The Annals of Statistics
2020, Vol. 48, No. 5, 2999–3020
https://doi.org/10.1214/19-AOS1919

SEMIPARAMETRIC BAYESIAN CAUSAL INFERENCE

BY KOLYAN RAY[1] AND AAD VAN DER VAART[2]

[1]*Department of Mathematics, Imperial College London, kolyan.ray@imperial.ac.uk*
[2]*Mathematical Institute, Leiden University, avdvaart@math.leidenuniv.nl*

We develop a semiparametric Bayesian approach for estimating the mean response in a missing data model with binary outcomes and a nonparametrically modelled propensity score. Equivalently, we estimate the causal effect of a treatment, correcting nonparametrically for confounding. We show that standard Gaussian process priors satisfy a semiparametric Bernstein–von Mises theorem under smoothness conditions. We further propose a novel propensity score-dependent prior that provides efficient inference under strictly weaker conditions. We also show that it is theoretically preferable to model the covariate distribution with a Dirichlet process or Bayesian bootstrap, rather than modelling its density.

1. Introduction. In many applications, one wishes to make inference concerning the causal effect of a treatment or condition. Examples include healthcare and assessing the impact of public policies amongst many others. The available data are often observational rather than the result of a carefully planned experiment or trial. The notion of "causal" then needs to be carefully defined and the statistical analysis must take into account other possible explanations for the observed outcomes.

A common framework for causal inference is the potential outcome setup [23, 32]. In this framework, every individual possesses two "potential outcomes", corresponding to the individual's outcomes with and without treatment. The treatment effect, which we wish to estimate, is thus the difference between these two potential outcomes. Since we only observe one out of each pair of outcomes, and not the corresponding "counterfactual" outcome, we do not directly observe samples of the treatment effect. Because in practice, particularly in observational studies, individuals are assigned treatments in a biased manner, a simple comparison of actual cases (i.e., treated individuals) and controls may be misleading due to selection bias. A typical way to overcome this is to gather the values of covariate variables that influence both outcome and treatment assignment ("confounders") and apply a correction based on the "propensity score", which is the conditional probability that a subject is treated as a function of the covariate values. Under the assumption that outcome and treatment assignment are independent given the covariates, the causal effect of treatment can be identified from the data. Popular estimation methods include "propensity score matching" [38, 40] and "double robust methods" [32, 37, 39, 41]. In this paper, we follow the approach of nonparametrically modelling the propensity score function and posing the estimation of the treatment effect as a problem of estimation of a functional on a semiparametric model [6, 43, 48]. Our methodological novelty is to follow a semiparametric Bayesian approach, putting nonparametric priors on the propensity score and/or on the unknown response function and the covariate distribution, possibly incorporating an initial estimator of the first function.

For notational simplicity, we in fact consider the missing data model which is mathematically equivalent to observing one arm of the causal setup. The model is also standard and

Received August 2018; revised August 2019.
MSC2020 subject classifications. Primary 62G20; secondary 62G15, 62G08.
Key words and phrases. Bernstein–von Mises, Gaussian processes, propensity score-dependent priors, causal inference, Dirichlet process.

widely studied on its own in biostatistical applications, where response variables are frequently missing, and is a template for a number of other models [33, 44]. For a recent review on estimating an average treatment effect over a (sub)population, a problem that has received considerable attention in the econometrics, statistics and epidemiology literatures; see Athey et al. [4].

Suppose that we observe n i.i.d. copies $X_1, \ldots, X_n$ of a random variable $X = (Z, R, RY)$, where R and Y take values in the two-point set $\{0, 1\}$ and are conditionally independent given Z. We think of Y as the outcome of a treatment and are interested in estimating its expected value $\mathbb{E}Y$. The problem is that the outcome Y is observed only if the indicator variable R takes value 1, as otherwise the third component of X is equal to 0. Whether the outcome is observed or not may well be dependent on its value, which precludes taking a simple average of the observed outcomes as an estimator for $\mathbb{E}Y$. The covariate Z is collected to correct for this problem; it is assumed to contain exactly the information that explains why the response Y is not observed except for purely random causes, so that the outcome Y and missingness indicator R are conditionally independent given Z, that is, the outcomes are *missing at random* (relative to Z).

The connection to causal inference is that we may think of Y as a "counterfactual" outcome if a treatment were assigned ($R = 1$) and its mean as "half" the treatment effect under the assumption of unconfoundedness. More precisely, if Y^1 and Y^0 denote the potential outcomes when treated or not treated, then in the causal model one would observe $(Z, R, Y^1R, Y^0(1-R))$ and be interested in estimating $\mathbb{E}Y^1 - \mathbb{E}Y^0$ under the assumption that Y^0, Y^1 are conditionally independent of R given Z. One can think of the missing data problem as simplifying this to observing (Z, R, Y^1R) and estimating $\mathbb{E}Y^1$. To estimate the causal effect, one could apply the missing data problem a second time, to the data $(Z, R, Y^0(1-R))$, to estimate $\mathbb{E}Y^0$, or do a simultaneous analysis. In the nonparametric setup, there will be no essential difference between the two.

The model for a single observation X can be described by the distribution of Z and the two conditional distributions of Y and R given Z. In this paper, we model these three components nonparametrically. We investigate a Bayesian approach, putting a nonparametric prior on the three components, in particular Gaussian process and Dirichlet process priors. We then consider the mean response $\mathbb{E}Y$ as a functional of the three components and study the induced marginal posterior distribution of $\mathbb{E}Y$ from a frequentist perspective. The aim is to derive conditions under which this marginal posterior distribution satisfies a Bernstein–von Mises theorem in the semiparametric sense, thus yielding recovery of the mean response at a $\sqrt{n}$-rate and asymptotic efficiency in the semiparametric sense.

In recent years, Bayesian approaches have become increasingly popular due to their excellent empirical performance for such problems [1, 2, 15, 19–22, 42, 54]. However, despite their increasing use in practice, there have been few corresponding theoretical results. Indeed, early work on semiparametric Bayesian approaches to this specific missing data problem produced negative results, proving that many common classes of priors, or more generally likelihood-based procedures, produce inconsistent estimates assuming no smoothness on the underlying parameters; see the results and discussion in [30, 36]. We attempt to shed light on this apparent gap between the excellent empirical performance observed in practice and the potentially disastrous theoretical performance.

The structured nature of the model, with three parameters (response function, propensity score and covariate distribution), requires careful consideration of prior distributions. As the likelihood factorizes over the three parameters, choosing these a priori independent will lead to a product posterior. We show that this can lead to efficient estimation of $\mathbb{E}Y$, but only under unnecessarily harsh smoothness requirements on the parameters. This is in agreement with the discussion in [30, 36], which applies to likelihood-based methods in general, including

semiparametric maximum likelihood [24]. Within our Bayesian setup, it is possible to correct this (partly) by modelling the response function and propensity score as a priori dependent, thus allowing the components to share information, despite the factorisation in the likelihood. In particular, we propose a novel Gaussian process prior that incorporates an estimate of the propensity score function, and show that it performs efficiently under strictly weaker conditions than for standard product priors (see [27] for an empirical investigation). Unlike for these latter priors, extra regularity of the binary regression function can compensate for low regularity of the propensity score, that is one direction of so-called "double robustness" [8, 39]. A related construction using Bayesian additive regression trees (BART) has been shown to work well empirically [20]. It can thus be both practically and theoretically advantageous to employ propensity score-dependent priors.

For the estimation of $\mathbb{E}Y$ at $\sqrt{n}$-rate, smoothness of the distribution of the covariate Z is not needed. In our main result, we therefore model this distribution by the standard nonparametric prior for a distribution: the Dirichlet process. In our concrete examples, the prior modelling thus consists of a combination of Gaussian and Dirichlet processes. In the Supplementary Material [28], we also consider modelling the covariate density, for instance by an exponentiated Gaussian process. Our result seems to indicate that even when the smoothness of the density is modelled correctly, this approach can induce a nonvanishing bias in the posterior distribution of $\mathbb{E}Y$, an effect that becomes more pronounced with increasing covariate dimension.

The papers [33, 35] consider estimation of $\mathbb{E}Y$ under minimal smoothness conditions on the parameters. Using estimating equations, the authors construct estimators that attain an optimal rate of convergence slower than $\sqrt{n}$ in cases where the component parameters have low smoothness. Furthermore, they construct estimators that attain a $\sqrt{n}$-rate under minimal smoothness conditions, less stringent than in earlier literature, using higher order estimating equations. It is unclear whether similar results can be obtained using a Bayesian approach. The constructions in the present paper can be compared to the estimators obtainable for linear (or first order) estimating equations. It remains to be seen whether Bayesian modelling is capable of performing the bias corrections necessary to handle true parameters of low smoothness levels in a similar manner as higher order estimating equations.

For smooth parametric models, the theoretical justification for posterior based inference is provided by the Bernstein–von Mises theorem or property (hereafter BvM). This property says that as the number of observations increases, the posterior distribution is approximately a Gaussian distribution centered at an efficient estimator of the true parameter and with covariance equal to the inverse Fisher information; see Chapter 10 of [48]. While such a result does not hold in full generality in infinite dimensions [14], semiparametric analogues can establish the BvM property for the marginal posterior of a finite-dimensional parameter in the presence of an infinite-dimensional nuisance parameter [7, 10, 11, 31]. In such cases, care is required in the choice of prior assigned to the nonparametric part, as oversmoothing may induce a bias in the posterior distribution of the finite-dimensional parameter.

Our main results are two theorems for general priors on the response function and/or propensity score, followed by corollaries for Gaussian process priors. In both cases, we combine these with a Dirichlet process prior on the covariate distribution. While the first theorem is in the spirit of earlier work, it is novel in its extension to a structured semiparametric model and its combination with the Dirichlet process, in both a modelling and a technical sense. The second theorem is innovative in its investigation of "half of double-robustness", as indicated in the preceding, and by showing that incorporating a prior perturbation in the least favourable direction can remove potential bias from the posterior. The latter device takes care of the usual "prior invariance condition" and has consequences beyond the model in this paper. The corollaries for Gaussian process priors illustrate the conditions of the main results,

and give concrete examples of inference. In the Supplementary Material [28], we present a third theorem, which covers the case that the covariate density, rather than the distribution, is modelled, which is again illustrated by Gaussian process priors.

An important consequence of the semiparametric BvM is that credible sets for the functional are asymptotically confidence regions with the same coverage level. The Bayesian approach thus automatically provides access to uncertainty quantification once one can sample from the posterior distribution. Obtaining confidence statements for average treatment effects is a current area of research and there has been recent progress in this direction, for example using random forests and regression trees [3, 53]. Our results show that Bayesian approaches can also yield valid frequentist uncertainty quantification in this setting.

The paper is structured as follows. In Section 2, we provide a review of the model, including the relevant semiparametric theory. Section 3 contains the two main theorems and their corollaries, with discussion in Section 4 and the main proofs in Section 5. The remaining sections are given in the Supplementary Material [28]. Section 6 gives the third theorem, with a joint prior on the propensity score, response function and covariate density. Technical results, auxiliary results and posterior contraction results are deferred to Sections 7, 8 and 9, respectively.

