Double Robust Bayesian Inference on Average Treatment Effects*

Christoph Breunig† Ruixuan Liu‡ Zhengfei Yu§

February 21, 2024

Abstract

We propose a double robust Bayesian inference procedure on the average treatment effect (ATE) under unconfoundedness. Our robust Bayesian approach involves two important modifications: first, we adjust the prior distributions of the conditional mean function; second, we correct the posterior distribution of the resulting ATE. Both adjustments make use of pilot estimators motivated by the semiparametric influence function for ATE estimation. We prove asymptotic equivalence of our Bayesian procedure and efficient frequentist ATE estimators by establishing a new semiparametric Bernstein-von Mises theorem under double robustness; i.e., the lack of smoothness of conditional mean functions can be compensated by high regularity of the propensity score and vice versa. Consequently, the resulting Bayesian credible sets form confidence intervals with asymptotically exact coverage probability. In simulations, our double robust Bayesian procedure leads to significant bias reduction of point estimation over conventional Bayesian methods and more accurate coverage of confidence intervals compared to existing frequentist methods. We illustrate our method in an application to the National Supported Work Demonstration.

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

*We thank Guido Imbens, the editor, and the anonymous reviewers, as well as Xiaohong Chen, Yanqin Fan, Essie Maasoumi, Yichong Zhang, and numerous seminar and conference participants for helpful comments and illuminating discussions. Yu gratefully acknowledges the support of JSPS KAKENHI Grant Number 21K01419.

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

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

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

1 Introduction

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

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

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

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

by considering an explicit posterior correction (in addition to the prior adjustment of Ray and van der Vaart [2020]). Not only the prior adjustment but also the posterior correction are based on functional forms that are closely related to the efficient influence function for the ATE, see Hahn [1998]. We show that the corrected posterior still satisfies the BvM Theorem under double robust smoothness assumptions. Our novel procedure combines the advantages of Bayesian methodology with the robustness features that are the strengths of frequentist procedures. Our credible intervals is Bayesianly justifiable, as the uncertainty quantification is made conditional on the observed data ([Rubin, 1984]) and can be also interpreted as frequentist confidence intervals with asymptotically exact coverage probability. Our procedure is inspired by the double machine learning (DML), as well as the bias-corrected matching approach from Abadie and Imbens [2011], as our robustification of an initial procedure removes some non-negligible bias and remains asymptotically valid under weaker regularity conditions. While the main part of our theoretical analysis focuses on the ATE of binary outcomes, also considered by Ray and van der Vaart [2020], we also outline extensions of our methodology to continuous and multinomial cases, as well as other causal parameters.

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

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

2 Setup and Implementation

This section provides the main setup of the average treatment effect (ATE) and motivates the new Bayesian methodology.

2.1 Setup

We consider a family of probability distributions $\{P_\eta : \eta \in \mathcal{H}\}$ for some parameter space $\mathcal{H}$, where the (possibly infinite dimensional) parameter η characterizes the probability model. Let η_0 be the true value of the parameter and denote $P_0 = P_{\eta_0}$, which corresponds to the frequentist distribution of observed data in the classical framework of causal inference. For individual i, consider a treatment indicator $D_i \in \{0, 1\}$. The observed outcome Y_i is determined by $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ where $(Y_i(1), Y_i(0))$ are the potential outcomes of individual i associated with $D_i = 1$ or 0. This paper focuses on the binary outcome case where both $Y_i(1)$ and $Y_i(0)$ take values of $\{1, 0\}$. The covariates for individual i are denoted by X_i, a vector of dimension p, with the distribution F_0 and the density f_0.[2] Let $\pi_0(x) = P_0(D_i = 1|X_i = x)$ denote the propensity score and $m_0(d, x) = P_0(Y_i = 1|D_i = d, X_i = x)$ the conditional mean. Suppose that the researcher observe an independent and identically distributed (i.i.d.) observations of $Z_i = (Y_i, D_i, X_i^\top)^\top$ for $i = 1, \ldots, n$. The joint density of Z_i is given by p_{π_0, m_0, f_0} where

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

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

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

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

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

by P_η consistent with the notation that describes the underlying statistical experiment in the first paragraph of this section, where

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

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

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

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

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

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

for the Riesz representor γ_η, which is given by

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

We assume throughout the paper that π_η is uniformly bounded away from zero and one. We write $\widetilde{\tau}_0 = \widetilde{\tau}_{\eta_0}$ and $\gamma_0 = \gamma_{\eta_0}$.

2.2 Double Robust Bayesian Point Estimators and Credible Sets

Our doubly robust inference procedure builds on a nonparametric Bayesian prior specification for m_η, which depends on a preliminary estimator for γ_0. We consider pilot estimators $\widehat{\pi}$ of the propensity score π_0 and $\widehat{m}$ of the conditional mean function m_0, which both are based on an auxiliary sample. We consider a plug-in estimator for the Riesz

representor γ_0 given by

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

The use of an auxiliary data for pilot estimators simplifies the technical analysis related to the propensity score adjusted priors; see Ray and van der Vaart [2020]. Also, it provides an effective way to control some negligible higher-order terms, see our Lemma C.2 in the online supplement; cf. related discussion on the sample splitting in the DML type methods on Page C6 of Chernozhukov et al. [2018]. In practice, we use the full data twice and do not split the sample, as we have not observed any over-fitting or loss of coverage thereby.

Our procedure builds on the following three steps that approximates the posterior distribution of τ_η, from which one can readily obtain the Bayesian point estimator and the credible set through Monte Carlo simulation draws.

1. Compute the adjusted prior on m:

$$m_\eta(d,x) = \Psi\left(\eta^m(d,x)\right) \qquad \text{and} \qquad \eta^m(d,x) = W^m(d,x) + \lambda\,\widehat{\gamma}(d,x), \tag{2.6}$$

   where $W^m(d,\cdot)$ is a continuous stochastic process independent of the random variable λ, which follows a normal prior $N(0,\sigma_n^2)$ for some $\sigma_n > 0$. The prior adjustment incorporates an initial estimator of the propensity score, with the variable λ determining the extent of this adjustment through its variance σ_n^2. Next, generate Monte Carlo samples from the posterior of $\eta^m(1,x)$ and $\eta^m(0,x)$; see Section 4 for more details. We denote a generic random function drawn from this posterior by $m_\eta^s(\cdot)$, for $s = 1,\ldots,B$.

2. Generate Bayesian bootstrap weights $M_{n1}^s,\ldots,M_{nn}^s$ where $M_{ni}^s = e_i^s/\sum_{i=1}^n e_i^s$ and e_i^s's are i.i.d. draws from the exponential distribution Exp(1) for $s = 1,\ldots,B$. A generic draw from the corrected posterior distribution for the ATE τ_η admits the following representation:

$$\check{\tau}_\eta^s = \tau_\eta^s - \widehat{b}_\eta^s, \quad s = 1,\ldots,B, \tag{2.7}$$

   where

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

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

3. Our $100 \cdot (1-\alpha)\%$ credible set $\mathcal{C}_n(\alpha)$ for the ATE parameter τ_0 is computed by

$$\mathcal{C}_n(\alpha) = \big\{\tau : q_n(\alpha/2) \leqslant \tau \leqslant q_n(1-\alpha/2)\big\}, \tag{2.9}$$

where $q_n(a)$ denotes the a quantile of $\{\check{\tau}_\eta^s : s = 1, \ldots, B\}$. Additionally, we get the Bayesian point estimator (the posterior mean) by averaging the simulation draws: $\overline{\tau}_\eta = \frac{1}{B}\sum_{s=1}^{B} \check{\tau}_\eta^s$.

In Section 4, we provide further guidance on the implementation of our double robust Bayesian method, using adjusted Gaussian process priors. Additionally, we provide recommendations on the implementation of the tuning parameter σ_n; for more details, refer to Section 5.1. Regarding the pilot estimator for the propensity score, we employ Lasso for logistic regression. As a pilot estimator of the conditional mean function, we use the posterior mean of uncorrected Gaussian process priors. More details are provided in Sections 4 and 5.1.

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

3 Main Theoretical Results

In this section, we derive the Bernstein-von Mises (BvM) theorem which establishes the asymptotic equivalance between our Bayesian procedure and frequentist-type efficient semiparametric estimation of the ATE. We consider asymptotically efficient estimators $\widehat{\tau}$

with the following linear representation:

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

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

3.1 Least Favorable Direction

Our prior correction through the Riesz representor γ_0 is motivated by the least favorable direction of Bayesian submodels. We first provide 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

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

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

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

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

Lemma 3.1. *Consider the submodel* (3.2)*. Under Assumption 1, the least favorable*

direction for estimating the ATE parameter in (2.3) *is:*

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

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

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

3.2 Assumptions for Inference

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

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

where $p_{\pi,m}$ denotes the conditional density of (Y_i, D_i) given X_i, given by (2.1) divided by the marginal density of X_i. We write $\mathcal{L}_\Pi(\sqrt{n}(\tau_\eta - \widehat{\tau})|Z^{(n)})$ for the marginal posterior distribution of $\sqrt{n}(\tau_\eta - \widehat{\tau})$. We focus on the case that η^π has a prior that is independent of the prior for (η^m, F). Because the factorization of the likelihood function (2.1) into (η^m, η^π, F) separately, so the posterior of η^π is also independent of the posterior for (η^m, F). Due to the fact that τ_η does not depend on η^π, it is unnecessary to further discuss a prior or posterior distribution on η^π. Also, see Theorem 6.1B in Little and Rubin [2019].

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

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

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

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

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

Assumption 3. [Complexity] For $\mathcal{G}_n = \{\bar{m}_\eta(\cdot) : \eta \in \mathcal{H}_n\}$ it holds $\sup_{\bar{m}_\eta \in \mathcal{G}_n} |(\mathbb{P}_n - P_0)\bar{m}_\eta| = o_{P_0}(1)$ and

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

Recall the propensity score-dependent prior on m given in (2.6), that is, $m(\cdot) = \Psi\left(W^m(\cdot) + \lambda\widehat{\gamma}(\cdot)\right)$. The restriction about λ is made through its hyperparameter $\sigma_n > 0$.

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

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

Assumption 3 restricts the functional class $\mathcal{G}_n$ to form a P_0-Glivenko-Cantelli class; see Section 2.4 of van der Vaart and Wellner [1996] and imposes a stochastic equicontinuity condition on a product structure involving $\widehat{\gamma}$ and m_η. The stochastic equicontinuity condition in (3.5) further relaxes the corresponding one, namely $\sup_{\eta \in \mathcal{H}_n^m} \mathbb{G}_n [m_\eta - m_0] = o_{P_0}(1)$, from Ray and van der Vaart [2020]. In the next section, we demonstrate that our formulation allows for double robustness under Hölder smoothness classes (see Remark 4.1). Hence, the complexity of the functional class $(m_\eta - m_0)$ can be compensated by sufficient regularity of the corresponding Riesz representor and vice versa. In essence, a condition similar to our Assumption 3 is also used in the frequentist literature; see Section 2 of Benkeser et al. [2017]. Nonetheless, the technical argument differs substantially from the frequentist's study, because we mainly need the condition (3.5) to control changes in the likelihood under perturbations along the estimated and true least favorable directions. This is unique to Bayesian analysis with nonparametric priors.

3.3 A Double Robust Bernstein-von Mises Theorem

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

To the best of our knowledge, this is the first BvM theorem that entails the double robustness. We discuss the distinction with Theorem 2 in Ray and van der Vaart [2020]. Their work laid the theoretical foundation that supports the usefulness of propensity score in Bayesian analysis of the ATE. They showed that propensity score adjustment via priors can allow for weak regularity conditions on the propensity score function, coining the corresponding property as the single robustness. Our analysis differs from Ray and van der

Vaart [2020] in two crucial ways. First, we improve on their Lemma 3 by showing that it is possible to verify the prior stability condition for propensity score-adjusted priors under the product structure in Assumption 3, modulo the "bias term" $b_{0,\eta}$. This separation is essential to identify the source of the restrictive condition, such as the Donsker property on m_η, which is mainly used to eliminate $b_{0,\eta}$. Second, our proposal introduces an explicit debiasing step, borrowing key insights from recent developments in the DML literature.

Remark 3.1 (Connection with frequentist robust estimation)**.** *In our BvM theorem, we do not restrict the centering estimator $\widehat{\tau}$, as long as it admits the linear representation as in* (3.1)*. A popular frequentist estimator for the ATE that achieves double robustness is*

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

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

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

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

4 Illustration with Gaussian Process Priors

We illustrate the general methodology by placing the Gaussian process prior on $\eta^m(d,\cdot)$ in relation to the conditional mean functions for $d \in \{0,1\}$. The Gaussian process regression has been extensively used among the machine learning community [Rassmusen and Williams, 2006], and started to gain popularity among economists [Kasy, 2018]. Our study further strengthened the appealing features of this modern Bayesian toolkit. We provide primitive conditions used in our main results in the previous section. In addition, we provide details on the implementation using Gaussian process priors and discuss the data-driven choices of tuning parameters.

4.1 Inference Based on Gaussian Process Priors

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

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

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

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

Proposition 4.1 (Squared Exponential Process Priors)**.** *The estimator $\widehat{\gamma}$ satisfies $\|\widehat{\gamma}\|_\infty =$*

$O_{P_0}(1)$ *and* $\|\widehat{\gamma} - \gamma_0\|_\infty = O_{P_0}\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$ *for some* $s_\pi > 0$. *Suppose* $m_0(d,\cdot) \in \mathcal{C}^{s_m}([0,1]^p)$ *for* $d \in \{0,1\}$ *and some* $s_m > 0$ *with* $\sqrt{s_\pi\, s_m} > p/2$. *Also,* $\|\widehat{m}(d,\cdot) - m_0(d,\cdot)\|_{2,F_0} = O_{P_0}\big((n/\log n)^{-s_m/(2s_m+p)}\big)$. *Consider the propensity score-dependent prior on* m *given by* $m(d,x) = \Psi\left(W_d^m(x) + \lambda\,\widehat{\gamma}(d,x)\right)$, *where* $W_d^m(x)$ *is the rescaled squared exponential process for* $d \in \{0,1\}$, *with its rescaling parameter* a_n *of the order in* (4.1) *and*

$$\left(\frac{n}{\log n}\right)^{-s_m/(2s_m+p)} \ll \sigma_n \lesssim 1. \tag{4.2}$$

Then, the corrected posterior distribution for the ATE satisfies Theorem 3.1.

