## The semiparametric Bernstein-von Mises theorem

Bickel, P.J.; Kleijn, B.J.K.

**DOI**
10.1214/11-AOS921
**Publication date**
2012
**Document Version**
Final published version
**Published in**
The Annals of Statistics

*The Annals of Statistics*
2012, Vol. 40, No. 1, 206–237
DOI: 10.1214/11-AOS921


# THE SEMIPARAMETRIC BERNSTEIN–VON MISES THEOREM

BY P. J. BICKEL AND B. J. K. KLEIJN[1]

*University of California, Berkeley and University of Amsterdam*

*Dedicated to the memory of David A. Freedman*

In a smooth semiparametric estimation problem, the marginal posterior for the parameter of interest is expected to be asymptotically normal and satisfy frequentist criteria of optimality if the model is endowed with a suitable prior. It is shown that, under certain straightforward and interpretable conditions, the assertion of Le Cam's acclaimed, but strictly parametric, Bernstein–von Mises theorem [*Univ. California Publ. Statist.* **1** (1953) 277–329] holds in the semiparametric situation as well. As a consequence, Bayesian point-estimators achieve efficiency, for example, in the sense of Hájek's convolution theorem [*Z. Wahrsch. Verw. Gebiete* **14** (1970) 323–330]. The model is required to satisfy differentiability and metric entropy conditions, while the nuisance prior must assign nonzero mass to certain Kullback–Leibler neighborhoods [Ghosal, Ghosh and van der Vaart *Ann. Statist.* **28** (2000) 500–531]. In addition, the marginal posterior is required to converge at parametric rate, which appears to be the most stringent condition in examples. The results are applied to estimation of the linear coefficient in partial linear regression, with a Gaussian prior on a smoothness class for the nuisance.

**1. Introduction.** The concept of efficiency has its origin in Fisher's 1920s claim of asymptotic optimality of the maximum-likelihood estimator in differentiable parametric models (Fisher [13]). In 1930s and 1940s, Fisher's ideas on optimality in differentiable models were sharpened and elaborated upon (see, e.g., Cramér [10]), until Hodges's 1951 discovery of a superefficient estimator indicated that a comprehensive understanding of optimality in differentiable estimation problems remained elusive. Further consideration directed attention to the property of *regularity* to delimit the class of estimators over which optimality is achieved. Hájek's convolution theorem (Hájek [17]) implies that within the class of regular estimates, asymptotic variance is lower-bounded by the Cramér–Rao bound in the limit experiment [29]. The asymptotic minimax theorem (Hájek [18]) underlines the central role of the concept of regularity. An estimator that is optimal among

Received October 2010; revised August 2011.

[1]Supported by a VENI-grant, Netherlands Organisation for Scientific Research (NWO).

*MSC2010 subject classifications.* Primary 62G86; secondary 62G20, 62F15.

*Key words and phrases.* Asymptotic posterior normality, posterior limit distribution, model differentiability, local asymptotic normality, semiparametric statistics, regular estimation, efficiency, Bernstein–Von Mises.

regular estimates is called *best-regular*; in a Hellinger differentiable model, an estimator $(\hat{\theta}_n)$ for $\theta$ is best-regular *if and only if* it is asymptotically linear, that is, for all $\theta$ in the model,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_\theta^{-1} \dot{\ell}_\theta(X_i) + o_{P_\theta}(1), \tag{1.1}$$

where $\dot{\ell}_\theta$ is the score for $\theta$ and $I_\theta$ the corresponding Fisher information. To address the question of efficiency in smooth parametric models from a Bayesian perspective, we turn to the Bernstein–von Mises theorem. In the literature many different versions of the theorem exist, varying both in (stringency of) conditions and (strength or) form of the assertion. Following Le Cam and Yang [31] (see also van der Vaart [43]), we state the theorem as follows. (For later reference, define a prior to be *thick* at $\theta_0$, if it has a Lebesgue density that is continuous and strictly positive at $\theta_0$.)

THEOREM 1.1 (Bernstein–von Mises, parametric). *Assume that $\Theta \subset \mathbb{R}^k$ is open and that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable and dominated. Suppose $X_1, X_2, \ldots$ forms an i.i.d. sample from $P_{\theta_0}$ for some $\theta_0 \in \Theta$. Assume that the model is locally asymptotically normal at $\theta_0$ with nonsingular Fisher information $I_{\theta_0}$. Furthermore, suppose that*:

(i) *the prior $\Pi_\Theta$ is thick at $\theta_0$;*
(ii) *for every $\varepsilon > 0$, there exists a test sequence $(\phi_n)$ such that*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta - \theta_0\| > \varepsilon} P_\theta^n (1 - \phi_n) \to 0.$$

*Then the posterior distributions converge in total variation,*

$$\sup_B \left| \Pi(\theta \in B \mid X_1, \ldots, X_n) - N_{\hat{\theta}_n, (n I_{\theta_0})^{-1}}(B) \right| \to 0$$

*in $P_{\theta_0}$-probability, where $(\hat{\theta}_n)$ denotes any best-regular estimator sequence.*

For a proof, the reader is referred to [31, 43] (or to Kleijn and van der Vaart [26], for a proof under model misspecification that has a lot in common with the proof of Theorem 5.1 below).

Neither the frequentist theory on asymptotic optimality nor Theorem 1.1 generalize fully to nonparametric estimation problems. Examples of the failure of the Bernstein–von Mises limit in infinite-dimensional problems (with regard to the *full* parameter) can be found in Freedman [14]. Freedman initiated a discussion concerning the merits of Bayesian methods in nonparametric problems as early as 1963, showing that even with a natural and seemingly innocuous choice of the nonparametric prior, posterior inconsistency may result [15]. This warning against instances of inconsistency due to ill-advised nonparametric priors was reiterated

in the literature many times over, for example, in Cox [9] and in Diaconis and Freedman [11, 12]. However, general conditions for Bayesian consistency were formulated by Schwartz as early as 1965 [37]; positive results on posterior rates of convergence in the same spirit were obtained in Ghosal, Ghosh and van der Vaart [16] (see also, Shen and Wasserman [40]). The combined message of negative and positive results appears to be that the choice of a nonparametric prior is a sensitive one that leaves room for unintended consequences unless due care is taken.

This lesson must also be taken seriously when one asks the question whether the posterior for the parameter of interest in a semiparametric estimation problem displays Bernstein–von Mises-type limiting behavior. Like in the parametric case, we estimate a finite-dimensional parameter $\theta \in \Theta$, but now in a model $\mathscr{P}$ that also leaves room for an infinite-dimensional nuisance parameter $\eta \in H$. We look for general sufficient conditions on model and prior such that the *marginal posterior for the parameter of interest* satisfies

$$\sup_B \bigl|\Pi\bigl(\sqrt{n}(\theta-\theta_0)\in B \mid X_1,\ldots,X_n\bigr) - N_{\tilde{\Delta}_n,\tilde{I}_{\theta_0,\eta_0}^{-1}}(B)\bigr| \to 0 \tag{1.2}$$

in $P_{\theta_0}$-probability, where

$$\tilde{\Delta}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{I}_{\theta_0,\eta_0}^{-1}\tilde{\ell}_{\theta_0,\eta_0}(X_i). \tag{1.3}$$

Here $\tilde{\ell}_{\theta,\eta}$ denotes the efficient score function and $\tilde{I}_{\theta,\eta}$ the efficient Fisher information [assumed to be nonsingular at $(\theta_0,\eta_0)$]. The sequence $\tilde{\Delta}_n$ also features on the r.h.s. of the semiparametric version of (1.1) (see Lemma 25.23 in [43]). Assertion (1.2) often implies efficiency of point-estimators like the posterior median, mode or mean (a first condition being that the estimate is a functional on $\mathbb{R}$, continuous in total-variation [24, 43]) and always leads to asymptotic identification of credible regions with efficient confidence regions. To illustrate, if $C$ is a credible set in $\Theta$, (1.2) guarantees that posterior coverage and coverage under the limiting normal for $C$ are (close to) equal. Because the limiting normals are *also* the asymptotic sampling distributions for efficient point-estimators, (1.2) enables interpretation of credible sets as asymptotically efficient confidence regions. From a practical point of view, the latter conclusion has an important implication: whereas it can be hard to compute optimal semiparametric confidence regions directly, simulation of a large sample from the marginal posterior (e.g., by MCMC techniques; see Robert [36]) is sometimes comparatively straightforward.

Instances of the Bernstein–von Mises limit have been studied in various semiparametric models: several papers have provided studies of asymptotic normality of posterior distributions for models from survival analysis. Particularly, Kim and Lee [22] show that the *infinite-dimensional* posterior for the cumulative hazard function under right-censoring converges at rate $n^{-1/2}$ to a Gaussian centered at the Aalen–Nelson estimator for a class of neutral-to-the-right process priors. In

Kim [21], the posterior for the baseline cumulative hazard function and regression coefficients in Cox's proportional hazard model are considered with similar priors. Castillo [6] considers marginal posteriors in Cox's proportional hazards model and Stein's symmetric location problem from a unified point of view. A general approach has been given in Shen [39], but his conditions may prove somewhat hard to verify in examples. Cheng and Kosorok [8] give a general perspective too, proving weak convergence of the posterior under sufficient conditions. Rivoirard and Rousseau [35] prove a version for linear functionals over the model, using a class of nonparametric priors based on infinite-dimensional exponential families. Boucheron and Gassiat [4] consider the Bernstein–von Mises theorem for families of discrete distributions. Johnstone [20] studies various marginal posteriors in the Gaussian sequence model.

*Notation and conventions.* The (frequentist) true distribution of the data is denoted $P_0$ and assumed to lie in $\mathscr{P}$, so that there exist $\theta_0 \in \Theta$, $\eta_0 \in H$ such that $P_0 = P_{\theta_0,\eta_0}$. We localize $\theta$ by introducing $h = \sqrt{n}(\theta - \theta_0)$ with inverse $\theta_n(h) = \theta_0 + n^{-1/2}h$. The expectation of a random variable $f$ with respect to a probability measure $P$ is denoted $Pf$; the sample average of $g(X)$ is denoted $\mathbb{P}_n g(X) = (1/n)\sum_{i=1}^n g(X_i)$ and $\mathbb{G}_n g(X) = n^{1/2}(\mathbb{P}_n g(X) - Pg(X))$ (for other conventions and nomenclature customary in empirical process theory, see [45]). If $h_n$ is stochastic, $P^n_{\theta_n(h_n),\eta} f$ denotes the integral $\int f(\omega)(dP^n_{\theta_n(h_n(\omega)),\eta}/dP^n_0)(\omega)\,dP^n_0(\omega)$. The Hellinger distance between $P$ and $P'$ is denoted $H(P, P')$ and induces a metric $d_H$ on the space of nuisance parameters $H$ by $d_H(\eta, \eta') = H(P_{\theta_0,\eta}, P_{\theta_0,\eta'})$, for all $\eta, \eta' \in H$. We endow the model with the Borel $\sigma$-algebra generated by the Hellinger topology and refer to [16] regarding issues of measurability.

**2. Main results.** Consider estimation of a functional $\theta : \mathscr{P} \to \mathbb{R}^k$ on a dominated nonparametric model $\mathscr{P}$ with metric $g$, based on a sample $X_1, X_2, \ldots$, i.i.d. according to $P_0 \in \mathscr{P}$. We introduce a prior $\Pi$ on $\mathscr{P}$ and consider the subsequent sequence of posteriors,

$$\Pi(A \mid X_1, \ldots, X_n) = \int_A \prod_{i=1}^n p(X_i)\,d\Pi(P) \Big/ \int_{\mathscr{P}} \prod_{i=1}^n p(X_i)\,d\Pi(P), \tag{2.1}$$

where $A$ is any measurable model subset. Typically, optimal (e.g., minimax) nonparametric posterior rates of convergence [16] are powers of $n$ (possibly modified by a slowly varying function) that converge to zero more slowly than the parametric $n^{-1/2}$-rate. Estimators for $\theta$ may be derived by "plugging in" a nonparametric estimate [cf. $\hat{\theta} = \theta(\hat{P})$], but optimality in rate or asymptotic variance cannot be expected to obtain generically in this way. This does not preclude efficient estimation of real-valued aspects of $P_0$: parametrize the model in terms of a finite-dimensional *parameter of interest* $\theta \in \Theta$ and a *nuisance parameter* $\eta \in H$ where $\Theta$ is open in $\mathbb{R}^k$ and $(H, d_H)$ an infinite-dimensional metric space:

$\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$. Assuming identifiability, there exist unique $\theta_0 \in \Theta$, $\eta_0 \in H$ such that $P_0 = P_{\theta_0,\eta_0}$. Assuming measurability of the map $(\theta, \eta) \mapsto P_{\theta,\eta}$, we place a product prior $\Pi_\Theta \times \Pi_H$ on $\Theta \times H$ to define a prior on $\mathscr{P}$. Parametric rates for the marginal posterior of $\theta$ are achievable because it is possible for contraction of the full posterior to occur anisotropically, that is, at rate $n^{-1/2}$ along the $\theta$-direction, but at a slower, nonparametric rate $(\rho_n)$ along the $\eta$-directions.