1.1. *Notation.* The notation $\lesssim$ denotes inequality up to a multiplicative constant that is fixed throughout and $\lfloor x \rfloor$ is the largest integer strictly smaller than x. The symbol Ψ is used for the logistic function given by $\Psi(x) = 1/(1+e^{-x})$. We abbreviate $\int f\,dP$ by Pf. For probability densities f and g with respect to some dominating measure ν, $h(f,g) = (\int (f^{1/2} - g^{1/2})^2 d\nu)^{1/2}$ is the Hellinger distance, $K(f,g)$ is the Kullback–Leibler divergence and $V(f,g) = \int (\log(f/g))^2\,dF$. We denote by $H^s = H^s([0,1]^d)$ and $C^s = C^s([0,1]^d)$ the L^2-Sobolev and Hölder spaces, respectively. For i.i.d. random variables $X_1, \ldots, X_n$ with common law P the notation $\mathbb{P}_n[h] = n^{-1}\sum_{i=1}^n h(X_i)$ and $\mathbb{G}_n[h] = \sqrt{n}(\mathbb{P}_n - Ph)$ are the empirical measure and process, respectively. The notation $\mathcal{L}(Z)$ denotes the law of a random element Z. We often drop the index n in the product measure P_η^n, writing P_η, and write P_0 instead of P_{η_0}, where η_0 is the true parameter for the data generating distribution. The ε-covering number of a set Θ for a semimetric d, denoted $N(\Theta, d, \varepsilon)$, is the minimal number of d-balls of radius ε needed to cover Θ and $N_{[]}(\Theta, d, \varepsilon)$ is the minimal number of brackets of size ε needed to cover a set of functions Θ.

2. Model details. Recall that we observe i.i.d. copies $X_1, \ldots, X_n$ of a random variable $X = (Z, R, RY)$, where R and Y take values in the two-point set $\{0,1\}$ and are conditionally independent given Z, which itself takes values in a given measurable space $\mathcal{Z}$. Denote the full sample by $X^{(n)} = (X_1, \ldots, X_n)$. This model can be parameterized via the marginal distribution F of Z and the conditional probabilities $a(z)^{-1} = P(R = 1|Z = z)$, called the *propensity score*, and $b(z) = P(Y = 1|Z = z)$, the regression of Y on Z. The distribution of an observation X is thus fully described by the triple (a, b, F). If F has a density f, then we may also use the triple (a, b, f).

For prior construction it will be useful to transform the parameters by a link function. Most smooth maps from $\mathbb{R}$ to $(0,1)$ may be used, but for definiteness we choose the logistic function $\Psi(t) = 1/(1+e^{-t})$, and consider the reparametrization

$$\eta^a = \Psi^{-1}(1/a), \qquad \eta^b = \Psi^{-1}(b), \tag{2.1}$$

and write $\eta = (\eta^a, \eta^b)$. If a density f of Z exists, then we define in addition

$$\eta^f = \log f,$$

and write by a slight abuse of notation $\eta = (\eta^a, \eta^b, \eta^f)$.

The density $p_{(a,b,f)} = p_\eta$ of X can now be given as

$$p_\eta(x) = \left(\frac{1}{a(z)}\right)^r \left(1 - \frac{1}{a(z)}\right)^{1-r} b(z)^{ry}(1 - b(z))^{r(1-y)} f(z). \tag{2.2}$$

Note that this factorizes over the parameters. If the covariate is not assumed to have a density and $\eta = (\eta^a, \eta^b)$, we use the same notation p_η, but then the factor $f(z)$ is understood to be 1, and the expression is the conditional density of (R, RY) given $Z = z$. Since p_η factorizes over the three (or two) parameters, the log-likelihood based on $X^{(n)}$ separates as

$$\ell_n(\eta) = \sum_{i=1}^n \log p_{(a,b,f)}(X_i) = \ell_n^a(\eta^a) + \ell_n^b(\eta^b) + \ell_n^f(\eta^f), \tag{2.3}$$

where each term is the logarithm of the factors involving only a or b or f, and $\ell_n^f(\eta^f)$ is understood to be absent when existence of a density f is not assumed. The functional of interest is the *mean response* $\mathbb{E}_\eta Y = \mathbb{E}_\eta b(Z)$, which can be expressed in the parameters as

$$\chi(\eta) = \int b\, dF = \int \Psi(\eta^b)(z) e^{\eta^f(z)}\, dz,$$

where the second representation is available if F has a density.

Estimators that are $\sqrt{n}$-consistent and asymptotically efficient for $\chi(\eta)$ have been constructed using various methods, but only if a or b (or both) are sufficiently smooth. In the present context, under the assumption that $a \in C^\alpha$ and $b \in C^\beta$, Robins et al. [35] have constructed estimators that are $\sqrt{n}$-consistent if $(\alpha + \beta)/2 \geq d/4$, where d is the dimension of the covariates. They have also shown that the latter condition is sharp: the minimax rate becomes slower than $1/\sqrt{n}$ when $(\alpha + \beta)/2 < d/4$ (see [34]). The estimators in [35] employ higher order estimating equations to obtain better control of the bias. First-order estimators, based on linear estimators or semiparametric maximum likelihood, have been shown to be $\sqrt{n}$-consistent only under the stronger condition

$$\frac{\alpha}{2\alpha + d} + \frac{\beta}{2\beta + d} \geq \frac{1}{2}; \tag{2.4}$$

see, for example, [37, 39]. In both cases, the conditions show a trade-off between the smoothness levels of a or b: higher α permits lower β and vice versa. This trade-off results from the multiplicative form of the bias of linear or higher-order estimators. So-called *double robust* estimators are able to exploit this structure, and work well if either a or b is sufficiently smooth. (More generally, it suffices that the parameters a and b can be estimated well enough, where the combined rates are relevant. The inequalities even remain valid with $\alpha = 0$ or $\beta = 0$ interpreted as the existence of $\sqrt{n}$-consistent estimators of a or b, as would be the case given a correctly specified finite-dimensional model.) We shall henceforth also assume that the parameters a and b are contained in Hölder spaces C^α and C^β, respectively. See [41] for a recent discussion of double robustness.

For estimation of $\mathbb{E}Y$ at $\sqrt{n}$-rate the covariate density f need not be smooth, which makes sense intuitively, as the functional can be written as an integral relative to the corresponding distribution F. (Counter to this intuition [34, 35] show this to be false for optimal estimation at slower than $\sqrt{n}$-rate.) This may motivate modelling F nonparametrically, in the Bayesian setting for instance with a Dirichlet process prior.

All these observations are valid only if the estimation problem is not affected by the parameters a, b or f taking values on the boundary of their natural ranges. For simplicity, we make the following assumption throughout.

ASSUMPTION. The true functions $1/a_0$ and b_0 are bounded away from 0 and 1 and f_0 is bounded away from 0 and ∞.

2.1. *Semiparametric information and least favourable direction.* We finish by reviewing the tangent space and information distance of the model, which is well known to play an important role in semiparametric estimation theory [5, 6, 43], and enters the Bayesian derivations through the "least favourable submodel". (See [10] or Chapter 12 of [16] for general reviews in the context of Bayesian estimation.)

With regards to the parametrization (2.1), consider the one-dimensional submodels $t \mapsto \eta_t$ induced by the paths

$$\frac{1}{a_t} = \Psi\big(\eta^a + t\mathfrak{a}\big), \qquad b_t = \Psi\big(\eta^b + t\mathfrak{b}\big), \qquad dF_t = dF e^{t\mathfrak{f}} \bigg(\int e^{t\mathfrak{f}}\, dF \bigg)^{-1}$$

for given directions $(\mathfrak{a}, \mathfrak{b}, \mathfrak{f})$ with $\int \mathfrak{f}\, dF = 0$, and given "starting" point $\eta = \eta_0$. Inserting these paths in the likelihood (2.2), and computing the derivative $\frac{d}{dt}_{|t=0} \log p_{\eta_t}(x)$ of the log likelihood, we obtain the "score function" at $\eta = \eta_0$ in the direction $(\mathfrak{a}, \mathfrak{b}, \mathfrak{f})$. This can be easily computed to be the sum of the score functions when varying the three parameters separately, which are given by

$$\begin{aligned} B_\eta^a \mathfrak{a}(X) &= \bigg(R - \tfrac{1}{a(Z)} \bigg) \mathfrak{a}(Z), \\ B_\eta^b \mathfrak{b}(X) &= R\big(Y - b(Z)\big) \mathfrak{b}(Z), \\ B_\eta^f \mathfrak{f}(X) &= \mathfrak{f}(Z). \end{aligned}$$

The operators B_η^a, B_η^b, B_η^f are the *score operators* for the three parameters. The overall score $B_\eta(\mathfrak{a}, \mathfrak{b}, \mathfrak{f})(X)$ when perturbing the three parameters simultaneously is the sum of the three terms in the previous display. The *efficient influence function* of the functional χ at the point η is known to take the form (see Example 25.43 of [48] with $\dot\chi_Q(y)$ the current $y - \chi(\eta)$ and $\phi(y, 0)$ the current (R, Z), or the derivation below)

$$\widetilde\chi_\eta(X) = Ra(Z)\big(Y - b(Z)\big) + b(Z) - \chi(\eta).$$

We can verify that this is the correct formula by verifying that this function has the two properties defining an efficient influence function ([48], p. 426). First, the derivative at $t = 0$ of the functional along a path $t \mapsto \eta_t = (a_t, b_t, f_t)$ as previously, is the inner product of the influence function with the score function of that path: $\frac{d}{dt}_{|t=0} \chi(\eta_t) = P_\eta \widetilde\chi_\eta(X) B_\eta(\mathfrak{a}, \mathfrak{b}, \mathfrak{f})(X)$ for every path $t \mapsto p_{\eta_t}$ of the above form. Second, the function $\widetilde\chi_\eta$ is contained in the closed linear span of the set of all score functions. Indeed, in the present case we have, for all x,

$$\widetilde\chi_\eta(x) = B_\eta \xi_\eta(x) = B_\eta^b a(x) + B_\eta^f \bigg(b - \int b\, dF \bigg)(x), \tag{2.5}$$

where ξ_η is the *least favourable direction* given by

$$\xi_\eta = \big(0, \xi_\eta^b, \xi_\eta^f\big) = \bigg(0, a, b - \int b\, dF\bigg).$$

The function ξ_η is the score function for the submodel $t \mapsto \eta_t$ corresponding to the perturbations in the directions of $(0, a, b - \int b\, dF)$ on (a, b, F). The latter submodel is called *least favourable*, since $t \mapsto p_{\eta_t}$ has the smallest information about the functional of interest at $t = 0$. According to semiparametric theory (e.g., Chapter 25 of [48], in particular formula (25.22)) a sequence of estimators $\widehat\chi_n = \widehat\chi_n(X^{(n)})$ is asymptotically efficient for estimating $\chi(\eta)$ at the true parameter η_0 if and only if

$$\widehat\chi_n = \chi(\eta_0) + \frac{1}{n} \sum_{i=1}^n \widetilde\chi_{\eta_0}(X_i) + o_{P_{\eta_0}}\big(n^{-1/2}\big). \tag{2.6}$$

The sequence $\sqrt{n}(\widehat{\chi}_n - \chi(\eta_0))$ is then asymptotically normal with mean zero and variance $P_{\eta_0}\widetilde{\chi}_{\eta_0}^2$, which is the smallest possible in a local minimax sense.