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

4.2 Implementation of Gaussian Process Priors

We provide details on the Gaussian process prior placed on $\eta^m(d,x)$ and its posterior computation. Following equation (2.6), the propensity score adjusted prior takes the form $\eta^m(d,x) = W^m(d,x) + \lambda\,\widehat{\gamma}(d,x)$: the first component $W^m(d,x)$ is a zero-mean Gaussian process with the commonly used squared exponential (SE) covariance function [Rassmusen and Williams, 2006, p.83]. That is, $K\left((d,x),(d',x')\right) :=$

$\nu^2 \exp\left(-a_{0n}^2(d-d')^2/2 - \sum_{l=1}^p a_{ln}^2(x_l - x_l')^2/2\right)$ where the hyperparameter ν^2 is the kernel variance and $a_{0n}, \ldots, a_{pn}$ are rescaling parameters that reflect the relevance of treatment and each covariate in predicting η^m. In practice, they can be obtained by maximizing the marginal likelihood.

Conditional on the data used to obtain the propensity score estimator $\widehat{\pi}$, the prior for η^m has zero mean and the covariance kernel K^c including an additional term based on the estimated Riesz representer $\widehat{\gamma}$ is given by $K^c\left((d,x),(d',x')\right) = K\left((d,x),(d',x')\right) + \sigma_n^2 \widehat{\gamma}(d,x)\,\widehat{\gamma}(d,x')$, cf. related constructions from Ray and Szabó [2019] and Ray and van der Vaart [2020]. The parameter σ_n, representing the standard deviation of λ, controls the weight of the prior correction. In the subsequent numerical exercise, we select σ_n such that the rate condition specified in Assumption 4 is satisfied. Our simulation results also suggest that the performance of our approach remains stable across various choices of σ_n.

Utilizing Gaussian process priors with zero mean and covariance function K^c, and incorporating the available data, we generate posterior draws of the vector $\left[\eta^m(d,X_1),\cdots,\eta^m(d,X_n)\right]^\top$ for $d \in \{0,1\}$. This can be achieved through the Laplace approximation method detailed in online Appendix G. When it comes to the pilot estimator $\widehat{m}$ required for our posterior correction in (2.8), we plug in the posterior mean of $m(d,x) = \Psi\left(\eta^m(d,x)\right)$, which is calculated using the unadjusted Gaussian process prior and the sample data. When the rescaling parameter a_n is as stated in Proposition 4.1, the convergence rate of $\widehat{m}$ is $O_{P_0}\big((n/\log n)^{-s_m/(2s_m+p)}\big)$. This can be shown by combining Theorems 11.22, 11.55 and 8.8 from Ghosal and Van der Vaart [2017].

5 Numerical Results

In this section, we apply our method to the well-known job-training data set that contains a treated sample of 185 men from the National Supported Work (NSW) experiment and a control sample of 2490 men from the Panel Study of Income Dynamics (PSID). The data has been used by LaLonde [1986], Dehejia and Wahba [1999], Abadie and Imbens [2011], and Armstrong and Kolesár [2021], among others. We also refer readers to Imbens [2004] and Imbens and Rubin [2015] for comprehensive reviews of the data.

5.1 Simulations

In this section, we consider a simulation study where the observations are randomly drawn from a large sample generated by the Wasserstein Generative Adversarial Networks

(WGAN) method from the the job-training real data, see Athey et al. [2021]. We view their simulated data as the population and repeatedly draw our simulation samples (each with 185 treated and 2490 control observations) for $1,000$ times of Monte Carlo replication. We slightly depart from previous studies by focusing on a binary outcome Y: the employment indicator for the year 1978, which is defined as an indicator for positive earnings. The treatment D is the participation in the NSW program. We are interested in the average treatment effect of the NSW program on the employment status. We consider three choices of covariates X: Spec I follows that of Abadie and Imbens [2011] and contains nine variables: age, education, black, Hispanic, married, earnings in 1974, earnings in 1975, unemployed in 1974, unemployed in 1975; Spec II follows Table 3 of Dehejia and Wahba [2002] that adds six variables to Spec I: the no degree indicator, quadratic terms of age, education, earnings in 1974 and 1975, and unemployed in 1974 $\times$ Hispanic; Spec III further adds the six interactions between the four continuous covariates (age,education, earnings in 1974 and 1975) and eight other interactions that are selected in Farrell [2015]: education $\times$ married, education $\times$ Hispanic, earnings in 1974 $\times$ married, earnings in 1974 $\times$ Hispanic, earnings in 1975 $\times$ unemployed in 1974, nodegree $\times$ unemployed in 1975, black $\times$ unemployed in 1975, unemployed in 1974 $\times$ unemployed in 1975.

Our double robust Bayesian method (DR Bayes) is implemented as given in $\check{\tau}_\eta^s$ in (2.7) using the adjusted Gaussian process prior, where the propensity score is estimated by Lasso for logisitic regression with the penalty parameter chosen by cross-validation [Friedman et al., 2010]. The posterior correction also builds on a pilot conditional mean estimator $\widehat{m}$, given here by the posterior mean of m_η using uncorrected Gaussian process priors. We set the tuning parameter σ_n that corresponds to the standard deviation of the adjusted prior by $\sqrt{\dim(X) n \log n}/\sum_{i=1}^n |\widehat{\gamma}(D_i, X_i)|$, which reflects the rate condition imposed in Assumption 4 (with probability approaching one). Online Appendix H presents additional simulation evidence, showing that the performance of DR Bayes is stable for different choices of σ_n, as long as the latter is not too small. We compare our method to the following two Bayesian procedures: First, we consider prior adjusted Bayesian method (PA Bayes) proposed by Ray and van der Vaart [2020] and implemented following τ_η^s in (2.8) with the same choice of estimated σ_n. Second, we consider an unadjusted Bayesian method (Bayes), following τ_η^s in (2.8) using Gaussian process priors. For further details on the implementation of the Gaussian process priors we refer to Section 4.2. All Bayesian methods are implemented based on $5,000$ posterior draws.

Table 1: Simulation results using WGAN-generated data for specifications I ($\dim(X) = 9$), II ($\dim(X) = 15$), and III ($\dim(X) = 29$). Trimming is based on $\hat{\pi} \in [t, 1-t]$ and $\bar{n}$ = the average sample size after trimming.

Spec	Methods	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
I		$t = 0.10(\bar{n} = 240)$			$t = 0.05(\bar{n} = 364)$			$t = 0.01(\bar{n} = 665)$		
	Bayes	-0.040	0.682	0.147	-0.010	0.845	0.148	-0.005	0.917	0.120
	PA Bayes	-0.002	0.982	0.274	0.037	0.940	0.260	0.051	0.875	0.310
	DR Bayes	-0.021	0.979	0.229	0.016	0.966	0.224	0.026	0.938	0.258
	Match	0.027	0.933	0.334	0.048	0.911	0.323	0.033	0.967	0.323
	Match BC	0.041	0.881	0.347	0.065	0.818	0.334	0.083	0.800	0.339
	DR TMLE	0.014	0.838	0.299	0.040	0.741	0.282	0.038	0.657	0.241
	DML	0.030	0.928	0.454	0.054	0.864	0.398	0.045	0.926	0.490
II		$t = 0.10(\bar{n} = 226)$			$t = 0.05(\bar{n} = 345)$			$t = 0.01(\bar{n} = 603)$		
	Bayes	-0.077	0.000	0.046	-0.078	0.000	0.032	-0.080	0.000	0.014
	PA Bayes	0.007	0.966	0.282	0.035	0.930	0.269	0.032	0.883	0.290
	DR Bayes	-0.013	0.964	0.233	0.012	0.957	0.230	0.011	0.930	0.258
	Match	0.005	0.956	0.319	0.032	0.923	0.301	0.018	0.963	0.285
	Match BC	0.108	0.764	0.388	0.174	0.584	0.454	0.246	0.537	0.635
	DR TMLE	0.016	0.860	0.292	0.035	0.755	0.280	0.033	0.716	0.243
	DML	0.020	0.942	0.424	0.042	0.868	0.364	0.032	0.918	0.410
III		$t = 0.10(\bar{n} = 212)$			$t = 0.05(\bar{n} = 321)$			$t = 0.01(\bar{n} = 613)$		
	Bayes	-0.077	0.015	0.047	-0.079	0.000	0.030	-0.080	0.000	0.011
	PA Bayes	0.005	0.962	0.296	0.029	0.934	0.277	0.035	0.890	0.290
	DR Bayes	-0.016	0.963	0.243	0.007	0.953	0.237	0.019	0.932	0.266
	Match	0.002	0.943	0.323	0.016	0.943	0.306	0.011	0.971	0.299
	Match BC	-0.024	0.937	0.457	-0.014	0.920	0.470	-0.025	0.941	0.532
	DR TMLE	-0.003	0.780	0.295	0.008	0.742	0.292	0.029	0.670	0.239
	DML	-0.001	0.932	0.370	0.020	0.890	0.373	0.027	0.924	0.385

We also compare our method to frequentist estimators. Match/Match BC corresponds to the nearest neighbor matching estimator and its bias-corrected version, which adjusts for differences in covariate values through regression Abadie and Imbens [2011]. DR TMLE corresponds to the doubly robust targeted maximum likelihood estimator by Benkeser et al. [2017]. DML corresponds to the double/debiased machine learning estimator of Chernozhukov et al. [2017], where the nuisance functions π_0 and m_0 are estimated using random forest. Since the job-training data contains a sizable proportion of units with propensity score estimates very close to 0 and 1, we follow Crump et al. [2009] and discard observations with the estimated propensity score outside the range $[t, 1-t]$, with the

trimming threshold $t \in \{0.10, 0.05, 0.01\}$.[4]

Table 1 presents the finite sample (mean) bias of the point estimator, coverage probability (CP) and the average length (CIL) of the 95% credible/confidence interval for the Bayesian and frequentist methods mentioned above. We use the full data twice in computing the prior/posterior adjustments and the posteriors of the conditional mean function. Online Appendix H reports the performance of DR Bayes using sample-split, which has similar coverage but larger credible interval length due to the halved sample.

Concerning the Bayesian methods for estimating the ATE, Table 1 reveals that unadjusted Bayes yields adequate coverage only in the low-dimensional case of Specification I and with a trimming constant $t = 0.01$. In all other cases, the coverage is highly inaccurate, and the bias increases significantly as more covariates are introduced. If the prior is corrected using the propensity score adjustment, then the results improve significantly. Nevertheless, our DR Bayes method demonstrates two further improvements: First, DR Bayes leads to smaller average confidence lengths in each case while simultaneously improving the coverage probability. For trimming thresholds $t \in \{0.05, 0.01\}$, this can be attributed to a reduction in bias, while for $t = 0.10$, DR Bayes also shows improvement in the CIL, which appears to stem from more accurate uncertainty quantification via our posterior correction. Second, when the trimming threshold is small, i.e., $t = 0.01$, propensity score estimators can be less accurate, leading to reduced coverage probabilities of PA Bayes. Our double robust Bayesian method, on the other hand, is still able to provide accurate coverage probabilities across all specifications considered. In other words, DR Bayes exhibits more stable performance than PA Bayes with respect to the trimming threshold. Take Specification I as an example: the empirical coverage probabilities for DR Bayes are 0.979, 0.966, and 0.938 when the trimming threshold t is set to 0.1, 0.05, and 0.01, respectively. In comparison, PA Bayes yields corresponding coverage probabilities of 0.982, 0.940, and 0.875.[5]

Our DR Bayes also exhibits encouraging performances when compared to frequentist methods. It provides a more accurate coverage than bias-corrected matching, DR TMLE and DML. Compared with the matching estimator that exhibits a similarly good coverage

[4]Crump et al. [2009] suggested a simple rule of thumb with a threshold of $t = 0.10$, while Athey et al. [2021] used $t = 0.05$. Applying the optimal trimming rule proposed by Crump et al. [2009] to our simulated samples yields an average optimal trimming threshold ranging between 0.072 and 0.074 across the three specifications.

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

performance, DR Bayes yields considerably shorter credible intervals in each specification considered.

5.2 An Empirical Illustration

We apply the Bayesian and frequentist methods considered above to the real job-training data. We report the estimation for the three different specification considered in the previous subsection and consider a varying choice of the threshold constant $t \in \{0.10, 0.05, 0.01\}$.[6] The results are presented in Table 2.

As a benchmark, the experimental data that uses both treated and control groups in NSW ($n = 445$) yields an ATE estimate (treated-control mean difference) equal to 0.111 with the 95% confidence interval $[0.026, 0.196]$. As we see from Table 2, the unadjusted Bayesian method yields large estimates under Spec I while very small ones under Spec II and Spec III. The adjusted Bayesian methods (PA and DR Bayes), on the other hand, produce estimates comparable to the experimental estimate. Taking $t = 0.05$ for example, PA Bayes finds that the job training program enhanced the employment by 11.2% to 16.8% across different specifications, and DR Bayes estimates the effect from 7.5% to 18.3%.

Consistent with our simulation results, bias-corrected matching and DR TMLE sometimes exhibit undesirable behavior: The bias-corrected matching produce large estimates (up to 39.2%) for Spec II and III. DR TMLE produces negative estimates for $t = 0.10$ when all other estimates are positive. In the case $t = 0.01$, where the overlapping condition is closer to violation for some units, adjusted Bayesian methods yield close-to-zero estimates under Spec II and III, while bias-corrected matching and DML yields tends to overestimate. The matching estimator, which performs best among frequentist methods in our simulations, produces similar estimates as PA and DR Bayes. In terms of estimation precision, the credible intervals based on DR Bayes are the shortest among the adjusted Bayesian and all the frequentist methods considered over all cases except for Spec II with $t = 0.01$. The credible intervals based on unadjusted Bayes are too short under Spec II and III to be expected to have a reasonable coverage.

[6] Applying the optimal trimming rule proposed by Crump et al. [2009] yields an optimal threshold of 0.064 for Spec I and II and 0.057 for Spec III.