2.1. *Method of proof.* The proof of (1.2) will consist of three steps: in Section 3, we show that the posterior concentrates its mass around so-called *least-favorable submodels* (see Stein [42] and [1, 43]). In the second step (see Section 4), we show that this implies local asymptotic normality (LAN) for integrals of the likelihood over $H$, with the efficient score determining the expansion. In Section 5, it is shown that these LAN integrals induce asymptotic normality of the marginal posterior, analogous to the way local asymptotic normality of parametric likelihoods induces the parametric Bernstein–von Mises theorem.

To see why asymptotic accumulation of posterior mass occurs around so-called least-favorable submodels, a crude argument departs from the observation that, according to (2.1), posterior concentration occurs in regions of the model with relatively high (log-)likelihood (barring inhomogeneities of the prior). Asymptotically, such regions are characterized by close-to-minimal Kullback–Leibler divergence with respect to $P_0$. To exploit this, let us assume that for each $\theta$ in a neighborhood $U_0$ of $\theta_0$, there exists a unique minimizer $\eta^*(\theta)$ of the Kullback–Leibler divergence,

$$-P_0 \log \frac{p_{\theta,\eta^*(\theta)}}{p_{\theta_0,\eta_0}} = \inf_{\eta \in H} \left( -P_0 \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}} \right) \tag{2.2}$$

giving rise to a submodel $\mathscr{P}^* = \{P^*_\theta = P_{\theta,\eta^*(\theta)} : \theta \in U_0\}$. As is well known [38], if $\mathscr{P}^*$ is smooth it constitutes a least-favorable submodel and scores along $\mathscr{P}^*$ are efficient. [In subsequent sections it is not required that $\mathscr{P}^*$ is defined by (2.2), only that $\mathscr{P}^*$ is least-favorable.] Neighborhoods of $\mathscr{P}^*$ are described with Hellinger balls in $H$ of radius $\rho > 0$ around $\eta^*(\theta)$, for all $\theta \in U_0$,

$$D(\theta, \rho) = \{\eta \in H : d_H(\eta, \eta^*(\theta)) < \rho\}. \tag{2.3}$$

To give a more precise argument for posterior concentration around $\eta^*(\theta)$, consider the posterior for $\eta$, *given* $\theta \in U_0$; unless $\theta$ happens to be equal to $\theta_0$, the submodel $\mathscr{P}_\theta = \{P_{\theta,\eta} : \eta \in H\}$ is misspecified. Kleijn and van der Vaart [27] show that the misspecified posterior concentrates asymptotically in any (Hellinger) neighborhood of the point of minimal Kullback–Leibler divergence with respect to the true distribution of the data. Applied to $\mathscr{P}_\theta$, we see that $D(\theta, \rho)$ receives asymptotic posterior probability one for any $\rho > 0$. For posterior concentration to occur [16, 27] sufficient prior mass must be present in certain Kullback–Leibler-type neigh-

borhoods. In the present context, these neighborhoods can be defined as

$$K_n(\rho, M) = \Biggl\{\eta \in H : P_0\Biggl(\sup_{\|h\|\le M} -\log\frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}}\Biggr) \le \rho^2, \tag{2.4}$$
$$P_0\Biggl(\sup_{\|h\|\le M} -\log\frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}}\Biggr)^2 \le \rho^2\Biggr\}$$

for $\rho > 0$ and $M > 0$. If this type of posterior convergence occurs with an appropriate form of uniformity over the relevant values of $\theta$ (see "consistency under perturbation," Section 3), one expects that the nonparametric posterior contracts into Hellinger neighborhoods of the curve $\theta \mapsto (\theta, \eta^*(\theta))$ (Theorem 3.1 and Corollary 3.3).

To introduce the second step, consider (2.1) with $A = B \times H$ for some measurable $B \subset \Theta$. Since the prior is of product form, $\Pi = \Pi_\Theta \times \Pi_H$, the marginal posterior for the parameter $\theta \in \Theta$ depends on the nuisance factor only through the integrated likelihood ratio,

$$S_n : \Theta \to \mathbb{R} : \theta \mapsto \int_H \prod_{i=1}^n \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta), \tag{2.5}$$

where we have introduced factors $p_{\theta_0,\eta_0}(X_i)$ in the denominator for later convenience; see (5.1). [The localized version of (2.5) is denoted $h \mapsto s_n(h)$; see (4.1).] The map $S_n$ is to be viewed in a role similar to that of the *profile likelihood* in semiparametric maximum-likelihood methods (see, e.g., Severini and Wong [38] and Murphy and van der Vaart [34]), in the sense that $S_n$ embodies the intermediate stage between nonparametric and semiparametric steps of the estimation procedure.

We impose smoothness through a form of Le Cam's local asymptotic normality: let $P \in \mathscr{P}$ be given, and let $t \mapsto P_t$ be a one-dimensional submodel of $\mathscr{P}$ such that $P_{t=0} = P$. Specializing to i.i.d. observations, we say that the model is *stochastically LAN* at $P \in \mathscr{P}$ along the direction $t \mapsto P_t$, if there exists an $L_2(P)$-function $g_P$ with $Pg_P = 0$ such that for all random sequences $(h_n)$ bounded in $P$-probability,

$$\log\prod_{i=1}^n \frac{p_{n^{-1/2}h_n}}{p}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h_n^T g_P(X_i) - \frac{1}{2}h_n^T I_P h_n + o_P(1). \tag{2.6}$$

Here $g_P$ is the score-function, and $I_P = P(g_P)^2$ is the Fisher information of the submodel at $P$. Stochastic LAN is slightly stronger than the usual LAN property [28, 31]. In examples, the proof of the ordinary LAN property often extends to stochastic LAN without significant difficulties.

Although formally only a convenience, the presentation benefits from an *adaptive* reparametrization (see Section 2.4 of Bickel et al. [1]): based on the least-favorable submodel $\eta^*$, we define, for all $\theta \in U_0$, $\eta \in H$,

$$(\theta, \eta(\theta, \zeta)) = \bigl(\theta, \eta^*(\theta) + \zeta\bigr), \qquad (\theta, \zeta(\theta, \eta)) = \bigl(\theta, \eta - \eta^*(\theta)\bigr), \tag{2.7}$$

and we introduce the notation $Q_{\theta,\zeta} = P_{\theta,\eta(\theta,\zeta)}$. With $\zeta = 0$, $\theta \mapsto Q_{\theta,0}$ describes the least-favorable submodel $\mathscr{P}^*$ and with a nonzero value of $\zeta$, $\theta \mapsto Q_{\theta,\zeta}$ describes a version thereof, translated over a nuisance direction (see Figure 2). Expressed in terms of the metric $r_H(\zeta_1, \zeta_2) = H(Q_{\theta_0,\zeta_1}, Q_{\theta_0,\zeta_2})$, the sets $D(\theta, \rho)$ are mapped to open balls $B(\rho) = \{\zeta \in H : r_H(\zeta, 0) < \rho\}$ centered at the origin $\zeta = 0$,

$$\{P_{\theta,\eta} : \theta \in U_0, \eta \in D(\theta, \rho)\} = \{Q_{\theta,\zeta} : \theta \in U_0, \zeta \in B(\rho)\}.$$

In the formulation of Theorem 2.1, we make use of a domination condition based on the quantities

$$U_n(\rho, h) = \sup_{\zeta \in B(\rho)} Q^n_{\theta_0,\zeta}\left(\prod_{i=1}^{n} \frac{q_{\theta_n(h),\zeta}}{q_{\theta_0,\zeta}}(X_i)\right)$$

for all $\rho > 0$ and $h \in \mathbb{R}^k$. Below, it is required that there exists a sequence $(\rho_n)$ with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$, such that, for every *bounded*, stochastic sequence $(h_n)$, $U(\rho_n, h_n) = O(1)$ (where the expectation concerns the stochastic dependence of $h_n$ as well; see *Notation and conventions*). For a single, fixed $\zeta$, the requirement says that the likelihood ratio remains integrable when we replace $\theta_n(h_n)$ by the maximum-likelihood estimator $\hat{\theta}_n(X_1, \ldots, X_n)$. Lemma 4.3 demonstrates that ordinary differentiability of the likelihood-ratio with respect to $h$, combined with a uniform upper bound on certain Fisher information coefficients, suffices to satisfy $U(\rho_n, h_n) = O(1)$ for all bounded, stochastic $(h_n)$ and every $\rho_n \downarrow 0$.

The second step of the proof can now be summarized as follows: assuming stochastic LAN of the model, contraction of the nuisance posterior as in Figure 1 and said domination condition are enough to turn LAN expansions for the integrand in (2.5) into a single LAN expansion for $S_n$. The latter is determined by the efficient score, because the locus of posterior concentration, $\mathscr{P}^*$, is a least-favorable submodel (see Theorem 4.2).

The third step is based on two observations: first, in a semiparametric problem, the integrals $S_n$ appear in the expression for the marginal posterior in exactly the same way as parametric likelihood ratios appear in the posterior for parametric problems. Second, the parametric Bernstein–von Mises proof depends on likelihood ratios *only* through the LAN property. As a consequence, local asymptotic normality for $S_n$ offers the possibility to apply Le Cam's proof of posterior asymptotic normality in semiparametric context. If, in addition, we impose contraction at parametric rate for the marginal posterior, the LAN expansion of $S_n$ leads to the conclusion that the marginal posterior satisfies the Bernstein–von Mises assertion (1.2); see Theorem 5.1.

2.2. *Main theorem.* Before we state the main result of this paper, general conditions imposed on models and priors are formulated:

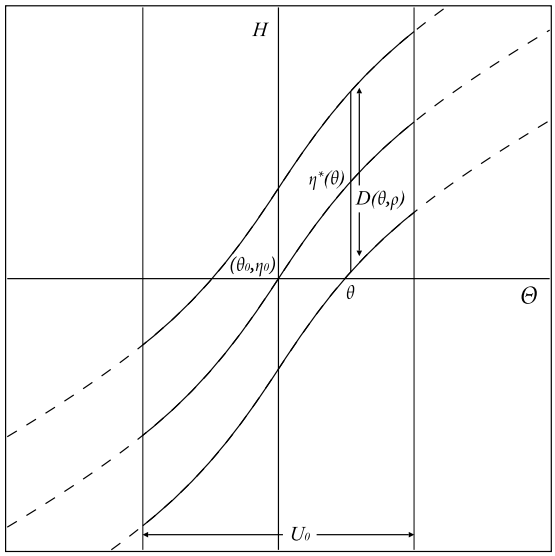


FIG. 1. *A neighborhood of* $(\theta_0, \eta_0)$. *Shown are the least-favorable curve* $\{(\theta, \eta^*(\theta)) : \theta \in U_0\}$ *and* (*for fixed* $\theta$ *and* $\rho > 0$) *the neighborhood* $D(\theta, \rho)$ *of* $\eta^*(\theta)$. *The sets* $D(\theta, \rho)$ *are expected to capture* ($\theta$*-conditional*) *posterior mass one asymptotically, for all* $\rho > 0$ *and* $\theta \in U_0$.

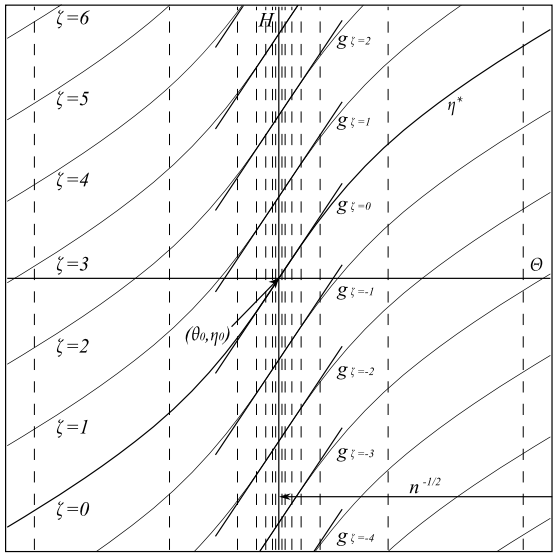


FIG. 2. *A neighborhood of* $(\theta_0, \eta_0)$. *Curved lines represent sets* $\{(\theta, \zeta) : \theta \in U_0\}$ *for fixed* $\zeta$. *The curve through* $\zeta = 0$ *parametrizes the least-favorable submodel. Vertical dashed lines delimit regions such that* $\|\theta - \theta_0\| \leq n^{-1/2}$. *Also indicated are directions along which the likelihood is expanded, with score functions* $g_\zeta$.