For a direction $v = (\mathfrak{a}, \mathfrak{b}, \mathfrak{f})$, the *information norm* corresponding to the score operator (or LAN norm in the language of [10, 11, 31]) equals

$$\begin{aligned}\|v\|_\eta^2 &:= P_\eta\big[(B_\eta v)\big]^2 = \int\left[\frac{1}{a}\left(1-\frac{1}{a}\right)\mathfrak{a}^2 + \frac{b(1-b)}{a}\mathfrak{b}^2 + (\mathfrak{f}-F\mathfrak{f})^2\right]dF \\ &=: \|\mathfrak{a}\|_a^2 + \|\mathfrak{b}\|_b^2 + \|\mathfrak{f}\|_F^2.\end{aligned}$$

It may be noted that the three components of the score operator are orthogonal, which is a consequence of the factorization of the likelihood. The minimal asymptotic variance $P_{\eta_0}\widetilde{\chi}_{\eta_0}^2$ for estimating $\chi(\eta)$ can be written in terms of the information norm as

$$\begin{aligned}\|\xi_{\eta_0}\|_{\eta_0}^2 &= P_{\eta_0}(B_{\eta_0}\xi_{\eta_0})^2 = P_{\eta_0}\widetilde{\chi}_{\eta_0}^2 \\ &= \int a_0 b_0(1-b_0)\,dF_0 + \int b_0^2\,dF_0 - \chi(\eta_0)^2.\end{aligned} \tag{2.7}$$

3. Results. We put a prior probability distribution Π on the parameter (η^a, η^b, F) or $\eta = (\eta^a, \eta^b, \eta^f)$, and consider the posterior distribution $\Pi(\cdot|X^{(n)})$ based on the observation $X^{(n)} = (X_1, \ldots, X_n)$. This induces posterior distributions on all measurable functions of η, including the functional of interest $\chi(\eta)$.

We write $\mathcal{L}_\Pi(\sqrt{n}(\chi(\eta) - \widehat{\chi}_n)|X^{(n)})$ for the marginal posterior distribution of $\sqrt{n}(\chi(\eta) - \widehat{\chi}_n)$, where $\widehat{\chi}_n$ is any random sequence satisfying (2.6). We shall be interested in proving that this distribution asymptotically looks like a centered normal distribution with variance $\|\xi_{\eta_0}\|_{\eta_0}^2$. For a precise statement of this approximation, let d_{BL} be the bounded Lipschitz distance on probability distributions on $\mathbb{R}$ (see Chapter 11 of [12]).

DEFINITION 1. Let $X^{(n)} = (X_1, \ldots, X_n)$ be i.i.d. observations with $X_i = (Z_i, R_i, R_i Y_i)$ arising from the density p_{η_0} in (2.2), whose distribution we denote by $P_0 = P_{\eta_0}$. We say that the posterior satisfies the *semiparametric Bernstein–von Mises (BvM)* if, for $\widehat{\chi}_n$ satisfying (2.6) and $\|\xi_{\eta_0}\|_{\eta_0}$ given by (2.7), as $n \to \infty$,

$$d_{\mathrm{BL}}\big(\mathcal{L}_\Pi\big(\sqrt{n}\big(\chi(\eta) - \widehat{\chi}_n\big)|X^{(n)}\big), N\big(0, \|\xi_{\eta_0}\|_{\eta_0}^2\big)\big) \to^{P_0} 0.$$

In Sections 3.2 and 3.3, we present two general results for priors on the parameters (a, b), combined with an independent Dirichlet process prior on F. In Section 3.2, the prior on the pair (a, b) is general, whereas in Section 3.3 we construct a prior on b using an estimator of the propensity score $1/a$, thus linking the two parameters. Following these general results we specialize to Gaussian process priors and obtain concrete results in Section 3.4.

An alternative to using the Dirichlet process on F is to put a prior on the triple (a, b, f), for f a density of F. A general result can be found in Section 6 below, but it requires stronger conditions for the BvM theorem to hold. Putting a prior on f introduces the additional bias term (6.6), whose vanishing becomes more restrictive as the covariate dimension increases and can be problematic in even moderate dimensions. Thus it appears preferable to directly model the distribution F.

3.1. *Posterior distribution relative to Dirichlet process prior.* Since the covariates $Z_1, \ldots, Z_n$ are fully observed and the functional of interest $\chi(\eta)$ is an integral relative to their distribution F, intuitively the estimation problem should not depend too much on properties of the covariate distribution. For $\sqrt{n}$-estimation, this intuition is shown to be correct in [35]. In our Bayesian setup, this suggests to put a prior on F that does not limit this distribution.

The standard "nonparametric prior" on the set of probability distributions on a (Polish) sample space is the Dirichlet process prior [13]. This distribution is characterized by a base measure ν, which can be any finite measure on the sample space. It is well known that in the model consisting of sampling F from the Dirichlet process prior and next sampling observations $Z_1, \ldots, Z_n$ from F, the posterior distribution of F given $Z_1, \ldots, Z_n$ is again a Dirichlet process with updated base measure $\nu + n\mathbb{F}_n$, where $\mathbb{F}_n$ is the empirical distribution of $Z_1, \ldots, Z_n$. (For full definitions and properties, see the review in Chapter 4 of [16].)

We utilize the Dirichlet process prior on F together with an independent prior on the remaining parameters (a, b), constructed from a prior on (η^a, η^b) using the logistic link function (2.1). Because the Dirichlet process prior does not give probability one to a dominated set of measures F, the resulting posterior distribution of (a, b, F) cannot be derived using Bayes's formula. However, we can obtain a representation as follows. The parameters and the data are generated through the hierarchical scheme:

- $F \sim \mathrm{DP}(\nu)$ independent from $\eta = (a, b) \sim \Pi$.
- Given (F, a, b) the covariates $Z_1, \ldots, Z_n$ are i.i.d. F.
- Given $(F, a, b, Z_1, \ldots, Z_n)$ the pairs (R_i, Y_i) are independent from products of binomial distributions with success probabilities $1/a(Z_i)$ and $b(Z_i)$.
- The observations are $X^{(n)} = (X_1, \ldots, X_n)$ with $X_i = (Z_i, R_i, R_i Y_i)$.

From this scheme, it follows that F and $(R^{(n)}, Y^{(n)})$ are independent given $(Z^{(n)}, a, b)$, and also that F and (a, b) are conditionally independent given $X^{(n)}$. We can then conclude that the posterior distribution of F given $X^{(n)}$ is the same as the posterior distribution of F given $Z^{(n)}$, which is the $\mathrm{DP}(\nu + n\mathbb{F}_n)$ distribution. Furthermore, the posterior distribution of (a, b) given $(F, X^{(n)})$ can be derived by Bayes's rule from the binomial likelihood of $(R^{(n)}, R^{(n)}Y^{(n)})$ given $Z^{(n)}$, which is dominated. Thus the posterior distribution is given by

$$\begin{aligned}&\Pi\big((a, b) \in A, F \in B | X^{(n)}\big)\\ &\quad= \int_B \frac{\int_A \prod_{i=1}^n p_{(a,b)}(R_i, R_i Y_i | Z_i)\, d\Pi(a, b)}{\int \prod_{i=1}^n p_{(a,b)}(R_i, R_i Y_i | Z_i)\, d\Pi(a, b)}\, d\Pi\big(F | Z^{(n)}\big),\end{aligned} \tag{3.1}$$

where $p_{(a,b)}$ is the conditional density of (R, RY) given Z, given by (2.2) with f deleted or taken equal to 1, and $\Pi(F \in \cdot | Z^{(n)})$ is the $\mathrm{DP}(\nu + n\mathbb{F}_n)$-distribution. This formula remains valid if $\nu = 0$, which yields the Bayesian bootstrap; see Chapter 4.7 of [16], and is also covered in the theorems below. We suspect that the theorems extend to other exchangeable bootstrap processes, as considered in [25] (see [47], Section 3.7.2).

3.2. *General prior on (a, b) and Dirichlet process prior on F.* Define $\eta_t(\eta) = \eta_t(\eta; n, \xi_{\eta_0})$ to be a perturbation of $\eta = (\eta^a, \eta^b)$ in the least favourable direction, restricted to the components corresponding to a and b:

$$\eta_t(\eta) = \left(\eta^a, \eta^b - \frac{t}{\sqrt{n}} \xi_{\eta_0}^b\right). \tag{3.2}$$

THEOREM 1. *Consider a prior Π consisting of an arbitrary prior on $\eta = (\eta^a, \eta^b)$ and an independent Dirichlet process prior on F. Assume that there exist measurable sets $\mathcal{H}_n$ of functions $\eta = (\eta^a, \eta^b)$ satisfying*

$$\Pi\big(\eta \in \mathcal{H}_n | X^{(n)}\big) \to^{P_0} 1, \tag{3.3}$$

$$\sup_{b = \Psi(\eta^b): \eta \in \mathcal{H}_n} \|b - b_0\|_{L^2(F_0)} \to 0, \tag{3.4}$$

$$\sup_{b = \Psi(\eta^b): \eta \in \mathcal{H}_n} \big|\mathbb{G}_n[b - b_0]\big| \to^{P_0} 0. \tag{3.5}$$

If for the path $\eta_t(\eta)$ in (3.2) and every t,

$$\frac{\int_{\mathcal{H}_n} \prod_{i=1}^n p_{\eta_t(\eta)}(R_i, R_i Y_i | Z_i)\, d\Pi(\eta)}{\int_{\mathcal{H}_n} \prod_{i=1}^n p_\eta(R_i, R_i Y_i | Z_i)\, d\Pi(\eta)} \to^{P_0} 1, \tag{3.6}$$

then the posterior distribution (3.1) satisfies the BvM theorem.

Conditions (3.3)–(3.5) permit to control the remainder terms in an expansion of the likelihood. They require that the posterior distribution of b concentrates on shrinking neighbourhoods about the true parameter b_0 (with no similar requirement for a), and hence mostly require consistency.

The uniformity in b required in (3.5) is unpleasant, as it will typically require that the class of b supported by the posterior distribution is not unduly large. The condition is linked to using the likelihood and similar conditions arise in maximum likelihood based estimation procedures, although (3.5) seems significantly weaker, as the uniformity is required only on the essential support of the posterior distribution, which might be much smaller than the full parameter space. The use of estimating equations can avoid uniformity conditions by sample splitting [35]. In the Bayesian framework, one might similarly base posterior distributions of different parameters on given subsamples, but this is unnatural so that we do not pursue this route here.

Under (3.4) a sufficient condition for (3.5) is that the class of functions b in the condition is contained in a fixed F_0-Donsker class (see Lemma 3.3.5 of [52]). In particular, it suffices that the posterior concentrates on a bounded set in H^s for $s > d/2$. While this condition is easy to establish for certain priors, such as uniform wavelet priors [17], for the Gaussian process priors considered below we employ relatively complicated arguments using metric entropy bounds to verify the condition.

Condition (3.6) measures the invariance of the prior for the full nuisance parameter under a shift in the least favourable direction $\xi^b_{\eta_0}$. It is a structural condition on the combination of prior and model, and if not satisfied may destroy the $\sqrt{n}$-rate in the BvM theorem (see [10] or [16] for further discussion). Although we shall verify the condition for several priors of interest below, this condition may impose smoothness conditions on the parameters, and prevent so-called "double robustness". We shall remove this condition for special priors in Theorem 2 below.

The invariance involves the component ξ^b_0 only, and not the other nonzero component ξ^f_0 of the least favourable direction. In contrast, in Theorem 3, which puts a prior on the covariate density f, the invariance involves the full least favourable direction (see (6.1)). Intuitively, the Dirichlet process is a fully nonparametric prior that never causes this type of bias.

Since $\xi^b_{\eta_0} = a_0$, Theorem 1 implicitly requires conditions on a_0 through (3.6), even though a does not appear in the functional $\chi(\eta)$. Such conditions become explicit for concrete priors below.

REMARK 1. If the quotient on the left-hand side of (3.6) is asymptotic to $e^{\mu_n t}(1 + o_{P_0}(1))$ for some possibly random sequence of real numbers μ_n, then the assertion of the BvM theorem is still true, but the normal approximation $N(0, \|\xi_{\eta_0}\|^2_{\eta_0})$ must be replaced by $N(\mu_n, \|\xi_{\eta_0}\|^2_{\eta_0})$. See [11, 31] for further discussion. The same is true for all other results in the following.

