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

Christoph Breunig* Ruixuan Liu† Zhengfei Yu‡

October 5, 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.

C Auxiliary Results

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

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

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

$$\mathrm{d}\nu(x,d,y) = (\pi_0(x))^d(1-\pi_0(x))^{1-d}\mathrm{d}\vartheta(d,y)\mathrm{d}F_0(x), \tag{C.1}$$

*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

where ϑ stands for the counting measure on $\{\{0,0\},\{0,1\},\{1,0\},\{1,1\}\}$. 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). Recall the notation $\rho^m(y,d,x) = y - m(d,x)$ used below.

Lemma C.1. *Let Assumption 1 be satisfied and $m_\eta = \Psi(\eta^m)$, then we have uniformly for $\eta^m \in \mathcal{H}_n^m$:*

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

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

Proof. The logistic distribution function Ψ satisfies $\Psi' = \Psi(1-\Psi)$ and $\Psi^{(2)} = \Psi(1-\Psi)(1-2\Psi)$. Recall the perturbation of η^m along the least favorable direction in (A.2) given by $\eta_t^m = \eta^m - t\xi_0^m/\sqrt{n}$ for $t \in \mathbb{R}$. Thus, $\log p_{\eta^m} - \log p_{\eta_t^m} = g(0) - g(1)$, where $g(u) = \log p_{\eta_u^m}$ for $u \in [0,1]$, as introduced at the beginning of Section B. We examine the following Taylor expansion uniformly for $\eta^m \in \mathcal{H}_n^m$:

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

where $\theta \leqslant \|g^{(3)}\|_\infty$.

We express the part of the log-likelihood involving η^m explicitly as follows.

$$\begin{aligned}
\log p_{\eta^m}(z) =& dy \log \frac{e^{\eta^m(1,x)}}{1+e^{\eta^m(1,x)}} + d(1-y)\log\frac{1}{1+e^{\eta^m(1,x)}} \\
&+ (1-d)y\log\frac{e^{\eta^m(0,x)}}{1+e^{\eta^m(0,x)}} + (1-d)(1-y)\log\frac{1}{1+e^{\eta^m(0,x)}} \\
=& d\left(y\eta^m(1,x) - \log(1+e^{\eta^m(1,x)})\right) + (1-d)\left(y\eta^m(0,x) - \log(1+e^{\eta^m(0,x)})\right).
\end{aligned} \tag{C.3}$$

Given equation (C.2), it remains to calculate the first three derivatives of the function g. Its first derivative is given by

$$g'(u) = -\frac{t}{\sqrt{n}}\gamma_0 \rho^{\Psi(\eta_u^m)},$$

where $\gamma_0\rho^{\Psi(\eta_u^m)}(y,d,x) = y - \Psi(\eta_u^m(d,x))$. The second and third derivative of g 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 γ_0, 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,

$$g'(0) = -\frac{t}{\sqrt{n}}\gamma_0\rho^{m_0} + \frac{t}{\sqrt{n}}\gamma_0(m_\eta - m_0), \tag{C.4}$$

and

$$g^{(2)}(0) = -\frac{t^2}{n}\gamma_0^2 m_\eta(1-m_\eta). \tag{C.5}$$

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*

$$\sqrt{n}\mathbb{P}_n[(\widehat{\gamma} - \gamma_0)\rho^{m_0}] = o_{P_0}(1). \tag{C.6}$$

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. Using the notation $\rho^{m_0}(Z_i) = Y_i - m_0(D_i, X_i)$, we evaluate for the conditional expectation that

$$\begin{aligned}
&\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\gamma_n-\gamma_0)(D_i,X_i)\rho^{m_0}(Z_i)\Big)^2 \mid (D_1,X_1),\ldots,(D_n,X_n)\right] \\
&= \frac{1}{n}\sum_{i\neq i'}(\gamma_n-\gamma_0)(D_i,X_i)(\gamma_n-\gamma_0)(D_{i'},X_{i'})\mathbb{E}_0\left[\rho^{m_0}(Z_i)\rho^{m_0}(Z_{i'}) \mid (D_i,X_i),(D_{i'},X_{i'})\right] \\
&= \frac{1}{n}\sum_{i=1}^n(\gamma_n-\gamma_0)^2(D_i,X_i)Var_0(Y_i|X_i).
\end{aligned}$$