Table 2: Estimates of ATE for the job-training data: trimming based on $\hat{\pi} \in [t, 1-t]$, $\bar{n}$ = sample size after trimming.

Spec I	$t = 0.10(\bar{n} = 245)$			$t = 0.05(\bar{n} = 398)$			$t = 0.01(\bar{n} = 740)$		
	ATE	95% CI	CIL	ATE	95% CI	CIL	ATE	95% CI	CIL
Bayes	0.214	[0.125, 0.299]	0.174	0.214	[0.130, 0.293]	0.163	0.197	[0.141, 0.251]	0.111
PA Bayes	0.151	[0.002, 0.283]	0.282	0.168	[0.037, 0.285]	0.248	0.090	[-0.075, 0.227]	0.302
DR Bayes	0.172	[0.051, 0.289]	0.238	0.183	[0.058, 0.299]	0.241	0.119	[-0.027, 0.250]	0.277
Match	0.188	[0.022, 0.355]	0.333	0.140	[-0.029, 0.309]	0.338	0.079	[-0.111, 0.269]	0.380
Match BC	0.157	[-0.006, 0.321]	0.327	0.145	[-0.021, 0.310]	0.331	0.180	[-0.004, 0.365]	0.369
DR TMLE	-0.022	[-0.173, 0.128]	0.301	0.084	[-0.067, 0.235]	0.302	0.037	[-0.202, 0.275]	0.477
DML	0.170	[0.013, 0.327]	0.314	0.126	[-0.054, 0.306]	0.360	0.338	[-0.143, 0.818]	0.962
Spec II	$t = 0.10(\bar{n} = 222)$			$t = 0.05(\bar{n} = 369)$			$t = 0.01(\bar{n} = 645)$		
	ATE	95% CI	CIL	ATE	95% CI	CIL	ATE	95% CI	CIL
Bayes	0.010	[-0.021, 0.043]	0.065	0.027	[-0.009, 0.063]	0.072	-0.005	[-0.024, 0.013]	0.038
PA Bayes	0.049	[-0.096, 0.187]	0.284	0.112	[-0.035, 0.232]	0.267	-0.004	[-0.139, 0.111]	0.249
DR Bayes	0.040	[-0.087, 0.158]	0.245	0.078	[-0.031, 0.174]	0.204	-0.006	[-0.200, 0.151]	0.352
Match	0.158	[-0.004, 0.320]	0.324	0.134	[-0.022, 0.290]	0.313	0.065	[-0.094, 0.223]	0.317
Match BC	0.250	[0.083, 0.417]	0.334	0.392	[0.194, 0.590]	0.396	0.352	[0.146, 0.558]	0.412
DR TMLE	0.029	[-0.111, 0.169]	0.280	0.130	[-0.050, 0.310]	0.360	0.107	[-0.106, 0.320]	0.426
DML	0.138	[-0.044, 0.321]	0.365	0.117	[-0.050, 0.284]	0.334	0.319	[-0.080, 0.718]	0.799
Spec III	$t = 0.10(\bar{n} = 234)$			$t = 0.05(\bar{n} = 390)$			$t = 0.01(\bar{n} = 712)$		
	ATE	95% CI	CIL	ATE	95% CI	CIL	ATE	95% CI	CIL
Bayes	0.006	[-0.019, 0.031]	0.051	0.025	[-0.019, 0.067]	0.086	-0.001	[-0.009, 0.006]	0.015
PA Bayes	0.096	[-0.058, 0.230]	0.288	0.117	[-0.034, 0.247]	0.281	-0.020	[-0.254, 0.122]	0.345
DR Bayes	0.068	[-0.050, 0.167]	0.218	0.075	[-0.040, 0.178]	0.219	-0.010	[-0.149, 0.099]	0.248
Match	0.192	[0.026, 0.358]	0.332	0.156	[-0.012, 0.325]	0.337	0.006	[-0.181, 0.192]	0.373
Match BC	0.173	[-0.005, 0.350]	0.355	0.280	[0.109, 0.451]	0.342	0.335	[0.114, 0.555]	0.441
DR TMLE	-0.038	[-0.217, 0.140]	0.358	0.232	[0.025, 0.438]	0.413	-0.054	[-0.236, 0.128]	0.364
DML	0.155	[-0.028, 0.338]	0.366	0.176	[-0.046, 0.398]	0.444	0.144	[-0.052, 0.339]	0.391

6 Extensions

This section extends the binary variable Y to encompass general cases, including continuous, counting, and multinomial outcomes. First, we examine the class of single-parameter exponential families, where the conditional density function is solely determined by the nonparmatric conditional mean function. This covers continuous outcomes and counting variables. Second, we consider the "vector" case of exponential families for multinomial outcomes. For both classes, we derive the novel correction to the Bayesian

procedure and delegate more technical discussions to the online Appendices D and F. Additionally, we outline extensions to other causal parameters of interest.

6.1 A Single-parameter Exponential Family

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

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

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

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

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

We consider the same reparametrization of (π, m, f) as in (2.2) except that now the second component of η uses the general link function q satisfying $\eta^m = q(m)$. We now state the least favorable direction for the exponential family case, which serves as motivation for the prior adjustment.

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

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

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

For the outcome family with $a = 1$, which includes Bernoulli, Poisson and exponential distributions, the least favorable direction for ATE estimation coincides with the one as given in Lemma 3.1. To implement the double robust Bayesian procedure for general outcomes, one can still follow the algorithm described in Section 2.2, with the logistic function Ψ in (2.6) replaced by the inverse link function q^{-1}. For the normal (homoscedastic) outcome where prior adjustment in (2.6) becomes $\lambda\widehat{\gamma}(d,x)/a$, the hyperparameter a can be determined together with other parameters of the Gaussian process by optimizing the marginal likelihood as in Ray and Szabó [2019]. In Proposition F.1, in the online supplementary appendix, we provide primitive conditions for the BvM Theorem to hold under double robust smoothness conditions.

6.2 Multinomial Outcomes

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

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

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

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

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

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

We emphasize that the least favorable direction calculation is not a trivial extension of Hahn [1998] or Ray and van der Vaart [2020]. This is because there are J nonparametric components involved in the conditional probability function of the multinomial outcomes

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

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

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

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

6.3 Other Causal Parameters

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

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

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

A Proofs of Main Results

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

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

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

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

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

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

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

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

show that the Laplace transform $I_n(t)$ given in (A.1) satisfies

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

for every t in a neighborhood of 0, where the limit at the right hand side of (A.3) is the Laplace transform of a $N(0, \mathrm{v}_0)$ distribution. Note that we can write $\tau_\eta = \int \bar{m}_\eta dF_\eta$. Further, let $\widehat{\tau} = \int \bar{m}_0 dF_0 + \mathbb{P}_n[\widetilde{\tau}_0]$, which satisfies (3.1).

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

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

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

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_0 \rho^{m_0}] + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] + t\sqrt{n}\int (\bar{m}_0 - \bar{m}_\eta) dF_0 + \frac{t^2}{2} P_0(B_0^m \xi_0^m)^2 + o_{P_0}(1),$$

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

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

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

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

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

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

display becomes

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

where $F_0\bar{m}_0 \equiv \int \bar{m}_0(x)dF_0(x)$ and $F_n\bar{m}_0 \equiv 1/n\sum_{i=1}^n \bar{m}_0(X_i)$. We take a closer examination about the empirical process term in the integral. Note that $dm(d,x) = dm(1,x)$ and $(1-d)m(d,x) = (1-d)m(0,x)$ for any $m(\cdot,\cdot)$ and x. Thus, we get

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

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

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

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

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

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

decomposition

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

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

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

where the last equation follows from the following derivation:

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

using the Cauchy-Schwarz inequality, Assumption 2, and Assumption 3. Consider the second summand on the right hand side of (A.5). From Lemma C.8 we infer

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

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

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

where the last equation is due to the equation C.7. □

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

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

Now the definition of the estimator $\widehat{\tau}$ given in (3.1) yields $\sqrt{n}\widehat{\tau} = \sqrt{n}\big(\tau_0 + \mathbb{P}_n\widetilde{\tau}_0\big) + o_{P_0}(1)$.

For any set A, we write $\mathbb{N}(A) := \int_A e^{-u^2/2}/\sqrt{2\pi}du$. Theorem 3.1 implies

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

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

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

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

Proof of Proposition 4.1. With slight abuse of notation, we stick to $\mathcal{H}_n^m$ for the set that receives the posterior mass going to 1 and $\eta^m(d,\cdot) \in \mathcal{H}_n^m$ for $d \in \{0,1\}$.[7] Note that $\widehat{\gamma}$ is based on an auxiliary sample and hence we can treat $\widehat{\gamma}$ below as a deterministic of functions denoted by γ_n satisfying the rate restrictions $\|\gamma_n\|_\infty = O(1)$ and $\|\gamma_n - \gamma_0\|_\infty = O\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$. We first verify Assumption 2 with $\varepsilon_n = (n/\log n)^{-s_m/(2s_m+p)}$. Let $\mathcal{H}_n^m := \{w_d + \lambda\gamma_n : (w_d,\lambda) \in \mathcal{W}_n\}$, where

$$\mathcal{W}_n := \left\{(w_d,\lambda) : w_d \in \mathcal{B}_n^m, |\lambda| \leqslant M\sigma_n\sqrt{n}\varepsilon_n\right\} \cap \left\{(w_d,\lambda) : \|\Psi(w_d(\cdot) + \lambda\gamma_n) - m_0(d,\cdot)\|_{2,F_0} \leqslant \varepsilon_n\right\},$$

where the sieve space $\mathcal{B}_n^m$ in the first restriction for the Gaussian process W_d is defined in the equation (C.8) with $d \in \{0,1\}$. Intuitively speaking, the bulk of the Gaussian process is contained in an ε_n-shell of a big multiple of the unit ball of the RKHS[8]. The second restriction concerns the posterior contraction rate and it is shown in our Lemma C.3. Referring to the condition $\|\widehat{\gamma}\|_\infty = O_{P_0}(1)$ and $\|\widehat{\gamma} - \gamma_0\|_\infty = O_{P_0}\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$, we write $r_n := C_r(n/\log n)^{-s_\pi/(2s_\pi+p)}$. Then $\sqrt{n}\varepsilon_n r_n = o(1)$ holds, if $2s_m/(2s_m+p) + 2s_\pi/(2s_\pi+p) > 1$, which can rewritten as $\sqrt{s_\pi\, s_m} > p/2$.

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

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

[7] When we write $\eta^m \in \mathcal{H}_n^m$, it means $(\eta^m(1,\cdot), \eta^m(0,\cdot)) \in \mathcal{H}_n^m \times \mathcal{H}_n^m$.

[8] We refer readers to the discussion leading to Lemma C.7 on the Reproducing Kernel Hilbert Space (RKHS) and related norms.

Consequently, from Lemma C.5 we infer

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

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

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

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

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

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

Finally, it remains to verify Assumption 4. By the univariate Gaussian tail bound, the prior mass of the set $\Lambda_n := \{\lambda : |\lambda| > u_n\sigma_n^2\sqrt{n}\}$ is bounded above by $e^{-u_n^2\sigma_n^2 n/2}$. Also, the Kullback-Leibler neighborhood around η_0^m has prior probability at least $e^{-n\varepsilon_n^2}$; see Lemma 4 in Ray and van der Vaart [2020]. By the assumption $\sigma_n \gg \varepsilon_n$ as imposed in the rate restriction (4.2), we have $\varepsilon_n^2 \lesssim u_n^2\sigma_n^2$, which means

$$\frac{\Pi(\lambda\in\Lambda_n)}{\Pi\left(\left\{(w,\lambda) : K \vee V(p_{\eta_0^m}, p_{w+\lambda\gamma}) \leqslant \varepsilon_n^2\right\}\right)} = o(e^{-n\varepsilon_n^2}).$$

The stated contraction $\Pi(\lambda \in \Lambda_n \mid Z^{(n)}) \to_{P_0} 0$ in Assumption 4(i) follows from Lemma 4 of Ray and van der Vaart [2020]. Regarding Assumption 4(ii), this set hardly differs from

the set $\mathcal{H}_n^m$ because $\sqrt{n}\varepsilon_n \to 0$ and $\|\gamma_n\|_\infty = O(1)$. Its posterior probability is seen to tend to 1 in probability by the same arguments as for $\mathcal{H}_n^m$, possibly after replacing ε_n with a multiple of itself. □

B Key Lemmas

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

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

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

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

$$\ell_n^m(\eta^m) - \ell_n^m(\eta_t^m) = t\mathbb{G}_n[\gamma_0\rho^{m_0}] + t\mathbb{G}_n[\gamma_0(m_0 - m_\eta)] + t\sqrt{n}\int(\bar{m}_0 - \bar{m}_\eta)dF_0 + \frac{t^2}{2}P_0(B_0^m\xi_0^m)^2 + o_{P_0}(1).$$

Proof. We start with the following decomposition:

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

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

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

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

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

The rest of the proof involves a standard Taylor expansion for the third term on the right hand side of the above equation. By the equation (C.5) in our Lemma C.1, we get

$$-nP_0g'(0) = t\sqrt{n}P_0[\gamma_0\rho^{m_0}] + t\sqrt{n}P_0[\gamma_0(m_0 - m_\eta)] = t\sqrt{n}\int(\bar{m}_0 - \bar{m}_\eta)dF_0,$$

by the fact that $P_0[\gamma_0 \rho^{m_0}] = 0$ and the definition of the Riesz representor γ_0 in (2.5). Regarding the second-order term in the Taylor expansion in the equation (C.6) of Lemma C.1, we get

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

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

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

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

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

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

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

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

$$\begin{aligned} nP_0[\log p_{\eta^m} - \log p_{\eta_t^m}] &= -n(P_0 g'(0) + P_0 g^{(2)}(0)) + o_{P_0}(1) \\ &= t^2 P_0(B_0^m \xi_0^m)^2 + t\sqrt{n}\int(\bar{m}_0 - \bar{m}_\eta)dF_0 + o_{P_0}(1), \end{aligned}$$

which leads to the desired result. □

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

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

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

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

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

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

where ϕ_{σ_n} denotes the probability density function of a $N(0, \sigma_n^2)$ random variable and the set B_n is defined by $B_n = \{(w, \lambda) : w + \lambda\gamma_n \in \mathcal{H}_n^m, |\lambda| \leqslant 2u_n\sigma_n^2\sqrt{n}\}$ where $u_n \to 0$ and $u_n n\sigma_n^2 \to \infty$. Considering the log likelihood ratio of two normal densities together with the constraint $|\lambda| \leqslant 2u_n\sigma_n^2\sqrt{n}$, it is shown on page 3015 of Ray and van der Vaart [2020] that

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

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

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

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

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

The desired result would follow from $\Pi(B_n|X^{(n)}) = 1 - o_{P_0}(1)$ and $\Pi(B_{n,t}|X^{(n)}) = 1 - o_{P_0}(1)$. The first convergence directly follows from Assumption 4. The set $B_{n,t}$ is the

intersection of these two conditions in Assumption 4, except that the restriction on λ in $B_{n,t}$ is $|\lambda + t/\sqrt{n}| \leqslant 2u_n\sqrt{n}\sigma_n^2$ instead of $|\lambda| \leqslant u_n\sqrt{n}\sigma_n^2$. By construction, we have $t/\sqrt{n} = o(u_n\sqrt{n}\sigma_n^2)$, so that $\Pi(B_{n,t}|X^{(n)}) = 1 - o_{P_0}(1)$.