(i) *Model assumptions.* Throughout the remainder of this article, $\mathscr{P}$ is assumed to be well specified and dominated by a $\sigma$-finite measure on the sample space and parametrized identifiably on $\Theta \times H$, with $\Theta \subset \mathbb{R}^k$ open and $H$ a subset of a metric vector-space with metric $d_H$. Smoothness of the model is required but mentioned explicitly throughout. We also assume that there exists an open neighborhood $U_0 \subset \Theta$ of $\theta_0$ on which a least-favorable submodel $\eta^* : U_0 \to H$ is defined.

(ii) *Prior assumptions.* With regard to the prior $\Pi$ we follow the product structure of the parametrization of $\mathscr{P}$, by endowing the parameterspace $\Theta \times H$ with a product-prior $\Pi_\Theta \times \Pi_H$ defined on a $\sigma$-field that includes the Borel $\sigma$-field generated by the product-topology. Also, it is assumed that the prior $\Pi_\Theta$ is thick at $\theta_0$.

With the above general considerations for model and prior in mind, we formulate the main result of this paper.

THEOREM 2.1 (Semiparametric Bernstein–von Mises). *Let $X_1, X_2, \ldots$ be distributed i.i.d.-$P_0$, with $P_0 \in \mathscr{P}$, and let $\Pi_\Theta$ be thick at $\theta_0$. Suppose that for large enough $n$, the map $h \mapsto s_n(h)$ is continuous $P_0^n$-almost-surely. Also assume that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN in the $\theta$-direction, for all $\zeta$ in an $r_H$-neighborhood of $\zeta = 0$ and that the efficient Fisher information $\tilde{I}_{\theta_0.\eta_0}$ is nonsingular. Furthermore, assume that there exists a sequence $(\rho_n)$ with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ such that:*

(i) *For all $M > 0$, there exists a $K > 0$ such that, for large enough $n$,*

$$\Pi_H(K_n(\rho_n, M)) \geq e^{-Kn\rho_n^2}.$$

(ii) *For all $n$ large enough, the Hellinger metric entropy satisfies*

$$N(\rho_n, H, d_H) \leq e^{n\rho_n^2}$$

*and, for every bounded, stochastic $(h_n)$.*

(iii) *The model satisfies the domination condition,*

$$U_n(\rho_n, h_n) = O(1). \tag{2.8}$$

(iv) *For all $L > 0$, Hellinger distances satisfy the uniform bound,*

$$\sup_{\{\eta \in H\,:\, d_H(\eta,\eta_0) \geq L\rho_n\}} \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} = o(1).$$

*Finally, suppose that*

(v) *for every $(M_n)$, $M_n \to \infty$, the posterior satisfies*

$$\Pi_n(\|h\| \leq M_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 1.$$

*Then the sequence of marginal posteriors for $\theta$ converges in total variation to a normal distribution,*

$$\sup_A \left|\Pi_n(h \in A \mid X_1, \ldots, X_n) - N_{\tilde{\Delta}_n, \tilde{I}_{\theta_0,\eta_0}^{-1}}(A)\right| \xrightarrow{P_0} 0, \tag{2.9}$$

*centered on $\tilde{\Delta}_n$ with covariance matrix $\tilde{I}_{\theta_0,\eta_0}^{-1}$.*

PROOF. The assertion follows from combination of Theorem 3.1, Corollary 3.3, Theorems 4.2 and 5.1. □

Let us briefly discuss some aspects of the conditions of Theorem 2.1. First, consider the required existence of a least-favorable submodel in $\mathscr{P}$. In many semiparametric problems, the efficient score function is *not* a proper score in the sense that it corresponds to a smooth submodel; instead, the efficient score lies in the $L_2$-closure of the set of all proper scores. So there exist sequences of so-called *approximately least-favorable* submodels whose scores converge to the efficient score in $L_2$ [43]. Using such approximations of $\mathscr{P}^*$, our proof will entail extra conditions, but there is no reason to expect problems of an overly restrictive nature. It may therefore be hoped that the result remains largely unchanged if we turn (2.7) into a sequence of reparametrizations based on suitably chosen approximately least-favorable submodels.

Second, consider the rate $(\rho_n)$, which must be slow enough to satisfy condition (iv) and is fixed at (or above) the minimax Hellinger rate for estimation of the nuisance with known $\theta_0$ by condition (ii), while satisfying (i) and (iii) as well. Conditions (i) and (ii) also arise when considering Hellinger rates for nonparametric posterior convergence and the methods of Ghosal et al. [16] can be applied in the present context with minor modifications. In addition, Lemma 4.3 shows that in a wide class of semiparametric models, condition (iii) is satisfied for *any* rate sequence $(\rho_n)$. Typically, the numerator in condition (iv) is of order $O(n^{-1/2})$, so that condition (iv) holds true for any $\rho_n$ such that $n\rho_n^2 \to \infty$. The above enables a rate-free version of the semiparametric Bernstein–von Mises theorem (Corollary 5.2), in which conditions (i) and (ii) above are weakened to become comparable to those of Schwartz [37] for nonparametric posterior consistency. Applicability of Corollary 5.2 is demonstrated in Section 7, where the linear coefficient in the partial linear regression model is estimated.

Third, consider condition (v) of Theorem 2.1: though it is necessary [as it follows from (2.9)], it is hard to formulate straightforward sufficient conditions to satisfy (v) in generality. Moreover, condition (v) involves the nuisance prior and, as such, imposes another condition on $\Pi_H$ besides (i). To lessen its influence on $\Pi_H$, constructions in Section 6 either work for all nuisance priors (see Lemma 6.1) or require only consistency of the nuisance posterior (see Theorem 6.2). The latter is based on the limiting behavior of posteriors in misspecified parametric models [24, 26] and allows for the tentative but general observation that a bias [cf. (6.6)] may ruin $n^{-1/2}$-consistency of the marginal posterior, especially if the rate $(\rho_n)$ is sub-optimal. In the example of Section 7, the "hard work" stems from condition (v) of Theorem 2.1: $\alpha > 1/2$ Hölder smoothness and boundedness of the family of regression functions in Corollary 7.2 are imposed in order to satisfy this condition. Since conditions (i) and (ii) appear quite reasonable and conditions (iii) and (iv) are satisfied relatively easily, condition (v) should be viewed as the most complicated in an essential way.

To conclude, consistency under perturbation (with appropriate rate) is one of the sufficient conditions, but it is by no means clear in how far it should also hold with necessity. One expects that in some situations where consistency under perturbation fails to hold fully, integral local asymptotic normality (see Section 4) is still satisfied in a weaker form. In particular, it is possible that (4.2) holds with a less-than-efficient score and Fisher information, a result that would have an interpretation analogous to suboptimality in Hájek's convolution theorem. What happens in cases where integral LAN fails more comprehensively is both interesting and completely mysterious from the point of view taken in this article.

**3. Posterior convergence under perturbation.** In this section, we consider contraction of the posterior around least-favorable submodels. We express this form of posterior convergence by showing that (under suitable conditions) the conditional posterior for the nuisance parameter contracts around the least-favorable submodel, conditioned on a sequence $\theta_n(h_n)$ for the parameter of interest with $h_n = O_{P_o}(1)$. We view the sequence of models $\mathscr{P}_{\theta_n(h_n)}$ as a random perturbation of the model $\mathscr{P}_{\theta_0}$ and generalize Ghosal et al. [16] to describe posterior contraction. Ultimately, random perturbation of $\theta$ represents the "appropriate form of uniformity" referred to just after definition (2.4). Given a rate sequence $(\rho_n)$, $\rho_n \downarrow 0$, we say that the conditioned nuisance posterior is *consistent under $n^{-1/2}$-perturbation at rate $\rho_n$*, if

$$\Pi_n\big(D^c(\theta, \rho_n) \mid \theta = \theta_0 + n^{-1/2}h_n; X_1, \ldots, X_n\big) \xrightarrow{P_0} 0 \tag{3.1}$$

for all bounded, stochastic sequences $(h_n)$.

THEOREM 3.1 (Posterior rate of convergence under perturbation). *Assume that there exists a sequence $(\rho_n)$ with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ such that for all $M > 0$ and every bounded, stochastic $(h_n)$:*

(i) *There exists a constant $K > 0$ such that for large enough $n$,*

$$\Pi_H(K_n(\rho_n, M)) \geq e^{-Kn\rho_n^2}. \tag{3.2}$$

(ii) *For $L > 0$ large enough, there exist $(\phi_n)$ such that for large enough $n$,*

$$P_0^n\phi_n \to 0, \qquad \sup_{\eta \in D^c(\theta_0, L\rho_n)} P^n_{\theta_n(h_n),\eta}(1 - \phi_n) \leq e^{-L^2n\rho_n^2/4}. \tag{3.3}$$

(iii) *The least-favorable submodel satisfies $d_H(\eta^*(\theta_n(h_n)), \eta_0) = o(\rho_n)$.*

*Then, for every bounded, stochastic $(h_n)$ there exists an $L > 0$ such that the conditional nuisance posterior converges as*

$$\Pi\big(D^c(\theta, L\rho_n) \mid \theta = \theta_0 + n^{-1/2}h_n; X_1, \ldots, X_n\big) = o_{P_0}(1) \tag{3.4}$$

*under $n^{-1/2}$-perturbation.*

PROOF. Let $(h_n)$ be a stochastic sequence bounded by $M$, and let $0 < C < 1$ be given. Let $K$ and $(\rho_n)$ be as in conditions (i) and (ii). Choose $L > 4\sqrt{1+K+C}$ and large enough to satisfy condition (ii) for some $(\phi_n)$. By Lemma 3.4, the events

$$A_n = \left\{ \int_H \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) \geq e^{-(1+C)n\rho_n^2} \Pi_H(K_n(\rho_n, M)) \right\}$$

satisfy $P_0^n(A_n^c) \to 0$. Using also the first limit in (3.3), we then derive

$$\begin{aligned} &P_0^n \Pi\big(D^c(\theta, L\rho_n) \mid \theta = \theta_n(h_n); X_1, \ldots, X_n\big) \\ &\quad \leq P_0^n \Pi\big(D^c(\theta, L\rho_n) \mid \theta = \theta_n(h_n); X_1, \ldots, X_n\big) 1_{A_n}(1-\phi_n) + o(1) \end{aligned}$$

[even with random $(h_n)$, the posterior $\Pi(\cdot|\theta = \theta_n(h_n); X_1, \ldots, X_n) \leq 1$, by definition (2.1)]. The first term on the r.h.s. can be bounded further by the definition of the events $A_n$,

$$\begin{aligned} &P_0^n \Pi\big(D^c(\theta, L\rho_n) \mid \theta = \theta_n; X_1, \ldots, X_n\big) 1_{A_n}(1-\phi_n) \\ &\quad \leq \frac{e^{(1+C)n\rho_n^2}}{\Pi_H(K_n(\rho_n, M))} P_0^n \left( \int_{D^c(\theta_n(h_n), L\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)(1-\phi_n)\, d\Pi_H \right). \end{aligned}$$

Due to condition (iii) it follows that

$$D\left(\theta_0, \frac{1}{2}L\rho_n\right) \subset \bigcap_{n\geq 1} D(\theta_n(h_n), L\rho_n) \tag{3.5}$$

for large enough $n$. Therefore,

$$\begin{aligned} &P_0^n \int_{D^c(\theta_n(h_n), L\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)(1-\phi_n)\, d\Pi_H(\eta) \\ &\quad \leq \int_{D^c(\theta_0, L\rho_n/2)} P^n_{\theta_n(h_n),\eta}(1-\phi_n)\, d\Pi_H(\eta). \end{aligned} \tag{3.6}$$

Upon substitution of (3.6) and with the use of the second bound in (3.3) and (3.2), the choice we made earlier for $L$ proves the assertion. □

We conclude from the above that besides sufficiency of prior mass, the crucial condition for consistency under perturbation is the existence of a test sequence $(\phi_n)$ satisfying (3.3). To find sufficient conditions, we follow a construction of tests based on the Hellinger geometry of the model, generalizing the approach of Birgé [2, 3] and Le Cam [30] to $n^{-1/2}$-perturbed context. It is easiest to illustrate their approach by considering the problem of testing/estimating $\eta$ when $\theta_0$ is known: we cover the nuisance model $\{P_{\theta_0,\eta} : \eta \in H\}$ by a minimal collection of Hellinger balls $B$ of radii $(\rho_n)$, each of which is convex and hence testable against $P_0$ with power

bounded by $\exp(-\frac{1}{4}nH^2(P_0, B))$, based on the minimax theorem [30]. The tests for the covering Hellinger balls are combined into a single test for the nonconvex alternative $\{P : H(P, P_0) \geq \rho_n\}$ against $P_0$. The order of the cover controls the power of the combined test. Therefore the construction requires an upper bound to Hellinger metric entropy numbers [45]

$$N(\rho_n, \mathscr{P}_{\theta_0}, H) \leq e^{n\rho_n^2}, \tag{3.7}$$

which is interpreted as indicative of the nuisance model's complexity in the sense that the lower bound to the collection of rates $(\rho_n)$ solving (3.7) is the Hellinger minimax rate for estimation of $\eta_0$. In the $n^{-1/2}$-perturbed problem, the alternative does not just consist of the complement of a Hellinger-ball in the nuisance factor $H$, but also has an extent in the $\theta$-direction shrinking at rate $n^{-1/2}$. Condition (3.8) below guarantees that Hellinger covers of $H$ like the above are large enough to accommodate the $\theta$-extent of the alternative, the implication being that the test sequence one constructs for the nuisance in case $\theta_0$ is known, can also be used when $\theta_0$ is known only up to $n^{-1/2}$-perturbation. Therefore, the entropy bound in Lemma 3.2 is (3.7). Geometrically, (3.8) requires that $n^{-1/2}$-perturbed versions of the nuisance model are contained in a narrowing sequence of metric cones based at $P_0$. In differentiable models, the Hellinger distance $H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})$ is typically of order $O(n^{-1/2})$ for all $\eta \in H$. So if, in addition, $n\rho_n^2 \to \infty$, limit (3.8) is expected to hold pointwise in $\eta$. Then only the uniform character of (3.8) truly forms a condition.