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

$$\mathbb{E}_0\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\gamma_n-\gamma_0)(D_i,X_i)\rho^{m_0}(Z_i)\Big)^2\right] \lesssim \|\pi_n - \pi_0\|_{2,F_0}^2 = o(1)$$

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 introduce a class of functions $\mathcal{B}_n^m$ which is shown to contain the Gaussian process W which sufficiently large probability, and is given by

$$\mathcal{B}_n^m := \varepsilon_n \mathbb{B}_1^{s_m,p} + M_n \mathbb{H}_1^{a_n}, \tag{C.7}$$

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

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

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 $m(d,x) = \Psi\left(W^m(d,x) + \lambda\, \gamma_n(d,x)\right)$ where $W^m(d,\cdot)$ 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, Ghosh, and van der Vaart (2000) to obtain the proper concentration rate for the posterior for the rescaled

squared exponential process:

$$\text{I. } \Pi((w,\lambda) : K \vee V(p_{\eta_0^m}, p_{w+\lambda\gamma_n}) \leqslant \varepsilon_n^2) \geqslant c_1 \exp(-c_2 n\varepsilon_n^2), \tag{C.8}$$

$$\text{II. } \Pi(\mathcal{P}_n^c) \leqslant \exp(-c_3 n\varepsilon_n^2), \tag{C.9}$$

$$\text{III. } \log N(\varepsilon_n, \mathcal{P}_n, \|\cdot\|_{L^2(\nu)}) \leqslant c_4 n\varepsilon_n^2, \tag{C.10}$$

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.14) 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.16), 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.1). Thus, we have

$$N(\varepsilon_n, \mathcal{P}_n, \|\cdot\|_{L^2(\nu)}) \leqslant N(\varepsilon_n/2, \mathcal{B}_n^m, \|\cdot\|_\infty) \times N(c\varepsilon_n, [0, 2M\sigma_n\sqrt{n}\varepsilon_n], |\cdot|) \lesssim n\varepsilon_n^2. \quad \text{(C.11)}$$

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),

$$\mathbb{E} \sup_h \left| \sum_{i=1}^n (\varphi(Z_i) h(Z_i) - P_0[\varphi h]) \right| \leqslant 2\|\varphi\|_\infty \mathbb{E} \sup_h \left| \sum_{i=1}^n \epsilon_i \frac{\varphi(Z_i)}{\|\varphi\|_\infty} h(Z_i) \right|. \tag{C.12}$$

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.12) 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). We introduce some simplifying notations by writing

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

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

$$\begin{aligned}\|p_{v^m} - p_{w^m}\|_{L^2(\nu)} &\leqslant \|\Psi(v^{m^1}) - \Psi(w^{m^1})\|_{L^2(F_0\pi_0)} \vee \|\Psi(v^{m^0}) - \Psi(w^{m^0})\|_{L^2(F_0(1-\pi_0))} \\ &\leqslant \|v^{m^1} - w^{m^1}\|_{2,F_0} \vee \|v^{m^0} - w^{m^0}\|_{2,F_0}. \end{aligned} \tag{C.13}$$

In addition, it holds that

$$K(p_{v^m}, p_{w^m}) \vee V(p_{v^m}, p_{w^m}) \leqslant \|v^{m^1} - w^{m^1}\|_{2,F_0}^2 \vee \|v^{m^0} - w^{m^0}\|_{2,F_0}^2. \tag{C.14}$$

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*

$$\log N(3\varepsilon_n, \mathcal{B}_n^m, \|\cdot\|_\infty) \lesssim n\varepsilon_n^2. \tag{C.15}$$

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

$$\phi_0^{a_n}(\varepsilon_n) \lesssim n\varepsilon_n^2, \tag{C.16}$$

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:

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

which is equivalent to the following lemma.

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

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

Proof. We start with the following identity:

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

where

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

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

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).

We consider

$$\mathcal{H}_n^m := \{w_d + \lambda\gamma_n : (w_d, \lambda) \in \mathcal{W}_n\}, \tag{C.17}$$

where

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

where the sieve space $\mathcal{B}_n^m$ in the first restriction for the Gaussian process W_d is defined in the equation (C.7) with $d \in \{0, 1\}$, and $\varepsilon_n = (n/\log n)^{-s_m/(2s_m+p)}$.