REMARK 2. If the supremum in (3.5), or similar variables below, is not measurable, then we interpret this statement in terms of outer probability.

Formula (3.1) shows that a draw from the posterior distribution of the functional of interest $\chi(\eta) = \int b\,dF$ is obtained by independently drawing b from its posterior distribution and F from the $\mathrm{DP}(\nu + n\mathbb{F}_n)$-distribution, and next forming the integral $\int b\,dF$. The posterior distribution of b is constructed from the conditional likelihood of $(R^{(n)}, R^{(n)}Y^{(n)})$ given $Z^{(n)}$ without involving F or its prior distribution. Instead of a Bayesian-motivated or bootstrap type choice for F, which requires randomization given $Z^{(n)}$, one could also directly plug in an estimator of F based on $Z^{(n)}$ and randomize only b from its posterior distribution. The empirical distribution $\mathbb{F}_n$ is an obvious choice. The proof of Theorem 1 suggests that for this choice, under the conditions of the theorem,

$$d_{\mathrm{BL}}\big(\mathcal{L}_\Pi\big(\sqrt{n}(\chi(\eta) - \widehat{\chi}_n)|X^{(n)}\big), N\big(0, \|\xi_{\eta_0}^{b_0}\|_{b_0}^2\big)\big) \to^{P_0} 0.$$

Compared to the BvM theorem, this suggests a normal approximation with the same centering, but a smaller variance, since the variance in the BvM theorem is the sum $\|\xi_{\eta_0}^{b_0}\|_{b_0}^2 + \|\xi_{\eta_0}^{f_0}\|_{F_0}^2$. The lack of posterior randomization of F thus results in an underestimation of the asymptotic variance. Using credible sets resulting from this "posterior" would give overconfident (wrong) uncertainty quantification. Since our focus is on the Bayesian approach, we do not purse such generalizations further.

3.3. *Propensity score-dependent priors*. To reduce unnecessary regularity conditions, it can be useful to use a preliminary estimate $\hat{a}_n$ of the inverse propensity score [35, 37, 39]. In a Bayesian setting, [20] suggest adding an estimate of the propensity score evaluated at the data as an additional covariate when using BART for causal inference [22]. In this section, we employ preliminary estimators $\hat{a}_n$ to augment the prior on b with the aim of weakening the conditions required for a semiparametric BvM.

Suppose we have a sequence of estimators $\hat{a}_n$ of the inverse propensity score satisfying

$$\|\hat{a}_n - a_0\|_{L^2(F_0)} = O_{P_0}(\rho_n) \tag{3.7}$$

for some sequence $\rho_n \to 0$. Since the propensity score is just a (binary) regression function of R onto Z, standard (adaptive) smoothing estimators satisfy this condition with rate $\rho_n = n^{-\alpha/(2\alpha+d)}$ if the propensity score is assumed to be contained in $C^\alpha([0,1]^d)$, which is the minimax rate over this space (note that $\hat{a}_n - a_0 = \hat{a}_n a_0(1/a_0 - 1/\hat{a}_n)$ will attain at least the rate of an estimator of the propensity score $1/a_0$ itself). Consider the following prior on b:

$$b(z) = \Psi\big(W_z^b + \lambda \hat{a}_n(z)\big), \tag{3.8}$$

where W^b is a continuous stochastic process independent of the random variable λ, which follows a prior $N(0, \sigma_n^2)$ distribution for given variance σ_n^2 (potentially varying with n, but fixed is allowed). The additional parameter λ has the role of making the prior link between the parameters b and a flexible; the variance σ_n^2 will be required not too small below.

We assume that $\hat{a}_n$ is based on observations that are independent of $X_1, \ldots, X_n$, the observations used in the likelihood to obtain the posterior distribution. Otherwise, the prior (3.8) becomes data-dependent, which significantly complicates the technical analysis. This independence seems, however, unnecessary in practice. The analogous prior to (3.8) for a continuous regression model is investigated empirically in the companion paper [27], where it performs well when $1/\hat{a}_n$ is trained on the same data as the posterior.

We may think of $\hat{a}_n$ as a degenerate prior on a, and then by the factorization of the likelihood the part of the likelihood involving a cancels from the posterior distribution (3.1) if marginalized to (b, F) (and hence $\chi(\eta)$). Of course, the same will happen if we assign an independent prior to a. Thus in both cases it is unnecessary to further discuss a prior on a.

THEOREM 2. *Given independent estimators $\hat{a}_n$ satisfying (3.7) and having $\|\hat{a}_n\|_\infty = O_{P_0}(1)$, consider the prior (3.8) for b with the stochastic process W^b and random variable $\lambda \sim N(0, \sigma_n^2)$ independent, and assign F an independent Dirichlet process prior. Assume that there exist measurable sets $\mathcal{H}_n^b$ of functions satisfying, for every $t \in \mathbb{R}$ and some numbers $u_n, \varepsilon_n^b \to 0$,*

$$\Pi\big(\lambda : |\lambda| \le u_n \sigma_n^2 \sqrt{n} | X^{(n)}\big) \to^{P_0} 1, \tag{3.9}$$

$$\Pi\big((w, \lambda) : w + \big(\lambda + tn^{-1/2}\big)\hat{a}_n \in \mathcal{H}_n^b | X^{(n)}\big) \to^{P_0} 1, \tag{3.10}$$

$$\sup_{b=\Psi(\eta^b):\eta^b \in \mathcal{H}_n^b} \|b - b_0\|_{L^2(F_0)} \le \varepsilon_n^b, \tag{3.11}$$

$$\sup_{b=\Psi(\eta^b):\eta^b \in \mathcal{H}_n^b} \big|\mathbb{G}_n[b - b_0]\big| \to^{P_0} 0. \tag{3.12}$$

If $n\sigma_n^2 \to \infty$ and $\sqrt{n}\rho_n \varepsilon_n^b \to 0$, then the posterior distribution satisfies the semiparametric BvM theorem.

The advantage of this theorem over Theorem 1 is that (3.6) does not appear in its conditions. (The theorem adds (3.9) and (3.10), but these are relatively mild.) As noted above, condition (3.6) requires a certain invariance of the prior of b in the the least favourable direction $\xi_{\eta_0}^b = a_0$, and typically leads to smoothness requirements on a. In contrast, we show below that Theorem 2 can yield the BvM theorem for propensity scores $1/a$ of arbitrarily low regularity. Thus the theorem is able to achieve what could be named *single robustness*. Whether "double robustness", the ability of also handling response functions b of arbitrarily low smoothness, is also achieved remains unclear. Specifically, we have not been able to verify condition (3.12) without assuming that the smoothness of b is above the usual threshold ($d/2$ in d dimensions).

The single robustness is achieved by perturbing the prior process for b in the least favourable direction using the auxiliary variable λ. Since the least favourable direction a_0 is unknown, this is replaced with an estimate $\hat{a}_n$.

Condition (3.9) puts a lower bound on the variability of the perturbation, that is, on the standard deviation σ_n of λ. An easy method to ascertain this condition is to show that the prior mass of the set λ in the left-hand side is exponentially small and next invoke Lemma 4. Specifically, by the univariate Gaussian tail bound the prior mass of $\{\lambda : |\lambda| > u_n \sigma_n^2 \sqrt{n}\}$ is bounded above by $e^{-u_n^2 \sigma_n^2 n/2}$. If the Kullback–Leibler neighbourhood in Lemma 4 has prior probability at least $e^{-n(\varepsilon_n^b)^2}$, then the lemma gives the sufficient condition $u_n^2 \sigma_n^2 \gtrsim (\varepsilon_n^b)^2$ for (3.9), that is, $\sigma_n \gg \varepsilon_n^b$.

3.4. *Specialization to Gaussian process priors.* In this section, we specialize Theorems 1 and 2 to Gaussian process priors. In all examples, the priors on the three parameters a, b and F are independent. Since a does not appear in $\chi(\eta)$ and the likelihood (2.2) factorizes over a, b and F, the a terms cancel from the marginal posterior distribution of $\chi(\eta)$. Thus the prior on a is irrelevant, and it is not necessary to consider it.

For simplicity, we take the covariate space to be the unit cube $\mathcal{Z} = [0, 1]^d$. Given a mean-zero Gaussian process $W^b = (W_z^b : z \in [0, 1]^d)$, we consider both the propensity score-dependent prior for b given by (3.8) and the more simple prior

$$b(z) = \Psi\big(W_z^b\big). \tag{3.13}$$

There are a great variety of Gaussian processes, and their success in nonparametric estimation is known to depend on their sample smoothness, as measured through their small ball probability (see [45, 46, 49, 51]). We derive a proposition on general Gaussian processes and consider the following specific examples.

EXAMPLE (Riemann–Liouville). In dimension $d = 1$, the *Riemann–Liouville* process released at zero of regularity $\bar{\beta} > 0$ is defined by

$$W_z^b = \sum_{k=0}^{\lfloor \bar{\beta} \rfloor + 1} g_k z^k + \int_0^z (z-s)^{\bar{\beta}-1/2}\, dB_s, \quad z \in [0,1], \tag{3.14}$$

where the (g_k) are i.i.d. standard normal random variables and B is an independent Brownian motion. This process is appropriate for nonparametric modelling of $C^{\bar{\beta}}([0,1])$-functions. We shall investigate the effect of the smoothness parameter $\bar{\beta}$ on the BvM theorem.

EXAMPLE (Gaussian series). Another commonly used Gaussian process prior consists of a finite series expansion with Gaussian coefficients. Let $\{\psi_{jk} : j \geq 1, k = 0, \ldots, 2^{jd} - 1\}$ denote a sufficiently regular boundary-adapted Daubechies wavelet basis of $L^2([0,1]^d)$. We assume it is regular enough for the decay of the wavelet coefficients to characterize all the relevant Besov $B^s_{\infty\infty}$-norms (which are equal to the C^s-Hölder norms for $s \notin \mathbb{N}$. For details on such wavelets and Besov spaces, see Chapter 4.3 of [18].) Consider the prior

$$W_z^b = \sum_{j=1}^{J_{\bar{\beta}}} \sum_{k=0}^{2^{jd}-1} \sigma_j g_{jk} \psi_{jk}(z), \quad g_{jk} \overset{\text{i.i.d.}}{\sim} N(0,1), \tag{3.15}$$

where $2^{J_{\bar{\beta}}} \sim n^{1/(2\bar{\beta}+d)}$, which tends to infinity with n, is the optimal dimension of a finite-dimensional model if the true parameter is known to be $\bar{\beta}$-smooth and $\sigma_j = 2^{-j(r+d/2)}$ for $r \geq 0$. Since we wish to perform the prior regularization via the truncation level $J_{\bar{\beta}}$ rather than the scaling coefficients σ_j, we restrict to considering $r \leq \beta \wedge \bar{\beta}$, for instance $r = 0$.

We can view both processes as Borel-measurable maps in the Banach space $C([0,1]^d)$, equipped with the uniform norm $\|\cdot\|_\infty$. In the following proposition, we consider a general zero-mean Gaussian process of this type. Such a process determines a so-called reproducing kernel Hilbert space (RKHS) $(\mathbb{H}^b, \|\cdot\|_{\mathbb{H}^b})$, and a "concentration function" at η_0^b, defined as, for $\varepsilon > 0$,

$$\phi_{\eta_0^b}(\varepsilon) = \inf_{h \in \mathbb{H}^b : \|h - \eta_0^b\|_\infty < \varepsilon} \|h\|^2_{\mathbb{H}^b} - \log P\big(\|W^b\|_\infty < \varepsilon\big). \tag{3.16}$$

For standard statistical models, the posterior contraction rate ε_n^b for such a Gaussian process prior is linked to the solution of the equation

$$\phi_{\eta_0^b}\big(\varepsilon_n^b\big) \sim n\big(\varepsilon_n^b\big)^2. \tag{3.17}$$

For details, see [50] and [49].

