt{n}} \sum_{i=1}^n \frac{1-D_i}{\pi_0} \frac{\pi_0(X_i)}{1-\pi_0(X_i)} (m_0 X_i) - m_\eta(X_i))}_{\text{ⓔ}} + o_{P_0}(1).
\end{aligned}$$

We immediately see that the term ⓑ cancels with ⓒ, and ⓐ + ⓔ cancels $t\sqrt{n}\, b_{0,\eta}$ in the expression for the Laplace transform $I_n(t)$. In addition, 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

$$\int_{\mathcal{H}_n^m} \exp\left(\frac{t^2}{2}\mathbb{E}_0\left[\frac{1-D}{\pi_0^2}\frac{\pi_0^2(X)}{(1-\pi_0(X))^2}(\Delta Y - m_0(X))^2\right] + \ell_n^m(\eta_t^m)\right)$$
$$\times \underbrace{\int \exp\left(t\sqrt{n}(P_\eta - \mathbb{P}_n)\left[\frac{d}{\pi_0}(\Delta y - m_\eta(x) - \tau_0)\right]\right) \mathrm{d}\Pi(F_\eta \mid Z^{(n)}_{\text{Treated}})}_{\text{(f)}} \mathrm{d}\Pi(\eta^m).$$

Note that the posterior law of F_η conditional on the observed data is equivalent to the Bayesian bootstrap measure. Using the envelope condition imposed in Assumption 3, we may apply Lemma C.3 to the (f) term so that for the conditional Laplace transform we have for all $t \in \mathbb{R}$:

$$\begin{aligned} I_n(t) &= \exp\left(\frac{t^2}{2}\left(Var_0\left[\frac{D}{\pi_0}(\Delta Y - m_0(X) - \tau_0)\right] + \mathbb{E}_0\left[\frac{1-D}{\pi_0^2}\frac{\pi_0^2(X)}{(1-\pi_0(X))^2}(\Delta Y - m_0(X))^2\right]\right)\right) \\ &\quad \times \frac{\int_{\mathcal{H}_n} e^{\ell_n^m(\eta_t^m)}\mathrm{d}\Pi(\eta^m)}{\int_{\mathcal{H}_n} e^{\ell_n^m(\eta^{m\prime})}\mathrm{d}\Pi(\eta^{m\prime})} \times \exp(o_{P_0}(1)) \\ &= \exp\left(\frac{t^2}{2}\mathrm{v}_0\right) + o_{P_0}(1), \end{aligned}$$

where the last line follows from the prior invariance property imposed in Assumption 4 and Lemma C.9.

We apply Lemma C.1 by taking $S_n = \sqrt{n}\upsilon_\eta(\tau_\eta - \widehat{\tau}) - b_{0,\eta} - R_{n,\eta}$ and the limiting law L as the normal distribution $N(0, \mathrm{v}_0)$. Thus, we have shown that the posterior distribution of $\sqrt{n}\upsilon_\eta(\tau_\eta - \widehat{\tau}) - b_{0,\eta} - R_{n,\eta}$ converges to $N(0, \mathrm{v}_0)$ in the bounded Lipschitz norm. Note that the bias term is asymptotically negligible given the stochastic equicontinuity in Assumption 2. We have also shown the negligibility of $R_{n,\eta}$ in Lemma C.8. Hence, we apply Lemma C.2 to show the conditional convergence of $\sqrt{n}(\tau_\eta - \widehat{\tau})$, as υ_η converges to 1, in P_0-probability conditional on the data, which concludes the proof. □

Proof of Theorem 5.1. 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). That is, it suffices to handle the ordinary posterior distribution with $\widehat{\gamma}$ set equal to a deterministic function γ_n. Consequently, for the analysis of the conditional Laplace transform $I_n(t)$, we can follow the proof of Theorem 4.1. Further, the prior stability condition is satisfied by Assumption 7 and the proof of Lemma B.2 from Breunig, Liu, and Yu (2025b).

In sum, we have shown that $\sqrt{n}\upsilon_\eta(\tau_\eta - \widehat{\tau}) - b_{0,\eta} - R_{n,\eta}$ converges to the normal

distribution $N(0, \mathrm{v}_0)$ in bounded Lipschitz norm by Lemma C.1. In Lemma C.8, we prove that $\sup_{\eta\in\mathcal{H}_n} \sqrt{n}(\upsilon_\eta - 1)b_{0,\eta} = o_{P_0}(1)$, which implies the conditional weak convergence of $\sqrt{n}\upsilon_\eta(\tau_\eta - \widehat{\tau} - b_{0,\eta})$ to the same normal distribution (under P_0). Finally, we establish this result for $\sqrt{n}(\tau_\eta - \widehat{\tau} - b_{0,\eta})$ by dropping the scaling factor υ_η, due to Lemma C.2. □

Proof of Theorem 5.2. 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 $b_{0,\eta} = \mathbb{P}_n[\gamma_0(m_0 - m_\eta)]$ and $\widehat{b}_\eta = \mathbb{P}_n[\widehat{\gamma}(\widehat{m} - m_\eta)]$. We make use of the decomposition

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

Consider the first summand on the right hand side of the previous equation. From Assumption 6 we infer

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

using the Cauchy-Schwarz inequality and Assumption 5. Consider the second summand on the right hand side of (A.3). Another application of the Cauchy-Schwarz inequality and Assumption 5 yields

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

Using Lemma C.10 we have $\mathbb{P}_n[\gamma_0(m_0 - \widehat{m})] = o_{P_0}(n^{-1/2})$ which completes the proof. □

For the exponential family, we have the conditional mean as follows:

$$\mathbb{E}_\eta[\Delta Y | D = 0, X = x] = \frac{(A' \circ q^{-1})(\eta^m(x))}{(q' \circ q^{-1})(\eta^m(x))}.$$

Now the operator under consideration is $\Upsilon := A \circ q^{-1}$ and its derivative is given by $\Upsilon' = (A' \circ q^{-1})/(q' \circ q^{-1})$. We can further simplify the expression to

$$\mathbb{E}_\eta[\Delta Y | D = 0, X = x] = \Upsilon'(\eta^m(x)),$$

which is used in the proofs below. We write L_n as some term which is a polynomial of $(\log n)$, whose exact value may change from line to line.

Proof of Proposition 6.1. Regarding the conditional mean function m_η, we consider the set $\mathcal{H}_n^m := \{w : w \in \mathcal{B}_n^m, \|\Upsilon'(w(\cdot)) - m_0(\cdot)\|_{2,F_0} \leqslant \varepsilon_n\}$, where $\mathcal{B}_n^m$ is the set defined in (C.4), that contains the Gaussian process with its posterior probability going to one. The posterior rate of contraction follows from the proof of Proposition 4.1 in Breunig, Liu, and Yu (2025a) without restricting the additional λ used in the prior adjustment. The Donsker property is satisfied, following the calculation on Page 561 in the same proof from Breunig, Liu, and Yu (2025a), if $s_m > p/2$.

We show the prior stability by verifying Conditions (3.18) in Proposition 1 from Ray and van der Vaart (2020). Recall the definition of the ball in $\mathbb{H}^m$ centered at the true Riesz representer γ_0 given by

$$\mathbb{H}^m(r_n) := \left\{h \in \mathbb{H}^m : \|h - \gamma_0\|_\infty \leqslant r_n \text{ and } \|h\|_{\mathbb{H}^m} \leqslant \sqrt{n} r_n\right\}$$

for some rate r_n. We need to verify Assumption 4, that is, there exists $\overline{\gamma}_n \in \mathbb{H}^m(\zeta_n)$ for a sequence $\zeta_n = o(1)$ with $\sqrt{n}\varepsilon_n\zeta_n = o(1)$ where $\varepsilon_n = n^{-s_m/(2s_m+d)}L_n$ throughout the analysis. We need to consider two cases separately.

(I) If $s_\pi \geqslant s_m$ (meaning the Riesz representer is more regular than the conditional mean, hence it also belongs to $\mathcal{C}^{s_m}([0,1]^p)$ itself), we can simply take $\overline{\gamma}_n = \gamma_0$ and $\zeta_n = n^{-1/2}\|\gamma_0\|_{\mathbb{H}^m}$. Because the Donsker property already enforces $s_m > p/2$, it is easy to see that the condition $\sqrt{n}\varepsilon_n\zeta_n \to 0$ is indeed satisfied.

(II) If $s_\pi < s_m$ (meaning the Riesz representer is less regular than the conditional mean), we apply Lemma C.4 with $\zeta_n = a_n^{-s_\pi} = n^{-s_\pi/(2s_m+p)}L_n$, so that $\|\overline{\gamma}_n\|_{\mathbb{H}^m} \leqslant Ca_n^p$. It is straightforward to check that $\|\overline{\gamma}_n\|_{\mathbb{H}^m} \leqslant \sqrt{n}\zeta_n$ automatically holds if $s_\pi < s_m$. Finally, $\sqrt{n}\varepsilon_n\zeta_n \to 0$ holds if and only if $n^{1/2-(s_\pi+s_m)/(2s_m+p)}L_n \to 0$. The aforementioned condition holds if $s_\pi > p/2$. □

Proof of Proposition 6.2. The posterior contraction follows from the proof of Proposition 4.1 in Breunig, Liu, and Yu (2025a), if one restricts to the control group only. 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 function m_η, we consider the set $\mathcal{H}_n^m := \{w + \lambda\gamma_n : (w,\lambda) \in \mathcal{W}_n\}$, where for some constant $C > 0$:

$$\mathcal{W}_n := \left\{(w,\lambda) : w \in \mathcal{B}_n^m, |\lambda| \leqslant C\varsigma_n\sqrt{n}\varepsilon_n\right\} \cap \left\{(w,\lambda) : \|\Upsilon'(w + \lambda\gamma_n) - m_0\|_{2,F_0} \leqslant \varepsilon_n\right\}, \qquad \text{(A.4)}$$

where $\mathcal{B}_n^m$ is the set defined in (C.4), that contains the Gaussian process with its posterior probability going to one.