LEMMA 3.2 (Testing under perturbation). *If $(\rho_n)$ satisfies $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ and the following requirements are met*:

(i) *For all $n$ large enough, $N(\rho_n, H, d_H) \leq e^{n\rho_n^2}$.*
(ii) *For all $L > 0$ and all bounded, stochastic $(h_n)$,*

$$\sup_{\{\eta \in H : d_H(\eta,\eta_0) \geq L\rho_n\}} \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} = o(1). \tag{3.8}$$

*Then for all $L \geq 4$, there exists a test sequence $(\phi_n)$ such that for all bounded, stochastic $(h_n)$,*

$$P_0^n \phi_n \to 0, \qquad \sup_{\eta \in D^c(\theta_0, L\rho_n)} P^n_{\theta_n(h_n),\eta}(1 - \phi_n) \leq e^{-L^2 n\rho_n^2/4} \tag{3.9}$$

*for large enough $n$.*

PROOF. Let $(\rho_n)$ be such that (i) and (ii) are satisfied. Let $(h_n)$ and $L \geq 4$ be given. For all $j \geq 1$, define $H_{j,n} = \{\eta \in H : jL\rho_n \leq d_H(\eta_0, \eta) \leq (j+1)L\rho_n\}$ and $\mathscr{P}_{j,n} = \{P_{\theta_0,\eta} : \eta \in H_{j,n}\}$. Cover $\mathscr{P}_{j,n}$ with Hellinger balls $B_{i,j,n}(\frac{1}{4}jL\rho_n)$, where

$$B_{i,j,n}(r) = \{P : H(P_{i,j,n}, P) \leq r\}$$

and $P_{i.j.n} \in \mathscr{P}_{j,n}$, that is, there exists an $\eta_{i,j,n} \in H_{j,n}$ such that $P_{i,j,n} = P_{\theta_0,\eta_{i,j,n}}$. Denote $H_{i,j,n} = \{\eta \in H_{j,n} : P_{\theta_0,\eta} \in B_{i,j,n}(\frac{1}{4} j L\rho_n)\}$. By assumption, the minimal number of such balls needed to cover $\mathscr{P}_{i,j}$ is finite; we denote the corresponding covering number by $N_{j,n}$, that is, $1 \le i \le N_{j,n}$.

Let $\eta \in H_{j,n}$ be given. There exists an $i$ $(1 \le i \le N_{j,n})$ such that $d_H(\eta, \eta_{i,j,n}) \le \frac{1}{4} j L\rho_n$. Then, by the triangle inequality, the definition of $H_{j,n}$ and assumption (3.8),

$$
\begin{aligned}
&H\bigl(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta_{i,j,n}}\bigr) \\
&\quad \le H\bigl(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}\bigr) + H\bigl(P_{\theta_0,\eta}, P_{\theta_0,\eta_{i,j,n}}\bigr) \\
&\quad \le \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} H(P_{\theta_0,\eta}, P_0) + \frac{1}{4} j L\rho_n \qquad (3.10) \\
&\quad \le \Biggl( \sup_{\{\eta \in H : d_H(\eta,\eta_0) \ge L\rho_n\}} \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} \Biggr) (j+1) L\rho_n + \frac{1}{4} j L\rho_n \\
&\quad \le \frac{1}{2} j L\rho_n
\end{aligned}
$$

for large enough $n$. We conclude that there exists an $N \ge 1$ such that for all $n \ge N$, $j \ge 1$, $1 \le i \le N_{j,n}$, $\eta \in H_{i,j,n}$, $P_{\theta_n(h_n),\eta} \in B_{i,j,n}(\frac{1}{2} j L\rho_n)$. Moreover, Hellinger balls are convex and for all $P \in B_{i,j,n}(\frac{1}{2} j L\rho_n)$, $H(P, P_0) \ge \frac{1}{2} j L\rho_n$. As a consequence of the minimax theorem (see Le Cam [30], Birgé [2, 3]), there exists a test sequence $(\phi_{i,j,n})_{n \ge 1}$ such that

$$
P_0^n \phi_{i,j,n} \vee \sup_P P^n (1 - \phi_{i,j,n}) \le e^{-n H^2(B_{i,j,n}(j L\rho_n/2), P_0)} \le e^{-n j^2 L^2 \rho_n^2/4},
$$

where the supremum runs over all $P \in B_{i,j,n}(\frac{1}{2} j L\rho_n)$. Defining, for all $n \ge 1$, $\phi_n = \sup_{j \ge 1} \max_{1 \le i \le N_{j,n}} \phi_{i,j,n}$, we find (for details, see the proof of Theorem 3.10 in [24]) that

$$
P_0^n \phi_n \le \sum_{j \ge 1} N_{j,n} e^{-L^2 j^2 n \rho_n^2/4}, \qquad P^n (1 - \phi_n) \le e^{-L^2 n \rho_n^2/4} \qquad (3.11)
$$

for all $P = P_{\theta_n(h_n),\eta}$ and $\eta \in D^c(\theta_0, L\rho_n)$. Since $L \ge 4$, we have for all $j \ge 1$,

$$
\begin{aligned}
N_{j,n} &= N\bigl(\tfrac{1}{4} L j \rho_n, \mathscr{P}_{j,n}, H\bigr) \le N\bigl(\tfrac{1}{4} L j \rho_n, \mathscr{P}, H\bigr) \\
&\le N(\rho_n, \mathscr{P}, H) \le e^{n\rho_n^2}
\end{aligned} \qquad (3.12)
$$

by assumption (3.7). Upon substitution of (3.12) into (3.11), we obtain the following bounds:

$$
P_0^n \phi_n \le \frac{e^{(1 - L^2/4) n \rho_n^2}}{1 - e^{-L^2 n \rho_n^2/4}}, \qquad \sup_{\eta \in D^c(\theta_0, L\rho_n)} P^n_{\theta_n(h_n),\eta} (1 - \phi_n) \le e^{-L^2 n \rho_n^2/4}
$$

for large enough $n$, which implies assertion (3.9). □

In preparation of Corollary 5.2, we also provide a version of Theorem 3.1 that only asserts consistency under $n^{-1/2}$-perturbation at *some* rate while relaxing bounds for prior mass and entropy. In the statement of the corollary, we make use of the family of Kullback–Leibler neighborhoods that would play a role for the posterior of the nuisance if $\theta_0$ were known [16].

$$K(\rho) = \left\{ \eta \in H : -P_0 \log \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} \le \rho^2, P_0 \left( \log \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} \right)^2 \le \rho^2 \right\} \tag{3.13}$$

for all $\rho > 0$. The proof below follows steps similar to those in the proof of Corollary 2.1 in [27].

COROLLARY 3.3 (Posterior consistency under perturbation). *Assume that for all* $\rho > 0$, $N(\rho, H, d_H) < \infty$, $\Pi_H(K(\rho)) > 0$ *and*:

(i) *For all* $M > 0$ *there is an* $L > 0$ *such that for all* $\rho > 0$ *and large enough* $n$, $K(\rho) \subset K_n(L\rho, M)$.

(ii) *For every bounded random sequence* $(h_n)$, $\sup_{\eta \in H} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})$ *and* $H(P_{\theta_0,\eta^*(\theta_n(h_n))}, P_{\theta_0,\eta_0})$ *are of order* $O(n^{-1/2})$.

*Then there exists a sequence* $(\rho_n)$, $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$, *such that the conditional nuisance posterior converges under* $n^{-1/2}$*-perturbation at rate* $(\rho_n)$.

PROOF. We follow the proof of Corollary 2.1 in Kleijn and van der Vaart [27] and add that, under condition (ii), (3.8) and condition (iii) of Theorem 3.1 are satisfied. We conclude that there exists a test sequence satisfying (3.3). Then the assertion of Theorem 3.1 holds. □

The following lemma generalizes Lemma 8.1 in Ghosal et al. [16] to the $n^{-1/2}$-perturbed setting.

LEMMA 3.4. *Let* $(h_n)$ *be stochastic and bounded by some* $M > 0$. *Then*

$$\begin{aligned} &P_0^n \left( \int_H \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) < e^{-(1+C)n\rho^2} \Pi_H(K_n(\rho, M)) \right) \\ &\qquad \le \frac{1}{C^2 n\rho^2} \end{aligned} \tag{3.14}$$

*for all* $C > 0$, $\rho > 0$ *and* $n \ge 1$.

PROOF. See the proof of Lemma 8.1 in Ghosal et al. [16] (dominating the $h_n$-dependent log-likelihood ratio immediately after the first application of Jensen's inequality). □

**4. Integrating local asymptotic normality.** The smoothness condition in the Le Cam's parametric Bernstein–von Mises theorem is a LAN expansion of the likelihood, which is replaced in semiparametric context by a stochastic LAN expansion of the integrated likelihood (2.5). In this section, we consider sufficient conditions under which the localized integrated likelihood

$$s_n(h) = \int_H \prod_{i=1}^n \frac{p_{\theta_0+n^{-1/2}h,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) \tag{4.1}$$

has the *integral LAN* property; that is, $s_n$ allows an expansion of the form

$$\log \frac{s_n(h_n)}{s_n(0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} h_n^T \tilde{\ell}_{\theta_0,\eta_0} - \frac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + o_{P_0}(1) \tag{4.2}$$

for every random sequence $(h_n) \subset \mathbb{R}^k$ of order $O_{P_0}(1)$, as required in Theorem 5.1. Theorem 4.2 assumes that the model is stochastically LAN and requires consistency under $n^{-1/2}$-perturbation for the nuisance posterior. Consistency not only allows us to restrict sufficient conditions to neighborhoods of $\eta_0$ in $H$, but also enables lifting of the LAN expansion of the integrand in (4.1) to an expansion of the integral $s_n$ itself; cf. (4.2). The posterior concentrates on the least-favorable submodel so that only the least-favorable expansion at $\eta_0$ contributes to (4.2) asymptotically. For this reason, the intergral LAN expansion is determined by the efficient score function (and not some other influence function). Ultimately, occurrence of the efficient score lends the marginal posterior (and statistics based upon it) properties of frequentist semiparametric optimality.

To derive Theorem 4.2, we reparametrize the model; cf. (2.7). While yielding adaptivity, this reparametrization also leads to $\theta$-dependence in the prior for $\zeta$, a technical issue that we tackle before addressing the main point of this section. We show that the prior mass of the relevant neighborhoods displays the appropriate type of stability, under a condition on local behavior of Hellinger distances in the least-favorable model. For smooth least-favorable submodels, typically $d_H(\eta^*(\theta_n(h_n)), \eta_0) = O(n^{-1/2})$ for all bounded, stochastic $(h_n)$, which suffices.