PROPOSITION 1. *Consider the prior* (3.13) *on b for a Gaussian process* W^b *with values in* $C([0,1]^d)$ *combined with an independent Dirichlet process prior on* F. *Let* $\varepsilon_n^b \to 0$ *satisfy* (3.17). *Suppose there exist sequences* $\xi_n \in \mathbb{H}^b$ *and* $\zeta_n^b \to 0$ *such that*

$$\big\|\xi_n^b - \xi_{\eta_0}^b\big\|_\infty \leq \zeta_n^b, \qquad \big\|\xi_n^b\big\|_{\mathbb{H}^b} \leq \sqrt{n}\zeta_n^b, \qquad \sqrt{n}\varepsilon_n^b\zeta_n^b \to 0. \tag{3.18}$$

Suppose further that there exist measurable sets $\mathcal{H}_n^b$ *of functions* η^b *such that* $\Pi(\eta^b \in (\mathcal{H}_n^b - t\xi_n^b/\sqrt{n})|X^{(n)}) \to^{P_0} 1$ *for every* $t \in \mathbb{R}$ *and* (3.5) *holds. Then the posterior distribution satisfies the semiparametric BvM theorem.*

For the examples of the Riemann–Liouville process and finite Gaussian series prior the preceding proposition implies the following.

COROLLARY 1. *Suppose* $a_0 \in C^\alpha([0,1]^d)$, $b_0 \in C^\beta([0,1]^d)$ *and consider the prior* (3.13) *on* b *with* W^b *the random series* (3.15) *combined with an independent Dirichlet process prior on* F. *If* $\alpha, \beta > d/2$ *and* $d/2 < \bar{\beta} < \alpha + \beta - d/2$, *then the posterior distribution satisfies the semiparametric BvM theorem. Moreover, when* $d = 1$ *the same result holds with* W^b *the Riemann–Liouville process* (3.14) *with parameter* $\bar{\beta}$.

For $\alpha, \beta > d/2$, the parameter $\bar{\beta}$ can always be chosen to satisfy the remaining condition in the corollary, in which case the BvM theorem holds. The values $\alpha, \beta > d/2$ are one particular pair satisfying (2.4). However, when using product priors, it does not seem possible to use extra smoothness in one parameter to offset low regularity in the other as in (2.4). To remedy this, we consider the propensity score-dependent prior (3.8).

COROLLARY 2. *Suppose* $a_0 \in C^\alpha([0,1]^d)$ *and* $b_0 \in C^\beta([0,1]^d)$. *Let* $\hat{a}_n$ *be an independent estimator satisfying* $\|\hat{a}_n\|_\infty = O_{P_0}(1)$ *and* (3.7) *for some* $\rho_n \to 0$. *Consider the prior* (3.8) *for* b, *where* W^b *is the random series* (3.15) *combined with an independent Dirichlet process prior on* F. *If* $\beta \wedge \bar{\beta} > d/2$ *and*

$$(n/\log n)^{-(\beta\wedge\bar{\beta})/(2\bar{\beta}+d)} \ll \sigma_n \lesssim 1, \qquad \sqrt{n}\rho_n (n/\log n)^{-(\beta\wedge\bar{\beta})/(2\bar{\beta}+d)} \to 0,$$

then the posterior distribution satisfies the semiparametric BvM. Moreover, when $d = 1$ *the same result holds with* W^b *the Riemann–Liouville process* (3.14) *with parameter* $\bar{\beta}$.

If $\bar{\beta} = \beta$ and $\rho_n = (\log n)^\kappa n^{-\alpha/(2\alpha+1)}$ is the minimax rate of estimation, possibly up to a logarithmic factor, then the above conditions reduce to $\beta > d/2$ and (2.4). If β is near the lower limit $d/2$, then the latter condition requires that α be bigger than nearly $d/2$ as well, but if β is large, then the latter condition will be satisfied for α close to zero. Thus the estimation method is able to exploit extra smoothness in b_0 to offset lower regularity in a_0, in particular if $0 < \alpha \le d/2$, unlike the standard product Gaussian process priors, where we required both $\alpha, \beta > d/2$. Since it is still needed that $\beta > d/2$, the preceding corollary does not give full "double robustness" in also taking advantage of extra regularity in a_0 if $0 < \beta \le d/2$. The technical reason is requirement (3.5), which is present in all our theorems, and used in the proofs to establish the LAN expansion of the model. Whether this is a fundamental limitation of the Bayesian approach or a purely technical artefact is unclear.

If W^b is a mean-zero Gaussian process with covariance kernel $K_{W^b}(z,z') = \mathbb{E}W_z^b W_{z'}^b$, then the term $W^b + \lambda\hat{a}_n$ in (3.8) is also a mean-zero Gaussian process with data-driven covariance

$$\mathbb{E}\big[W_z^b + \lambda\hat{a}_n(z)\big]\big[W_{z'}^b + \lambda\hat{a}_n(z')\big] = K_{W^b}(z,z') + \sigma_n^2 \hat{a}_n(z)\hat{a}_n(z').$$

In this case, the propensity score-dependent prior corresponds to an easy to implement correction to the prior covariance function. In particular, one can use standard methods for Gaussian process posterior computation, such as Laplace or sparse approximations [26]. In practice, we would also suggest to truncate the estimator $1/\hat{a}_n$ away from 0 for numerical stability. Computational and empirical aspects of this new prior are investigated in the continuous regression model in a companion paper [27], where it is found that incorporating an estimator of the propensity score in this way significantly improves the performance of Gaussian process priors.

4. Discussion. A key technical difficulty for establishing semiparametric BvM results is controlling the ratio (3.6) (or (6.7)). While one can use the Cameron–Martin theorem for Gaussian priors, such results are typically more involved outside the Gaussian setting. The hyper parameter λ in the prior (3.8) removes this obstacle, allowing results for a much wider

class of priors. For instance, one may select W^b in (3.8) to be a truncated prior or sieve prior, without having to establish (3.6) directly for those priors.

Such a prior construction generalizes to other models and functionals. Consider a model $\mathcal{P} = (P_\eta : \eta \in \mathcal{H})$ and a parameter $\chi(\eta)$. For a prior of the form $\eta = W + \lambda\hat{\xi}_n$, where W is a continuous stochastic process, $\lambda \sim N(0, \sigma_n^2)$ and $\hat{\xi}_n$ is an estimate of the least favourable direction ξ_{η_0} of χ at η_0 in the model $\mathcal{P}$, similar results to the above should hold. We emphasize, however, that such a prior is designed for semiparametric estimation of the specific functional χ and will not perform any better for any other functional. It is thus suitable for estimating a functional of interest in the presence of a high or infinite-dimensional nuisance parameter that can have a significant impact, as in the model we study here.

5. Proofs of the main results.

5.1. *Proof of Theorem 1: General prior on b and Dirichlet process prior.* PROOF OF THEOREM 1. The total variation distance between the posterior distributions based on the prior Π and the prior $\Pi_n(\cdot) := \Pi(\cdot \cap \mathcal{H}_n)/\Pi(\mathcal{H}_n)$, which is Π conditioned to $\mathcal{H}_n$, is bounded above by $2\Pi(\mathcal{H}_n^c|X^{(n)})$ (e.g., p. 142 of [48]). Since this tends to zero in probability by assumption and the total variation topology is stronger than the weak topology, it suffices to show the desired result for the conditioned prior Π_n instead of Π.

Let $\widehat{\chi}_n = \chi(\eta_0) + \mathbb{P}_n\widetilde{\chi}_{\eta_0}$, so that it satisfies (2.6) with the remainder term identically zero. The posterior Laplace transform of the variable $\sqrt{n}(\chi(\eta) - \widehat{\chi}_n)$ is given by, for $t \in \mathbb{R}$,

$$\begin{aligned} I_n(t) &= \mathbb{E}^{\Pi_n}\big[e^{t\sqrt{n}(\chi(\eta)-\widehat{\chi}_n)}|X^{(n)}\big] \\ &= \int\int_{\mathcal{H}_n} \frac{e^{t\sqrt{n}\int(b\,dF - b_0\,dF_0) - t\mathbb{G}_n[\widetilde{\chi}_{\eta_0}] + \ell_n^b(\eta) - \ell_n^b(\eta_t)} e^{\ell_n^b(\eta_t)}}{\int_{\mathcal{H}_n} e^{\ell_n^b(\eta')}\,d\Pi(\eta')}\,d\Pi(\eta)\,d\Pi\big(F|X^{(n)}\big), \end{aligned}$$

in view of (3.1) and the factorization of the likelihood over a and b. This is (obviously) true for any η_t, in particular for the path $\eta_t = \eta_t(\eta)$ defined in (3.2). We shall show that $I_n(t)$ tends in probability to $\exp(t^2\|\xi_{\eta_0}\|_{\eta_0}^2/2)$, which is the Laplace transform of a $N(0, \|\xi_{\eta_0}\|_{\eta_0}^2)$ distribution, for every t in a neighbourhood of 0. Since convergence of conditional Laplace transforms in probability implies conditional convergence in distribution in probability (see Lemma 14 below), this would complete the proof.

At the end of the proof, we shall show that, uniformly in $\eta \in \mathcal{H}_n$,

$$\ell_n^b(\eta) - \ell_n^b(\eta_t) = t\mathbb{G}_n\big[\widetilde{\chi}_{\eta_0}^b\big] + t\sqrt{n}\int (b_0 - b)\,dF_0 + \frac{t^2}{2}\big\|\xi_{\eta_0}^b\big\|_{b_0}^2 + o_{P_0}(1), \tag{5.1}$$

where $\widetilde{\chi}_\eta^b = B_\eta^b a$ is the component of the efficient influence function in the b direction (see (2.5)). Inserting this Taylor expansion in the preceding display, we see that

$$\begin{aligned} I_n(t) = \int\int_{\mathcal{H}_n} &\frac{e^{t\sqrt{n}\int(b\,dF - b_0\,dF_0) + t\sqrt{n}\int(b_0-b)\,dF_0} e^{\ell_n^b(\eta_t)}}{\int_{\mathcal{H}_n} e^{\ell_n^b(\eta')}\,d\Pi(\eta')}\,d\Pi(\eta)\,d\Pi\big(F|X^{(n)}\big) \\ &\times e^{-t\mathbb{G}_n[\widetilde{\chi}_{\eta_0}^f] + \frac{t^2}{2}\|\xi_{\eta_0}^b\|_{b_0}^2 + o_{P_0}(1)}, \end{aligned}$$

where $\widetilde{\chi}_{\eta_0}^f = \widetilde{\chi}_{\eta_0} - \widetilde{\chi}_{\eta_0}^b = b_0 - \chi(\eta_0)$. Note that the integral in the denominator is a constant relative to η and F, since all variables are integrated out. By Fubini's theorem, the double integral without the normalizing constant equals

$$\int_{\mathcal{H}_n} e^{\ell_n^b(\eta_t)} \int e^{t\sqrt{n}\int b\,d(F-F_0)}\,d\Pi\big(F|X^{(n)}\big)\,d\Pi(\eta).$$

Let $\mathbb{F}_n = n^{-1}\sum_{i=1}^n \delta_{Z_i}$ denote the empirical distribution of the covariates. By assumption (3.5), we certainly have that $\sup\{|(\mathbb{F}_n - F_0)b| : b = \Psi(\eta^b), \eta \in \mathcal{H}_n\}$ tends to zero in probability. Therefore, Lemma 1 below yields that for every t in a neighbourhood of zero, the preceding display equals

$$e^{o_{P_0}(1)} \int_{\mathcal{H}_n} e^{\ell_n^b(\eta_t)} e^{t\sqrt{n}\int bd(\mathbb{F}_n - F_0)} e^{\frac{t^2}{2}\|b - F_0 b\|^2_{L^2(F_0)}}\, d\Pi(\eta).$$

Since $\|b - b_0\|_{L^2(F_0)} \to 0$ uniformly on $\mathcal{H}_n$ and $\sqrt{n}\int bd(\mathbb{F}_n - F_0) = \mathbb{G}_n[b_0] + o_{P_0}(1)$ by assumption (3.5), the previous display equals

$$e^{t\mathbb{G}_n[b_0] + \frac{t^2}{2}\|b_0 - F_0 b_0\|^2_{L^2(F_0)} + o_{P_0}(1)} \int_{\mathcal{H}_n} e^{\ell_n^b(\eta_t)}\, d\Pi(\eta).$$

We insert this in the expression for $I_n(t)$, combine the two exponential terms using that $\widetilde{\chi}^f_{\eta_0} = b_0 - \chi(\eta_0)$ and $\|b_0 - F_0 b_0\|_{L^2(F_0)} = \|\xi^f_{\eta_0}\|_{F_0}$, and invoke assumption (3.6), to see that $I_n(t)$ tends to $e^{t^2\|\xi_{\eta_0}\|^2_{\eta_0}/2}$ in probability. The theorem then follows by the convergence of Laplace transforms.