We finish the proof by establishing the following result:

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

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

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

Next, we apply third-order Taylor expansions in Lemma C.1 separately to the two terms in the brackets of the above display:

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

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

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

In order to control T_1, we evaluate

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

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

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

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

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

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

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

References

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

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

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

S. Athey, G. W. Imbens, J. Metzger, and E. Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 2021.

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

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

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

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

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

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

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

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

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

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

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

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

R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

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

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

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

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

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

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

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

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

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

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

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

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

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

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

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

R. J. Little and D. B. Rubin. *Statistical analysis with missing data (3rd Edition)*. John Wiley & Sons, 2019.

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

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

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

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

Y. Ritov, P. J. Bickel, A. C. Gamst, and B. J. K. Kleijn. The bayesian analysis of complex, high-dimensional models: Can it be coda? *Statistical Science*, 29(4):619–639, 2014.

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

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

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

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

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

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

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

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

A. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

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

A. W. van der Vaart and J. H. van Zanten. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37:2655–2675, 2009.

L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

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

Supplement to "Double Robust Bayesian Inference on Average Treatment Effects"

Christoph Breunig Ruixuan Liu Zhengfei Yu

February 21, 2024

This online supplementary appendix contains materials to support our main paper. Appendix C collects some auxiliary results. Appendix D collects the proofs for lemmas in Section 6 of the main paper. Appendix E provides least favorable directions for other causal parameters of interest besides the ATE. Appendix F states and proves the BvM theorem for outcome variables belonging to one-parameter exponential family described in Section 6 of the main paper. Appendix G describes how to draw the posterior of the conditional mean function using the Laplace approximation. Appendix H 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. For two sequences a_n, b_n, we write $a_n \lesssim b_n$, if $a_n \leqslant C b_n$.

C Auxiliary Results

The likelihood associated with the component $\eta^m = \Psi^{-1}(m)$ is given by

$$p_{\eta^m}(z) = m(d,x)^y (1 - m(d,x))^{1-y}, \tag{C.1}$$

with the corresponding log-likelihood $\ell_n^m(\eta^m) = \sum_{i=1}^n \log p_{\eta^m}(Z_i)$. In other words, $p_{\eta^m}(\cdot)$ is the density with respect to the dominating measure

$$d\nu(x,d,y) = (\pi_0(x))^d (1 - \pi_0(x))^{1-d} d\vartheta(d,y) dF_0(x), \tag{C.2}$$

where ϑ stands for the counting measure on $\{\{0,0\},\{0,1\},\{1,0\},\{1,1\}\}$. We introduce some simplifying notations by writing

$$m^1(\cdot) = m(1,\cdot) \quad \text{and} \quad m^0(\cdot) = m(0,\cdot).$$

For two generic probability densities p and q, we denote the Kullback-Leibler (KL) divergence by $K(p,q)$ and the square KL variation by $V(p,q)$; see Appendix B in Ghosal and Van der Vaart [2017].

Lemma C.1. *Let Assumption 1 be satisfied and $m_\eta = \Psi(\eta^m)$, then we have*

$$\log p_{\eta^m} - \log p_{\eta_t^m} = \frac{t}{\sqrt{n}}\gamma_0\rho^{m_\eta} + \frac{t^2}{2n}\gamma_0^2 m_\eta(1-m_\eta) + R_n,$$

where $\|R_n\|_\infty \lesssim n^{-3/2}$.

Proof. First of all, the link function Ψ satisfies that $\Psi' = \Psi(1-\Psi)$ and $\Psi^{(2)} = \Psi(1-\Psi)(1-2\Psi)$ by straightforward calculus. Thus, $\log p_{\eta^m} - \log p_{\eta_t^m} = g(0) - g(1)$, where $g(u) = \log p_{\eta_u^m}$. We examine the following Taylor expansion:

$$g(0) - g(1) = -g'(0) - g^{(2)}(0)/2 - \theta, \tag{C.3}$$

where $\theta \leqslant \|g^{(3)}\|_\infty$. We express the part of the log-likelihood involving η^m explicitly as follows.

$$\begin{aligned}\log p_{\eta^m}(z) =& dy \log \frac{e^{\eta^m(1,x)}}{1+e^{\eta^m(1,x)}} + d(1-y)\log\frac{1}{1+e^{\eta^m(1,x)}} \\ &+(1-d)y\log\frac{e^{\eta^m(0,x)}}{1+e^{\eta^m(0,x)}} + (1-d)(1-y)\log\frac{1}{1+e^{\eta^m(0,x)}} \\ =& d(y\eta^{m^1} - \psi(\eta^{m^1})) + (1-d)(y\eta^{m^0} - \psi(\eta^{m^0}))\end{aligned} \tag{C.4}$$

where $\psi(\eta) = \log(1+e^\eta)$.

Recall the least favorable direction $\xi_0^m(d,x) = \gamma_0(d,x) = d/\pi_0(x) - (1-d)/(1-\pi_0(x))$. Also, note that $d(1-d) = 0$. These derivatives can be calculated by splitting the right hand side of the equation (C.4) into these d and $(1-d)$ terms separately. For instance, the first-order derivative is

$$g'(u) = -\frac{t}{\sqrt{n}}\left[\frac{d}{\pi_0(x)}(y - \Psi(\eta_u^{m^1}))\right] + \frac{t}{\sqrt{n}}\left[\frac{1-d}{1-\pi_0(x)}(y-\Psi(\eta_u^{m^0}))\right] = -\frac{t}{\sqrt{n}}\gamma_0(y-\Psi(\eta_u^m)).$$

The other two can be computed along the same lines:

$$g^{(2)}(u) = -\frac{t^2}{n}\gamma_0^2\Psi'(\eta_u^m), \quad g^{(3)}(u) = -\frac{t^3}{n^{3/2}}\gamma_0^3\Psi^{(2)}(\eta_u^m).$$

In the above expression involving the Riesz representor, we have

$$\gamma_0^2(d,x) = \frac{d}{\pi_0^2(x)} + \frac{1-d}{(1-\pi_0(x))^2} \quad \text{and} \quad \gamma_0^3(d,x) = \frac{d}{\pi_0^3(x)} - \frac{1-d}{(1-\pi_0(x))^3},$$

again because of $d(1-d) = 0$. Evaluating at $u = 0$, we have $\Psi(\eta_u^m) = \Psi(\eta^m) = m_\eta$ and consequently,

$$g'(0) = -\frac{t}{\sqrt{n}}\gamma_0\rho^{m_0} + \frac{t}{\sqrt{n}}\gamma_0(m_\eta - m_0), \tag{C.5}$$

and

$$g^{(2)}(0) = -\frac{t^2}{n}\gamma_0^2(m_\eta(1-m_\eta)). \tag{C.6}$$

For the remainder term, we have $\|g^{(3)}\|_\infty \lesssim n^{-3/2}$, given the uniform boundedness of $\Psi^{(2)}(\cdot)$. □

Lemma C.2. *Let Assumptions 1 and 2 be satisfied. Then, we have*

$$\sqrt{n}\mathbb{P}_n[(\widehat{\gamma} - \gamma_0)\rho^{m_0}] = o_{P_0}(1). \tag{C.7}$$

Proof. Since $\widehat{\gamma}$ is based on an auxiliary sample, it is sufficient to consider deterministic functions γ_n with the same rates of convergence as $\widehat{\gamma}$. We also write the corresponding propensity score as π_n, which is associated with γ_n. Denoting $U_i = Y_i - m_0(D_i, X_i)$, we evaluate for the conditional expectation that

$$\begin{aligned}
&\mathbb{E}_0\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\gamma_n - \gamma_0)(D_i,X_i)U_i\right)^2 \mid (D_1,X_1),\ldots,(D_n,X_n)\right] \\
&= \frac{1}{n}\sum_{i\neq i'}(\gamma_n - \gamma_0)(D_i,X_i)(\gamma_n - \gamma_0)(D_{i'},X_{i'})\mathbb{E}_0\left[U_iU_{i'} \mid (D_i,X_i),(D_{i'},X_{i'})\right] \\
&= \frac{1}{n}\sum_{i=1}^n(\gamma_n - \gamma_0)^2(D_i,X_i)Var_0(Y_i|X_i).
\end{aligned}$$

We have $Var_0(Y_i|X_i) \leqslant 1$ since $Y_i \in \{0,1\}$ and thus we obtain for the unconditional squared

expectation that

$$\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\gamma_n-\gamma_0)(D_i,X_i)U_i\Big)^2\right]\lesssim\|\pi_n-\pi_0\|_{2,F_0}^2=o(1)$$

by Assumption 2, which implies the desired result. □

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 take the sieve space to be

$$\mathcal{B}_n^m:=\varepsilon_n\mathbb{B}_1^{s_m,p}+M_n\mathbb{H}_1^{a_n},\tag{C.8}$$

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 $M_n=-2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$. The addition of the small ball $\varepsilon_n\mathbb{B}_1^{s_m,p}$ creates an ε_n-cushion around the multiple $M_n\mathbb{H}_1^{a_n}$. This is necessary to create enough mass of the sieve space for the Gaussian process W. For notational simplicity, we suppress the dependence of the rescaled Gaussian process on the rescaling parameter a_n in the following proofs.

Lemma C.3. *Under the conditions of Proposition 4.1, the posterior distributions of the conditional mean functions contract at rate ε_n, i.e.,*

$$\Pi\left(\|m_\eta(d,\cdot)-m_0(d,\cdot)\|_{2,F_0}\geqslant M\varepsilon_n\mid Z^{(n)}\right)\to_{P_0}0$$

for $d\in\{0,1\}$ and every sufficiently large M, as $n\to\infty$.

Proof. By the assumed stochastic independence between the pair $Z^{(n)}$ and $\widehat{\gamma}$, we can proceed by studying the ordinary posterior distribution relative to the prior with $\widehat{\gamma}$ set equal to a deterministic function γ_n and (w,λ) following their prior. In other words, it is sufficient to consider the prior on m given by $m(d,x)=\Psi\left(W_d^m(x)+\lambda\,\gamma_n(d,x)\right)$ where $W_d^m(\cdot)$ is the rescaled squared exponential process independent of $\lambda\sim N(0,\sigma_n^2)$ and γ_n a

sequence of functions $\|\gamma_n\|_\infty = O(1)$. It suffices to examine two conditional means $m_\eta(1,\cdot)$ and $m_\eta(0,\cdot)$ separately. We focus on the treatment arm with $d = 1$, and leave d off the notations in W^m or η^m as understood.

We verify the following generic results in Theorem 2.1 of Ghosal et al. [2000] to obtain the proper concentration rate for the posterior for the rescaled squared exponential process:

$$\text{I.}\quad \Pi((w,\lambda) : K \vee V(p_{\eta_0^m}, p_{w+\lambda\gamma_n}) \leqslant \varepsilon_n^2) \geqslant c_1 \exp(-c_2 n\varepsilon_n^2), \tag{C.9}$$

$$\text{II.}\quad \Pi(\mathcal{P}_n^c) \leqslant \exp(-c_3 n\varepsilon_n^2), \tag{C.10}$$

$$\text{III.}\quad \log N(\varepsilon_n, \mathcal{P}_n, \|\cdot\|_{L^2(\nu)}) \leqslant c_4 n\varepsilon_n^2, \tag{C.11}$$

for positive constant terms $c_1, \cdots, c_4$ and for the set:

$$\mathcal{P}_n = \left\{p_{w+\lambda\gamma_n} : w \in \mathcal{B}_n^m, |\lambda| \leqslant M\sigma_n\sqrt{n}\varepsilon_n\right\}.$$

(I). The inequality (C.15) in Lemma C.6 yields

$$\{(w,\lambda) : \|w - \eta_0^m\|_\infty \leqslant c\varepsilon_n, |\lambda| \leqslant c\varepsilon_n\} \subset \left\{(w,\lambda) : K \vee V(p_{\eta_0^m}, p_{w+\lambda\gamma_n}) \leqslant \varepsilon_n^2\right\}.$$

Given that we have independent priors of W^m and λ, the prior probability of the set on the left of the above display can be lower bounded by $\Pi(\|W^m - \eta_0^m\|_\infty \leqslant c\varepsilon_n)\Pi(|\lambda| \leqslant c\varepsilon_n)$. By Proposition 11.19 of Ghosal and Van der Vaart [2017] regarding the small exponent function $\phi_0^{a_n}$ and together with the upper bound (C.17), we infer

$$\Pi(\|W^m - \eta_0^m\|_\infty \leqslant c\varepsilon_n) \geqslant \exp\left(-\phi_0^{a_n}(\varepsilon_n/2)\right) \geqslant \exp\left(-cn\varepsilon_n^2\right),$$

for some positive constant c. The second term is lower bounded by $C\varepsilon_n/\sigma_n$, which is of order $O(\varepsilon_n)$ for $\sigma_n = O(1)$. Therefore, we have ensured that the prior assigns enough mass around a Kullback-Leibler neighborhood of the truth.