We first verify Assumption 5 with $\varepsilon_n = (n/\log n)^{-s_m/(2s_m+p)}$. The posterior contraction rate is shown in Lemma C.3 of Breunig, Liu, and Yu (2025b). 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 6. It is sufficient to deal with the resulting empirical process $\mathbb{G}_n$. From Lemma C.5 in Breunig, Liu, and Yu (2025b) we infer

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

where the last equation follows from the proof of Proposition 4.1 in Breunig, Liu, and Yu (2025a). Assumption 7 (prior stability) follows from the proof on pages 561-562 in Breunig, Liu, and Yu (2025a). □

Supplement to "Semiparametric Bayesian Difference-in-Differences"

Christoph Breunig Ruixuan Liu Zhengfei Yu

June 14, 2025

This supplementary appendix contains materials to support our main paper. Appendix B derives the least favorable direction for the ATT. Appendix C collects some auxiliary results used for the derivations of our Bernstein-von Mises Theorems. Appendix D provides details regarding the implementation. Appendix E presents additional simulation evidence.

In this supplement, $C > 0$ 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.

B Least Favorable Direction

Our prior correction through the Riesz representer γ_0 is motivated by the least favorable direction of Bayesian submodels. As we show below, this correction is indeed sufficient for our double-robust BvM theorem. We first provide least favorable calculations of Bayesian submodels, which are closely linked to semiparametric efficiency derivations. Consider the one-dimensional submodel $t \mapsto \eta_t$ defined by the path

$$m_t(x) = q^{-1}(\eta^m + t\mathfrak{m})(x) \quad \text{and} \quad f_t(z) = f(z)e^{t\mathfrak{f}(z)} \left(\int e^{t\mathfrak{f}(z)} f(z) \mathrm{d}z \right)^{-1}, \tag{B.1}$$

for the given direction $(\mathfrak{m}, \mathfrak{f})$ with $\int f(z)\mathrm{d}z = 0$. The difficulty of estimating the parameter τ_{η_t} for the submodels depends on the direction $(\mathfrak{m}, \mathfrak{f})$. Among them, let $\xi_\eta = (\xi_\eta^m, \xi_\eta^f)$ be the *least favorable direction* that is associated with the most difficult submodel, i.e., 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 $\eta_t := (m_t, f_t)$. Taking derivative of the logarithmic density $\log p_{\eta_t}(z)$ with respect to t and evaluating at $t = 0$ gives a score operator B_η, which

we derive explicitly in the following proof. 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), where $\widetilde{\tau}_\eta$ is the efficient influence function for estimation of the ATT is given in (3.5).

Lemma B.1. *Let Assumption 1 be satisfied, then the least favorable direction for estimating the ATT parameter in* (2.3) *is:*

$$\xi_\eta(y,d,x) = \left(\frac{\gamma_\eta(0,x)}{a}, \frac{d\big(y - m_\eta(x) - \tau_\eta\big)}{\pi_\eta} \right)$$

where the Riesz representer γ_η *is given in* (3.6).

Proof. For the submodel defined in (B.1), the definition of the joint density p_η given (2.2) evaluated at the perturbation η_t for the control arm yields

$$\begin{aligned} \log p_{\eta_t}(y,0,x) = \log c(y) + ay(\eta^m + t\mathfrak{m})(d,x) - A(q^{-1}(\eta^m + t\mathfrak{m}))(d,x) \\ + t\mathfrak{f}(x) - \log \mathbb{E}[e^{t\mathfrak{f}(X)}] + \log f(x). \end{aligned}$$

Taking derivative with respect to t and evaluating at $t = 0$ gives the score operator:

$$B_\eta(\mathfrak{m},\mathfrak{f})(Z) = B_\eta^m \mathfrak{m}(Z) + B_\eta^f \mathfrak{f}(Z), \tag{B.2}$$

where $B_\eta^f \mathfrak{f}(Z) = \mathfrak{f}(Z)$ and

$$\begin{aligned} B_\eta^m \mathfrak{m}(Z) &= (1-D)\left[a\Delta Y - \frac{A'(m_\eta(X))}{q'(m_\eta(X))} \right] \mathfrak{m}(X), \\ &= a(1-D)\left(\Delta Y - m_\eta(X)\right)\mathfrak{m}(X). \end{aligned}$$

In the last equation, we made use of the relation (explicitly given here for continuous outcomes):

$$\begin{aligned} A'(m_\eta(x)) &= q'(m_\eta(x)) \int ayc(y) \exp\left[q(m_\eta(x))ay - A(m_\eta(x))\right] \mathrm{d}y \\ &= q'(m_\eta(x))\, \mathbb{E}_\eta\left[a\Delta Y | D = 0, X = x\right], \end{aligned}$$

which follows from the exponential family assumption. In this case, there is a one-to-one correspondence between the conditional density function and the conditional mean function of the outcome given covariates. The efficient influence function for estimation of the ATT

parameter τ_η in (3.5) is given by $\widetilde{\tau}_\eta(\Delta Y, D, X) = \gamma_\eta(D, X)(\Delta Y - m_\eta(X)) - \frac{D}{\pi_\eta}\tau_\eta$. Now the score operator B_η given in (B.2) applied to

$$\xi_\eta(y, d, x) = \left(\frac{\gamma_\eta(0, x)}{a}, \frac{d\big(y - m_\eta(x) - \tau_\eta\big)}{\pi_\eta} \right)$$

yields $B_\eta \xi_\eta = \widetilde{\tau}_\eta$.

It remains to formally check the pathwise differentiability of the ATT (van der Vaart, 1998), in order to justify that the influence function is indeed of the same form as obtained by Hahn (1998). This involves verifying that

$$\frac{\partial}{\partial t}\tau_{\eta_t}\Big|_{t=0} = \mathbb{E}_\eta \left[\left(\frac{D}{\pi_\eta}(\Delta Y - m_\eta(X) - \tau_\eta) - \frac{(1-D)\pi_\eta(X)}{(1-\pi_\eta(X))\pi_\eta}(\Delta Y - m_\eta(X)) \right) S_\eta(Z) \right],$$

i.e., the pathwise derivative of the parameter of interest can be expressed as the inner product of the influence function $\widetilde{\tau}_\eta$ and the score function $S_\eta := B_\eta^m m_\eta + B_\eta^f f_\eta$. For simplicity of notation, below we write $S_\eta^0 := B_\eta^f f_\eta$ and $S_\eta^1 := B_\eta^m m_\eta$.

We decompose the score function into two parts, S_η^0 and S_η^1, which correspond to the score associated with the conditional density function $f_{\eta,(\Delta Y|D,X)}(y|0,x)$ and the remaining part of the likelihood. Then, we apply the chain rule to

$$\frac{\partial}{\partial t}\tau_{\eta_t}\Big|_{t=0} = \frac{\partial}{\partial t}\left\{ \frac{\mathbb{E}_{\eta_t}[D(\Delta Y - m_{\eta_t}(X))]}{\mathbb{E}_{\eta_t}[D]} \right\}\Big|_{t=0},$$

with some perturbed likelihood p_{η_t}.

Considering the derivatives with respect to the expectation sign on the numerator and denominator, we have

$$\begin{aligned} \frac{1}{\pi_\eta}\mathbb{E}_\eta[D(\Delta Y - m_\eta(X))S_\eta^1(Z)] - \frac{\mathbb{E}_\eta[D(\Delta Y - m_\eta(X))]}{\pi_\eta^2}\mathbb{E}_\eta[DS_\eta^1(Z)] \\ = \mathbb{E}_\eta \left[\frac{D}{\pi_\eta}(\Delta Y - m_\eta(X) - \tau_\eta)S_\eta^1(Z) \right]. \end{aligned} \tag{B.3}$$

Referring to the derivative with respect to the conditional mean of the control group, it suffices to compute

$$\frac{\mathbb{E}_\eta\left[D\frac{\partial}{\partial t}m_{\eta_t}(X)\Big|_{t=0} \right]}{\mathbb{E}_\eta[D]} = \mathbb{E}_\eta \left[\frac{\pi_\eta(X)}{\pi_\eta}\frac{\partial}{\partial t}m_{\eta_t}(X)\Big|_{t=0} \right].$$

Now one can apply the similar calculus in Example 2 of Jonathan (2019) to the conditional mean to obtain

$$\mathbb{E}_\eta\left[\frac{\pi_\eta(X)}{\pi_\eta}\frac{\partial}{\partial t}m_{\eta_t}(X)\Big|_{t=0}\right] = \mathbb{E}_\eta\left[\frac{\pi_\eta(X)}{\pi_\eta}\frac{1-D}{1-\pi_\eta(X)}(\Delta Y - m_\eta(X))S^0_\eta(Z)\right]. \tag{B.4}$$

The desired conclusion follows from combining the identities (B.3) and (B.4). □

C Auxiliary Results

C.1 Useful Lemmas

C.1.1 Results on Conditional Weak Convergence

We first present a useful result from part of Theorem 1.13.1 in van der Vaart and Wellner (2023) concerning conditional weak convergence. To do so, we introduce a sequence of random variables S_n, a subfield $\mathcal{H}_n$ of their associated σ-algebra, and a Borel probability measure L.