LEMMA 4.1 (Prior stability). *Let $(h_n)$ be a bounded, stochastic sequence of perturbations, and let $\Pi_H$ be any prior on $H$. Let $(\rho_n)$ be such that $d_H(\eta^*(\theta_n(h_n)), \eta_0) = o(\rho_n)$. Then the prior mass of radius-$\rho_n$ neighborhoods of $\eta^*$ is stable, that is,*

$$\Pi_H(D(\theta_n(h_n), \rho_n)) = \Pi_H(D(\theta_0, \rho_n)) + o(1). \tag{4.3}$$

PROOF. Let $(h_n)$ and $(\rho_n)$ be such that $d_H(\eta^*(\theta_n(h_n)), \eta_0) = o(\rho_n)$. Denote $D(\theta_n(h_n), \rho_n)$ by $D_n$ and $D(\theta_0, \rho_n)$ by $C_n$ for all $n \geq 1$. Since

$$|\Pi_H(D_n) - \Pi_H(C_n)| \leq \Pi_H\big((D_n \cup C_n) \setminus (D_n \cap C_n)\big),$$

we consider the sequence of symmetric differences. Fix some $0 < \alpha < 1$. Then for all $\eta \in D_n$ and all $n$ large enough, $d_H(\eta, \eta_0) \le d_H(\eta, \eta^*(\theta_n(h_n))) + d_H(\eta^*(\theta_n(h_n)), \eta_0) \le (1+\alpha)\rho_n$, so that $D_n \cup C_n \subset D(\theta_0, (1+\alpha)\rho_n)$. Furthermore, for large enough $n$ and any $\eta \in D(\theta_0, (1-\alpha)\rho_n)$, $d_H(\eta, \eta^*(\theta_n(h_n))) \le d_H(\eta, \eta_0) + d_H(\eta_0, \eta^*(\theta_n(h_n))) \le \rho_n + d_H(\eta_0, \eta^*(\theta_n(h_n))) - \alpha\rho_n < \rho_n$, so that $D(\theta_0, (1-\alpha)\rho_n) \subset D_n \cap C_n$. Therefore,

$$(D_n \cup C_n) \setminus (D_n \cap C_n) \subset D\big(\theta_0, (1+\alpha)\rho_n\big) \setminus D\big(\theta_0, (1-\alpha)\rho_n\big) \to \varnothing,$$

which implies (4.3). □

Once stability of the nuisance prior is established, Theorem 4.2 hinges on stochastic local asymptotic normality of the submodels $t \mapsto Q_{\theta_0+t,\zeta}$, for all $\zeta$ in an $r_H$-neighborhood of $\zeta = 0$. We assume there exists a $g_\zeta \in L_2(Q_{\theta_0,\zeta})$ such that for every random $(h_n)$ bounded in $Q_{\theta_0,\zeta}$-probability,

$$\log \prod_{i=1}^n \frac{q_{\theta+n^{-1/2}h_n,\zeta}}{q_{\theta_0,0}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h_n^T g_\zeta(X_i) - \frac{1}{2} h_n^T I_\zeta h_n + R_n(h_n, \zeta), \tag{4.4}$$

where $I_\zeta = Q_{\theta_0,\zeta} g_\zeta g_\zeta^T$ and $R_n(h_n, \zeta) = o_{Q_{\theta_0,\zeta}}(1)$. Equation (4.4) specifies the (minimal) tangent set (van der Vaart [43], Section 25.4) with respect to which differentiability of the model is required. Note that $g_0 = \tilde{\ell}_{\theta_0,\eta_0}$.

THEOREM 4.2 (Integral local asymptotic normality). *Suppose that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN for all $\zeta$ in an $r_H$-neighborhood of $\zeta = 0$. Furthermore, assume that posterior consistency under $n^{-1/2}$-perturbation obtains with a rate $(\rho_n)$ also valid in (2.8). Then the integral LAN-expansion (4.2) holds.*

PROOF. Throughout this proof $G_n(h, \zeta) = \sqrt{n} h^T \mathbb{P}_n g_\zeta - \frac{1}{2} h^T I_\zeta h$, for all $h$ and all $\zeta$. Furthermore, we abbreviate $\theta_n(h_n)$ to $\theta_n$ and omit explicit notation for $(X_1, \dots, X_n)$-dependence in several places.

Let $\delta, \varepsilon > 0$ be given, and let $\theta_n = \theta_0 + n^{-1/2} h_n$ with $(h_n)$ bounded in $P_0$-probability. Then there exists a constant $M > 0$ such that $P_0^n(\|h_n\| > M) < \frac{1}{2}\delta$ for all $n \ge 1$. With $(h_n)$ bounded, the assumption of consistency under $n^{-1/2}$-perturbation says that

$$P_0^n\big(\log \Pi\big(D(\theta, \rho_n) \mid \theta = \theta_n; X_1, \dots, X_n\big) \ge -\varepsilon\big) > 1 - \tfrac{1}{2}\delta$$

for large enough $n$. This implies that the posterior's numerator and denominator are related through

$$P_0^n \Bigg( \int_H \prod_{i=1}^n \frac{p_{\theta_n,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) \le e^{\varepsilon} 1_{\{\|h_n\| \le M\}} \int_{D(\theta_n,\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta) \Bigg) > 1 - \delta. \tag{4.5}$$

We continue with the integral over $D(\theta_n, \rho_n)$ under the restriction $\|h_n\| \le M$ and parametrize the model locally in terms of $(\theta, \zeta)$ [see (2.7)]

$$\int_{D(\theta_n,\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n,\eta}}{p_{\theta_0,\eta_0}}(X_i)\,d\Pi_H(\eta) = \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi(\zeta \mid \theta = \theta_n), \tag{4.6}$$

where $\Pi(\cdot|\theta)$ denotes the prior for $\zeta$ given $\theta$, that is, $\Pi_H$ translated over $\eta^*(\theta)$. Next we note that by Fubini's theorem and the domination condition (2.8), there exists a constant $L > 0$ such that

$$\begin{aligned}&\left| P_0^n \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\big(d\Pi(\zeta \mid \theta = \theta_n) - d\Pi(\zeta \mid \theta = \theta_0)\big)\right| \\ &\qquad \le L\big|\Pi\big(B(\rho_n) \mid \theta = \theta_n\big) - \Pi\big(B(\rho_n) \mid \theta = \theta_0\big)\big|\end{aligned}$$

for large enough $n$. Since the least-favorable submodel is stochastically LAN, Lemma 4.1 asserts that the difference on the r.h.s. of the above display is $o(1)$, so that

$$\begin{aligned}&\int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi(\zeta \mid \theta = \theta_n) \\ &\qquad = \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi(\zeta) + o_{P_0}(1),\end{aligned} \tag{4.7}$$

where we use the notation $\Pi(A) = \Pi(\zeta \in A|\theta = \theta_0)$ for brevity. We define for all $\zeta, \varepsilon > 0$, $n \ge 1$ the events $F_n(\zeta, \varepsilon) = \{\sup_h |G_n(h, \zeta) - G_n(h, 0)| \le \varepsilon\}$. With (2.8) as a domination condition, Fatou's lemma and the fact that $F_n^c(0, \varepsilon) = \varnothing$ lead to

$$\begin{aligned}&\limsup_{n\to\infty} \int_{B(\rho_n)} Q_{\theta_n,\zeta}^n(F_n^c(\zeta, \varepsilon))\,d\Pi(\zeta) \\ &\qquad \le \int \limsup_{n\to\infty} 1_{B(\rho_n)\setminus\{0\}}(\zeta) Q_{\theta_n,\zeta}^n(F_n^c(\zeta, \varepsilon))\,d\Pi(\zeta) = 0\end{aligned} \tag{4.8}$$

[again using (2.8) in the last step]. Combined with Fubini's theorem, this suffices to conclude that

$$\int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i)\,d\Pi(\zeta) = \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) 1_{F_n(\zeta,\varepsilon)}\,d\Pi(\zeta) + o_{P_0}(1), \tag{4.9}$$

and we continue with the first term on the right-hand side. By stochastic local asymptotic normality for every $\zeta$, expansion (4.4) of the log-likelihood implies that

$$\prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) = \prod_{i=1}^n \frac{q_{\theta_0,\zeta}}{q_{\theta_0,0}}(X_i) e^{G_n(h_n,\zeta)+R_n(h_n,\zeta)}, \tag{4.10}$$

where the rest term is of order $o_{Q_{\theta_0,\zeta}}(1)$. Accordingly, we define, for every $\zeta$, the events $A_n(\zeta,\varepsilon) = \{|R_n(h_n,\zeta)| \le \frac{1}{2}\varepsilon\}$, so that $Q^n_{\theta_0,\zeta}(A_n^c(\zeta,\varepsilon)) \to 0$. Contiguity then implies that $Q^n_{\theta_n,\zeta}(A_n^c(\zeta,\varepsilon)) \to 0$ as well. Reasoning as in (4.9) we see that

$$\begin{aligned}&\int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) 1_{F_n(\zeta,\varepsilon)}\, d\Pi(\zeta)\\ &\quad = \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) 1_{A_n(\zeta,\varepsilon)\cap F_n(\zeta,\varepsilon)}\, d\Pi(\zeta) + o_{P_0}(1).\end{aligned} \tag{4.11}$$

For fixed $n$ and $\zeta$ and for all $(X_1,\dots,X_n) \in A_n(\zeta,\varepsilon)\cap F_n(\zeta,\varepsilon)$,

$$\left|\log \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) - G_n(h_n,0)\right| \le 2\varepsilon,$$

so that the first term on the right-hand side of (4.11) satisfies the bounds

$$\begin{aligned}&e^{G_n(h_n,0)-2\varepsilon} \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_0,\zeta}}{q_{\theta_0,0}}(X_i) 1_{A_n(\zeta,\varepsilon)\cap F_n(\zeta,\varepsilon)}\, d\Pi(\zeta)\\ &\le \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_n,\zeta}}{q_{\theta_0,0}}(X_i) 1_{A_n(\zeta,\varepsilon)\cap F_n(\zeta,\varepsilon)}\, d\Pi(\zeta)\\ &\le e^{G_n(h_n,0)+2\varepsilon} \int_{B(\rho_n)} \prod_{i=1}^n \frac{q_{\theta_0,\zeta}}{q_{\theta_0,0}}(X_i) 1_{A_n(\zeta,\varepsilon)\cap F_n(\zeta,\varepsilon)}\, d\Pi(\zeta).\end{aligned} \tag{4.12}$$

The integral factored into lower and upper bounds can be relieved of the indicator for $A_n \cap F_n$ by reversing the argument that led to (4.9) and (4.11) (with $\theta_0$ replacing $\theta_n$), at the expense of an $e^{o_{P_0}(1)}$-factor. Substituting in (4.12) and using, consecutively, (4.11), (4.9), (4.7) and (4.5) for the bounded integral, we find

$$e^{G_n(h_n,0)-3\varepsilon+o_{P_0}(1)} s_n(0) \le s_n(h_n) \le e^{G_n(h_n,0)+3\varepsilon+o_{P_0}(1)} s_n(0).$$

Since this holds with arbitrarily small $0 < \varepsilon' < \varepsilon$ for large enough $n$, it proves (4.2). □

With regard to the nuisance rate $(\rho_n)$, we first note that our proof of Theorem 2.1 fails if the slowest rate required to satisfy (2.8) vanishes *faster* then the optimal rate for convergence under $n^{-1/2}$-perturbation [as determined in (3.7) and (3.2)].

However, the rate $(\rho_n)$ does not appear in assertion (4.2), so if said contradiction between conditions (2.8) and (3.7)/(3.2) do not occur, the sequence $(\rho_n)$ can remain entirely internal to the proof of Theorem 4.2. More particularly, if condition (2.8) holds for *any* $(\rho_n)$ such that $n\rho_n^2 \to \infty$, integral LAN only requires consistency under $n^{-1/2}$-perturbation at *some* such $(\rho_n)$. In that case, we may appeal to Corollary 3.3 instead of Theorem 3.1, thus relaxing conditions on model

entropy and nuisance prior. The following lemma shows that a first-order Taylor expansion of likelihood ratios combined with a boundedness condition on certain Fisher information coefficients is enough to enable use of Corollary 3.3 instead of Theorem 3.1.