We conclude by a proof of (5.1). This entails an expansion of the log likelihood $\ell_n^b(\eta) - \ell_n^b(\eta_t)$ along the submodel η_t. We can decompose

$$\begin{aligned}\ell_n^b(\eta) - \ell_n^b(\eta_t) &= t\mathbb{G}_n[\widetilde{\chi}^b_{\eta_0}] + \sqrt{n}\mathbb{G}_n\left[\log p_\eta - \log p_{\eta_t} - \frac{t}{\sqrt{n}}\widetilde{\chi}^b_{\eta_0}\right] \\ &\quad + nP_{\eta_0}[\log p_\eta - \log p_{\eta_t}].\end{aligned} \tag{5.2}$$

We shall show that the second term on the right tends to zero in probability, while the third term tends to the quadratic $t^2\|a_0\|^2_{b_0}/2$, where $a_0 = \xi^b_{\eta_0}$.

The definition $\eta_u := (\eta^a, \eta^b_u)$ with $\eta^b_u = \eta^b - tu\xi^b_{\eta_0}/\sqrt{n}$, for $u \in [0,1]$, gives a path from $\eta_{u=0} = \eta$ (not η_0!) to $\eta_{u=1} = \eta_t$, so that $\log p_\eta - \log p_{\eta_t} = g(0) - g(1)$ for $g(u) = \log p_{\eta_u}$. We shall replace this difference in both terms on the right of (5.2) by the Taylor expansion $g(0) - g(1) = -g'(0) - g''(0)/2 - \theta$, where $|\theta| \le \|g'''\|_\infty$. The expansion will be uniform in $\eta \in \mathcal{H}_n$, although the dependence of g and θ on η is not indicated in the notation.

By explicit calculations, the derivatives of g can be seen to be

$$\begin{aligned} g'(u) &= -\frac{t}{\sqrt{n}} B^b_{\eta_u} a_0 = -\frac{t}{\sqrt{n}} r\big(y - \Psi\big(\eta^b_u\big)\big)a_0, \\ g''(u) &= -\frac{t^2}{n} r\Psi'\big(\eta^b_u\big)a_0^2, \qquad g'''(u) = \frac{t^3}{n^{3/2}} r\Psi''\big(\eta^b_u\big)a_0^3, \end{aligned}$$

where we have omitted the function arguments (r, y, z). Since $|\theta| \le \|g'''\|_\infty \lesssim n^{-3/2}$, it follows that both $\sqrt{n}\mathbb{G}_n\theta$ and $nP_0\theta$ tend to zero in probability, uniformly in $\eta \in \mathcal{H}_n$. Since $B^b_{\eta_0} a_0 = \widetilde{\chi}^b_{\eta_0}$,

$$\begin{aligned} g'(0) &= -\frac{t}{\sqrt{n}} B^b_\eta a_0 = -\frac{t}{\sqrt{n}}\widetilde{\chi}^b_{\eta_0} + \frac{t}{\sqrt{n}} r(b - b_0)a_0, \\ g''(0) &= -\frac{t^2}{n} r\Psi'\big(\eta^b\big)a_0^2 = -\frac{t^2}{n} r\Psi'\big(\eta^b_0\big)a_0^2 - \frac{t^2}{n} r\big(b(1-b) - b_0(1-b_0)\big)a_0^2 \end{aligned}$$

for $b = \Psi(\eta^b)$, since $\Psi' = \Psi(1 - \Psi)$.

By assumption (3.5) and Lemma 11, applied with $\mathcal{H}_{n,1}$ the set of functions $\sqrt{n}(b - b_0)$ and $\mathcal{H}_{n,2} = \{r\}$, we have that $\mathbb{G}_n[r(b - b_0)a_0] \to 0$ in probability, uniformly in $\{b = \Psi(\eta^b) : \eta \in \mathcal{H}_n\}$, whence $\sqrt{n}\mathbb{G}_n g'(0) = -t\mathbb{G}_n[\widetilde{\chi}^b_{\eta_0}] + o_{P_0}(1)$, uniformly in $\eta \in \mathcal{H}_n$. By again assumption (3.5) and Lemma 11, $\mathbb{G}_n[r(b(1-b) - b_0(1-b_0))a_0^2] \to 0$ in probability, whence

$\sqrt{n}\mathbb{G}_n g''(0) = O_{P_0}(n^{-1/2}) \to 0$ in probability. We conclude that the second term on the right in (5.2) tends to zero in probability, uniformly in $\eta \in \mathcal{H}_n$.

Since $\Psi'(\eta_0^b) = b_0(1-b_0)$ and $\int b_0(1-b_0)a_0\,dF_0 = \|\xi_{\eta_0}^b\|_{b_0}^2$,

$$-nP_{\eta_0}g'(0) = t\sqrt{n}\int (b_0 - b)\,dF_0,$$

$$\begin{aligned}-nP_{\eta_0}g''(0) - t^2\|\xi_{\eta_0}^b\|_{b_0}^2 &= t^2 P_{\eta_0}\big[r\big(b(1-b) - b_0(1-b_0)\big)a_0^2\big] \\ &\lesssim P_{\eta_0}\big[r|b-b_0|a_0^2\big] \le \|b-b_0\|_{L^1(F_0)}\|a_0\|_\infty.\end{aligned}$$

Therefore, $nP_{\eta_0}[-g'(0) - g''(0)/2]$ is equal to $t\sqrt{n}\int (b_0-b)\,dF_0 + t^2\|\xi_{\eta_0}^b\|_{b_0}^2/2 + o_{P_0}(1)$. The third term on the right of (5.2) is equivalent to the same expression. This concludes the proof of (5.1). □

The preceding proof makes use of the following lemma, which can be considered a BvM theorem for the Laplace transform of the Dirichlet posterior process. A proof of the lemma can be found in [29].

Let $\mathbb{F}_n$ be the empirical distribution of an i.i.d. sample $Z_1, \ldots, Z_n$ from a distribution F_0 on a Polish sample space $(\mathcal{Z}, \mathcal{C})$, and given $Z_1, \ldots, Z_n$ let F_n be the distribution of a draw from the Dirichlet process with base measure $\nu + n\mathbb{F}_n$. Thus ν is a finite measure on $(\mathcal{Z}, \mathcal{C})$, and $F_n|Z_1, \ldots, Z_n \sim \mathrm{DP}(\nu + n\mathbb{F}_n)$ is the posterior distribution obtained when equipping the distribution of the observations $Z_1, Z_2, \ldots, Z_n$ with a Dirichlet process prior with base measure ν. The case $\nu = 0$ is allowed.

LEMMA 1. *Suppose $\mathcal{G}_n$ are separable classes of measurable functions such that $\sup_{g\in\mathcal{G}_n}|\mathbb{F}_n g - F_0 g| \to 0$ in probability and have envelope functions G_n satisfying $\nu G_n = O(1)$ and $F_0 G_n^{2+\delta} = O(1)$ for some $\delta > 0$. Then for every t in a sufficiently small neighbourhood of 0, in probability,*

$$\sup_{g\in\mathcal{G}_n}\big|\mathbb{E}\big[e^{t\sqrt{n}(F_n g - \mathbb{F}_n g)}|Z_1, \ldots, Z_n\big] - e^{t^2 F_0(g-F_0 g)^2/2}\big| \to 0.$$

5.2. *Proof of Theorem 2: Propensity score-dependent prior.* PROOF OF THEOREM 2. For the propensity score-dependent prior (3.8), the posterior distribution for $\sqrt{n}(\chi(\eta) - \hat{\chi}_n)$ is dependent both on the data $X^{(n)}$ and the estimator $\hat{a}_n$, and hence the bounded Lipschitz distance between this posterior distribution and the approximating normal distribution in Definition 1 is a function $H(X^{(n)}, \hat{a}_n)$ of this pair of stochastic variables. By the assumed stochastic independence of $X^{(n)}$ and $\hat{a}_n$, the expectation of this distance can be disintegrated as $\mathbb{E}H(X^{(n)}, \hat{a}_n) = \int \mathbb{E}H(X^{(n)}, a)\,dP^{\hat{a}_n}(a)$, where the expectation inside the integral is relative to $X^{(n)}$ only and concerns the "ordinary" posterior distribution relative to the prior (3.8) with $\hat{a}_n$ set equal to the deterministic function a, that is, the posterior distribution for the prior of the form $\Psi(w + \lambda a)$ on b, for a fixed function a and (w, λ) following their prior. Since the bounded Lipschitz distance is bounded, $\mathbb{E}H(X^{(n)}, \hat{a}_n)$ certainly tends to zero if for every $\eta > 0$ there exist sets $\mathcal{A}_n$ with $\Pr(\hat{a}_n \in \mathcal{A}_n) > 1 - \eta$ such that $\mathbb{E}H(X^{(n)}, a) \to 0$, uniformly in $a \in \mathcal{A}_n$.

In view of (3.10), there exist sets A_n with $\Pr(\hat{a}_n \in A_n) \to 1$ and $\mathbb{E}\Pi((w, \lambda) : w + (\lambda + tn^{-1/2})a \in \mathcal{H}_n^b|X^{(n)}) \to 1$, uniformly in $a \in A_n$ (see the lemma below for details). Since we assume that $\|\hat{a}_n\|_\infty = O_{P_0}(1)$ and (3.7), we can further reduce these sets to $\mathcal{A}_n = \{a \in A_n : \|a\|_\infty \le M, \|a - a_0\|_{L^2(F_0)} \le M\rho_n\}$, and then show that $\mathbb{E}H(X^{(n)}, a) \to 0$ uniformly in $a \in \mathcal{A}_n$, for (every) fixed $M > 0$. Thus in the remainder of the proof we fix $\hat{a}_n$ to be a deterministic sequence a_n in $\mathcal{A}_n$.

We verify the conditions of Theorem 1. By (3.9)–(3.12), conditions (3.3)–(3.5) are met by $\mathcal{H}_n = \{\eta : \eta^b = w + \lambda a_n, (w, \lambda) \in B_n\}$, for

$$B_n = \big\{(w, \lambda) : w + \lambda a_n \in \mathcal{H}_n^b, |\lambda| \le 2u_n\sigma_n^2\sqrt{n}\big\}.$$

It therefore remains only to control the change of measure (3.6). We need only consider the b part of the integrals, as the a part cancels. Because the assumptions become "more true" if u_n is replaced by a bigger sequence and $n\sigma_n^2 \to \infty$, we may assume that $u_n \to 0$ and $u_n n\sigma_n^2 \to \infty$.