Lemma C.1. *The following two statements are equivalent: (i)* $d_{BL}(\mathcal{L}(S_n \mid \mathcal{H}_n), L) \xrightarrow{P_0} 0$*; (ii) for every* t *in some neighborhood of* 0*,*

$$\mathbb{E}[e^{tS_n} \mid \mathcal{H}_n] \xrightarrow{P_0} \int e^{tx} dL(x) < \infty.$$

We now state a conditional Slutsky result, which coincides with Lemma 10 in Yiu, Fong, Holmes, and Rousseau (2023) and is included here for completeness.

Lemma C.2. *Let* $Z^{(n)} = (Z_1, \ldots, Z_n)$ *be i.i.d. variables from a distribution* P_0 *on a Polish sample space* $(\mathcal{Z}, \mathcal{A})$*. Suppose that* $(P_n)_n$ *is a sequence of random probability measures on* $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ *such that* P_n *os* $\sigma(Z^{(n)})$*-measurable for each* n*. Let* (U_n, V_n) *be variables each taking values in* $\mathbb{R}$ *with* $(U_n, V_n)|P_n \sim P_n$ *and denote the marginals by* P_n^U *and* P_n^V *for* U_n *and* V_n *respectively. Suppose that*

$$d_{BL}(P_n^U, P^U) \to^{P_0} 0$$
$$d_{BL}(P_n^V, \delta_{\{c\}}) \to^{P_0} 0,$$

where P^U *is a fixed probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, and* c *is a fixed constant in* $\mathbb{R}$*. Then*

we have

$$d_{BL}(\mathcal{L}(U_n + V_n|P_n), \mathcal{L}(U_n + c|P_n)) \to^{P_0} 0$$
$$d_{BL}(\mathcal{L}(U_n V_n|P_n), \mathcal{L}(cU_n|P_n)) \to^{P_0} 0.$$

We now state the following generalization of Lemma 1 from Ray and van der Vaart (2020), where the function $g(\cdot)$ may depend on random variables beyond the covariates X. A close inspection of their proof shows that the argument remains valid when $g(\cdot)$ is a function of $Z = (Y, D, X^\top)^\top$.

Lemma C.3. *Suppose $\mathcal{G}_n$ is a sequence of separable classes of measurable functions, such that*

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

and there exists an envelope function G_n such that $\mathbb{E}_0[G_n^{2+\delta}] = O(1)$, for some $\delta > 0$. Then for every t in a sufficiently small neighborhood of 0,

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

C.1.2 Results on Gaussian Processes

Consider a mean-zero Gaussian random element W in a separable Banach space $\mathbb{B}$ defined on a probability space $(\Omega, \mathcal{U}, P)$ and $\mathbb{H}^m$ its RKHS. The dual space $\mathbb{B}^*$ of the Banach space $\mathbb{B}$ consists of all continuous and linear maps $b^* : \mathbb{B} \mapsto \mathbb{R}$. Define a map U by $U(Sb^*) = b^*(W)$, $b^* \in \mathbb{B}^*$. By the definition of RKHS the map $S\mathbb{B}^* : \mathbb{H} \mapsto \mathbb{L}_2(\Omega, \mathcal{U}, P)$ is an isometry. Let $U : \mathbb{H} \mapsto \mathbb{L}_2(\Omega, \mathcal{U}, P)$ be its extension to the full RKHS. If W is a mean-zero Gaussian random element in a separable Banach space and h is an element of its RKHS, then by the Cameron-Martin Theorem, the distributions P^{W+h} and P^W of $W + h$ and W on $\mathbb{B}$ are equivalent with Radon-Nikodym density

$$\frac{dP^{W+h}}{dP^W}(W) = \exp\left(Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2 \right), \quad \text{almost surely.} \tag{C.1}$$

Regarding the uncorrected prior, we consider the Gaussian process prior W^m for the conditional mean as Borel-measurable maps in the Banach space $C([0,1]^d)$, equipped with the uniform norm $\|\cdot\|_\infty$. Such a process also determines a reproducing kernel Hilbert space

(RKHS) $(\mathbb{H}^m, \|\cdot\|_{\mathbb{H}^m})$ and a so-called concenctration function η_0^m, defined as, for $\varepsilon > 0$,

$$\phi_{\eta_0^m}(\varepsilon) := \inf_{h \in \mathbb{H}^m : \|h - \eta_0^m\|_\infty < \varepsilon} \|h\|_{\mathbb{H}^m}^2 - \log \Pr(\|W^m\|_\infty < \varepsilon). \tag{C.2}$$

The posterior contraction rate ε_n^m for such a Gaussian process prior is determined by the solution of the equation:

$$\phi_{\eta_0^m}(\varepsilon_n^m) \sim n(\varepsilon_n^m)^2. \tag{C.3}$$

Each Gaussian process comes with an intrinsic Hilbert space determined by its covariance kernel. This space is critical in analyzing the rate of contraction for its induced posterior. Consider a Hilbert space $\mathbb{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and associated norm $\|\cdot\|_{\mathbb{H}}$. $\mathbb{H}$ is an Reproducing Kernel Hilbert Space (RKHS) if there exists a symmetric, positive definite function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, called a kernel, that satisfies two properties: (i) $k(\cdot, \boldsymbol{x}) \in \mathbb{H}$ for all $\boldsymbol{x} \in \mathcal{X}$ and; (ii) $f(\boldsymbol{x}) = \langle f, k(\cdot, \boldsymbol{x}) \rangle_{\mathbb{H}}$ for all $\boldsymbol{x} \in \mathcal{X}$ and $f \in \mathbb{H}$. It is well-known that every kernel defines a RKHS and every RKHS admits a unique reproducing kernel.

Let $\mathbb{H}_1^{a_n}$ be the unit ball of the RKHS for the rescaled squared exponential process and let $\mathbb{B}_1^{s_m,p}$ be the unit ball of the Hölder class $\mathcal{C}^{s_m}([0,1]^p)$ in terms of the supremum norm $\|\cdot\|_\infty$. We introduce a class of functions $\mathcal{B}_n^m$ which is shown to contain the Gaussian process W which sufficiently large probability, and is given by

$$\mathcal{B}_n^m := \varepsilon_n \mathbb{B}_1^{s_m,p} + M_n \mathbb{H}_1^{a_n}, \tag{C.4}$$

where $a_n = n^{1/(2s_m+p)} (\log n)^{-(1+p)/(2s_m+p)}$, $\varepsilon_n = n^{-s_m/(2s_m+p)} \log^{p+1}(n)$, and $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$. For notational simplicity, we suppress the dependence of the rescaled Gaussian process on the rescaling parameter a_n.

Lemma C.4 (Lemma 11.56 in Ghosal and Van der Vaart (2017))**.** *Consider the rescaled squared exponential process with rescaling factor a. For any $s > 0$ and $w \in \mathcal{C}^s([0,1]^p)$, there exist constants C_1 and C_2 (depending only on w) such that*

$$\inf_{h : \|h - w\|_\infty \leqslant C_1 a^{-s}} \|h\|_{\mathbb{H}^a}^2 \leqslant C_2 a^p. \tag{C.5}$$

C.2 Expansions

Recall the definition of the score operator

$$B_\eta^m \mathfrak{m}(Z) = (1 - D)(\Delta Y - m_\eta(X)) \mathfrak{m}(X).$$

The least favorable direction for the conditional mean in the control group is $\gamma_\eta(0,x) = -\frac{1}{\pi_\eta}\frac{\pi_\eta(x)}{1-\pi_\eta(x)}$. To simplify the notation in the following derivation, we also write $\gamma_{c,\eta}(d,x) = (1-d)\gamma_\eta(0,x)$ to signify this relationship to the control group. Given any η^m, the perturbation we consider is as follows:

$$\eta_t^m(x) := \eta^m(x) - t\gamma_{\eta_0}(0,x)/\sqrt{n}.$$

Below we denote the conditional density function $p_{c,\eta}(y,d,x) = f^{1-d}_{(\Delta Y|D,X),\eta}(y,0,x)$. From the proof of Lemma B.1 we observe

$$\mathbb{E}_\eta[\Delta Y \mid D=0, X=x] = \frac{(A'\circ q^{-1})(\eta^m(x))}{(q'\circ q^{-1})(\eta^m(x))}.$$

Now the operator under consideration is $\Upsilon = A\circ q^{-1}$ and its derivative is given by $\Upsilon' = (A'\circ q^{-1})/(q'\circ q^{-1})$. For the exponential family under consideration, the first and second order cumulants (conditional on covariates) are:

$$\mathbb{E}_\eta[\Delta Y \mid D=0, X=x] = \Upsilon'(\eta^m(x)), \quad Var_\eta(\Delta Y \mid D=0, X=x) = \Upsilon^{(2)}(\eta^m(x)). \quad \text{(C.6)}$$

The conditional variance formula also shows the convexity of $\Upsilon(\cdot)$. Related proofs can be found page 19 in Appendix F of Breunig, Liu, and Yu (2025b).