LEMMA 4.3. *Let $\Theta$ be one-dimensional. Assume that there exists a $\rho > 0$ such that for every $\zeta \in B(\rho)$ and all $x$ in the samplespace, the map $\theta \mapsto \log(q_{\theta,\zeta}/q_{\theta_0,\zeta})(x)$ is continuously differentiable on $[\theta_0 - \rho, \theta_0 + \rho]$ with Lebesgue-integrable derivative $g_{\theta,\zeta}(x)$ such that*

$$\sup_{\zeta \in B(\rho)} \sup_{\{\theta : |\theta - \theta_0| < \rho\}} Q_{\theta,\zeta} g_{\theta,\zeta}^2 < \infty. \tag{4.13}$$

*Then, for every $\rho_n \downarrow 0$ and all bounded, stochastic $(h_n)$, $U_n(\rho_n, h_n) = O(1)$.*

PROOF. Let $(h_n)$ be stochastic and upper-bounded by $M > 0$. For every $\zeta$ and all $n \geq 1$,

$$\begin{aligned} Q_{\theta_0,\zeta}^n \left| \prod_{i=1}^n \frac{q_{\theta_n(h_n),\zeta}}{q_{\theta_0,\zeta}}(X_i) - 1 \right| &= Q_{\theta_0,\zeta}^n \left| \int_{\theta_0}^{\theta_n(h_n)} \sum_{i=1}^n g_{\theta',\zeta}(X_i) \prod_{j=1}^n \frac{q_{\theta',\zeta}}{q_{\theta_0,\zeta}}(X_j)\, d\theta' \right| \\ &\leq \int_{\theta_0 - M/\sqrt{n}}^{\theta_0 + M/\sqrt{n}} Q_{\theta',\zeta}^n \left| \sum_{i=1}^n g_{\theta',\zeta}(X_i) \right| d\theta' \\ &\leq \sqrt{n} \int_{\theta_0 - M/\sqrt{n}}^{\theta_0 + M/\sqrt{n}} \sqrt{Q_{\theta',\zeta} g_{\theta',\zeta}^2}\, d\theta', \end{aligned}$$

where the last step follows from the Cauchy–Schwarz inequality. For large enough $n$, $\rho_n < \rho$ and the square-root of (4.13) dominates the difference between $U(\rho, h_n)$ and 1. □

**5. Posterior asymptotic normality.** Under the assumptions formulated before Theorem 2.1, the marginal posterior density $\pi_n(\cdot|X_1, \ldots, X_n) : \Theta \to \mathbb{R}$ for the parameter of interest with respect to the prior $\Pi_\Theta$ equals

$$\pi_n(\theta|X_1, \ldots, X_n) = S_n(\theta) \Big/ \int_\Theta S_n(\theta')\, d\Pi_\Theta(\theta'), \tag{5.1}$$

$P_0^n$-almost-surely. One notes that this form is equal to that of a *parametric* posterior density, but with the parametric likelihood replaced by the integrated likelihood $S_n$. By implication, the proof of the parametric Bernstein–von Mises theorem can be applied to its semiparametric generalization, if we impose sufficient conditions for the parametric likelihood on $S_n$ instead. Concretely, we replace the smoothness requirement for the likelihood in Theorem 1.1 by (4.2). Together with a condition expressing marginal posterior convergence at parametric rate, (4.2) is sufficient to derive asymptotic normality of the posterior; cf. (1.2).

THEOREM 5.1 (Posterior asymptotic normality). *Let $\Theta$ be open in $\mathbb{R}^k$ with a prior $\Pi_\Theta$ that is thick at $\theta_0$. Suppose that for large enough $n$, the map $h \mapsto s_n(h)$ is continuous $P_0^n$-almost-surely. Assume that there exists an $L_2(P_0)$-function $\tilde{\ell}_{\theta_0,\eta_0}$ such that for every $(h_n)$ that is bounded in probability, (4.2) holds, $P_0\tilde{\ell}_{\theta_0,\eta_0} = 0$ and $\tilde{I}_{\theta_0,\eta_0}$ is nonsingular. Furthermore suppose that for every $(M_n)$, $M_n \to \infty$, we have*

$$\Pi_n(\|h\| \le M_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 1. \tag{5.2}$$

*Then the sequence of marginal posteriors for $\theta$ converges to a normal distribution in total variation,*

$$\sup_A \left|\Pi_n(h \in A \mid X_1, \ldots, X_n) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,\eta_0}}(A)\right| \xrightarrow{P_0} 0,$$

*centered on $\tilde{\Delta}_n$ with covariance matrix $\tilde{I}^{-1}_{\theta_0,\eta_0}$.*

PROOF. The proof is identical to that of Theorem 2.1 in [26] upon replacement of parametric likelihoods with integrated likelihoods. □

There is room for relaxation of the requirements on model entropy and minimal prior mass, if the limit (2.8) holds in a fixed neighborhood of $\eta_0$. The following corollary applies whenever (2.8) holds for *any rate* $(\rho_n)$. The simplifications are such that the entropy and prior mass conditions become comparable to those for Schwartz's posterior consistency theorem [37], rather than those for posterior rates of convergence following Ghosal, Ghosh and van der Vaart [16].

COROLLARY 5.2 (Semiparametric Bernstein–von Mises, rate-free). *Let $X_1$, $X_2, \ldots$ be i.i.d.-$P_0$, with $P_0 \in \mathscr{P}$, and let $\Pi_\Theta$ be thick at $\theta_0$. Suppose that for large enough $n$, the map $h \mapsto s_n(h)$ is continuous $P_0^n$-almost-surely. Also assume that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN in the $\theta$-direction, for all $\zeta$ in an $r_H$-neighborhood of $\zeta = 0$ and that the efficient Fisher information $\tilde{I}_{\theta_0.\eta_0}$ is nonsingular. Furthermore, assume that*:

(i) *For all $\rho > 0$, the Hellinger metric entropy satisfies, $N(\rho, H, d_H) < \infty$ and the nuisance prior satisfies $\Pi_H(K(\rho)) > 0$.*

(ii) *For every $M > 0$, there exists an $L > 0$ such that for all $\rho > 0$ and large enough $n$, $K(\rho) \subset K_n(L\rho, M)$.*

*Assume also that for every bounded, stochastic $(h_n)$*:

(iii) *There exists an $r > 0$ such that, $U_n(r, h_n) = O(1)$.*

(iv) *Hellinger distances satisfy, $\sup_{\eta \in H} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) = O(n^{-1/2})$, and that*

(v) *For every $(M_n)$, $M_n \to \infty$, the posterior satisfies,*

$$\Pi_n(\|h\| \le M_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 1.$$

*Then the sequence of marginal posteriors for $\theta$ converges in total variation to a normal distribution*,

$$\sup_A \left|\Pi_n(h \in A \mid X_1, \ldots, X_n) - N_{\tilde{\Delta}_n, \tilde{I}^{-1}_{\theta_0,\eta_0}}(A)\right| \xrightarrow{P_0} 0,$$

*centered on $\tilde{\Delta}_n$ with covariance matrix $\tilde{I}^{-1}_{\theta_0,\eta_0}$.*

PROOF. Under conditions (i), (ii), (iv) and the stochastic LAN assumption, the assertion of Corollary 3.3 holds. Due to condition (iii), condition (2.8) is satisfied for large enough $n$. Condition (v) then suffices for the assertion of Theorem 5.1. □

A critical note can be made regarding the qualification "rate-free" of Corollary 5.2: although the nuisance rate does not make an explicit appearance, rate restrictions may arise upon further analysis of condition (v). Indeed this is the case in the example of Section 7, where smoothness requirements on the regression family are interpretable as restrictions on the nuisance rate. However, semiparametric models exist, in which no restrictions on nuisance rates arise in this way: if $H$ is a convex subspace of a linear space, and the dependence $\eta \mapsto P_{\theta,\eta}$ is linear (a so-called *convex-linear* model, e.g., mixture models, errors-in-variables regression and other information-loss models), the construction of suitable tests (cf. Le Cam [30], Birgé [2, 3]) does not involve Hellinger metric entropy numbers or restrictions on nuisance rates of convergence. Consequently there exists a class of semiparametric examples for which Corollary 5.2 stays rate-free even after further analysis of its condition (v).

As shown in [26], the particular form of the limiting posterior in Theorem 5.1 is a consequence of local asymptotic normality, in this case imposed through (4.2). The marginal posterior converges exactly to the asymptotic sampling distribution of a frequentist best-regular estimator as a consequence. Other expansions (e.g., in LAN models for non-i.i.d. data or under the condition of *local asymptotic exponentiality* (Ibragimov and Has'minskii [19])) can be dealt with in the same manner if we adapt the limiting form of the posterior accordingly, giving rise to other (e.g., one-sided exponential) limit distributions (see Kleijn and Knapik [25]).

**6. Marginal posterior convergence at parametric rate.** Condition (5.2) in Theorem 5.1 requires that the posterior measures of a sequence of model subsets of the form

$$\Theta_n \times H = \left\{(\theta, \eta) \in \Theta \times H : \sqrt{n}\|\theta - \theta_0\| \leq M_n\right\} \tag{6.1}$$

converge to one in $P_0$-probability, for every sequence $(M_n)$ such that $M_n \to \infty$. Essentially, this condition enables us to restrict the proof of Theorem 5.1 to the shrinking domain in which (4.2) applies. In this section, we consider two distinct

approaches: the first (Lemma 6.1) is based on bounded likelihood ratios (see also condition (B3) of Theorem 8.2 in Lehmann and Casella [32]). The second is based on the behavior of misspecified parametric posteriors (Theorem 6.2). The latter construction illustrates the intricacy of this section's subject most clearly and provides some general insight. Methods proposed here are neither compelling nor exhaustive; we simply put forth several possible approaches and demonstrate the usefulness of one of them in Section 7.

LEMMA 6.1 [Marginal parametric rate (I)]. *Let the sequence of maps $\theta \mapsto S_n(\theta)$ be $P_0$-almost-surely continuous and such that (4.2) is satisfied. Furthermore, assume that there exists a constant $C > 0$ such that for any $(M_n)$, $M_n \to \infty$,*

$$P_0^n\left(\sup_{\eta\in H}\sup_{\theta\in\Theta_n^c}\mathbb{P}_n \log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} \le -\frac{CM_n^2}{n}\right) \to 1. \tag{6.2}$$

*Then, for any nuisance prior $\Pi_H$ and parametric prior $\Pi_\Theta$, thick at $\theta_0$,*

$$\Pi(n^{1/2}\|\theta-\theta_0\| > M_n \mid X_1,\dots,X_n) \xrightarrow{P_0} 0 \tag{6.3}$$

*for any $(M_n)$, $M_n \to \infty$.*

PROOF. Let $(M_n)$, $M_n \to \infty$ be given. Define $(A_n)$ to be the events in (6.2) so that $P_0^n(A_n^c) = o(1)$ by assumption. In addition, let

$$B_n = \left\{\int_\Theta S_n(\theta)\,d\Pi_\Theta(\theta) \ge e^{-CM_n^2/2}S_n(\theta_0)\right\}.$$

By (4.2) and Lemma 6.3, $P_0^n(B_n^c) = o(1)$ as well. Then

$$\begin{aligned}
&P_0^n\Pi(\theta\in\Theta_n^c \mid X_1,\dots,X_n)\\
&\quad\le P_0^n\Pi(\theta\in\Theta_n^c \mid X_1,\dots,X_n)1_{A_n\cap B_n} + o(1)\\
&\quad\le e^{CM_n^2/2}P_0^n\left(S_n(\theta_0)^{-1}\int_H\int_{\Theta_n^c}\prod_{i=1}^n\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i)\prod_{i=1}^n\frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i)\,d\Pi_\Theta\,d\Pi_H 1_{A_n}\right)\\
&\qquad + o(1)\\
&\quad= o(1),
\end{aligned}$$

which proves (6.3). □

Although applicable directly in the model of Section 7, most other examples would require variations. Particularly, if the full, nonparametric posterior is known to concentrate on a sequence of model subsets $(V_n)$, then Lemma 6.1 can be preceded by a decomposition of $\Theta \times H$ over $V_n$ and $V_n^c$, reducing condition (6.2) to a supremum over $V_n^c$ (see Section 2.4 in Kleijn [24] and the discussion following the following theorem).

Our second approach assumes such concentration of the posterior on model subsets, for example, deriving from nonparametric consistency in a suitable form. Though the proof of Theorem 6.2 is rather straightforward, combination with results in misspecified parametric models [26] leads to the observation that marginal parametric rates of convergence can be ruined by a bias.