(II). Referring to the sieve space for the Gaussian process, we apply Borell's inequality from Proposition 11.17 of Ghosal and Van der Vaart [2017]:

$$\Pr\{W^m \notin \mathcal{B}_n^m\} \leqslant 1 - \Phi(\iota_n + M_n),$$

where $\Phi(\cdot)$ is the c.d.f. of a standard normal random variable and the sequence ι_n is given by $\Phi(\iota_n) = \Pr\{W \in \varepsilon_n \mathbb{B}_1^{s_m,p}\} = e^{-\phi_0^{a_n}(\varepsilon_n)}$. Since our choice of ε_n leads to $\phi_0^{a_n}(\varepsilon_n) \leqslant n\varepsilon_n^2$, we have $\iota_n \geqslant -M_n/2$ if $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$ for some $C > 1$. In this case, $\Pi(\mathcal{B}_n^{mc}) \leqslant$

$1 - \Phi(M_n/2) \leqslant e^{-Cn\varepsilon_n^2}$. Next, we apply the univariate Gaussian tail inequality for λ:

$$\Pr\{|\lambda| \geqslant u_n \sigma_n \sqrt{n}\} \leqslant 2e^{-u_n^2 n\sigma_n^2/2},$$

which is bounded above by $e^{-Cn\varepsilon_n^2}$ for $u_n \to 0$ sufficiently slowly, given our assumption $\varepsilon_n = o(\sigma_n)$. Hence, by the union bound, we have $\Pi(\mathcal{P}_n^c) \lesssim e^{-Cn\varepsilon_n^2}$.

(III). To bound the entropy number of the functional class $\mathcal{P}_n$, consider the inequality

$$\|p_{w+\lambda\gamma} - p_{\bar{w}+\bar{\lambda}\gamma_n}\|_{L^2(\nu)} \lesssim \|w - \bar{w}\|_{2,F_0} + |\lambda - \bar{\lambda}|\|\gamma_n\|_\infty,$$

where the dominating measure ν is (C.2). Thus, we have

$$N(\varepsilon_n, \mathcal{P}_n, \|\cdot\|_{L^2(\nu)}) \leqslant N(\varepsilon_n/2, \mathcal{B}_n^m, \|\cdot\|_\infty) \times N(c\varepsilon_n, [0, 2M\sigma_n\sqrt{n}\varepsilon_n], |\cdot|) \lesssim n\varepsilon_n^2. \quad \text{(C.12)}$$

Note that the logarithm of the second term grows at the rate of $O(\log n)$, and it is the first term that dominates. Because Ψ is monotone and Lipschitz, a set of ϵ-brackets in $L^2(F_0)$ for $\mathcal{B}_n^m$ translates into a set of ε-brackets in $L^2(\nu)$ for $\mathcal{P}_n$. Thus, Lemma C.7 gives us $\log N(3\varepsilon_n, \mathcal{B}_n^m, \|\cdot\|) \lesssim n\varepsilon_n^2$.

By Lemma 15 of Ray and van der Vaart [2020], this delivers the posterior contraction rate for $m_\eta(1,\cdot)$ in terms of the $L^2(F_0\pi_0)$-norm, which is equivalent to the $L^2(F_0)$-norm weighted by the propensity score π_0. Analogous arguments lead to the desired result for the conditional mean $m_\eta(0,\cdot)$ for the control group. □

Let $M_{ni} = e_i / \sum_{i=1}^n e_i$, where e_i's are independently and identically drawn from the exponential distribution Exp(1). We also denote $X^{(n)} = (X_i)_{i=1}^n$. We adopt the following notations: $\mathbb{F}_n^* \bar{m}_\eta = \sum_{i=1}^n M_{ni} \bar{m}_\eta(X_i)$, $\mathbb{F}_n \bar{m}_\eta = n^{-1} \sum_{i=1}^n \bar{m}_\eta(X_i)$ and $F_0 \bar{m}_\eta = \int \bar{m}_\eta(x) dF_0(x)$. Let $X^{(n)} = (X_i)_{i=1}^n$.

Lemma C.4. *Let the functional class $\{\bar{m}_\eta : \eta \in \mathcal{H}_n\}$ be a P_0-Glivenko-Cantelli class. Then for every t in a sufficiently small neighborhood of 0, in P_0-probability,*

$$\sup_{\bar{m}_\eta : \eta \in \mathcal{H}_n} \left| \mathbb{E}\left[e^{t\sqrt{n}((\mathbb{F}_n^* - \mathbb{F}_n)\bar{m}_\eta)} \mid X^{(n)} \right] - e^{t^2 F_0(\bar{m}_\eta - F_0\bar{m}_\eta)^2/2} \right| \to 0.$$

Proof. We verify the conditions from Lemma 1 in Ray and van der Vaart [2020]. First, the Bayesian bootstrap law $\mathbb{F}_n^*$ is the same as the posterior law for F, when its prior is a Dirichlet process with its base measure taken to be zero. Second, the assumed P_0-Glivenko-Cantelli

class entails

$$\sup_{\eta \in \mathcal{H}_n} |(\mathbb{F}_n - F_0)\bar{m}_\eta| = o_{P_0}(1).$$

Last, the required moment condition on the envelope function for the class involving $\bar{m}_\eta$ is automatically satisfied because of $\|\bar{m}_\eta\|_\infty \leqslant 1$. □

The following lemma is in the same spirit of Lemma 9 in Ray and van der Vaart [2020] with one important difference. That is, we do not restrict the range of the function φ to be $[0, 1]$. As we apply this lemma by taking $\varphi = \gamma_n - \gamma_0$, it can take on negative values. We apply the more general contraction principle from Theorem 4.12 of Ledoux and Talagrand [1991] instead of Proposition A.1.10 of van der Vaart and Wellner [1996]. This allows us to relax the positive range restriction in Ray and van der Vaart [2020].

Lemma C.5. *Consider a set $\mathcal{H}$ of measurable functions $h : \mathcal{Z} \mapsto \mathbb{R}$ and a bounded measurable function φ. We have*

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

Proof. We start with $\mathbb{G}_n(\varphi h) = \mathbb{G}_n(\varphi(h - P_0 h)) + P_0 h \mathbb{G}_n(\varphi)$. The expectation of $P_0 h \mathbb{G}_n(\varphi)$ is bounded by the second term on the right hand side of the inequality in the stated lemma. It suffices to bound $\mathbb{G}_n(\varphi(h - P_0 h))$ for any function h such that $P_0 h = 0$.

Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables independent of observations $Z^{(n)}$. By Lemma 2.3.6 of van der Vaart and Wellner [1996],

$$\mathbb{E} \sup_h \left| \sum_{i=1}^n (\varphi(Z_i)h(Z_i) - P_0[\varphi h]) \right| \leqslant 2\|\varphi\|_\infty \mathbb{E} \sup_h \left| \sum_{i=1}^n \epsilon_i \frac{\varphi(Z_i)}{\|\varphi\|_\infty} h(Z_i) \right|. \tag{C.13}$$

Because $-1 \leqslant \varphi(Z_i)/\|\varphi\|_\infty \leqslant 1$ for all $i = 1, \ldots, n$, we can apply the contraction principle as in Theorem 4.12 on page 112 of Ledoux and Talagrand [1991]. The contraction mapping is understood to be $h \mapsto \frac{\varphi}{\|\varphi\|_\infty} \times h$ herein. Hence, the above inequality (C.13) remains true if the variables $\frac{\varphi(Z_i)}{\|\varphi\|_\infty}$ on the right hand side are removed. Another application by the symmetrization inequality from Lemma 2.3.6 of van der Vaart and Wellner [1996] that decouples the Rademacher variables leads to the desired result. □

The next lemma upper bounds the L^2 distance and Kullback-Leibler divergence of the probability density functions by the L^2 distance of the reparametrized function η^m, cf. Lemma 2.8 of Ghosal and Van der Vaart [2017] or Lemma 15 of Ray and van der Vaart [2020].

Lemma C.6. *For any measurable functions $v^m, w^m : [0,1] \mapsto \mathbb{R}$, we have*

$$\begin{aligned}\|p_{v^m} - p_{w^m}\|_{L^2(\nu)} &\leqslant \|\Psi(v^{m^1}) - \Psi(w^{m^1})\|_{L^2(F_0\pi_0)} \vee \|\Psi(v^{m^0}) - \Psi(w^{m^0})\|_{L^2(F_0(1-\pi_0))} \\ &\leqslant \|v^{m^1} - w^{m^1}\|_{2,F_0} \vee \|v^{m^0} - w^{m^0}\|_{2,F_0}. \end{aligned} \tag{C.14}$$

In addition, it holds that

$$K(p_{v^m}, p_{w^m}) \vee V(p_{v^m}, p_{w^m}) \leqslant \|v^{m^1} - w^{m^1}\|_{2,F_0}^2 \vee \|v^{m^0} - w^{m^0}\|_{2,F_0}^2. \tag{C.15}$$

The small ball exponent function for the associated Gaussian process prior is

$$\phi_0(\varepsilon) := -\log \Pr(\|W\|_\infty < \varepsilon);$$

see equation (11.10) in Ghosal and Van der Vaart [2017]. In the above display, $\|\cdot\|_\infty$ is the uniform norm of $\mathcal{C}([0,1]^p)$, the Banach space in which the Gaussian process sits. $\mathbb{H}$ is the reproducing kernel Hilbert space (RKHS) of the process with its RKHS norm $\|\cdot\|_{\mathbb{H}}$. To abuse the notation a bit, we denote the small ball exponent of the rescaled process $W(at)$ by $\phi_0^a(\varepsilon)$. Lemma 11.55 in Ghosal and Van der Vaart [2017] gives this bound for the (rescaled) squared exponential process:

$$\phi_0^a(\varepsilon) \lesssim a^p(\log(a/\varepsilon))^{1+p}.$$

Lemma C.7. *Assume that $\varepsilon_n = n^{-s_m/(2s_m+p)}(\log n)^{s_m(1+p)/(2s_m+p)}$ and $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$ for a positive constant $C > 1$. Also, let $a_n \asymp n^{1/(2s_m+p)}(\log n)^{-(1+p)/(2s_m+p)}$. Then, for the sieve space $\mathcal{B}_n^m = \varepsilon_n \mathbb{B}_1^{s_m,p} + M_n \mathbb{H}_1^{a_n}$, we have*

$$\log N(3\varepsilon_n, \mathcal{B}_n^m, \|\cdot\|_\infty) \lesssim n\varepsilon_n^2. \tag{C.16}$$

Proof. The argument is similar as in Lemma 11.20 of Ghosal and Van der Vaart [2017]. We provide the proof for completeness. Let $h_1, \cdots, h_N \in M_n \mathbb{H}_1^{a_n}$ be $2\varepsilon_n$-separated functions in terms of the Banach space norm. Then, the ε_n-balls $h_1 + \varepsilon_n \mathbb{B}_1^{s_m,p}, \cdots, h_N + \varepsilon_n \mathbb{B}_1^{s_m,p}$ are disjoint. Therefore, we have

$$1 \geqslant \sum_{j=1}^N \Pr\{W \in h_j + \varepsilon_n \mathbb{B}_1^{s_m,p}\} \geqslant \sum_{j=1}^N e^{-\|h_j\|_{\mathbb{H}}^2/2} \Pr\{W \in \varepsilon_n \mathbb{B}_1^{s_m,p}\} \geqslant n e^{-M_n^2/2} e^{-\phi_0^{a_n}(\varepsilon_n)},$$

where the second inequality follows from Lemma 11.18 of Ghosal and Van der Vaart [2017]

and the last inequality makes use of the fact that $h_1, \cdots, h_N \in M_n\mathbb{H}_1$, as well as the definition of the small ball exponent function.

For a maximal $2\varepsilon_n$-separated set $h_1, \cdots, h_N$, the balls around $h_1, \cdots, h_N$ of radius $2\varepsilon_n$ cover the set $M_n\mathbb{H}_1^{a_n}$. Thus, we have $\log N(2\varepsilon_n, M_n\mathbb{H}_1^{a_n}, \|\cdot\|_\infty) \leqslant \log N \leqslant M_n^2/2 + \phi_0^{a_n}(\varepsilon_n)$. Referring to the inequality (iii) of Lemma K.6 of Ghosal and Van der Vaart [2017] for the quantile function of a standard normal distribution, we have $M_n^2 \lesssim n\varepsilon_n^2$ by the choice of M_n stated in the lemma. It is straightforward yet tedious to verify that

$$\phi_0^{a_n}(\varepsilon_n) \lesssim n\varepsilon_n^2, \tag{C.17}$$

for the specified a_n and ε_n. Since any point of $\mathcal{B}_n^m$ is within ε_n of an element of $M_n\mathbb{H}_1^{a_n}$, this also serves as a bound on $\log N(3\varepsilon_n, \mathcal{B}_n^m, \|\cdot\|_\infty)$. □

A key step in showing the validity of the debiasing step is the following:

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

which is equivalent to the following lemma.

Lemma C.8. *Under Assumption 2 for the pilot estimators, the following result holds:*

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

Proof. We start with the following identity:

$$\mathbb{P}_n[\widehat{\gamma}\rho^{\widehat{m}} + \widehat{\bar{m}}] = \mathbb{P}_n[\gamma_0\rho^{m_0} + \bar{m}_0] + R_{n1} + R_{n2}.$$

where

$$R_{n1} = \frac{1}{n}\sum_{D_i}(Y_i - \widehat{m}(1, X_i))\left(\frac{1}{\widehat{\pi}(X_i)} - \frac{1}{\pi_0(X_i)}\right) + \frac{1}{n}\sum_{1-D_i}(Y_i - \widehat{m}(0, X_i))\left(\frac{1}{1-\widehat{\pi}(X_i)} - \frac{1}{1-\pi_0(X_i)}\right),$$

$$R_{n2} = \frac{1}{n}\sum_i(\widehat{m}(1, X_i) - m_0(1, X_i))\left(1 - \frac{D_i}{\pi_0(X_i)}\right) + \frac{1}{n}\sum_i(\widehat{m}(0, X_i) - m_0(0, X_i))\frac{D_i - \pi_0(X_i)}{1-\pi_0(X_i)}.$$

Referring to the first term R_{n1}, we have

$$
\begin{aligned}
R_{n1} = & \frac{1}{n}\sum_{D_i}(m_0(1,X_i)-\widehat{m}(1,X_i))\left(\frac{1}{\widehat{\pi}(X_i)}-\frac{1}{\pi_0(X_i)}\right)+\frac{1}{n}\sum_{D_i}(Y_i-m_0(1,X_i))\left(\frac{1}{\widehat{\pi}(X_i)}-\frac{1}{\pi_0(X_i)}\right) \\
& -\frac{1}{n}\sum_{1-D_i}(m_0(0,X_i)-\widehat{m}(0,X_i))\left(\frac{1}{1-\widehat{\pi}(X_i)}-\frac{1}{1-\pi_0(X_i)}\right) \\
& -\frac{1}{n}\sum_{1-D_i}(Y_i-m_0(0,X_i))\left(\frac{1}{1-\widehat{\pi}(X_i)}-\frac{1}{1-\pi_0(X_i)}\right).
\end{aligned}
$$

The negligibility of the first and third terms in R_{n1} follows from the Cauchy-Schwarz inequality and the rate conditions imposed in Assumption 2. The second and fourth terms can be combined together so that the negligibility can be shown as in Lemma C.2.