Lemma C.5. *Let Assumptions 1 and 5 hold. Then, we have uniformly for $\eta \in \mathcal{H}_n$:*

$$\log p_{c,\eta^m} - \log p_{c,\eta_t^m} = \frac{t}{\sqrt{n}}[\gamma_{c,0}(m_0 - m_\eta)] + \frac{t^2}{2n}\left[\Upsilon^{(2)}(\eta_0^m)\gamma_{c,0}^2\right] + O_{P_0}(n^{-3/2}).$$

Proof. For this purpose, we use the notation $g(u) = \log p_{c,\eta_{c,u}^m}$ for $u\in[0,1]$. Specifically, in the one-parameter exponential family case, we have

$$\log p_{c,\eta_u^m}(y,d,x) = (1-d)\left[y\eta_u^m(x) - \Upsilon(\eta_u^m(x)) + \log c(y)\right].$$

By the definition of $\Upsilon(\cdot)$, we can obtain the first to third order derivatives of g as

$$\begin{aligned} g'(0) &= \frac{t}{\sqrt{n}}\gamma_{c,0}\rho^{\Upsilon'(\eta^m)} = \frac{t}{\sqrt{n}}\gamma_{c,0}\rho^{m_\eta}, \\ g^{(2)}(0) &= \frac{t^2}{n}\gamma_{c,0}^2\Upsilon^{(2)}(\eta^m), \quad g^{(3)}(\tilde{u}) = \frac{t^3}{n^{3/2}}\gamma_{c,0}^3\Upsilon^{(3)}(\eta_{\tilde{u}}^m), \end{aligned}$$

where $\tilde{u}$ is some intermediate value between 0 and 1. In the above calculation, we have

made use of (C.6). □

Lemma C.6. *Let Assumptions 1 and 5 hold. Then, we have uniformly for $\eta \in \mathcal{H}_n$:*

$$\begin{aligned}\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) =& \frac{t}{\sqrt{n}} \sum_{i=1}^n \frac{1-D_i}{\pi_0} \frac{\pi_0(X_i)}{1-\pi_0(X_i)} (\Delta Y_i - m_0(X_i)) \\ &+ \frac{t^2}{2} \mathbb{E}_0 \left[\frac{1-D}{\pi_0^2} \frac{\pi_0^2(X)}{(1-\pi_0(X))^2} (\Delta Y - m_0(X))^2 \right] \\ &- \frac{t}{\sqrt{n}} \sum_{i=1}^n \frac{1-D_i}{\pi_0} \frac{\pi_0(X_i)}{1-\pi_0(X_i)} (m_0(X_i) - m_\eta(X_i)) + o_{P_0}(1).\end{aligned}$$

Proof. We start with the following decomposition:

$$\begin{aligned}\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) =& t\mathbb{G}_n[\gamma_{c,0}\rho^{m_0}] + \sqrt{n}\mathbb{G}_n[\log p_{c,\eta^m} - \log p_{c,\eta_t^m} - \frac{t}{\sqrt{n}}\gamma_{c,0}\rho^{m_0}] \\ &+ nP_0[\log p_{c,\eta^m} - \log p_{c,\eta_t^m}],\end{aligned}$$

where $\gamma_{c,0}(d,x) = -\frac{1-d}{\pi_0}\frac{\pi_0(x)}{1-\pi_0(x)}$ and $\rho^{m_0}(Z) = \Delta Y - m_0(X)$. Then, we apply the expansion in Lemma C.5 so that

$$\begin{aligned}&\sqrt{n}\mathbb{G}_n[\log p_{c,\eta^m} - \log p_{c,\eta_t^m} - \frac{t}{\sqrt{n}}\gamma_{c,0}\rho^{m_0}] \\ =& t\mathbb{G}_n[\gamma_{c,0}(m_\eta - m_0)] + \frac{t^2}{2}(\mathbb{P}_n - P_0)[\gamma_{c,0}^2 \Upsilon^{(2)}(\eta^m)] + o_{P_0}(n^{-1/2}),\end{aligned}$$

uniformly in $\eta^m \in \mathcal{H}_n^m$. The second term on the right hand side vanishes because of the P_0-Glivenko-Cantelli (GC) property and the permanence GC theorem, i.e., Theorem 2.10.5 in van der Vaart and Wellner (2023). Then, we infer for the stochastic equicontinuity term that

$$\sqrt{n}\mathbb{G}_n[\log p_{c,\eta^m} - \log p_{c,\eta_t^m} - \frac{t}{\sqrt{n}}\gamma_{c,0}\rho^{m_0}] = t\mathbb{G}_n[\gamma_{c,0}(m_\eta - m_0)] + o_{P_0}(1),$$

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

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_{c,0}\rho^{m_0}] + t\mathbb{G}_n[\gamma_{c,0}(m_0 - m_\eta)] + nP_0[\log p_{c,\eta^m} - \log p_{c,\eta_t^m}] + o_{P_0}(1),$$

uniformly in $\eta^m \in \mathcal{H}_n^m$ and we control $nP_0[\log p_{\eta^m} - \log p_{\eta_t^m}]$ in the remainder of the proof. Specifically, we apply Lemma C.7 and obtain uniformly for $\eta \in \mathcal{H}_n$,

$$nP_0[\log p_{c,\eta^m} - \log p_{c,\eta_t^m}] = P_0[\gamma_{c,0}(m_0 - m_\eta)] + t^2 P_0\left[\Upsilon^{(2)}(\eta_0^m)\gamma_{c,0}^2\right] + o_{P_0}(1).$$

Now using that

$$\frac{1}{\sqrt{n}}\sum_i \gamma_{c,0}(D_i,X_i)(m_0(X_i)-m_\eta(X_i)) = \mathbb{G}_n[\gamma_{c,0}(m_0-m_\eta)] - \sqrt{n}P_0[\gamma_{c,0}(m_0-m_\eta)]$$

the result follows. □

Lemma C.7. *Let Assumptions 1 and 5 hold. Then, we have uniformly for $\eta \in \mathcal{H}_n$:*

$$nP_0 \log\left(\frac{p_{c,\eta^m}}{p_{c,\eta_t^m}}\right) = t\sqrt{n}P_0[\gamma_{c,0}(m_0-m_\eta)] + t^2 P_0\left[\Upsilon^{(2)}(\eta_0^m)\gamma_{c,0}^2\right] + o_{P_0}(1).$$

Proof. First, we note that $P_0[\gamma_{c,0}\rho^{m_0}] = 0$ and

$$\begin{aligned} P_0\left(B_{\eta_0}^m \gamma_{c,0}\right)^2 &= \mathbb{E}_0\left[\left(B_{\eta_0}^m\left(-\frac{1-D}{\pi_0}\frac{\pi_0(x)}{1-\pi_0(x)}\right)\right)^2\right] \\ &= \mathbb{E}_0\left[(\Delta Y - m_0(X))^2 \frac{1-D}{\pi_0^2}\frac{\pi_0^2(X)}{(1-\pi_0(X))^2}\right] \\ &= P_0\left[\Upsilon^{(2)}(\eta_0^m)\gamma_{c,0}^2\right] \end{aligned}$$

using $Var_\eta(\Delta Y \mid D=0, X=x) = \Upsilon^{(2)}(\eta^m(x))$ as in (C.6). Recall the function $g(u) = \log p_{c,\eta_{c,u}^m}$ for $u \in [0,1]$ in Lemma C.5. Based on the expansion therein, and the posterior convergence of η^m, it can be expressed as

$$\begin{aligned} nP_0 g^{(2)}(0) =& t^2 \mathbb{E}_0[\gamma_0^2(0,X)\Upsilon^{(2)}(\eta_0^m(0,X))] + o_{P_0}(1) \\ =& t^2 \mathbb{E}_0[\gamma_0^2(0,X)(\Delta Y - m_0(X))] + o_{P_0}(1) = t^2 P_0(B_{c,0}^m \gamma_{c,0})^2 + o_{P_0}(1), \end{aligned}$$

where the score operator $B_0^m = B_{\eta_0}^m$ is given in the proof of Lemma B.1. Consequently, we obtain, uniformly for $\eta \in \mathcal{H}_n$,

$$\begin{aligned} nP_0[\log p_{c,\eta^m} - \log p_{c,\eta_t^m}] &= -n(P_0 g'(0) + P_0 g^{(2)}(0)) + o_{P_0}(1) \\ &= t\sqrt{n}P_0[\gamma_{c,0}(m_0-m_\eta)] + t^2 P_0(B_{c,0}^m\gamma_{c,0})^2 + o_{P_0}(1), \end{aligned}$$

which leads to the desired result. □

The next lemma is about smaller order terms in the proof of our BvM theorems. Note that the posterior law of F_η coincides with the Bayesian bootstrap. Here, the negligibility of those terms refers to the randomness with respect to the Bayesian bootstrap weights (for which we use P_M to highlight this dependence), conditional on the observed data.