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

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

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

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

which follows from the exponential family assumption.

In this case, there is a one-to-one correspondence between the conditional density function and the conditional mean function of the outcome given covariates. 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. □

We emphasize that the least favorable direction calculation in the multinomial outcome case 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.

Proof of Lemma 6.2. Consider the log transformation of the joint density of $Z = (Y, D, X^\top)^\top$ given by

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

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. 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

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

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

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

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:

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

Given that $\mathbb{1}_{\{y=0\}} = 1 - \sum_{j=1}^{J} \mathbb{1}_{\{y=j\}}$, the previous equation simplifies to

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

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 aligns with the requirement of the score operator.

From our verification of differentiability, we infer 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

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

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

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

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

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

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:

$$\iint y\frac{\partial}{\partial t}p_t(y|1,x)\Big|_{t=0}\mathrm{d}\nu(y)f_\eta(x)\mathrm{d}\mu(x)+\iint yp_\eta(y|1,x)\mathrm{d}\nu(y)\frac{\partial}{\partial t}f_t(x)\Big|_{t=0}\mathrm{d}\mu(x). \tag{D.2}$$

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

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

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

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

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.2), 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

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

Referring to the second part of (D.2), we immediately obtain

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

Similarly for the control arm, we derive

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

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

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

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:

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

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

The efficient influence function for estimation of the policy effect parameter τ_η^P is given by

$$\widetilde{\tau}_\eta^P(z) = \gamma_\eta^P(x)(y - m_\eta(x))$$

where $\gamma_\eta^P(x) = \frac{g_1(x) - g_0(x)}{f(x)}$. Now the score operator B_η given in (E.3) applied to $\xi_\eta^P(x) = \big(\gamma_\eta^P(x), 0\big)$, yields $B_\eta \xi_\eta^P = \widetilde{\tau}_\eta^P$. Thus, ξ_η^P 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,

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

for some known functions $c(\cdot)$, $q(\cdot)$, and $A(\cdot)$. We reparameterize the model using the link function q: We consider the reparametrization using the link function q:

$$\eta^m(d, x) = q(m(d, x))$$

and we define the mapping $\Upsilon(\cdot) := A \circ q^{-1}(\cdot)$, used below. Because the generalization of the binary outcome case to the above exponantial family involves some change of the likelihood function related to the conditional mean, we outline the necessary modifications.

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 $\Upsilon = A \circ q^{-1}$ *is three time differentiable with* $\|\Upsilon^{(\ell)}\|_\infty < \infty$ *for* $\ell = 2, 3$. *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* $m(d,x) = q^{-1}\left(W^m(d,x) + \lambda\,\widehat{\gamma}(d,x)\right)$, *where* $W^m(d,x)$ *is the rescaled squared exponential process for* $d \in \{0,1\}$, *with its rescaling parameter* a_n *of the order in* (4.1), $(n/\log n)^{-s_m/(2s_m+p)} \lesssim u_n\sigma_n$ *for some deterministic sequence* $u_n \to 0$, *and* $\sigma_n \lesssim 1$. *Then, the posterior distribution satisfies Theorem 3.1.*

Proof. A close inspection shows that there are mainly three parts that we need to adapt the argument due to the change of p_{η^m} in the likelihood function. The first one is about the Kullback-Leibler (KL) divergence and related metrics in showing the posterior contraction rate. The second one concerns the local asymptotic normality (LAN) expansion used in the conditional Laplace transform, where we show the connection of its second-order term to part of the variance of the influence function. Finally, we make use of the imposed smoothness assumption on Υ to show the negligibility of third order terms, which is needed for verifying the prior stability. We proceed in three steps.

Step 1. First, in deriving the posterior contraction rate or determining the proper localized set $\mathcal{H}_n^m$, we need proper upper bounds for the Hellinger distance and Kullback-Leibler (KL) divergence between two probability density functions (p_{η^m}, p_{v^m}) by the L^2 distance of the reparametrized functions (η^m, v^m). Recall that $p_{\eta^m}(y,d,x) = c(y)\exp\left[q(m(d,x))ay - A(m(d,x))\right]$. To abuse the notation a bit, we denote the corresponding probability densities by p_{η^m} and p_{v^m}. From the proof of Lemma 6.1 we observe

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

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

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

The conditional variance formula also shows the convexity of $\Upsilon(\cdot)$; see Brown (1986).