Consider R_{n2}. To bound its first summand, we condition on $(X_1,\ldots,X_n)$, as well as the pilot estimators $\widehat{m}$ and $\widehat{\pi}$, which are computed over the external sample. We use the fact that $(D_i-\pi_0(X_i))$ has a conditional zero mean. Specifically, this leads to

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

using that $Var_0(D_i|X_i)=\pi_0(X_i)(1-\pi_0(X_i))$. By the overlapping condition as imposed in Assumption 1, i.e., $\bar{\pi}<\pi_0(X_i)$ for all $1\leqslant i\leqslant n$ and the uniform convergence of $\widehat{\pi}$ to π_0, we obtain

$$
\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i-\pi_0(X_i)}{\widehat{\pi}(X_i)}(\widehat{m}(1,X_i)-m_0(1,X_i))\Big)^2\,\Big|\,\widehat{m},\widehat{\pi}\right]\lesssim\|\widehat{m}(1,\cdot)-m_0(1,\cdot)\|_{2,F_0}^2=o_{P_0}(1),
$$

where the last equation is due to the convergence rate for the pilot estimator $\widehat{m}$ in Assumption 3. The negligibility of the second term in R_{n2} is proved in a similar fashion. □

The following lemma shows the stochastic equicontinuity when the true conditional mean function belongs to a Hölder space, which is P_0-Donsker, i.e., $s_m>p/2$. The main complication is that the sieve space related to the Gaussian process prior is not a fixed P_0-Donsker class, as it changes with sample size n and the envelope function is also slowly diverging, cf. the comments in the third paragraph on Page 2007 of Ray and van der Vaart [2020]. More specifically, for the rescaled squared exponential process priors, we rely on the metric entropy bounds in van der Vaart and van Zanten [2009]. With this important

modification, the proof is along similar lines with the proof of Lemma 7 of Ray and van der Vaart [2020] for the Riemann-Lioville process; also, see Lemma 5 of Ray and van der Vaart [2020].

Lemma C.9. *Recall that the sieve space related to the Gaussian process is* $\mathcal{B}_n^m = \varepsilon_n \mathbb{B}_1^{s_m,p} + M_n \mathbb{H}_1^{a_n}$. *For* $s_m > p/2$, *we have* $\mathbb{E}_0 \sup_{\eta \in \mathcal{H}_n^m} \mathbb{G}_n [m_\eta - m_0] = o(1)$.

Proof. Because the link function $\Psi(\cdot)$ is monotone and Lipschitz continuous, separate sets of brackets for the two constituents of the set $\varepsilon_n \mathbb{B}_1^{s_m,p} + M_n \mathbb{H}_1^{a_n}$, as well as the bracket for $\{\lambda : |\lambda| \leqslant M\sigma_n\sqrt{n}\varepsilon_n\}$ can be combined into brackets for the sum space.

$$\log N_{[]}(\varepsilon, \mathcal{H}_n^m, L^2(P_0)) \leqslant \log N(\varepsilon, \varepsilon_n \mathbb{B}_1^{s_m,p}, \|\cdot\|_\infty) + \log N(\varepsilon, M_n \mathbb{H}_1^{a_n}, \|\cdot\|_\infty) + \log N(c\varepsilon, [0, 2M\sigma_n\sqrt{n}\varepsilon_n], |\cdot|).$$

The last term is of strictly smaller order than the second one. The bound for the first component attached to the Hölder space can be found in Proposition C.5 of Ghosal and Van der Vaart [2017]:

$$\log N(\varepsilon, \varepsilon_n \mathbb{B}_1^{s_m,p}, \|\cdot\|_\infty) \lesssim \left(\frac{\varepsilon_n}{\varepsilon}\right)^{s_m/p},$$

which is bounded if we take $\varepsilon = \varepsilon_n$. The entropy bound for the first component is given in Lemma C.7, which states that $\log N(\varepsilon, M_n \mathbb{H}_1^{a_n}, \|\cdot\|_\infty) \lesssim n\varepsilon_n^2 \lesssim \varepsilon_n^{-2\upsilon}$, with $\upsilon = p/(2s_m)$ modulo some $\log n$ term on the right hand of the bound. In this case, the empirical process bound of [Han, 2021, p.2644] yields

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

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

D Proofs of Section 6

Proof of Lemma 6.1. For the submodel $t \to \eta_t$ defined in 6.1, we evaluate

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

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

$$B_\eta(\mathfrak{p}, \mathfrak{m}, \mathfrak{f})(Z) = B_\eta^\pi \mathfrak{p}(Z) + B_\eta^m \mathfrak{m}(Z) + B_\eta^f \mathfrak{f}(Z), \tag{D.1}$$

where $B_\eta^\pi \mathfrak{p}(Z) = (D - \pi_\eta(X))\mathfrak{p}(X)$, $B_\eta^f \mathfrak{f}(Z) = \mathfrak{f}(X)$, and

$$\begin{aligned} B_\eta^m \mathfrak{m}(O) &= \left[aY - \frac{A'(m_\eta(D,X))}{q'(m_\eta(D,X))} \right] \mathfrak{m}(D,X), \\ &= a\left(Y - m_\eta(D,X)\right) \mathfrak{m}(D,X), \end{aligned} \tag{D.2}$$

where the second equality follows from the property of the moments of exponential family, see e.g., Theorem 9.47 of Wasserman [2004] with $T(y) = ay$:

$$\mathbb{E}_\eta \left[aY | D, X \right] = \frac{A'(m_\eta(D,X))}{q'(m_\eta(D,X))}.$$

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. One can easily verify the differentiability of the ATE parameter in the sense of van der Vaart [1998] and show that the efficient influence function remains the same as in Hahn [1998] and Ray and van der Vaart [2020]. Given the particular form of the efficient influence function $\widetilde{\tau}_\eta$ in (2.4), the function $\xi_\eta = (\xi_\eta^\pi, \xi_\eta^m, \xi_\eta^f)$ defined in (3.4) satisfies $B_\eta \xi_\eta = \widetilde{\tau}_\eta$, and hence, ξ_η defines the least favorable direction. □

Proof of Lemma 6.2. We emphasize that the least favorable direction calculation is not a trivial extension of Hahn [1998] or Ray and van der Vaart [2020], because there are J nonparametric components involved in the conditional probabilities of the multinomial outcomes given covariates, and we need to consider the perturbation of all J components together.

Consider the log transformation of the joint density of $Z = (Y, D, X^\top)^\top$ given by

$$\log p_\eta(z) = d \log(\pi_\eta(x)) + (1-d) \log(1 - \pi_\eta(x)) + \sum_{j=0}^{J} 1_{\{y_i = j\}} \log\left(m_{j,\eta}\left(d,x\right)\right) + \log f(x)$$

Following the proof of Lemma 3.1, it is sufficient to consider the perturbations for $j = 1, \ldots, J$:

$$\Psi_j(\eta^{m_1} + t\mathfrak{m}_1, \cdots, \eta^{m_J} + t\mathfrak{m}_J)(d,x) = \frac{\exp((\eta^{m_j} + t\mathfrak{m}_j)(d,x))}{1 + \sum_{l=1}^{J} \exp((\eta^{m_l} + t\mathfrak{m}_l)(d,x))}$$

or

$$\log \Psi_j(\eta^{m_1} + t\mathfrak{m}_1, \cdots, \eta^{m_J} + t\mathfrak{m}_J)(d,x) = (\eta^{m_j} + t\mathfrak{m}_j)(d,x) - \log\left(1 + \sum_{l=1}^{J} \exp((\eta^{m_l} + t\mathfrak{m}_l)(d,x))\right).$$

Taking derivatives

$$\left.\frac{d\log\Psi_j(\eta^{m_1}+t\mathfrak{m}_1,\cdots,\eta^{m_J}+t\mathfrak{m}_J)(d,x)}{dt}\right|_{t=0} = \mathfrak{m}_j(d,x) - \frac{\sum_{l=1}^J \exp(\eta^{m_l}(d,x))\mathfrak{m}_l(d,x)}{1+\sum_{l=1}^J \exp(\eta^{m_l}(d,x))}$$
$$= \mathfrak{m}_j(d,x) - \sum_{l=1}^J m_{\eta,l}(d,x)\mathfrak{m}_l(d,x)$$

by the definition of $m_{\eta,l}$. Likewise, we also obtain

$$\left.\frac{d\log\Psi_0(\eta^{m_1}+t\mathfrak{m}_1,\cdots,\eta^{m_J}+t\mathfrak{m}_J)(d,x)}{dt}\right|_{t=0} = -\sum_{l=1}^J m_{\eta,l}(d,x)\mathfrak{m}_l(d,x).$$

We need to verify the differentiability of the ATE parameter in the sense of van der Vaart [1998]. Due to its technical feature, we leave this to the end of the proof. From there, we can see that the score operator of the vector of conditional means $(m_1,\ldots,m_J)$ is as follows:

$$B_\eta^m(\mathfrak{m}_1,\ldots,\mathfrak{m}_J)(z) = \sum_{j=0}^J 1_{\{y=j\}}\left.\frac{d\log\Psi_j(\eta^{m_1}+t\mathfrak{m}_1,\cdots,\eta^{m_J}+t\mathfrak{m}_J)(d,x)}{dt}\right|_{t=0}$$
$$= \sum_{j=1}^J 1_{\{y=j\}}\left(\mathfrak{m}_j(d,x) - \sum_{l=1}^J m_{\eta,l}(d,x)\mathfrak{m}_l(d,x)\right) + 1_{\{y=0\}}\left(-\sum_{l=1}^J m_{\eta,l}(d,x)\mathfrak{m}_l(d,x)\right).$$

By the fact that $1_{\{y=0\}} = 1 - \sum_{j=1}^J 1_{\{y=j\}}$, it simplifies to

$$B_\eta^m(\mathfrak{m}_1,\ldots,\mathfrak{m}_J)(z) = \sum_{j=1}^J \left(1_{\{y=j\}} - m_{\eta,j}(d,x)\right)\mathfrak{m}_j(d,x).$$

Note that the conditional mean of $B_\eta^m(\mathfrak{m}_1,\ldots,\mathfrak{m}_J)(z)$ is zero for any $\mathfrak{m}_j(d,x)$, which agrees with the requirement of the score operator.

From our verification of the differentiability, we confirm that the influence function is of the generic form given in Hahn [1998] and Ray and van der Vaart [2020]. Also, it is contained in the closed linear span of the set of all score functions. Now, if we choose $\mathfrak{m}_j = j\gamma_\eta$, $1 \leqslant j \leqslant J$, we obtain

$$B_\eta^m(\gamma_\eta, 2\gamma_\eta,\ldots,J\gamma_\eta)(z) = \left(\underbrace{\sum_{j=1}^J 1_{\{y=j\}}j}_{=y} - \underbrace{\sum_{j=1}^J j\, m_{\eta,j}(d,x)}_{=m_\eta(d,x)}\right)\gamma_\eta(d,x) = (y - m_\eta(d,x))\,\gamma_\eta(d,x),$$

which shows the results.