We refer readers to Page 2891 in Cheng and Huang (2010) for comprehensive discussion about disentangling the sources of randomness coming from the observed data and the Bayesian bootstrap weights. Formally, we define $\Delta_n = o_{P_M}(1)$ in P_0-probability, if for any small positive ϵ and δ, it holds $P_0\left(P_{M|Z^{(n)}}(|\Delta_n| > \epsilon) > \delta\right) \to 0$ In addition, we define $\Delta_n = O_{P_M}(1)$ in P_0-probability, if for any small positive δ, there exists a large enough C such that $P_0\left(P_{M|Z^{(n)}}(|\Delta_n| > C) > \delta\right) \to 0$. For the next result, recall the definition he remainder term $R_{n,\eta}$ given in the proof of Theorem 4.1 by

$$R_{n,\eta} = \sqrt{n}\mathbb{P}_n\left[\left(\frac{D-\pi_0(X)}{1-\pi_0(X)}\right)(\Delta Y - m_0(X)) - D\tau_0\right](1-\upsilon_\eta),$$

where $\upsilon_\eta = \mathbb{E}_\eta[D]/\pi_0$.

Lemma C.8. *Under Assumption 3(i), it holds*

(i) $\sup_{\eta\in\mathcal{H}_n} R_{n,\eta} = o_{P_M}(1)$ *in* P_0*-probability and*

(ii) $sup_{\eta\in\mathcal{H}_n}\sqrt{n}(\upsilon_\eta - 1)b_{0,\eta} = o_{P_M}(1)$ *in* P_0*-probability.*

Proof. The uniformity of $\eta \in \mathcal{H}_n$ related to the term υ_η is innocuous, as the posterior law of F_η is equivalent to the Bayesian bootstrap measure, which no longer depends on η. That is, we can write

$$\mathbb{E}_\eta[D] = \int \mathrm{d}F_\eta(y,1,x) = \sum_{i=1}^n M_{ni}D_i, \quad \text{with} \quad M_{ni} = e_i/\sum_{i=1}^n e_i, \quad \text{for} \quad e_i \overset{iid}{\sim} \mathrm{Exp}(1),$$

conditional on the observed data $Z^{(n)}$.

Proof of (i). Because that $P_0\left[\left(\frac{D-\pi_0(X)}{1-\pi_0(X)}\right)(\Delta Y - m_0(X)) - D\tau_0\right] = 0$, we can write

$$R_{n,\eta} = \mathbb{G}_n\left[\left(\frac{D-\pi_0(X)}{1-\pi_0(X)}\right)(\Delta Y - m_0(X)) - D\tau_0\right] \times \left(1 - \frac{\mathbb{E}_\eta[D]}{\pi_0}\right).$$

By the moment condition for the envelope function in Assumption 3 (i), The first term is $O_{P_0}(1)$ and the second term is $O_{P_M}(1)$ in P_0-probability. By the relationship in (71) of Cheng and Huang (2010), the remainder term $R_{n,\eta}$ converges to zero in P_Z-probability, conditional on the data.

Proof of (ii). For the second part, conditional on the observed data $Z^{(n)}$, we have

$$\begin{aligned}\sqrt{n}(\upsilon_\eta - 1) &= \frac{\sqrt{n}}{\pi_0}\left(\sum_{i=1}^n M_{ni}D_i - \pi_0\right) \\ &= \frac{1}{\pi_0}\left(\mathbb{G}_n^*[D] + \mathbb{G}_n[D]\right) = O_{P_M}(1) \quad \text{in} \;\; P_0 - \text{probability}, \end{aligned} \tag{C.7}$$

where $\mathbb{G}_n^*$ denotes the Bayesian bootstrap weighted analog of $\mathbb{G}_n$. In addition, the definition of the bias term $b_{0,\eta}$ yields

$$b_{0,\eta} = \frac{1}{n}\sum_{i=1}^n \gamma_0(D_i, X_i)[m_0(X_i) - m_\eta(X_i)] = (\mathbb{P}_n - P_0)[\gamma_0(m_0 - m_\eta)],$$

where the second equation follows from $\mathbb{E}_0[\gamma_0(D, X) \mid X] = 0$. By the P_0-Glivenko-Cantelli property of the conditional mean function imposed in Assumption 3(i), we have $\sup_{\eta\in\mathcal{H}_n}|b_{0,\eta}| = o_{P_Z}(1)$, which, combined with (C.7), concludes the proof. □

C.3 Prior Stability of GP Priors

In this section, we verify the prior stability using standard Gaussian process priors, which is used in the proof of Theorem 4.1. Here we follow the strategy in Section 5.3 of Ray and van der Vaart (2020). We first approximate η_t^m by an element in the RKHS $\mathbb{H}$ and then apply the Cameron-Martin theorem in (C.1); see Proposition I.20 in (Ghosal and Van der Vaart, 2017).

Lemma C.9. *Under the conditions in Assumption 4, we have*

$$\frac{\int_{\mathcal{H}_n} e^{\ell_n^m(\eta_t^m)}\mathrm{d}\Pi(\eta^m)}{\int_{\mathcal{H}_n} e^{\ell_n^m(\eta^{m\prime})}\mathrm{d}\Pi(\eta^{m\prime})} \to^{P_0} 1. \tag{C.8}$$

Proof. Let $\overline{\gamma}_n \in \mathbb{H}^m(\zeta_n)$, as stated in our Assumption 4. Also, we set $\eta_{n,t} = \eta^m - t\overline{\gamma}_n/\sqrt{n}$. By the Cameron-Martin theorem, the distribution $\Pi_{n,t}$ of $\eta_{n,t}$ if η^m is distributed according to the prior Π has the Radon-Nikodym density

$$\frac{d\Pi_{n,t}}{d\Pi}(\eta^m) = \exp\left(tU_n(\eta^m)/\sqrt{n} - t^2\|\overline{\gamma}_n\|_{\mathbb{H}^m}^2/(2n)\right), \tag{C.9}$$

where $U_n(\cdot)$ is a centered Gaussian variable with variance $\|\overline{\gamma}_n\|_{\mathbb{H}^m}^2$.

By the Gaussian tail bound, we have

$$\Pi(\eta^m : |U_n(\eta^m)| > M\sqrt{n}\varepsilon_n\|\overline{\gamma}_n\|_{\mathbb{H}^m}) \leqslant 2\exp(-M^2 n\varepsilon_n^2/2). \tag{C.10}$$

As a result, the posterior measure of the set in the display tends to 0 in probability for large enough M by Lemma 4 of Ray and van der Vaart (2020). Hence the set

$$B_n := \left\{\eta^m : |U_n(\eta^m)| \leqslant M\sqrt{n}\varepsilon_n\|\overline{\gamma}_n\|_{\mathbb{H}^m}\right\} \cap \mathcal{H}_n$$

also satisfies $\Pi(B_n|Z^{(n)}) \to 1$ in probability. Considering (C.9) on the set B_n and using Assumption 4, we have

$$\left|\log \frac{d\Pi_{n,t}}{d\Pi}(\eta^m)\right| \leqslant M|t|\sqrt{n}\varepsilon_n\zeta_n + \frac{t^2\zeta_n^2}{2} \to 0.$$

By applying Lemma 3 in Ray and van der Vaart (2020) with $A_n = B_n$, $\xi_0 = \gamma_0$ and w_n a sufficiently large fixed constant, we have

$$\sup_{\eta^m \in B_n} |\ell_n^m(\eta_{n,t}) - \ell_n^m(\eta_t^m)| = o_{P_0}(1)$$

By the change of variable $\eta^m - t\overline{\gamma}_n/\sqrt{n} \mapsto v$, we have

$$\frac{\int_{B_n} e^{\ell_n^m(\eta_t^m)} d\Pi(\eta^m)}{\int_{B_n} e^{\ell_n^m(\eta^m)} d\Pi(\eta^m)} = \frac{\int_{B_n} e^{\ell_n^m(\eta_{n,t})} d\Pi(\eta^m)}{\int_{B_n} e^{\ell_n^m(\eta^m)} d\Pi(\eta^m)} e^{o_{P_0}(1)} = \frac{\int_{B_{n,t}} e^{\ell_n^m(v)} d\Pi_{n,t}(v)}{\int_{B_n} e^{\ell_n^m(\eta^m)} d\Pi(\eta^m)} e^{o_{P_0}(1)},$$

where $B_{n,t} = B_n - t\overline{\gamma}_n/\sqrt{n}$. We can replace $\Pi_{n,t}$ in the numerator by Π at the cost of anther multiplicative $1 + o_{P_0}(1)$ term. This makes the quotient into the ratio $\Pi(B_{n,t}|Z^{(n)})/\Pi(B_n|Z^{(n)})$. It has already been proved that $\Pi(B_n|Z^{(n)}) = 1 - o_{P_0}(1)$, so it is sufficient to prove the same result for the numerator, i.e., $\Pi(B_{n,t}|Z^{(n)}) = 1 - o_{P_0}(1)$. Note that

$$\begin{aligned} B_{n,t}^c = &\left\{v : v + t\overline{\gamma}_n/\sqrt{n} \notin \mathcal{H}_n^m\right\} \cap \left\{v : \|\Upsilon'(v + t\overline{\gamma}_n/\sqrt{n}) - m_0(v)\|_{2,F_0} > \varepsilon_n\right\} \\ &\cap \left\{v : |U_n(v + t\overline{\gamma}_n/\sqrt{n})| > M\sqrt{n}\varepsilon_n\|\xi_n\|_{\mathbb{H}^m}\right\}. \end{aligned}$$

The posterior probability of the first set tends to zero in probability by assumption.