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

compute

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

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

$$\begin{aligned}
K(p_{\eta^m}, p_{v^m}) &= \int (\Upsilon'(\eta^m(d,x))(\eta^m(d,x) - v^m(d,x)) - [\Upsilon(\eta^m(d,x)) - \Upsilon(v^m(d,x))]) \\
&\qquad \times \pi^d(x)(1-\pi(x))^{1-d} \mathrm{d}\vartheta(d) \mathrm{d}F_\eta(x) \\
&= \int \Upsilon^{(2)}(\tilde{\eta}^m(d,x))(\eta^m(d,x) - v^m(d,x))^2 \pi^d(x)(1-\pi(x))^{1-d} \mathrm{d}\vartheta(d) \mathrm{d}F_\eta(x) \\
&\lesssim \|v^{m^1} - \eta^{m^1}\|^2_{2,F_\eta} \vee \|v^{m^0} - \eta^{m^0}\|^2_{2,F_\eta},
\end{aligned}$$

where the last inequality follows from the condition $\|\Upsilon^{(2)}\|_\infty < \infty$. Recall that

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

Therefore, we continue with the right hand side inequality of (F.2) and calculate

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

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

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

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

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

Because the posterior contraction holds in terms of the Hellinger distance under general conditions, the above upper bound allows us to translate this contraction to the corresponding L_2 distance for the η^m function.

Step 2. We examine the changes to the LAN expansion in Lemma B.1 as follows. For this purpose, we use the notation $g(u) = \log p_{\eta_u^m}$ for $u \in [0,1]$, as introduced at the beginning of Section B. Specifically, in the one-parameter exponential family case, we have

$$\log p_{\eta_u^m}(z) = y\eta_u^m(d,x) - \Upsilon(\eta_u^m(d,x)) + \log c(y).$$

By the definition of $\Upsilon(\cdot)$, we know that it is a convex function, given that $\Upsilon^{(2)}(\eta^m(d,x)) = Var_\eta(Y|D=d, X=x)$. Thereafter, we can obtain the first to third order derivatives of g as

$$\begin{aligned} g'(0) &= \frac{t}{\sqrt{n}}\gamma_0 \rho^{\Upsilon'(\eta^m)} = \frac{t}{\sqrt{n}}\gamma_0 \rho^{m_\eta}, \\ g^{(2)}(0) &= \frac{t^2}{n}\gamma_0^2 \Upsilon^{(2)}(\eta^m), \quad g^{(3)}(\tilde{u}) = \frac{t^3}{n^{3/2}}\gamma_0^3 \Upsilon^{(3)}(\eta_{\tilde{u}}^m), \end{aligned}$$

where $\tilde{u}$ is some intermediate value between 0 and 1.

Following the lines of the proof of Lemma B.1, the second moment $P_0 g^{(2)}(0)$ must be derived for the exponential family case. Based on the previous calculations and the posterior convergence of η^m, it can be expressed as

$$\begin{aligned} nP_0 g^{(2)}(0) =& t^2 \mathbb{E}_0[\gamma_0^2(D,X)\Upsilon^{(2)}(\eta_0^m(D,X))] + o_{P_0}(1) = t^2 \mathbb{E}_0[\gamma_0^2(D,X) Var_0(Y|D,X)] + o_{P_0}(1) \\ =& t^2 \mathbb{E}_0[\gamma_0^2(D,X)(Y - m_0(D,X))] + o_{P_0}(1) = t^2 P_0(B_0^m(\xi_0^m))^2 + o_{P_0}(1), \end{aligned}$$

where the score operator $B_0^m = B_{\eta_0}^m$ is given in the proof of Lemma 6.1.