Now we check the pathwise differentiability of the ATE. To avoid the long display of various formulas, we consider the following decomposition

$$\frac{d}{dt}\tau_{\eta_t}\Big|_{t=0} = \frac{d}{dt}\int \mathbb{E}_{\eta_t}[Y|D=1,X=x]dF_{\eta_t}(x) - \frac{d}{dt}\int \mathbb{E}_{\eta_t}[Y|D=0,X=x]dF_{\eta_t}(x),$$

and we focus on the first derivative involving the treatment group, as the other one can be handled analogously. We start with

$$\frac{d}{dt}\mathbb{E}_{\eta_t}[\mathbb{E}_{\eta_t}[Y|D=1,X]] = \iint y\frac{d}{dt}p_t(y|1,x)f_t(x)\Big|_{t=0}d\nu(y)d\mu(x),$$

where $p_t(y|1,x)$ and $f_t(x)$ are the perturbed conditional density of outcome and marginal density of covariates, respectively. In addition, ν stands for the counting measure and μ is the Lebesgue measure. By the chain rule, we need to compute the following sum:

$$\iint y\frac{d}{dt}p_t(y|1,x)\Big|_{t=0}d\nu(y)f_\eta(x)d\mu(x) + \iint yp_\eta(y|1,x)d\nu(y)\frac{d}{dt}f_t(x)\Big|_{t=0}d\mu(x). \tag{D.3}$$

Regarding the first part of the above sum, we follow the outline in Example 2 of Jonathan [2019] to compute

$$\frac{d}{dt}p_t(y|d,x) = \frac{d}{dt}\left[\prod_{j=0}^{J} m_{t,j}(d,x)^{1\{y=j\}}\right] = \sum_{j=0}^{J} 1_{\{y=j\}}\frac{\partial}{\partial t}m_{t,j}(d,x)\prod_{k\neq j} m_{t,k}^{1\{y=k\}}(d,x).$$

We thus evaluate for the derivatives of the conditional mean functions

$$\frac{\partial}{\partial t}m_{t,j}(d,x)\Big|_{t=0} = m_{\eta,j}(d,x)\left(\mathfrak{m}_j(d,x) - \sum_{l=1}^{J} m_{\eta,l}(d,x)\mathfrak{m}_l(d,x)\right), \quad \text{for} \quad j=1,\cdots,J,$$

and

$$\frac{\partial}{\partial t}m_{t,0}(d,x)\Big|_{t=0} = m_{\eta,0}(d,x)\left(-\sum_{l=1}^{J} m_{\eta,l}(d,x)\mathfrak{m}_l(d,x)\right).$$

Thereafter, derivative of the conditional density can be written as

$$\begin{aligned}\frac{d}{dt}p_t(y|d,x)\Big|_{t=0} &= \left[\sum_{j=1}^{J} 1\{y=j\}\left(\mathfrak{m}_j(d,x) - \sum_{l=1}^{J} m_{\eta,l}(d,x)\mathfrak{m}_l(d,x)\right)\right]\prod_{j=0}^{J} m_{\eta,j}(d,x)^{1\{y=j\}} \\ &= \left(B_\eta^m(\mathfrak{m}_1,\ldots,\mathfrak{m}_J)(z) - \mathbb{E}_\eta[B_\eta^m(\mathfrak{m}_1,\ldots,\mathfrak{m}_J)(Z)|D=d,X=x]\right)p_\eta(y|d,x),\end{aligned}$$

where the last equality follows from the fact that the conditional mean of the score given (D, X) is zero. To simplify the notation, we denote this conditional score function by

$$S_\eta(z) = B_\eta^m(\mathfrak{m}_1, \ldots, \mathfrak{m}_J)(z).$$

Referring to the first term in the summation (D.3), we resort to the technique in Example 2 of Jonathan [2019] by converting the conditional argument from $d = 1$ to $d \in \{0, 1\}$. Similar to the first two terms in the long display on Page 15 of Jonathan [2019], we obtain

$$\iint y \frac{d}{dt} p_t(y|1, x)\Big|_{t=0} d\nu(y) f(x) d\mu(x) = \mathbb{E}_\eta \left[\frac{D}{\pi_\eta(X)} (Y - m_\eta(D, X)) S_\eta(Z) \right].$$

Referring to the second part of (D.3), we immediately obtain

$$\iint y p_t(y|1, x) d\nu(y) \frac{d}{dt} f_t(x)\Big|_{t=0} d\mu(x) = \mathbb{E}_\eta \left[(m_\eta(1, X) - \mathbb{E}_\eta[m_\eta(1, X)]) S_\eta(Z) \right]$$

Similarly for the control arm, we derive

$$\begin{aligned} &\frac{d}{dt} \int \mathbb{E}_{\eta_t}[Y|D = 0, X = x] dF_{\eta_t}(x)\Big|_{t=0} \\ =&\mathbb{E}_\eta \left[\left(m_\eta(0, X) - \mathbb{E}_\eta[m(0, X)] + \frac{1 - D}{1 - \pi_\eta(X)} (Y - m_\eta(D, X)) \right) S_\eta(Z) \right]. \end{aligned}$$

The remaining part boils down to the existence of a vector-valued function $\tilde{\tau}_{P_\eta}$ such that

$$\begin{aligned} \frac{d}{dt} \tau(\eta_t)\Big|_{t=0} &= \mathbb{E}_\eta \left[\tilde{\tau}_\eta(Z) B_\eta^m(\mathfrak{m}_1, \ldots, \mathfrak{m}_J)(Z) \right] \\ &= \mathbb{E}_\eta \left[\left((\bar{m}_\eta(X) - \tau_\eta) + \left(\frac{D}{\pi_\eta(X)} - \frac{1 - D}{1 - \pi_\eta(X)} \right) (Y - m_\eta(D, X)) \right) S_\eta(Z) \right]. \end{aligned}$$

Consequently, we can take the solution as $\tilde{\tau}_\eta(z) = \bar{m}_\eta(x) - \tau_\eta + \gamma_\eta(d, x)(y - m_\eta(d, x))$, which concludes the proof. □

E Least Favorable Directions for Other Causal Parameters

In this part, we provide details on the least favorable directions for the first two examples in Section 6.3. We properly address the binary outcome Y and the reparameterization through the logistic type link function $\Psi(\cdot)$.

E.1 Average Policy Effects

The joint density of $Z_i = (Y_i, X_i)$ can be written as

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

The observed data Z_i can be described by (m, f). It proves to be more convenient to consider the reparametrization of (m, f) given by $\eta = (\eta^m, \eta^f)$, where

$$\eta^m = \Psi^{-1}(m), \quad \eta^f = \log f. \tag{E.2}$$

Consider the one-dimensional submodel $t \mapsto \eta_t$ defined by the path

$$m_t(x) = \Psi(\eta^m + t\mathfrak{m})(x), \quad f_t(x) = f(x)e^{t\mathfrak{f}(x)}/\mathbb{E}[e^{t\mathfrak{f}(X)}],$$

for the given direction $(\mathfrak{m}, \mathfrak{f})$ with $\mathbb{E}[\mathfrak{f}(X)] = 0$. For this submodel, we further evaluate

$$\begin{aligned}\log p_{\eta_t}(z) = {} & y \log \Psi(\eta^m + t\mathfrak{m})(x) + (1 - y) \log(1 - \Psi(\eta^m + t\mathfrak{m}))(x) \\ & + t\mathfrak{f}(x) - \log \mathbb{E}[e^{t\mathfrak{f}(X)}] + \log f(x).\end{aligned}$$

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

$$B_\eta(\mathfrak{m}, \mathfrak{f})(Z) = B^m_\eta \mathfrak{m}(Z) + B^f_\eta \mathfrak{f}(Z), \tag{E.3}$$

where $B^m_\eta \mathfrak{m}(Z) = (Y - m_\eta(X))\mathfrak{m}(X)$ and $B^f_\eta \mathfrak{f}(Z) = \mathfrak{f}(X)$.

The efficient influence function for estimation of the policy effect parameter τ^P_η is given by

$$\widetilde{\tau}^P_\eta(z) = \gamma^P_\eta(x)(y - m_\eta(x))$$

where $\gamma^P_\eta(x) = \frac{g_1(x) - g_0(x)}{f(x)}$. Now the score operator B_η given in (E.3) applied to $\xi^P_\eta(x) = \big(\gamma^P_\eta(x), 0\big)$, yields $B_\eta \xi^P_\eta = \widetilde{\tau}^P_\eta$. Thus, ξ^P_η defines the least favorable direction for this policy effect parameter.

E.2 Average Derivative

The joint density of $Z_i = (Y_i, D_i, X_i)$ can be written as

$$p_{m,f}(z) = m(d, x)^y(1 - m(d, x))^{(1-y)} f(d, x). \tag{E.4}$$

The observed data Z_i can be described by (m, f). It proves to be more convenient to consider the reparametrization of (m, f) given by $\eta = (\eta^m, \eta^f)$, where

$$\eta^m = \Psi^{-1}(m), \quad \eta^f = \log f. \tag{E.5}$$

Consider the one-dimensional submodel $t \mapsto \eta_t$ defined by the path

$$m_t(d,x) = \Psi(\eta^m + t\mathfrak{m})(d,x), \quad f_t(d,x) = f(d,x)e^{t\mathfrak{f}(d,x)}/\mathbb{E}[e^{t\mathfrak{f}(D,X)}],$$

for the given direction $(\mathfrak{m}, \mathfrak{f})$ with $\mathbb{E}[\mathfrak{f}(D,X)] = 0$. For this submodel defined in (E.5), we further evaluate

$$\begin{aligned}\log p_{\eta_t}(z) = {} & y \log \Psi(\eta^m + t\mathfrak{m})(d,x) + (1-y)\log(1 - \Psi(\eta^m + t\mathfrak{m}))(d,x) \\ & + t\mathfrak{f}(d,x) - \log \mathbb{E}[e^{t\mathfrak{f}(D,X)}] + \log f(d,x).\end{aligned}$$

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

$$B_\eta(\mathfrak{m}, \mathfrak{f})(Z) = B_\eta^m \mathfrak{m}(Z) + B_\eta^f \mathfrak{f}(Z), \tag{E.6}$$

where $B_\eta^m \mathfrak{m}(Z) = (Y - m_\eta(D,X))\mathfrak{m}(D,X)$ and $B_\eta^f \mathfrak{f}(Z) = \mathfrak{f}(D,X)$. The efficient influence function for estimation of the AD parameter $\tau_\eta^{AD} = \mathbb{E}\left[\partial_d m_\eta(D,X)\right]$ is given by

$$\widetilde{\tau}_\eta^{AD}(z) = \partial_d m_\eta(d,x) - \mathbb{E}[\partial_d m_\eta(d,x)] + \gamma_\eta^{AD}(d,x)(y - m_\eta(d,x))$$

where $\gamma_\eta^{AD}(d,x) = \partial_d \pi_\eta(d,x)/\pi_\eta(d,x)$. Now the score operator B_η given in (E.6) applied to

$$\xi_\eta^{AD}(d,x) = \left(\gamma_\eta^{AD}(d,x), \partial_d m_\eta(d,x) - \mathbb{E}[\partial_d m_\eta(D,X)]\right),$$

yields $B_\eta \xi_\eta^{AD} = \widetilde{\tau}_\eta^{AD}$. Thus, ξ_η^{AD} defines the least favorable direction for the AD.

F Theory for One-parameter Exponential Family

We take $a = 1$ in the exponential family for simplicity, that is,

$$f_{Y|D,X}(y; m(d,x)) = c(y)\exp\left[q(m(d,x))Y - A(m(d,x))\right], \tag{F.1}$$

and for some known functions $c(y), q(m)$ and $A(m)$. We reparameterize the model using the link function q:

$$\eta^m(d,x) = q(m(d,x)), \quad m(d,x) = q^{-1}(\eta^m(d,x)),$$

and we define the mapping $B := A \circ q^{-1}$.

Proposition F.1 (One-parameter exponential family)**.** *Consider the one-parameter exponential family for the conditional distribution specified by* (F.1)*. Assume that the function $B = A \circ q^{-1}$ is three time differentiable with $\|B^{(3)}\|_\infty < \infty$. The estimator $\widehat{\gamma}$ satisfies $\|\widehat{\gamma}\|_\infty = O_{P_0}(1)$ and $\|\widehat{\gamma} - \gamma_0\|_\infty = O_{P_0}\big((n/\log n)^{-s_\pi/(2s_\pi+p)}\big)$ for some $s_\pi > 0$. Suppose $m_0(d,\cdot) \in \mathcal{C}^{s_m}([0,1]^p)$ for $d \in \{0,1\}$ and some $s_m > 0$ with $\sqrt{s_\pi\, s_m} > p/2$. Also, $\|\widehat{m}(d,\cdot) - m_0(d,\cdot)\|_{2,F_0} = O_{P_0}\big((n/\log n)^{-s_m/(2s_m+p)}\big)$. Consider the propensity score-dependent prior on m given by $m(d,x) = q^{-1}\left(W_d^m(x) + \lambda\,\widehat{\gamma}(d,x)\right)$, where $W_d^m(x)$ is the rescaled squared exponential process for $d \in \{0,1\}$, with its rescaling parameter a_n of the order in* (4.1) *and $(n/\log n)^{-s_m/(2s_m+p)} \ll \sigma_n \lesssim 1$. Then, the posterior distribution satisfies Theorem 3.1.*

Proof. Because our analysis for the binary outcome has served as the template, we only outline the necessary modifications. Due to the change of the likelihood function form in the conditional probability density of the outcome, we need to adapt the argument in showing the contraction rate of the posterior and the local asymptotic normality (LAN) expansion used in the conditional Laplace transform, as well as verifying the prior stability.

First, in deriving the rate of posterior contraction or determining the proper localized set $\mathcal{H}_n^m$, we need proper upper bounds for the L^2 distance and Kullback-Leibler (KL) divergence of the probability density functions by the L^2 distance of the reparametrized functions η^m, v^m. To abuse the notation a bit, we denote the corresponding probability densities by p_{η^m} and p_{v^m}. For the exponential family under consideration, the first and second order cumulants (conditional on covariates) are:

$$\mathbb{E}_\eta[Y|D=d,X=x] = B'(\eta^m(d,x)), \quad Var_\eta(Y|D=d,X=x) = B^{(2)}(\eta^m(d,x)).$$

Considering the KL divergence $K(p_{\eta^m}, p_{v^m}) = \int \log\left(p_{\eta^m}(z)/p_{v^m}(z)\right) p_{\eta^m}(z)dz$, we first compute

$$\log \frac{p_{\eta^m}(z)}{p_{v^m}(z)} = (\eta^m(d,x) - v^m(d,x))y - [B(\eta^m(d,x)) - B(v^m(d,x))].$$

Integrating over the conditional density for any given (d, x) and utilizing the fact that the conditional mean is $m_\eta(d,x) = B'(\eta_\eta^m(d,x))$, we proceed for some intermediate value $\tilde{\eta}^m$:

$$\begin{aligned} K(p_{\eta^m}, p_{v^m}) &= \int \left(B'(\eta^m)(\eta^m - v^m) - [B(\eta^m) - B(v^m)]\right) \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ &= \int B^{(2)}(\tilde{\eta}^m)(\eta^m - v^m)^2 \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ &\lesssim \|v^{m^1} - \eta^{m^1}\|_{2,F_\eta}^2 \vee \|v^{m^0} - \eta^{m^0}\|_{2,F_\eta}^2. \end{aligned}$$

Recall that

$$V(p_{\eta^m}, p_{v^m}) = \int \left[\log \frac{p_{\eta^m}(z)}{p_{v^m}(z)} - K(p_{\eta^m}, p_{v^m})\right]^2 p_{\eta^m}(z) dz \leqslant \int \left[\log \frac{p_{\eta^m}(z)}{p_{v^m}(z)}\right]^2 p_{\eta^m}(z) dz. \tag{F.2}$$

Therefore, we continue with the right hand side inequality of F.2.