Considering the second term, we make use of the smoothness of the link function to get

$$\|\Upsilon'(\eta^m + t\overline{\gamma}_n\sqrt{n}) - \Upsilon'(\eta^m)\|_{2,F_0} \lesssim \frac{\|\overline{\gamma}_n\|_{2,F_0}}{\sqrt{n}} \lesssim \frac{1}{\sqrt{n}}.$$

Therefore, the second set is contained in $\{\eta^m : \|\Upsilon'(\eta^m) - m_0\|_{2,F_0} > \varepsilon_n - C/\sqrt{n}\}$, which has posterior porbability $o_{P_0}(1)$.

When it comes to the third term, note that

$$U_n(\eta^m + t\overline{\gamma}_n/\sqrt{n}) \sim \mathbb{N}(-t\|\overline{\gamma}_n\|^2_{\mathbb{H}^m}/\sqrt{n}, \|\overline{\gamma}_n\|^2_{\mathbb{H}^m}),$$

if η^m is distributed according to the GP prior. Because the mean $t\|\overline{\gamma}_n\|^2_{\mathbb{H}^m}/\sqrt{n}$ of this Gaussian variable is negligible relative to its standard deviation, we can utilize the Gaussian tail bound to show $\Pi(|U_n(\eta^m + t\overline{\gamma}_n/\sqrt{n})| > M\sqrt{n}\varepsilon_n\|\overline{\gamma}_n\|_{\mathbb{H}^m})$ is exponentially small. □

Lemma C.10. *Let Assumptions 1 and 5 be satisfied. Then, we have*

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

Proof. The estimator $\widehat{m}$ is based on an auxiliary sample and hence it is sufficient to consider deterministic functions m_n with the same rates of convergence as $\widehat{m}$. We compute

$$\begin{aligned}
&\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \gamma_0(D_i, X_i)(m_n - m_0)(X_i)\Big)^2 \mid X_1, \ldots, X_n\right] \\
&= \frac{1}{n}\sum_{i,i'}(m_n - m_0)(X_i)(m_n - m_0)(X_{i'})\mathbb{E}_0\left[\gamma_0(D_i, X_i)\gamma_0(D_{i'}, X_{i'}) \mid X_i, X_{i'}\right] \\
&= \frac{1}{n}\sum_{i=1}^n (m_n - m_0)^2(D_i, X_i)Var_0(\gamma_0(D_i, X_i)|X_i),
\end{aligned}$$

using that

$$\mathbb{E}_0[\gamma_0(D, X) \mid X] = \frac{\pi_0(X)}{\pi_0} - \frac{1 - \pi_0(X)}{\pi_0}\frac{\pi_0(X)}{1 - \pi_0(X)} = 0$$

Now overlap as imposed in Assumption 1(iii) implies

$$Var_0(\gamma_0(D_i, X_i)|X_i) = \frac{\pi_0(X)}{\pi_0^2} + \frac{\pi_0^2(X)}{(1 - \pi_0(X))\pi_0^2} \lesssim 1.$$

Consequently, we obtain for the unconditional squared expectation that

$$\mathbb{E}_0\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_0(D_i,X_i)(m_n-m_0)(X_i)\right)^2\right]\lesssim \|m_n-m_0\|_{2,F_0}^2=o(1)$$

by Assumption 5, which implies the desired result. □

D Computational Details in Algorithms 1 and 2

Recall that the prior placed on $m(x)$ in Algorithm 1 is a Gaussian process with mean μ used squared exponential (SE) covariance function (Rassmusen and Williams, 2006, p.83) $K(x,x') := \nu^2\exp(-\sum_{l=1}^{p}a_{ln}^2(x_l-x_l')^2/2)$. In implementation, hyperparameters μ, ν^2, $a_{0n},\ldots,a_{pn}$ and σ^2 (the variance of the noise ϵ) are determined by maximizing the marginal likelihood. In Algorithm 2, the adjusted prior placed on $m(x)$ is given by $K_c(x,x') = K(x,x')+\varsigma_n^2\widehat{\gamma}(0,x)\,\widehat{\gamma}(0,x')$, cf. related constructions from Ray and Szabó (2019), Ray and van der Vaart (2020), and Breunig, Liu, and Yu (2025a). 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 $\varsigma_n = \nu\log n_c/(\sqrt{n_c}\,\Gamma_n)$ in Algorithm 2 satisfies the rate condition in Assumption 7 with probability approaching one. It is similar to that suggested by Ray and Szabó (2019, page 6), which is proportional to $1/(\sqrt{n}\,\Gamma_n)$. The factor Γ_n normalizes the second term (the adjustment term) of K_c to have the same scale as the unadjusted covariance K.

We describe how Step (a) of Posterior Computation in Algorithm 2 is conducted. The corresponding step in Algorithm 1 immediately follows by replacing the adjusted kernel function K_c by the original kernel function K. Let $\boldsymbol{y}_0$ be the vector of $\{\Delta Y_i : D_i=0\}$, $\boldsymbol{X}_0\in\mathbb{R}^{n_c\times p}$ be the matrix of data $\{X_i : D_i=0\}$ and $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ be the matrix of data $\{X_i : i=1,\cdots,n\}$. Let $\boldsymbol{m}_{n_c}$ and $\boldsymbol{m}_n$ be the n_c-vector and n-vector of the function $m(x)$ evaluated at $\boldsymbol{X}_0$ and $\boldsymbol{X}$ respectively:

$$\boldsymbol{m}_{n_c}=[m(X_i),\cdots,D_i=0]^\top,\text{and}\quad \boldsymbol{m}_n=[m(X_1),\ldots,m(X_n)]^\top.$$

For matrices $\boldsymbol{X}$ and $\boldsymbol{X}_0$, we define $K_c(\boldsymbol{X},\boldsymbol{X}_0)$ as a $n\times n_c$ matrix whose (i,j)-th element is $K_c(X_i,X_{0j})$, where X_i is the i-th row of $\boldsymbol{X}$ and X_{0j} is the j-th row of $\boldsymbol{X}_0$. Analogously, $K_c(\boldsymbol{X}_0,\boldsymbol{X}_0)$ is an $n_c\times n_c$ matrix with the (i,j)-th element being $K_c(X_{0i},X_{0j})$, and $K_c(\boldsymbol{X},\boldsymbol{X})$ is a $n\times n$ matrix with the (i,j)-th element being $K_c(X_i,X_j)$.

Given the GP prior with mean μ and covariance kernel K_c, the posterior of $\boldsymbol{m}_n$ has

a Gaussian distribution with the mean $\bar{\boldsymbol{m}}_n$ and covariance $\mathrm{V}(\boldsymbol{m}_n)$ specified as follows (Rassmusen and Williams, 2006, p.16):

$$\begin{aligned}\bar{\boldsymbol{m}}_n &= \mu \boldsymbol{I}_n + K_c(\boldsymbol{X}, \boldsymbol{X}_0)\left[K_c(\boldsymbol{X}_0, \boldsymbol{X}_0) + \sigma^2 \boldsymbol{I}_{n_c}\right]^{-1}(y_0 - \mu \boldsymbol{I}_{n_c}),\\ \mathrm{V}(\boldsymbol{m}_n) &= K_c(\boldsymbol{X}, \boldsymbol{X}) - K_c(\boldsymbol{X}, \boldsymbol{X}_0)\left[K_c(\boldsymbol{X}_0, \boldsymbol{X}_0) + \sigma^2 \boldsymbol{I}_{n_c}\right]^{-1} K_c^{\top}(\boldsymbol{X}, \boldsymbol{X}_0).\end{aligned}$$

We use the Matlab toolbox `GPML` for implementation.[10]

For the implementation of the pilot estimator $\widehat{\gamma}$ given in (3.7), we recommend a Logit regression for estimating the propensity score $\pi(x)$. As a pilot estimator $\widehat{m}$ in Algorithm 2 for posterior adjustment, we use the uncorrected posterior mean $\sum_{s=1}^{B} m_\eta^s/B$, where m_η^s follows Algorithm 1 (Posterior Computation, Step (a)). When the rescaling parameter a_n is as stated in Proposition 6.2, 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).

E Additional Simulation Results

This section provides additional simulation results. Section E.1 presents finite sample results of DR Bayes under varying values of ς_n and when employing sample splitting. Herein, we also examine the case when the propensity score has a non-negligible probability of taking extreme values (close to one), leading to a near violation of the overlap condition. Section E.2 presents finite-sample results for cases where the conditional distribution of ΔY given $(D, X^{\top})$ does not belong to the natural exponential family.

E.1 Sensitivity with respect to implementation details

Tables A1 evaluates the sensitivity of finite sample performance of DR Bayes with respect to the variance ς_n that determines influence strength of the prior correction term. We set $\varsigma_n = c_\varsigma \times \nu(\log n_c)/(\sqrt{n_c}\,\Gamma_n)$ with $c_\varsigma \in \{1/5, 1/2, 1, 2, 5\}$. Note that $c_\varsigma = 1$ corresponds to the simulation results of DR Bayes in Table 1. Table A1 shows that the performance of DR Bayes is not sensitive to the choice of c_ς.