For the b term, (3.6) equals

$$\frac{\int_{B_n} e^{\ell_n^b(w+\lambda a_n - ta_0/\sqrt{n})}\phi_{\sigma_n}(\lambda)\,d\lambda\,d\Pi(w)}{\int_{B_n} e^{\ell_n^b(w+\lambda a_n)}\phi_{\sigma_n}(\lambda)\,d\lambda\,d\Pi(w)}, \tag{5.3}$$

where ϕ_σ denotes the probability density function of a $N(0, \sigma^2)$ random variable. By Lemma 3, applied with $A_n = \{w + \lambda a_n : (w, \lambda) \in B_n\}$, $\xi_n = a_n$, $\xi_0 = a_0$, $\zeta_n = M\rho_n$, w_n the constant M in the definition of $\mathcal{A}_n$ and $\varepsilon_n = \varepsilon_n^b$,

$$\sup_{(w,\lambda)\in B_n}\left|\ell_n^b\left(w + \lambda a_n - \frac{t}{\sqrt{n}}a_0\right) - \ell_n^b\left(w + \left(\lambda - \frac{t}{\sqrt{n}}\right)a_n\right)\right| = o_{P_0}(1).$$

Furthermore, for $|\lambda| \le 2u_n\sigma_n^2\sqrt{n}$, we have for the log likelihood ratio of two normal densities

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

Consequently, the numerator of (5.3) equals

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

By the change of variables $\lambda - t/\sqrt{n} \rightsquigarrow \lambda'$ the ratio (5.3) therefore equals, for $B_{n,t} = \{(w, \lambda) : (w, \lambda + t/\sqrt{n}) \in B_n\}$,

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

Since $\Pi(B_n|X^{(n)}) = 1 - o_{P_0}(1)$, it remains to show that $\Pi(B_{n,t}|X^{(n)}) = 1 - o_{P_0}(1)$.

The set $B_{n,t}$ is the intersection of the sets in assumptions (3.10) (with $\hat{a}_n = a_n$) and (3.9), except that the restriction on λ in $B_{n,t}$ is $|\lambda + t/\sqrt{n}| \le 2u_n\sqrt{n}\sigma_n^2$, whereas in (3.9) the restriction is $|\lambda| \le u_n\sqrt{n}\sigma_n^2$. Since $t/\sqrt{n} \ll u_n\sqrt{n}\sigma_n^2$ by construction, the latter restriction implies the former, and hence $\Pi(B_{n,t}|X^{(n)}) = 1 - o_{P_0}(1)$ by assumption. □

LEMMA 2. *For given v define $A_n(v)$ to be the set of all a such that $\mathbb{E}\Pi((w, \lambda) : w + (\lambda + tn^{-1/2})a \in \mathcal{H}_n^b|X^{(n)}) > 1 - v$. If (3.10) holds, then there exists $v_n \downarrow 0$ such that $\Pr(\hat{a}_n \in A_n(v_n)) \to 1$.*

PROOF. For given a and x, define

$$G_n(a, x) = \Pi\big((w, \lambda) : w + \big(\lambda + tn^{-1/2}\big)a \in \mathcal{H}_n^b|X^{(n)} = x\big).$$

Then the given expectation is $H_n(a) := \mathbb{E}G_n(a, X^{(n)})$ and $A_n(v) = \{a : H_n(a) > 1 - v\}$. By (3.10), the dominated convergence theorem and the independence of $\hat{a}_n$ and $X^{(n)}$, we have $\mathbb{E}H_n(\hat{a}_n) = \mathbb{E}G_n(\hat{a}_n, X^{(n)}) \to 1$. Since $0 \le H_n(a) \le 1$, this implies that $H_n(\hat{a}_n) \to^P 1$. Then $\Pr(H_n(\hat{a}_n) > 1 - v_n) \to 1$, for $v_n \downarrow 0$ sufficiently slowly by a standard argument. □

5.3. *Proofs for Section 3.4: Gaussian process priors.* PROOF OF PROPOSITION 1. We verify the conditions of Theorem 1. By Lemma 16 with norm $\|\cdot\|_\infty$, the posterior distribution of b contracts about b_0 at rate ε_n^b in $L^2(F_0)$. For $\mathcal{H}_n^b$ the sets as in the statement of the proposition, define

$$\mathcal{H}_n = \{(\eta^a, \eta^b) : \eta^b \in \mathcal{H}_n^b, \|\Psi(\eta^b) - b_0\|_{L^2} \le \varepsilon_n^b\}.$$

Then $\Pi(\mathcal{H}_n|X^{(n)}) \to^{P_0} 1$ as $n \to \infty$, by assumption. It follows that $\mathcal{H}_n$ satisfies conditions (3.3)–(3.4), while (3.5) is satisfied by assumption.

It remains to verify (3.6). Following [10], we first approximate the perturbation η_t^b by an element in the RKHS $\mathbb{H}^b$ and then apply the Cameron–Martin theorem. Let $\xi_n^b \in \mathbb{H}^b$ satisfy (3.18), and set $\eta_{n,t} = \eta_{n,t}(\eta^b) = \eta^b - t\xi_n^b/\sqrt{n}$. By the Cameron–Martin theorem (see Lemma 13), the distribution $\Pi_{n,t}$ of $\eta_{n,t}$ if η^b is distributed according to the prior Π has Radon–Nikodym density

$$\frac{d\Pi_{n,t}}{d\Pi}(\eta^b) = e^{tU_n(\eta^b)/\sqrt{n} - t^2\|\xi_n^b\|_{\mathbb{H}^b}^2/(2n)},$$

where $U_n(\eta^b)$ is a centered Gaussian variable with variance $\|\xi_n^b\|_{\mathbb{H}^b}^2$ if $\eta^b \sim \Pi$, and $\|\cdot\|_{\mathbb{H}^b}$ is the RKHS norm of the Gaussian process η^b. By the univariate Gaussian tail bound,

$$\Pi(\eta^b : |U_n(\eta^b)| > M\sqrt{n}\varepsilon_n^b\|\xi_n^b\|_{\mathbb{H}^b}) \le 2e^{-M^2n(\varepsilon_n^b)^2/2}. \tag{5.4}$$

Consequently, by Lemma 4 the posterior measure of the set in the display tends to 0 in probability, for large enough M. Hence the sets

$$B_n = \{\eta^b : |U_n(\eta^b)| \le M\sqrt{n}\varepsilon_n^b\|\xi_n^b\|_{\mathbb{H}^b}\} \cap \mathcal{H}_n^b$$

also satisfy $\Pi(B_n|X^{(n)}) \to 1$ in probability. On the sets B_n, in view of (3.18),

$$\left|\log\frac{d\Pi_{n,t}}{d\Pi}(\eta^b)\right| \le M|t|\sqrt{n}\varepsilon_n^b\zeta_n^b + \frac{t^2}{2}(\zeta_n^b)^2 \to 0. \tag{5.5}$$

Furthermore, by Lemma 3 applied with $A_n = B_n$, $\xi_0 = \xi_{\eta_0}^b$, $\varepsilon_n = \varepsilon_n^b$, $\zeta_n = \zeta_n^b$ and w_n a sufficiently large fixed constant, we have

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

(Note that condition (7.1) holds by assumption (3.12) and Lemma 10.) By the last display followed by the change of integration variable $\eta^b - t\xi_n^b/\sqrt{n} \rightsquigarrow v$,

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

where $B_{n,t} = B_n - t\xi_n^b/\sqrt{n}$. By (5.5), we can next replace $\Pi_{n,t}$ in the numerator by Π at the cost of another multiplicative $1 + o_{P_0}(1)$ term. This turns the quotient into the ratio $\Pi(B_{n,t}|X^{(n)})/\Pi(B_n|X^{(n)})$. We have already shown that $\Pi(B_n|X^{(n)}) = 1 - o_{P_0}(1)$, so it suffices to show the same result holds true for the numerator. Now

$$\begin{aligned} B_{n,t}^c = {} & \{v : v + t\xi_n^b/\sqrt{n} \notin \mathcal{H}_n^b\} \cup \{v : \|\Psi(v + t\xi_n^b/\sqrt{n}) - b_0\|_{L^2(F_0)} > \varepsilon_n^b\} \\ & \cup \{v : |U_n(v + t\xi_n^b/\sqrt{n})| > M\sqrt{n}\varepsilon_n^b\|\xi_n^b\|_{\mathbb{H}^b}\}. \end{aligned}$$

The posterior probability of the first set tends to zero in probability by assumption. Since $\|\Psi(\eta^b + t\xi_n^b/\sqrt{n}) - \Psi(\eta^b)\|_{L^2(F_0)} \lesssim \|\xi_n^b/\sqrt{n}\|_{L^2(F_0)} \lesssim 1/\sqrt{n}$, the second set is contained in $\{\eta^b : \|\Psi(\eta^b) - b_0\|_{L^2(F_0)} > \varepsilon_n^b - C/\sqrt{n}\}$, which has posterior probability $o_{P_0}(1)$ by

Lemma 16, possibly after replacing ε_n^b by a multiple of itself. For the third set, we use that $U_n(\eta^b + t\xi_n^b/\sqrt{n}) \sim N(-t\|\xi_n^b\|_{\mathbb{H}}^2/\sqrt{n}, \|\xi_n^b\|_{\mathbb{H}}^2)$ if η^b is distributed according to the prior, by Lemma 13. Since the mean $t\|\xi_n^b\|_{\mathbb{H}}^2/\sqrt{n}$ of this variable is negligible relative to its standard deviation, $\Pi(|U_n(\eta^b + t\xi_n^b/\sqrt{n})| > M\sqrt{n}\varepsilon_n^b\|\xi_n^b\|_{\mathbb{H}^b})$ differs not substantially from the left-hand side of (5.4), whence it is also exponentially small, so that again Lemma 4 applies to see that the posterior probability tends to zero. □

PROOF OF COROLLARY 1. The proof follows by verifying the conditions of Proposition 1, separately for the two prior processes.

Series prior (3.15): Using the form of the concentration function in the proof of Theorem 4.5 of [49], we see that (3.17) is satisfied for

$$\varepsilon_n^b = n^{-\frac{\beta\wedge\bar\beta}{2\bar\beta+d}}\log n.$$

Condition (3.5) is verified in Lemma 6, under the assumption $\beta \wedge \bar\beta > d/2$.

It thus remains only to establish (3.18), the approximation by elements of the RKHS. Write $J = J_{\bar\beta}$ and define $V_J = \operatorname{span}(\psi_{jk} : j \le J, k)$. Recall that the RKHS of the Gaussian series prior (3.15) equals

$$\mathbb{H}^b = \Big\{ w \in V_J : \|w\|_{\mathbb{H}^b}^2 := \sum_{j\le J}\sum_k \sigma_j^{-2}\big|\langle w, \psi_{jk}\rangle_{L^2}\big|^2 < \infty \Big\}. \tag{5.6}$$

From the computations in Theorem 4.5 of [49], one gets that for $\xi_{\eta_0}^b = a_0 \in C^\alpha$ and any $\zeta_n^b \gtrsim n^{-\alpha/(2\bar\beta+d)}$,

$$\inf_{\xi:\|\xi-a_0\|_\infty\le\zeta_n^b}\|\xi\|_{\mathbb{H}^b} \lesssim \begin{cases} (\zeta_n^b)^{-\frac{r-\alpha+d/2}{\alpha}\wedge 0} & \text{if } r-\alpha+d/2 \ne 0, \\ \log(1/\zeta_n^b) & \text{if } r-\alpha+d/2 = 0. \end{cases} \tag{5.7}$$

If follows that (3.18) is satisfied if we can choose $\zeta_n^b \to 0$ so that the right-hand side of the display is bounded above by $\sqrt{n}\zeta_n^b$ and $\sqrt{n}\varepsilon_n^b\zeta_n^b \to 0$.