$$\begin{aligned} & V(p_{\eta^m}, p_{v^m}) \\ \leqslant & \int \left\{(\eta^m(d,x) - v^m(d,x))y - [B(\eta^m(d,x)) - B(v^m(d,x))]\right\}^2 p_{\eta^m}(z) dz \\ = & \int (\eta^m(d,x) - v^m(d,x))^2 [B^{(2)}(\eta^m(d,x)) + (B'(\eta^m(d,x)))^2] \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ & -2 \int (B(\eta^m(d,x)) - B(v^m(d,x)))(\eta^m(d,x) - v^m(d,x)) B'(\eta^m(d,x)) \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ & + \int (B(\eta^m(d,x)) - B(v^m(d,x)))^2 \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ = & \int B^{(2)}(\eta^m(d,x))(\eta^m(d,x) - v^m(d,x))^2 \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ & + \int \left\{(\eta^m(d,x) - v^m(d,x)) B'(\eta^m(d,x)) - [B(\eta^m(d,x)) - B(v^m(d,x))]\right\}^2 \pi^d(x)(1-\pi(x))^{1-d} f_\eta(x) dx \\ \lesssim & \|v^{m^1} - \eta^{m^1}\|_{2,F_\eta}^2 \vee \|v^{m^0} - \eta^{m^0}\|_{2,F_\eta}^2, \end{aligned}$$

where in the first equality we have made use of the fact that

$$\mathbb{E}_\eta[Y^2 | D = d, X = x] = B^{(2)}(\eta^m(d,x)) + (B'(\eta^m(d,x)))^2.$$

In sum, we have

$$K(p_{\eta^m}, p_{v^m}) \vee V(p_{\eta^m}, p_{v^m}) \leqslant \|v^{m^1} - \eta^{m^1}\|_{2,F_\eta}^2 \vee \|v^{m^0} - \eta^{m^0}\|_{2,F_\eta}^2.$$

In addition, the squared Hellinger distance can be upper bounded by the KL divergence

from Lemma B.1 in Ghosal and Van der Vaart [2017], so we have

$$\|\sqrt{p_{v^m}} - \sqrt{p_{\eta^m}}\|_{L^2(\nu)} \leqslant \|v^{m^1} - \eta^{m^1}\|_{2,F_\eta} \vee \|v^{m^0} - \eta^{m^0}\|_{2,F_\eta}. \tag{F.3}$$

Second, we outline the changes to the LAN expansion as follows. For this purpose, we also define

$$g(u) = \log p_u(z) = y\eta_u(d,x) - B(\eta_u(d,x)) + \log c(y). \tag{F.4}$$

By the property of the one-parameter exponential family, we know $B(\cdot)$ is a convex function under our smoothness assumption for $B(\cdot)$. Thereafter, we can obtain the first to third order derivatives as

$$\begin{aligned} g'(0) &= \frac{t}{\sqrt{n}}\gamma_0(y - B'(\eta^m(d,x))) = \frac{t}{\sqrt{n}}\gamma_0(y - m(d,x)), \\ g^{(2)}(0) &= \frac{t^2}{n}\gamma_0^2 B^{(2)}(\eta^m(d,x)), \quad g^{(3)}(\tilde{u}) = \frac{t^3}{n^{3/2}}\gamma_0^3 B^{(3)}(\eta^m_{\tilde{u}}(d,x)). \end{aligned}$$

In a key step to show the prior stability condition, we need to establish the following log-likelihood expansion:

$$\sup_{\eta^m \in \mathcal{H}_n^m} \left|\ell_n^m(\eta^m - t\gamma_n/\sqrt{n}) - \ell_n^m(\eta^m - t\gamma_0/\sqrt{n})\right| = o_{P_0}(1), \tag{F.5}$$

where $\eta^m_{n,t} = \eta^m - t\gamma_n/\sqrt{n}$ and $\eta^m_t = \eta^m - t\gamma_0/\sqrt{n}$. Consider the following decomposition of the log-likelihood:

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

Then, we apply third-order Taylor expansions for the one-parameter exponential family separately to the two terms in the brackets of the above display:

$$\begin{aligned} n\mathbb{P}_n[\log p_{\eta^m_{n,t}} - \log p_{\eta^m}] &= -t\sqrt{n}\mathbb{P}_n\left[\gamma_n(y - m_\eta)\right] - \frac{t^2}{2}\mathbb{P}_n\left[\gamma_n^2 B^{(2)}(\eta^m(d,x))\right] - \frac{t^3}{\sqrt{n}}\mathbb{P}_n\left[\gamma_n^3 B^{(3)}(\eta^m_{u^*})\right], \\ n\mathbb{P}_n[\log p_{\eta^m} - \log p_{\eta^m_t}] &= t\sqrt{n}\mathbb{P}_n\left[\gamma_0(y - m_\eta)\right] + \frac{t^2}{2}\mathbb{P}_n\left[\gamma_0^2 B^{(2)}(\eta^m(d,x))\right] + \frac{t^3}{\sqrt{n}}\mathbb{P}_n\left[\gamma_0^3 B^{(3)}(\eta^m_{u^{**}})\right], \end{aligned}$$

for some intermediate points $u^*, u^{**} \in (0,1)$, cf. the equation (B.1). The rest of the proof follows similar lines to our proof of Proposition 4.1. □

G Algorithm for drawing the posterior of η^m

We describe the Laplace approximation method that is used to draw the posterior of $\eta^m(d,x)$; see Rassmusen and Williams [2006, Chapters 3.3 to 3.5] for more details on properties of the Laplace approximation. as follows. Let $\boldsymbol{W} = [\boldsymbol{D}, \boldsymbol{X}] \in \mathbb{R}^{n\times(p+1)}$ be the matrix of (D,X) in the data, $\boldsymbol{W}^* \in \mathbb{R}^{2n\times(p+1)}$ the evaluation points $(1,X)$ and $(0,X)$

$$\boldsymbol{W}^* = \begin{bmatrix} \mathbf{1}_n, & \boldsymbol{X} \\ \mathbf{0}_n, & \boldsymbol{X} \end{bmatrix},$$

and $\boldsymbol{\eta}_n^*$ a $2n$-vector that gives the latent function $\eta^m(d,x)$ evaluated at $\boldsymbol{W}^*$:

$$\boldsymbol{\eta}^* = [\eta^m(1,X_1), \cdots, \eta^m(1,X_n), \eta^m(0,X_1), \cdots, \eta^m(0,X_n)]^\top.$$

Let $\boldsymbol{\eta} = [\eta^m(D_1,X_1), \ldots, \eta^m(D_n,X_n)]^\top$ denote the n-vector of the latent function at $\boldsymbol{W}$. For matrices $\boldsymbol{W}^*$ and $\boldsymbol{W}$, we define $K_c(\boldsymbol{W}^*, \boldsymbol{W})$ as a $2n \times n$ matrix whose (i,j)-th element is $K_c(W_i^*, W_j)$, where W_i^* is the i-th row of $\boldsymbol{W}^*$ and W_j is the j-th row of $\boldsymbol{W}$. Analogously, $K_c(\boldsymbol{W}, \boldsymbol{W})$ is an $n \times n$ matrix with the (i,j)-th element being $K_c(W_i, W_j)$, and $K_c(\boldsymbol{W}^*, \boldsymbol{W}^*)$ is a $2n \times 2n$ matrix with the (i,j)-th element being $K_c(W_i^*, W_j^*)$.

Given the mean-zero GP prior with its covariance kernel K_c, the posterior of $\boldsymbol{\eta}^*$ is approximated by a Gaussian distribution with the mean $\bar{\boldsymbol{\eta}}^*$ and covariance $\mathrm{V}(\boldsymbol{\eta}^*)$ using the Laplace approximation. To be specific, let

$$\begin{aligned} \bar{\boldsymbol{\eta}}^* &= K_c(\boldsymbol{W}^*, \boldsymbol{W}) K_c^{-1}(\boldsymbol{W}, \boldsymbol{W})\, \widehat{\boldsymbol{\eta}}, \\ \mathrm{V}(\boldsymbol{\eta}^*) &= K_c(\boldsymbol{W}^*, \boldsymbol{W}^*) - K_c(\boldsymbol{W}^*, \boldsymbol{W}) \left(K_c(\boldsymbol{W}, \boldsymbol{W}) + \boldsymbol{\nabla}^{-1}\right)^{-1} K_c^\top(\boldsymbol{W}^*, \boldsymbol{W}), \end{aligned}$$

where $\widehat{\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\boldsymbol{W}, \boldsymbol{Y})$ maximizes the posterior $p(\boldsymbol{\eta}|\boldsymbol{W}, \boldsymbol{Y})$ on the latent $\boldsymbol{\eta}$ and $\boldsymbol{\nabla} = -\frac{\partial^2 \log p(\boldsymbol{Y}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^\top}$ is a $n \times n$ diagonal matrix with the i-th diagonal entry being $-\frac{\partial^2 \log p(\boldsymbol{Y}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_i^2}$. We use the Matlab toolbox `GPML` for the implementation.[9] In sum, we get the posterior draws of the vectors $[\eta^m(1,X_1), \cdots, \eta^m(1,X_n)]^\top$ and $[\eta^m(0,X_1), \cdots, \eta^m(0,X_n)]^\top$ from the above approximating Gaussian distribution with the mean $\bar{\boldsymbol{\eta}}^*$ and covariance $\mathrm{V}(\boldsymbol{\eta}^*)$.We then obtain the posterior draws of the ATE by equation (2.8) via $m(d,X_i) = \Psi(\eta^m(d,X_i))$ for $d \in \{0,1\}$.

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

H Additional Simulation Results

Appendix H presents additional simulation results for adjusted Bayesian inference methods. The design is the same as that in Section 5.1. Tables A1 evaluates the sensitivity of finite sample performance with respect to the variance σ_n that determines influence strength of the prior correction term. We set $\sigma_n = c_\sigma \times \sqrt{\dim(X)\, n \log n} / \sum_{i=1}^n |\widehat{\gamma}(D_i, X_i)|$ with $c_\sigma \in \{0.5, 1, 10\}$. Note that $c_\sigma = 1$ corresponds to the simulation results reported in the main text. The performance of PA and DR Bayes, especially the latter, appears stable with respect to the choice of c_σ.

Table A1: The effect of σ_n on adjusted Bayesian inference methods:trimming based on $\hat{\pi} \in [t, 1-t]$, $\bar{n}$ = the average sample size after trimming.

	Methods	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
Spec I		$t = 0.10(\bar{n} = 240)$			$t = 0.05(\bar{n} = 364)$			$t = 0.01(\bar{n} = 665)$		
$t_\sigma = 0.5$	PA Bayes	-0.018	0.979	0.231	0.022	0.960	0.233	0.037	0.956	0.296
	DR Bayes	-0.029	0.983	0.213	0.008	0.970	0.212	0.017	0.982	0.251
$t_\sigma = 1$	PA Bayes	-0.002	0.982	0.274	0.037	0.940	0.260	0.051	0.875	0.310
	DR Bayes	-0.021	0.979	0.229	0.016	0.966	0.224	0.026	0.938	0.258
$t_\sigma = 10$	PA Bayes	0.015	0.967	0.313	0.048	0.920	0.277	0.059	0.833	0.314
	DR Bayes	-0.014	0.977	0.244	0.022	0.958	0.232	0.031	0.899	0.260
Spec II		$t = 0.10(\bar{n} = 226)$			$t = 0.05(\bar{n} = 345)$			$t = 0.01(\bar{n} = 603)$		
$c_\sigma = 0.5$	PA Bayes	-0.005	0.972	0.263	0.025	0.956	0.259	0.026	0.909	0.285
	DR Bayes	-0.022	0.967	0.220	0.004	0.962	0.222	0.006	0.946	0.253
$c_\sigma = 1$	PA Bayes	0.007	0.966	0.282	0.035	0.930	0.269	0.032	0.883	0.290
	DR Bayes	-0.013	0.964	0.233	0.012	0.957	0.230	0.011	0.930	0.258
$c_\sigma = 10$	PA Bayes	0.012	0.959	0.289	0.038	0.919	0.273	0.034	0.876	0.292
	DR Bayes	-0.009	0.964	0.238	0.015	0.954	0.233	0.013	0.923	0.259
Spec III		$t = 0.10(\bar{n} = 212)$			$t = 0.05(\bar{n} = 321)$			$t = 0.01(\bar{n} = 613)$		
$c_\sigma = 0.5$	PA Bayes	-0.003	0.971	0.282	0.023	0.946	0.271	0.032	0.906	0.287
	DR Bayes	-0.022	0.966	0.235	0.002	0.953	0.232	0.016	0.945	0.263
$c_\sigma = 1$	PA Bayes	0.005	0.962	0.296	0.029	0.934	0.277	0.035	0.890	0.290
	DR Bayes	-0.016	0.963	0.243	0.007	0.953	0.237	0.019	0.932	0.266
$c_\sigma = 10$	PA Bayes	0.008	0.960	0.303	0.031	0.930	0.279	0.036	0.888	0.290
	DR Bayes	-0.014	0.961	0.246	0.008	0.950	0.238	0.020	0.934	0.266

Table A2 reports the finite sample performance of DR Bayes using sample-split. We use one half of the sample (92 treated and 1245 control observations) to estimate the prior and posterior adjustments, and then draw the posterior of the conditional mean $m(d, x)$ using the other half of the sample (93 treated and 1245 control observations). The effective

sample size $\bar{n}$ in Table A2 corresponds to the after-trimming size of 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. The credible interval length increases as a result of halving the sample size.

Table A2: Double robust Bayesian (DR Bayes) inference methods using sample-split: trimming based on $\hat{\pi} \in [t, 1-t]$, $\bar{n}$ = the average sample size after trimming.

Spec	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
I	$t = 0.10(\bar{n} = 124)$			$t = 0.05(\bar{n} = 185)$			$t = 0.01(\bar{n} = 340)$		
	-0.009	0.969	0.332	0.020	0.965	0.324	0.020	0.932	0.389
II	$t = 0.10(\bar{n} = 118)$			$t = 0.05(\bar{n} = 177)$			$t = 0.01(\bar{n} = 317)$		
	-0.009	0.938	0.333	0.024	0.924	0.330	0.025	0.898	0.416
III	$t = 0.10(\bar{n} = 115)$			$t = 0.05(\bar{n} = 172)$			$t = 0.01(\bar{n} = 331)$		
	-0.013	0.943	0.350	0.012	0.926	0.340	0.028	0.892	0.427