[10]The `GPML` toolbox can be downloaded from http://gaussianprocess.org/gpml/code/matlab/doc/.

Table A1: The effect of ς_n on DR Bayes, normal errors.

Design	c_ς	n	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
			$p=5$			$p=10$			$p=20$		
I	1/5	500	0.012	0.958	0.626	0.014	0.934	0.655	0.030	0.923	0.688
		1000	0.005	0.948	0.437	-0.001	0.950	0.447	0.007	0.949	0.463
	1/2	500	0.010	0.960	0.628	0.013	0.934	0.657	0.029	0.925	0.690
		1000	0.005	0.949	0.438	-0.001	0.950	0.448	0.006	0.948	0.463
	1	500	0.010	0.960	0.628	0.012	0.934	0.657	0.028	0.926	0.691
		1000	0.004	0.948	0.438	-0.001	0.950	0.448	0.006	0.948	0.464
	2	500	0.009	0.960	0.628	0.012	0.934	0.657	0.028	0.926	0.691
		1000	0.004	0.948	0.438	-0.001	0.950	0.448	0.006	0.948	0.464
	5	500	0.009	0.960	0.628	0.012	0.934	0.657	0.028	0.927	0.691
		1000	0.004	0.948	0.438	-0.001	0.950	0.448	0.006	0.948	0.464
II	1/5	500	0.045	0.930	0.631	0.053	0.901	0.659	0.070	0.902	0.691
		1000	0.021	0.937	0.439	0.021	0.934	0.450	0.031	0.922	0.464
	1/2	500	0.036	0.929	0.635	0.045	0.903	0.664	0.064	0.908	0.696
		1000	0.016	0.942	0.441	0.015	0.936	0.452	0.026	0.929	0.466
	1	500	0.034	0.933	0.637	0.043	0.904	0.665	0.063	0.908	0.697
		1000	0.015	0.942	0.442	0.014	0.937	0.453	0.026	0.931	0.467
	2	500	0.033	0.934	0.637	0.043	0.905	0.665	0.063	0.909	0.697
		1000	0.015	0.945	0.442	0.014	0.937	0.453	0.025	0.931	0.467
	5	500	0.033	0.934	0.637	0.043	0.905	0.665	0.063	0.909	0.697
		1000	0.014	0.945	0.442	0.014	0.937	0.453	0.025	0.931	0.467

Table A2 reports the finite sample performance of DR Bayes using sample-split and compare it to and compares it to the results in Tables 1 and 2 that use the full sample twice. Sample-split uses one half of the sample to estimate $\widehat{\pi}$ and $\widehat{m}$, and then draw the posterior of the conditional mean m using the other half of the sample. The effective sample size n_e corresponds to the subsample used for drawing posteriors. As Table A2 shows, DR Bayes using sample-split yields similar coverage probabilities as its counterpart in Table 1 that uses the full sample twice.

Table A2: DR Bayes using sample-split, normal errors, n_e= sample size used for drawing the posterior (half of the full sample for the sample-split approach).

Design	n_e	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
Sample-split		$p = 5$			$p = 10$			$p = 20$		
I	500	-0.004	0.947	0.652	-0.009	0.944	0.693	-0.026	0.921	0.786
	1000	0.002	0.939	0.444	-0.008	0.954	0.463	-0.002	0.945	0.492
II	500	-0.013	0.910	0.660	-0.021	0.904	0.701	-0.051	0.889	0.789
	1000	0.003	0.927	0.448	-0.011	0.937	0.467	-0.011	0.920	0.494
III	500	0.013	0.937	0.570	0.012	0.927	0.605	0.011	0.876	0.684
	1000	0.010	0.937	0.393	-0.000	0.958	0.409	0.011	0.923	0.441
IV	500	0.058	0.907	0.578	0.065	0.886	0.614	0.062	0.828	0.686
	1000	0.046	0.906	0.397	0.041	0.927	0.413	0.054	0.896	0.444
Full sample		$p = 5$			$p = 10$			$p = 20$		
I	500	0.010	0.960	0.628	0.012	0.934	0.657	0.028	0.926	0.691
	1000	0.004	0.948	0.438	-0.001	0.950	0.448	0.006	0.948	0.464
II	500	0.034	0.933	0.637	0.043	0.904	0.665	0.063	0.908	0.697
	1000	0.015	0.942	0.442	0.014	0.937	0.453	0.026	0.931	0.467
III	500	0.019	0.925	0.559	0.020	0.924	0.577	0.047	0.871	0.610
	1000	0.010	0.938	0.390	0.007	0.926	0.398	0.010	0.921	0.419
IV	500	0.075	0.901	0.567	0.082	0.862	0.585	0.135	0.796	0.614
	1000	0.045	0.891	0.393	0.049	0.882	0.402	0.064	0.859	0.423

In practice, data sometimes yield estimated propensity scores with extreme values close to 1, bringing the overlap condition close to being violated. A common remedy is to trim the sample based on the estimated propensity scores, discarding observations where the scores exceed a certain threshold (Crump, Hotz, Imbens, and Mitnik, 2009). We evaluate the performance of our Bayesian methods in such scenarios.

To generate the data with extreme propensity scores, we generate simulated data following Designs I and II in Section 7.1, but with a larger function g, defined as $g(x) = \sum_{j=1}^{p} x_j/j$. As a result, $P(\pi(X) > 0.95) \approx 0.1$ and $P(\pi(X) > 0.99) \approx 0.01$. We discard the units whose estimated propensity score exceeds $1 - t$, where $t = 0.05$ and 0.01. Table A3 shows that the relative performance among various methods remains largely the same as in Table 1. In particular, DR Bayes delivers a more stable coverage performance than nonparametric Bayes, DR, IPW and DML , especially under Design II and/or small trimming ($t = 0.01$).

Table A3: Simulation results with extreme propensity scores, trimming based on the estimated propensity score within $(0, 1-t]$, $n = 1000$, $\bar{n}$ = sample size after trimming.

Design		Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
I	$t = 0.05$	$p = 5$ ($\bar{n} = 903$)			$p = 10$ ($\bar{n} = 902$)			$p = 20$ ($\bar{n} = 885$)		
	Bayes	0.033	0.940	0.536	0.033	0.943	0.531	0.059	0.921	0.539
	DR Bayes	0.003	0.952	0.577	-0.001	0.959	0.587	0.004	0.941	0.598
	DR	-0.007	0.937	0.673	-0.007	0.934	0.601	-0.007	0.900	0.545
	OR	0.004	0.944	0.498	-0.005	0.952	0.494	0.000	0.950	0.501
	IPW^{HT}	0.001	0.939	1.669	-0.016	0.937	1.727	-0.016	0.951	1.789
	$\text{IPW}^{\text{Hájek}}$	0.009	0.933	0.841	-0.002	0.938	0.869	-0.001	0.934	0.877
	TWFE	2.925	0.000	0.736	3.167	0.000	0.742	3.223	0.000	0.750
	DML	-0.014	0.961	0.776	-0.008	0.980	0.970	0.065	0.970	1.193
	$t = 0.01$	$p = 5$ ($\bar{n} = 989$)			$p = 10$ ($\bar{n} = 986$)			$p = 20$ ($\bar{n} = 981$)		
	Bayes	0.052	0.926	0.602	0.050	0.935	0.593	0.081	0.914	0.603
	DR Bayes	0.008	0.919	0.705	0.004	0.942	0.727	0.007	0.938	0.751
	DR	-0.018	0.901	0.733	-0.015	0.890	0.647	-0.021	0.840	0.563
	OR	0.005	0.934	0.527	-0.006	0.949	0.523	-0.001	0.947	0.531
	IPW^{HT}	0.004	0.883	3.144	-0.021	0.866	3.320	-0.018	0.869	3.429
	$\text{IPW}^{\text{Hájek}}$	0.030	0.875	1.429	0.020	0.860	1.493	0.026	0.854	1.495
	TWFE	3.511	0.000	0.752	3.787	0.000	0.756	3.906	0.000	0.763
	DML	-0.016	0.920	1.142	0.013	0.914	1.472	0.139	0.882	1.802
II	$t = 0.05$	$p = 5$			$p = 10$			$p = 20$		
	Bayes	0.093	0.886	0.553	0.102	0.884	0.551	0.122	0.857	0.557
	DR Bayes	0.013	0.938	0.580	0.013	0.939	0.590	0.024	0.920	0.598
	DR	0.003	0.932	0.677	0.006	0.925	0.620	0.020	0.895	0.573
	OR	0.359	0.581	0.801	0.359	0.290	0.568	0.349	0.338	0.575
	IPW^{HT}	0.004	0.942	1.915	-0.019	0.942	2.060	-0.012	0.951	2.212
	$\text{IPW}^{\text{Hájek}}$	0.012	0.930	0.946	-0.003	0.938	0.966	0.007	0.924	0.962
	TWFE	2.453	0.000	0.777	2.633	0.000	0.787	2.665	0.000	0.795
	DML	-0.013	0.957	0.778	-0.010	0.968	0.943	0.053	0.962	1.168
	$t = 0.01$	$p = 5$			$p = 10$			$p = 20$		
	Bayes	0.154	0.825	0.637	0.168	0.798	0.630	0.193	0.757	0.639
	DR Bayes	0.035	0.895	0.713	0.036	0.913	0.734	0.044	0.914	0.753
	DR	0.016	0.898	0.739	0.033	0.881	0.672	0.050	0.822	0.598
	OR	0.592	0.047	0.641	0.575	0.053	0.635	0.566	0.065	0.641
	IPW^{HT}	0.008	0.875	3.772	-0.023	0.864	4.068	-0.011	0.864	4.307
	$\text{IPW}^{\text{Hájek}}$	0.039	0.865	1.722	0.025	0.848	1.741	0.041	0.829	1.707
	TWFE	3.159	0.000	0.855	3.351	0.000	0.851	3.430	0.000	0.852
	DML	-0.012	0.897	1.254	0.013	0.882	1.554	0.135	0.856	1.882