THEOREM 6.2 [Marginal parametric rate (II)]. *Let $\Pi_\Theta$ and $\Pi_H$ be given. Assume that there exists a sequence $(H_n)$ of subsets of $H$, such that the following two conditions hold*:

(i) *The nuisance posterior concentrates on $H_n$ asymptotically,*

$$\Pi(\eta \in H \setminus H_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 0. \tag{6.4}$$

(ii) *For every $(M_n)$, $M_n \to \infty$,*

$$P_0^n \sup_{\eta \in H_n} \Pi(n^{1/2}\|\theta - \theta_0\| > M_n \mid \eta, X_1, \ldots, X_n) \to 0. \tag{6.5}$$

*Then the marginal posterior for $\theta$ concentrates at parametric rate, that is,*

$$\Pi(n^{1/2}\|\theta - \theta_0\| > M_n \mid \eta, X_1, \ldots, X_n) \xrightarrow{P_0} 0$$

*for every sequence $(M_n)$, $M_n \to \infty$.*

PROOF. Let $(M_n)$, $M_n \to \infty$ be given, and consider the posterior for the complement of (6.1). By assumption (i) of the theorem and Fubini's theorem,

$$\begin{aligned} &P_0^n \Pi(\theta \in \Theta_n^c \mid X_1, \ldots, X_n) \\ &\quad \le P_0^n \int_{H_n} \Pi(\theta \in \Theta_n^c \mid \eta, X_1, \ldots, X_n)\, d\Pi(\eta \mid X_1, \ldots, X_n) + o(1) \\ &\quad \le P_0^n \sup_{\eta \in H_n} \Pi(n^{1/2}\|\theta - \theta_0\| > M_n \mid \eta, X_1, \ldots, X_n) + o(1), \end{aligned}$$

the first term of which is $o(1)$ by assumption (ii) of the theorem. □

Condition (ii) of Theorem 6.2 has an interpretation in terms of misspecified parametric models (Kleijn and van der Vaart [26] and Kleijn [24]). For fixed $\eta \in H$, the $\eta$-conditioned posterior on the parametric model $\mathscr{P}_\eta = \{P_{\theta,\eta} : \theta \in \Theta\}$ is required to concentrate in $n^{-1/2}$-neighborhoods of $\theta_0$ under $P_0$. However, this misspecified posterior concentrates around $\Theta^*(\eta) \subset \Theta$, the set of points in $\Theta$ where the Kullback–Leibler divergence of $P_{\theta,\eta}$ with respect to $P_0$, is minimal. Assuming that $\Theta^*(\eta)$ consists of a unique minimizer $\theta^*(\eta)$, the dependence of the Kullback–Leibler divergence on $\eta$ must be such that

$$\sup_{\eta \in H_n} \|\theta^*(\eta) - \theta_0\| = o(n^{-1/2}) \tag{6.6}$$

in order for posterior concentration to occur on the strips (6.1). In other words, minimal Kullback–Leibler divergence may bias the (points of convergence of) $\eta$-conditioned parametric posteriors to such an extent that consistency of the marginal posterior for $\theta$ is ruined.

The occurrence of this bias is a property of the semiparametric model rather than a peculiarity of the Bayesian approach: when (point-)estimating with solutions to score equations, for example, the same bias occurs (see, e.g., Theorem 25.59 in [43] and subsequent discussion). Frequentist literature also offers some guidance toward mitigation of this circumstance. First of all, it is noted that the bias indicates the existence of a better (i.e., bias-less) choice of parametrization to ask the relevant semiparametric question. If the parametrization is fixed, alternative point-estimation methods may resolve bias, for example, through replacement of score equations by general estimating equations (see, e.g., Section 25.9 in [43]), loosely equivalent to introducing a suitable penalty in a likelihood maximization procedure.

For a so-called *curve-alignment model* with Gaussian prior, the no-bias problem has been addressed and resolved in a fully Bayesian manner by Castillo [5]: like a penalty in an ML procedure, Castillo's (rather subtle choice of) prior guides the procedure away from the biased directions and produces Bernstein–von Mises efficiency of the marginal posterior. A most interesting question concerns generalization of Castillo's intricate construction to more general Bayesian context.

Recalling definitions (2.5) and (4.1), we conclude this section with a lemma used in the proof of Lemma 6.1 to lower-bound the denominator of the marginal posterior.

LEMMA 6.3. *Let the sequence of maps $\theta \mapsto S_n(\theta)$ be $P_0$-almost-surely continuous and such that (4.2) is satisfied. Assume that $\Pi_\Theta$ is thick at $\theta_0$ and denoted by $\Pi_n$ in the local parametrization in terms of $h$. Then*

$$P_0^n\left(\int s_n(h)\,d\Pi_n(h) < a_n s_n(0)\right) \to 0 \tag{6.7}$$

*for every sequence $(a_n)$, $a_n \downarrow 0$.*

PROOF. Let $M > 0$ be given, and define $C = \{h : \|h\| \le M\}$. Denote the rest-term in (4.2) by $h \mapsto R_n(h)$. By continuity of $\theta \mapsto S_n(\theta)$, $\sup_{h\in C}|R_n(h)|$ converges to zero in $P_0$-probability. If we choose a sequence $(\kappa_n)$ that converges to zero slowly enough, the corresponding events $B_n = \{\sup_C |R_n(h)| \le \kappa_n\}$, satisfy $P_0^n(B_n) \to 1$. Next, let $(K_n)$, $K_n \to \infty$ be given. There exists a $\pi > 0$ such that $\inf_{h\in C} d\Pi_n/d\mu(h) \ge \pi$, for large enough $n$. Combining, we find

$$\begin{aligned} &P_0^n\left(\int \frac{s_n(h)}{s_n(0)}\,d\Pi_n(h) \le e^{-K_n^2}\right) \\ &\quad\le P_0^n\left(\left\{\int_C \frac{s_n(h)}{s_n(0)}\,d\mu(h) \le \pi^{-1}e^{-K_n^2}\right\} \cap B_n\right) + o(1). \end{aligned} \tag{6.8}$$

On $B_n$, the integral LAN expansion is lower bounded so that, for large enough $n$,

$$\begin{aligned} &P_0^n\left(\left\{\int_C \frac{s_n(h)}{s_n(0)}\,d\mu(h) \le \pi^{-1}e^{-K_n^2}\right\}\cap B_n\right) \\ &\quad\le P_0^n\left(\int_C e^{h^T\mathbb{G}_n\tilde{\ell}_{\theta_0,\eta_0}}\,d\mu(h) \le \pi^{-1}e^{-K_n^2/4}\right) \end{aligned} \tag{6.9}$$

since $\kappa_n \le \frac{1}{2}K_n^2$ and $\sup_{h\in C}|h^T\tilde{I}_{\theta_0,\eta_0}h| \le M^2\|\tilde{I}_{\theta_0,\eta_0}\| \le \frac{1}{4}K_n^2$, for large enough $n$. Conditioning $\mu$ on $C$, we apply Jensen's inequality to note that, for large enough $n$,

$$\begin{aligned} &P_0^n\left(\int_C e^{h^T\mathbb{G}_n\tilde{\ell}_{\theta_0,\eta_0}}\,d\mu(h) \le \pi^{-1}e^{-K_n^2/4}\right) \\ &\quad\le P_0^n\left(\int h^T\mathbb{G}_n\tilde{\ell}_{\theta_0,\eta_0}\,d\mu(h|C) \le -\frac{1}{8}K_n^2\right) \end{aligned}$$

since $-\log\pi\mu(C) \le \frac{1}{8}K_n^2$, for large enough $n$. The probability on the right is bounded further by Chebyshev's and Jensen's inequalities and can be shown to be of order $O(K_n^{-4})$. Combining with (6.8) and (6.9) then proves (6.7). □

**7. Semiparametric regression.** The *partial linear regression* model describes the observation of an i.i.d. sample $X_1, X_2, \ldots$ of triplets $X_i = (U_i, V_i, Y_i) \in \mathbb{R}^3$, each assumed to be related through the regression equation

$$Y = \theta_0 U + \eta_0(V) + e, \tag{7.1}$$

where $e \sim N(0,1)$ is independent of $(U,V)$. Interpreting $\eta_0$ as a nuisance parameter, we wish to estimate $\theta_0$. It is assumed that $(U,V)$ has an unknown distribution $P$, Lebesgue absolutely continuous with density $p:\mathbb{R}^2 \to \mathbb{R}$. The distribution $P$ is assumed to be such that $PU = 0$, $PU^2 = 1$ and $PU^4 < \infty$. At a later stage, we also impose $P(U - \mathrm{E}[U|V])^2 > 0$ and a smoothness condition on the conditional expectation $v \mapsto \mathrm{E}[U|V=v]$.

As is well known [1, 7, 33, 43], penalized ML estimation in a smoothness class of regression functions leads to a consistent estimate of the nuisance and efficient point-estimation of the parameter of interest. The necessity of a penalty signals that the choice of a prior for the nuisance is a critical one. Kimeldorf and Wahba [23] assume that the regression function lies in the Sobolev space $H^k[0,1]$ (see [44] for definition), and define the nuisance prior through the Gaussian process

$$\eta(t) = \sum_{i=0}^{k} Z_i\frac{t^i}{i!} + (I_{0+}^k W)(t), \tag{7.2}$$

where $W = \{W_t : t \in [0,1]\}$ is Brownian motion on $[0,1]$, $(Z_0, \ldots, Z_k)$ form a $W$-independent, $N(0,1)$-i.i.d. sample and $I_{0+}^k$ denotes $(I_{0+}^1 f)(t) = \int_0^t f(s)\,ds$ or $I_{0+}^{i+1}f = I_{0+}^1 I_{0+}^i f$ for all $i \ge 1$. The prior process $\eta$ is zero-mean Gaussian of (Hölder-)smoothness $k + 1/2$ and the resulting posterior mean for $\eta$ concentrates

asymptotically on the smoothing spline that solves the penalized ML problem [39, 46]. MCMC simulations based on Gaussian priors have been carried out by Shively, Kohn and Wood [41].

Here, we reiterate the question of how frequentist sufficient conditions are expressed in a Bayesian analysis based on Corollary 5.2. We show that with a nuisance of known (Hölder-)smoothness greater than $1/2$, the process (7.2) provides a prior such that the marginal posterior for $\theta$ satisfies the Bernstein–von Mises limit. To facilitate the analysis, we think of the regression function and the process (7.2) as elements of the Banach space $(C[0,1], \|\cdot\|_\infty)$. At a later stage, we relate to Banach subspaces with stronger norms to complete the argument.

THEOREM 7.1. *Let $X_1, X_2, \ldots$ be an i.i.d. sample from the partial linear model (7.1) with $P_0 = P_{\theta_0,\eta_0}$ for some $\theta_0 \in \Theta$, $\eta_0 \in H$. Assume that $H$ is a subset of $C[0,1]$ of finite metric entropy with respect to the uniform norm and that $H$ forms a $P_0$-Donsker class. Regarding the distribution of $(U,V)$, suppose that $PU = 0$, $PU^2 = 1$ and $PU^4 < \infty$, as well as $P(U - \mathrm{E}[U|V])^2 > 0$, $P(U - \mathrm{E}[U|V])^4 < \infty$ and $v \mapsto \mathrm{E}[U|V = v] \in H$. Endow $\Theta$ with a prior that is thick at $\theta_0$ and $C[0,1]$ with a prior $\Pi_H$ such that $H \subset \mathrm{supp}(\Pi_H)$. Then the marginal posterior for $\theta$ satisfies the Bernstein–von Mises limit,*

$$\sup_{B\in\mathscr{B}} \left|\Pi\big(\sqrt{n}(\theta-\theta_0)\in B \mid X_1,\ldots,X_n\big) - N_{\tilde\Delta_n,\tilde I^{-1}_{\theta_0,f_0}}(B)\right| \xrightarrow{P_0} 0, \tag{7.3}$$

*where $\tilde\ell_{\theta_0,\eta_0}(X) = e(U - \mathrm{E}[U|V])$ and $\tilde I_{\theta_0,\eta_0} = P(U - \mathrm{E}[U|V])^2$.*

PROOF. For any $\theta$ and $\eta$, $-P_{\theta_0,\eta_0}\log(p_{\theta,\eta}/p_{\theta_0,\eta_0}) = \frac{1}{2}P_{\theta_0,\eta_0}((\theta-\theta_0)U + (\eta-\eta_0)(V))^2$, so that for fixed $\theta$, minimal KL-divergence over $H$ obtains at $\eta^*(\theta) = \eta_0 - (\theta-\theta_0)\mathrm{E}[U|V]$, $P$-almost-surely. For fixed $\zeta$, the submodel $\theta \mapsto Q_{\theta,\zeta}$ satisfies

$$\begin{aligned}
&\log\prod_{i=1}^n \frac{p_{\theta_0+n^{-1/2}h_n,\eta^*(\theta_0+n^{-1/2}h_n)+\zeta}}{p_{\theta_0,\eta_0+\zeta}}(X_i)\\
&\quad= \frac{h_n}{\sqrt{n}}\sum_{i=1}^n g_\zeta(X_i) - \frac{1}{2}{h_n}^2 P_{\theta_0,\eta_0+\zeta}{g_\zeta}^2\\
&\qquad+ \frac{1}{2}{h_n}^2(\mathbb{P}_n - P)(U - \mathrm{E}[U|V])^2
\end{aligned} \tag{7.4}$$

for all stochastic $(h_n)$, with $g_\zeta(X) = e(U - \mathrm{E}[U|V])$, $e = Y - \theta_0 U - (\eta_0 + \zeta)(V) \sim N(0,1)$ under $P_{\theta_0,\eta_0+\zeta}$. Since $PU^2 < \infty$, the last term on the right is $o_{P_{\theta_0,\eta_0+\zeta}}(1)$ if $(h_n)$ is bounded in probability. We conclude that $\theta \mapsto Q_{\theta,\zeta}$ is stochastically LAN. In addition, (7.4) shows that $h \mapsto s_n(h)$ is continuous for ev-

ery $n \geq 1$. By assumption, $\tilde{I}_{\theta_0,\eta_0} = P_0 g_0^2 = P(U - \mathrm{E}[U|V])^2$ is strictly positive. We also observe at this stage that $H$ is totally bounded in $C[0,1]$, so that there exists a constant $D > 0$ such that $\|H\|_\infty \leq D$.