Step 3. Finally, we need to establish the following expansion in a key step to show the

prior stability condition:

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

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

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

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:

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

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 Posterior Computation in Algorithm 1

We describe the Laplace approximation method used in Algorithm 1 (Posterior computation, Step (a)) for drawing the posterior of η^m and thus m_η; see Rassmusen and Williams (2006, Chapters 3.3 to 3.5) for more details on the Laplace approximation. 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 $\boldsymbol{\eta}^*$ a $2n$-vector that gives the latent function η^m evaluated at $\boldsymbol{W}^*$:

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

Let $\boldsymbol{\eta} = \left[\eta^m(D_1, X_1), \ldots, \eta^m(D_n, X_n)\right]^\top$ 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.[1] In sum, we get the posterior draws of the vectors $[\eta^m(1, X_1), \ldots, \eta^m(1, X_n)]^{\top}$ and $[\eta^m(0, X_1), \ldots, \eta^m(0, X_n)]^{\top}$ 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 $d \in \{0, 1\}$.

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 $\sigma_n = c_\sigma \times \log n / (\sqrt{n}\, \widehat{\Gamma})$ with $c_\sigma \in \{1/5, 1/2, 1, 2, 5\}$. Note that $c_\sigma = 1$ corresponds to the simulation results reported in the main text. The performance of DR Bayes appears stable with respect to the choice of c_σ. The performance of PA Bayes, on the other hand, deteriorates when σ_n takes relatively small or large values, such as the cases with $c_\sigma = 1/5$, $t = 0.10$ and $c_\sigma = 5$, $t = 0.01$.

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

Table A1: The effect of c_σ on adjusted Bayesian inference methods: trimming based on $\widehat{\pi} \in [t, 1-t]$, $\bar{n}$ = the average sample size after trimming. CP = coverage probability, CIL = average length of the 95% credible/confidence interval.

c_σ	Methods	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
		$t = 0.10(\bar{n} = 240)$			$t = 0.05(\bar{n} = 363)$			$t = 0.01(\bar{n} = 664)$		
1/5	PA Bayes	-0.036	0.794	0.169	-0.003	0.937	0.176	0.009	0.992	0.220
	DR Bayes	-0.038	0.919	0.193	-0.006	0.961	0.190	0.001	0.989	0.214
1/2	PA Bayes	-0.024	0.962	0.215	0.016	0.968	0.222	0.032	0.968	0.289
	DR Bayes	-0.031	0.976	0.207	0.005	0.971	0.207	0.014	0.985	0.248
1	PA Bayes	-0.008	0.981	0.260	0.033	0.949	0.254	0.047	0.897	0.308
	DR Bayes	-0.024	0.983	0.223	0.014	0.970	0.221	0.023	0.952	0.258
2	PA Bayes	0.005	0.979	0.294	0.043	0.933	0.270	0.055	0.848	0.312
	DR Bayes	-0.018	0.980	0.236	0.019	0.961	0.229	0.028	0.922	0.260
5	PA Bayes	0.013	0.971	0.311	0.047	0.928	0.276	0.058	0.836	0.313
	DR Bayes	-0.015	0.980	0.242	0.022	0.958	0.232	0.030	0.907	0.261

Table A2 reports the finite sample performance of PA and DR Bayes using sample-split and compares it to the results in Table 1 that uses the full sample twice. Sample-split uses 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 $\bar{n}$ 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: Adjusted Bayesian inference methods using sample-split: trimming based on $\widehat{\pi} \in [t, 1-t]$, $\bar{n}$ = the average sample size after trimming. CP = coverage probability, CIL = average length of the 95% credible interval.

	Bias	CP	CIL	Bias	CP	CIL	Bias	CP	CIL
Sample-split	$t = 0.10(\bar{n} = 124)$			$t = 0.05(\bar{n} = 185)$			$t = 0.01(\bar{n} = 339)$		
PA Bayes	-0.024	0.986	0.339	0.009	0.977	0.342	0.017	0.950	0.410
DR Bayes	-0.013	0.968	0.321	0.017	0.962	0.317	0.016	0.934	0.385
Full sample	$t = 0.10(\bar{n} = 240)$			$t = 0.05(\bar{n} = 363)$			$t = 0.01(\bar{n} = 664)$		
PA Bayes	-0.008	0.981	0.260	0.033	0.949	0.254	0.047	0.897	0.308
DR Bayes	-0.024	0.983	0.223	0.014	0.970	0.221	0.023	0.952	0.258

References

Brown, L. D. (1986): *Fundamentals of statistical exponential families: with applications in statistical decision theory*. IMS.

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

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

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

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

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

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

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

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

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

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

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