E.2 Sensitivity with respect to error distributions

This section consider the scenarios when the error terms ϵ_{i1}, $\epsilon_{i2}(0)$ and $\epsilon_{i2}(1)$ in our designs deviate from the standard normal distribution. Table A4 presents the results for the designs where the error terms ϵ_{i1}, $\epsilon_{i2}(0)$, and $\epsilon_{i2}(1)$ take χ^2–distribution with 3 degrees of freedom (normalized to have a mean of zero and unit variance). Table A5 considers the case of heteroskedastic errors where $\epsilon_{i2}(d) \sim N(0, e(x))$ and $e(x) = \sum_{j=1}^{p}(x_j - (-1)^{(j-1)})^2/2p$, for $d \in \{0, 1\}$. In Tables A4 and A5, both Bayesian and frequentist methods demonstrate a performance similar to that observed in Table 1, which considers normal errors. Thus, the results in Tables A4 and A5 suggest that our Bayesian methods, which assume normal errors, can still deliver strong finite-sample performance even when the underlying error distribution deviates from standard normality. This observation aligns with the theoretical findings on misspecification in nonparametric Bayesian inference by Kleijn and van der Vaart (2006).

Table A4: Simulation results for designs with $\chi^2(3)$ errors.

Design		Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
I	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.026	0.940	0.613	0.051	0.927	0.630	0.067	0.908	0.645
	DR Bayes	0.001	0.938	0.633	0.017	0.933	0.652	0.024	0.932	0.684
	DR	-0.009	0.931	0.729	-0.002	0.922	0.665	-0.001	0.897	0.616
	OR	-0.006	0.945	0.584	0.003	0.944	0.594	0.002	0.931	0.605
	IPW^{HT}	-0.034	0.944	1.787	0.023	0.929	1.920	-0.021	0.928	2.256
	$\text{IPW}^{\text{Hájek}}$	-0.012	0.933	1.104	0.026	0.921	1.172	0.003	0.911	1.287
	TWFE	2.301	0.000	1.115	2.575	0.000	1.142	2.674	0.000	1.153
	DML	-0.013	0.967	0.921	0.080	0.947	1.046	0.212	0.895	1.255
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.008	0.946	0.428	0.018	0.940	0.434	0.036	0.939	0.444
	DR Bayes	-0.005	0.948	0.435	-0.001	0.944	0.447	0.008	0.943	0.463
	DR	-0.011	0.948	0.524	-0.006	0.933	0.477	0.002	0.947	0.450
	OR	-0.008	0.940	0.413	-0.003	0.947	0.416	-0.001	0.948	0.423
	IPW^{HT}	0.008	0.930	1.199	-0.001	0.932	1.362	0.005	0.916	1.493
	$\text{IPW}^{\text{Hájek}}$	0.007	0.935	0.768	0.003	0.936	0.861	0.008	0.917	0.912
	TWFE	2.299	0.000	0.790	2.563	0.000	0.810	2.682	0.000	0.819
	DML	-0.001	0.963	0.586	0.036	0.960	0.718	0.144	0.902	0.871
II	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.078	0.908	0.632	0.115	0.863	0.652	0.135	0.854	0.670
	DR Bayes	0.022	0.924	0.642	0.050	0.915	0.661	0.065	0.900	0.691
	DR	0.005	0.937	0.769	0.030	0.913	0.723	0.043	0.876	0.681
	OR	0.248	0.725	0.712	0.278	0.666	0.726	0.270	0.691	0.737
	IPW^{HT}	-0.043	0.943	2.241	0.032	0.924	2.415	-0.028	0.921	2.873
	$\text{IPW}^{\text{Hájek}}$	-0.016	0.921	1.437	0.036	0.913	1.458	0.003	0.890	1.549
	TWFE	2.147	0.000	1.263	2.360	0.000	1.283	2.433	0.000	1.287
	DML	-0.022	0.960	1.065	0.078	0.933	1.166	0.202	0.884	1.392
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.040	0.927	0.440	0.057	0.923	0.448	0.079	0.896	0.459
	DR Bayes	0.009	0.936	0.439	0.013	0.936	0.452	0.028	0.921	0.466
	DR	0.005	0.946	0.557	0.010	0.926	0.526	0.031	0.913	0.501
	OR	0.250	0.508	0.506	0.263	0.491	0.513	0.271	0.462	0.519
	IPW^{HT}	0.017	0.929	1.509	-0.002	0.934	1.737	0.008	0.917	1.926
	$\text{IPW}^{\text{Hájek}}$	0.015	0.925	1.003	0.003	0.932	1.094	0.012	0.904	1.128
	TWFE	2.152	0.000	0.896	2.348	0.000	0.910	2.449	0.000	0.917
	DML	0.003	0.946	0.651	0.032	0.947	0.793	0.138	0.893	0.970

Table A5: Simulation results for designs with heteroskedastic error: $\epsilon_{i2}(d) \sim N(0, h(x))$, where $h(x) = \sum_{j=1}^{p}(x_j - (-1)^{(j-1)})^2/2p$.

Design		Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
I	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.031	0.948	0.536	0.041	0.932	0.548	0.062	0.909	0.567
	DR Bayes	0.011	0.949	0.544	0.012	0.921	0.568	0.024	0.924	0.596
	DR	-0.011	0.936	0.677	-0.005	0.902	0.596	-0.002	0.904	0.547
	OR	0.001	0.957	0.504	0.003	0.940	0.511	0.004	0.938	0.526
	IPW^{HT}	0.026	0.936	1.692	-0.020	0.928	1.964	-0.017	0.922	2.288
	$\text{IPW}^{\text{Hájek}}$	0.026	0.933	1.036	0.003	0.909	1.163	0.008	0.908	1.274
	TWFE	2.306	0.000	1.085	2.575	0.000	1.114	2.688	0.000	1.127
	DML	0.014	0.959	0.835	0.078	0.937	1.021	0.212	0.907	1.243
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.016	0.948	0.373	0.015	0.939	0.378	0.029	0.940	0.387
	DR Bayes	0.004	0.940	0.379	-0.000	0.934	0.388	0.006	0.943	0.400
	DR	-0.002	0.937	0.487	-0.005	0.933	0.432	-0.003	0.925	0.398
	OR	0.003	0.951	0.357	-0.003	0.943	0.359	0.001	0.951	0.366
	IPW^{HT}	-0.002	0.942	1.218	0.002	0.950	1.327	0.004	0.940	1.490
	$\text{IPW}^{\text{Hájek}}$	0.004	0.926	0.761	0.004	0.940	0.823	0.009	0.930	0.895
	TWFE	2.309	0.000	0.770	2.574	0.000	0.789	2.693	0.000	0.801
	DML	0.003	0.961	0.561	0.041	0.959	0.682	0.146	0.924	0.852
II	$n = 500$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.074	0.900	0.555	0.095	0.867	0.569	0.121	0.858	0.591
	DR Bayes	0.034	0.923	0.552	0.041	0.899	0.576	0.060	0.892	0.602
	DR	0.014	0.937	0.720	0.024	0.887	0.661	0.045	0.904	0.619
	OR	0.263	0.635	0.648	0.274	0.629	0.660	0.277	0.636	0.673
	IPW^{HT}	0.039	0.924	2.124	-0.024	0.913	2.486	-0.018	0.919	2.916
	$\text{IPW}^{\text{Hájek}}$	0.039	0.924	1.360	0.005	0.900	1.468	0.014	0.897	1.540
	TWFE	2.165	0.000	1.238	2.359	0.000	1.257	2.448	0.000	1.265
	DML	0.015	0.940	0.964	0.072	0.908	1.152	0.203	0.892	1.380
	$n = 1000$		$p = 5$			$p = 10$			$p = 20$	
	Bayes	0.039	0.922	0.383	0.049	0.906	0.391	0.067	0.896	0.402
	DR Bayes	0.015	0.927	0.383	0.014	0.936	0.392	0.025	0.921	0.404
	DR	0.008	0.934	0.523	0.012	0.930	0.483	0.028	0.915	0.457
	OR	0.256	0.418	0.462	0.265	0.375	0.467	0.274	0.376	0.475
	IPW^{HT}	-0.002	0.937	1.549	0.005	0.934	1.694	0.012	0.932	1.922
	$\text{IPW}^{\text{Hájek}}$	0.006	0.925	1.014	0.007	0.929	1.052	0.018	0.920	1.113
	TWFE	2.161	0.000	0.878	2.357	0.000	0.892	2.457	0.000	0.901
	DML	0.000	0.935	0.633	0.039	0.947	0.751	0.142	0.902	0.948