- If $r-\alpha+d/2 > 0$, then (5.7) is bounded by $\sqrt{n}\zeta_n^b$ for $\zeta_n^b \gtrsim n^{-\alpha/(2r+d)}$. Since we also require $\zeta_n^b \gtrsim n^{-\alpha/(2\bar\beta+d)}$, we may take $\zeta_n^b \sim n^{-\alpha/(2\bar\beta+d)} \vee n^{-\alpha/(2r+d)} = n^{-\alpha/(2\bar\beta+d)}$ since $r \le \beta\wedge\bar\beta$ by assumption. Then $\sqrt{n}\varepsilon_n^b\zeta_n^b \to 0$ for $\beta\wedge\bar\beta > d/2 + \bar\beta - \alpha$.
- If $r-\alpha+d/2 < 0$, then (5.7) is bounded by $\sqrt{n}\zeta_n^b$ and also $\zeta_n^b \gtrsim n^{-\alpha/(2\bar\beta+d)}$ for the choice $\zeta_n^b \sim n^{-1/2} \vee n^{-\alpha/(2\bar\beta+d)}$. If $1/2 \le \alpha/(2\bar\beta+d)$, then $\sqrt{n}\varepsilon_n^b\zeta_n^b \sim \varepsilon_n^b \to 0$. If $1/2 > \alpha/(2\bar\beta+d)$, then $\sqrt{n}\varepsilon_n^b\zeta_n^b \sim n^{(\bar\beta+d/2-\beta\wedge\bar\beta-\alpha)/(2\bar\beta+d)}(\log n) \to 0$ for $\beta\wedge\bar\beta > d/2 + \bar\beta - \alpha$.
- If $r-\alpha+d/2 = 0$, then one takes $\zeta_n \sim [(\log n)^{1/2}n^{-1/2}] \vee n^{-\alpha/(2\bar\beta+d)}$. This is the same as the previous case apart from the extra logarithmic factor, so $\sqrt{n}\varepsilon_n^b\zeta_n^b \to 0$ under exactly the same conditions.

Examining all the cases, the above can be summarized as (3.18) holds if $\beta\wedge\bar\beta > [\bar\beta - \alpha + d/2] \vee 0$. Together with the condition $\beta\wedge\bar\beta > d/2$ needed to verify (3.5) above, this is equivalent to $\alpha, \beta > d/2$ and $d/2 < \bar\beta < \alpha + \beta - d/2$.

Riemann–Liouville prior (3.14): The proof follows in much the same way. Using the form of the concentration function in Theorem 4 of Castillo [9], we see that (3.17) is satisfied for

$$\varepsilon_n^b = n^{-\frac{\beta\wedge\bar\beta}{2\bar\beta+1}}(\log n)^\kappa,$$

where κ is function of $(\beta, \bar\beta)$, given explicitly in [9]. Condition (3.5) is verified in Lemma 5, under the assumption $\beta\wedge\bar\beta > 1/2$.

It thus remains to establish (3.18). Recall that the RKHS of the Riemann–Liouville process is the Sobolev space $H^{\bar\beta+1/2}$. From the computations in Theorem 4 of [9], one gets that for $\xi^b_{\eta_0} = a_0 \in C^\alpha$, as $\zeta^b_n \to 0$,

$$\inf_{\xi:\|\xi-a_0\|_\infty\le\zeta^b_n} \|\xi\|_{\mathbb{H}^b} \lesssim (\zeta^b_n)^{-\frac{\bar\beta-\alpha+1/2}{\alpha}\wedge 0}.$$

If follows that (3.18) is satisfied if we can choose ζ^b_n so that the right-hand side of the display is bounded above by $\sqrt{n}\zeta^b_n$ and $\sqrt{n}\varepsilon^b_n\zeta^b_n \to 0$. If $\bar\beta \le \alpha - 1/2$, simply set $\xi^b_n = \xi^b_{\eta_0}$ and $\zeta^b_n = n^{-1/2}\|\xi^b_{\eta_0}\|_{\mathbb{H}^b}$. If $\bar\beta > \alpha - 1/2$, take $\zeta^b_n = n^{-\frac{\alpha}{2\bar\beta+1}}$, so that $\sqrt{n}\varepsilon^b_n\zeta^b_n \to 0$ for $\beta\wedge\bar\beta > 1/2 + \bar\beta - \alpha$. A careful analysis of all cases shows that these inequalities, together with the requirement $\beta\wedge\bar\beta > 1/2$, are equivalent to $\alpha, \beta > 1/2$ and $1/2 < \bar\beta < \alpha + \beta - 1/2$. □

PROOF OF COROLLARY 2. We verify the conditions of Theorem 2, where we replace $\hat a_n$ by a deterministic sequence with $\|a_n\|_\infty = O(1)$ as explained in the proof of Theorem 2. Since $\varepsilon^b_n = n^{-(\beta\wedge\bar\beta)/(2\bar\beta+d)}(\log n)^\kappa$ solves (3.17) (see proof of Corollary 1), the contraction rate follows from Lemma 17. Together with Lemmas 7 and 8 for the Riemann–Liouville and series priors, respectively, this verifies conditions (3.10)–(3.12). To verify (3.9), we use the Gaussian tail inequality to see that $\Pi(|\lambda| \ge u_n\sigma_n\sqrt{n}) \le 2e^{-u_n^2 n\sigma_n^2/2}$. This is bounded above by $e^{-Ln(\varepsilon^b_n)^2}$ for $u_n \to 0$ sufficiently slowly, since $\varepsilon^b_n = o(\sigma_n)$ by assumption. Lemma 4 now implies (3.9). □

Acknowledgements. We would like to thank two referees for helpful comments and for drawing several references to our attention. The first author would also like to thank Richard Nickl for helpful conversations on symmetrization. Most of this work was done while Kolyan Ray was at Leiden University and King's College London.

The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric Bayesian causal inference" (DOI: 10.1214/19-AOS1919SUPP; .pdf). In the supplement, we present an additional theorem, putting a general prior on (a, b, f), and we provide the missing proofs. We linearly continue the numbering scheme for sections, lemmas, etc., from the main document in the supplement, and items referred to which do not appear in the main article can be found in the supplement (e.g., Lemma 3).

REFERENCES

[1] ALAA, A. and VAN DER SCHAAR, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *Proceedings of the* 35*th International Conference on Machine Learning* 129–138.

[2] ALAA, A. M. and VAN DER SCHAAR, M. (2017). Deep multi-task Gaussian processes for survival analysis with competing risks. In *Advances in Neural Information Processing Systems* 30 2329–2337.

[3] ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

[4] ATHEY, S., IMBENS, G., PHAM, T. and WAGER, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *Am. Econ. Rev.* **107** 278–281. https://doi.org/10.1257/aer.p20171042

[5] BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **11** 432–452. MR0696057 https://doi.org/10.1214/aos/1176346151

[6] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

[7] BICKEL, P. J. and KLEIJN, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *Ann. Statist.* **40** 206–237. MR3013185 https://doi.org/10.1214/11-AOS921

[8] BICKEL, P. J. and KWON, J. (2001). Inference for semiparametric models: Some questions and an answer. *Statist. Sinica* **11** 863–960. MR1867326

[9] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287 https://doi.org/10.1214/08-EJS273

[10] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753 https://doi.org/10.1007/s00440-010-0316-5

[11] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

[12] DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358 https://doi.org/10.1017/CBO9780511755347

[13] FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629. MR0438568

[14] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119 https://doi.org/10.1214/aos/1017938917

[15] FUTOMA, J., HARIHARAN, S. and HELLER, K. (2017). Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *Proceedings of the 34th International Conference on Machine Learning* 1174–1182.

[16] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[17] GINÉ, E. and NICKL, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395 https://doi.org/10.1214/11-AOS924

[18] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[19] HAHN, P. R., CARVALHO, C. M., PUELZ, D. and HE, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* **13** 163–182. MR3737947 https://doi.org/10.1214/16-BA1044

[20] HAHN, P. R., MURRAY, J. S. and CARVALHO, C. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Available at arXiv:1706.09523.

[21] HECKMAN, J. J., LOPES, H. F. and PIATEK, R. (2014). Treatment effects: A Bayesian perspective. *Econometric Rev.* **33** 36–67. MR3170840 https://doi.org/10.1080/07474938.2013.807103

[22] HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

[23] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences*: *An introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[24] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386

[25] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301

[26] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[27] RAY, K. and SZABÓ, B. T. (2019). Debiased Bayesian inference for average treatment effects. In *Advances in Neural Information Processing Systems* 33 11952–11962.

[28] RAY, K. and VAN DER VAART, A. (2020). Supplement to "Semiparametric Bayesian causal inference." https://doi.org/10.1214/19-AOS1919SUPP

[29] RAY, K. and VAN DER VAART, A. W. (2020). On the Bernstein–von Mises theorem for the Dirichlet process. Available at arXiv:2008.01130.

[30] RITOV, Y., BICKEL, P. J., GAMST, A. C. and KLEIJN, B. J. K. (2014). The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statist. Sci.* **29** 619–639. MR3300362 https://doi.org/10.1214/14-STS483

[31] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033 https://doi.org/10.1214/12-AOS1004

[32] ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758 https://doi.org/10.1016/0270-0255(86)90088-6

[33] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman*. *Inst. Math. Stat.* (*IMS*) *Collect.* **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527

[34] ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

[35] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

[36] ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat. Med.* **16** 285–319.

[37] ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

[38] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

[39] ROTNITZKY, A. and ROBINS, J. M. (1995). Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand. J. Stat.* **22** 323–333. MR1363216

[40] RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

[41] SEAMAN, S. R. and VANSTEELANDT, S. (2018). Introduction to double robust methods for incomplete data. *Statist. Sci.* **33** 184–197. MR3797709 https://doi.org/10.1214/18-STS647

[42] TADDY, M., GARDNER, M., CHEN, L. and DRAPER, D. (2016). A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *J. Bus. Econom. Statist.* **34** 661–672. MR3548002 https://doi.org/10.1080/07350015.2016.1172013

[43] VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845 https://doi.org/10.1214/aos/1176347976

[44] VAN DER VAART, A. (2014). Higher order tangent spaces and influence functions. *Statist. Sci.* **29** 679–686. MR3300365 https://doi.org/10.1214/14-STS478

[45] VAN DER VAART, A. and VAN ZANTEN, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448. MR2357712 https://doi.org/10.1214/07-EJS098

[46] VAN DER VAART, A. and VAN ZANTEN, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12** 2095–2119. MR2819028

[47] VAN DER VAART, A. and WELLNER, J. A. (2000). Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli classes. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 115–133. Birkhäuser, Boston, MA. MR1857319

[48] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[49] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

[50] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh*. *Inst. Math. Stat.* (*IMS*) *Collect.* **3** 200–222. IMS, Beachwood, OH. MR2459226 https://doi.org/10.1214/074921708000000156

[51] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442 https://doi.org/10.1214/08-AOS678

[52] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[53] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

[54] ZIGLER, C. M. and DOMINICI, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *J. Amer. Statist. Assoc.* **109** 95–107. MR3180549 https://doi.org/10.1080/01621459.2013.869498