For any $x \in \mathbb{R}^3$ and all $\zeta$, the map $\theta \mapsto \log q_{\theta,\zeta}/q_{\theta_0,\zeta}(x)$ is continuously differentiable on all of $\Theta$, with score $g_{\theta,\zeta}(X) = e(U - \mathrm{E}[U|V]) + (\theta - \theta_0)(U - \mathrm{E}[U|V])^2$. Since $Q_{\theta,\zeta} g_{\theta,\zeta}^2 = P(U - \mathrm{E}[U|V])^2 + (\theta - \theta_0)^2 P(U - \mathrm{E}[U|V])^4$ does not depend on $\zeta$ and is bounded over $\theta \in [\theta_0 - \rho, \theta_0 + \rho]$, Lemma 4.3 says that $U(\rho_n, h_n) = O(1)$ for all $\rho_n \downarrow 0$ and all bounded, stochastic $(h_n)$. So for this model, we can apply the rate-free version of the semiparametric Bernstein–von Mises theorem, Corollary 5.2, and its condition (iii) is satisfied.

Regarding condition (ii) of Corollary 5.2, we first note that, for $M > 0$, $n \geq 1$, $\eta \in H$,

$$\sup_{\|h\| \leq M} -\log \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}} = \frac{M^2}{2n} U^2 + \frac{M}{\sqrt{n}} \bigl| U\bigl(e - (\eta - \eta_0)(V)\bigr) \bigr| - e(\eta - \eta_0)(V) + \frac{1}{2}(\eta - \eta_0)^2(V),$$

where $e \sim N(0,1)$ under $P_{\theta_0,\eta_0}$. With the help of the boundedness of $H$, the independence of $e$ and $(U, V)$ and the assumptions on the distribution of $(U, V)$, it is then verified that condition (ii) of Corollary 5.2 holds. Turning to condition (i), it is noted that for all $\eta_1, \eta_2 \in H$, $d_H(\eta_1, \eta_2) \leq -P_{\theta_0,\eta_2} \log(p_{\theta_0,\eta_1}/p_{\theta_0,\eta_2}) = \frac{1}{2}\|\eta_1 - \eta_2\|_{2,P}^2 \leq \frac{1}{2}\|\eta_1 - \eta_2\|_\infty^2$. Hence, for any $\rho > 0$, $N(\rho, \mathscr{P}_{\theta_0}, d_H) \leq N((2\rho)^{1/2}, H, \|\cdot\|_\infty) < \infty$. Similarly, one shows that for all $\eta$ both $-P_0 \log(p_{\theta_0,\eta}/p_{\theta_0,\eta_0})$ and $P_0(\log(p_{\theta_0,\eta}/p_{\theta_0,\eta_0}))^2$ are bounded by $(\frac{1}{2} + D^2)\|\eta - \eta_0\|_\infty^2$. Hence, for any $\rho > 0$, $K(\rho)$ contains a $\|\cdot\|_\infty$-ball. Since $\eta_0 \in \mathrm{supp}(\Pi_H)$, we see that condition (i) of Corollary 5.2 holds. Noting that $(p_{\theta_n(h),\eta}/p_{\theta_0,\eta}(X))^{1/2} = \exp((h/2\sqrt{n})eU - (h^2/4n)U^2)$, one derives the $\eta$-independent upper bound,

$$H^2\bigl(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}\bigr) \leq \frac{M^2}{2n} P U^2 + \frac{M^3}{6n^2} P U^4 = O(n^{-1})$$

for all bounded, stochastic $(h_n)$, so that condition (iv) of Corollary 5.2 holds.

Concerning condition (v), let $(M_n)$, $M_n \to \infty$ be given and define $\Theta_n$ as in Section 6. Rewrite $\sup_{\eta \in H} \sup_{\theta \in \Theta_n^c} \mathbb{P}_n \log(p_{\theta,\eta}/p_{\theta_0,\eta}) = \sup_{\theta \in \Theta_n^c}((\theta - \theta_0) \times (\sup_\zeta \mathbb{P}_n ZW) - \frac{1}{2}(\theta - \theta_0)^2 \mathbb{P}_n W^2)$, where $Z = e_0 - \zeta(V)$, $W = U - \mathrm{E}[U|V]$. The maximum-likelihood estimate $\hat{\theta}_n$ for $\theta$ is therefore of the form $\hat{\theta}_n = \theta_0 + R_n$, where $R_n = \sup_\zeta \mathbb{P}_n ZW / \mathbb{P}_n W^2$. Note that $P_0 ZW = 0$ and that $H$ is assumed to be $P_0$-Donsker, so that $\sup_\zeta \mathbb{G}_n ZW$ is asymptotically tight. Since, in addition, $\mathbb{P}_n W^2 \to P_0 W^2$ almost surely and the limit is strictly positive by assumption,

$P_0^n(\sqrt{n}|R_n| > \frac{1}{4}M_n) = o(1)$. Hence,

$$P_0^n\Big(\sup_{\eta\in H}\sup_{\theta\in\Theta_n^c}\mathbb{P}_n\log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} > -\frac{CM_n^2}{n}\Big)$$
$$\le P_0^n\Big(\sup_{\theta\in\Theta_n^c}\Big(\frac{1}{4}|\theta-\theta_0|\frac{M_n}{n^{1/2}} - \frac{1}{2}(\theta-\theta_0)^2\Big)\mathbb{P}_n W^2 > -\frac{CM_n^2}{n}\Big) + o(1)$$
$$\le P_0^n(\mathbb{P}_n W^2 < 4C) + o(1).$$

Since $P_0 W^2 > 0$, there exists a $C > 0$ small enough such that the first term on the right-hand side is of order $o(1)$ as well, which shows that condition (6.2) is satisfied. Lemma 6.1 asserts that condition (v) of Corollary 5.2 is met as well. Assertion 7.3 now holds. □

In the following corollary we choose a prior by picking a suitable $k$ in (7.2) and conditioning on $\|\eta\|_\alpha < M$. The resulting prior is shown to be well defined below and is denoted $\Pi^k_{\alpha,M}$.

COROLLARY 7.2. *Let $\alpha > 1/2$ and $M > 0$ be given; choose $H = \{\eta \in C^\alpha[0,1] : \|\eta\|_\alpha < M\}$ and assume that $\eta_0 \in C^\alpha[0,1]$. Suppose the distribution of the covariates $(U, V)$ is as in Theorem 7.1. Then, for any integer $k > \alpha - 1/2$, the conditioned prior $\Pi^k_{\alpha,M}$ is well defined and gives rise to a marginal posterior for $\theta$ satisfying (7.3).*

PROOF. Choose $k$ as indicated; the Gaussian distribution of $\eta$ over $C[0,1]$ is based on the RKHS $H^{k+1}[0,1]$ and denoted $\Pi^k$. Since $\eta$ in (7.2) has smoothness $k + 1/2 > \alpha$, $\Pi^k(\eta \in C^\alpha[0,1]) = 1$. Hence, one may also view $\eta$ as a Gaussian element in the Hölder class $C^\alpha[0,1]$, which forms a separable Banach space even with strengthened norm $\|\cdot\| = \|\eta\|_\infty + \|\cdot\|_\alpha$, without changing the RKHS. The trivial embedding of $C^\alpha[0,1]$ into $C[0,1]$ is one-to-one and continuous, enabling identification of the prior induced by $\eta$ on $C^\alpha[0,1]$ with the prior $\Pi^k$ on $C[0,1]$. Given $\eta_0 \in C^\alpha[0,1]$ and a sufficiently smooth kernel $\phi_\sigma$ with bandwidth $\sigma > 0$, consider $\phi_\sigma \star \eta_0 \in H^{k+1}[0,1]$. Since $\|\eta_0 - \phi_\sigma \star \eta_0\|_\infty$ is of order $\sigma^\alpha$, and a similar bound exists for the $\alpha$-norm of the difference [44], $\eta_0$ lies in the closure of the RKHS both with respect to $\|\cdot\|_\infty$ and to $\|\cdot\|$. Particularly, $\eta_0$ lies in the support of $\Pi^k$, in $C^\alpha[0,1]$ with norm $\|\cdot\|$. Hence, $\|\cdot\|$-balls centered on $\eta_0$ receive nonzero prior mass, that is, $\Pi^k(\|\eta - \eta_0\| < \rho) > 0$ for all $\rho > 0$. Therefore, $\Pi^k(\|\eta - \eta_0\|_\infty < \rho, \|\eta\|_\alpha < \|\eta_0\|_\alpha + \rho) > 0$, which guarantees that $\Pi^k(\|\eta - \eta_0\|_\infty < \rho, \|\eta\|_\alpha < M) > 0$, for small enough $\rho > 0$. This implies that $\Pi^k(\|\eta\|_\alpha < M) > 0$, and

$$\Pi^k_{\alpha,M}(B) = \Pi^k(B \mid \|\eta\|_\alpha < M)$$

is well defined for all Borel-measurable $B \subset C[0,1]$. Moreover, it follows that $\Pi^k_{\alpha,M}(\|\eta - \eta_0\|_\infty < \rho) > 0$ for all $\rho > 0$. We conclude that $k$ times integrated Brownian motion started at random, conditioned to be bounded by $M$ in $\alpha$-norm, gives rise to a prior that satisfies $\operatorname{supp}(\Pi^k_{\alpha,M}) = H$. As is well-known [45], the entropy numbers of $H$ with respect to the uniform norm satisfy, for every $\rho > 0$, $N(\rho, H, \|\cdot\|_\infty) \leq K\rho^{-1/\alpha}$, for some constant $K > 0$ that depends only on $\alpha$ and $M$. The associated bound on the bracketing entropy gives rise to finite bracketing integrals, so that $H$ universally Donsker. Then, if the distribution of the covariates $(U, V)$ is as assumed in Theorem 7.1, the Bernstein–von Mises limit (7.3) holds. □

**Acknowledgments.** The authors would like to thank D. Freedman, A. Gamst, C. Klaassen, B. Knapik and A. van der Vaart for valuable discussions and suggestions. B. J. K. Kleijn thanks U.C. Berkeley's Statistics Dept. and Cambridge's Isaac Newton Institute for their hospitality.

## REFERENCES

[1] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*, 2nd ed. Springer, New York. MR1623559

[2] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129

[3] BIRGÉ, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282. MR0764150

[4] BOUCHERON, S. and GASSIAT, E. (2009). A Bernstein–von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588

[5] CASTILLO, I. (2011). Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. Preprint, CNRS.

[6] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99.

[7] CHEN, H. and SHIAU, J. J. H. (1991). A two-stage spline smoothing method for partially linear models. *J. Statist. Plann. Inference* **27** 187–201. MR1096678

[8] CHENG, G. and KOSOROK, M. R. (2008). General frequentist properties of the posterior profile distribution. *Ann. Statist.* **36** 1819–1853. MR2435457

[9] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[10] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. *Princeton Mathematical Series* **9**. Princeton Univ. Press, Princeton, NJ. MR0016588

[11] DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.

[12] DIACONIS, P. W. and FREEDMAN, D. (1998). Consistency of Bayes estimates for nonparametric regression: Normal theory. *Bernoulli* **4** 411–444. MR1679791

[13] FISHER, R. A. (1959). *Statistical Methods and Scientific Inference*, 2nd ed. Oliver and Boyd, London.

[14] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119

[15] FREEDMAN, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403. MR0158483

[16] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[17] HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330. MR0283911

[18] HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 175–194. Univ. California Press, Berkeley, CA. MR0400513

[19] IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. *Applications of Mathematics* **16**. Springer, New York. MR0620321

[20] JOHNSTONE, I. (2010). High dimensional Bernstein–von Mises: Simple examples. In *Borrowing Strength*: *Theory Powering Applications—A Festschrift for Lawrence D. Brown* (J. Berger, T. Cai and I. Johnstone, eds.) 87–98. IMS, Beachwood, OH.

[21] KIM, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. MR2283713

[22] KIM, Y. and LEE, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32** 1492–1512. MR2089131

[23] KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41** 495–502. MR0254999

[24] KLEIJN, B. (2003). Bayesian asymptotics under misspecification. Ph.D. thesis, Free Univ. Amsterdam.

[25] KLEIJN, B. and KNAPIK, B. (2012). Semiparametric posterior limits under local asymptotic exponentiality. Preprint, Korteweg-de Vries Institute, Amsterdam.

[26] KLEIJN, B. and VAN DER VAART, A. (2008). The Bernstein–von Mises theorem under misspecification. Preprint.

[27] KLEIJN, B. J. K. and VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395

[28] LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277–329. MR0054913

[29] LE CAM, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 245–261. Univ. California Press, Berkeley, CA. MR0415819

[30] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411

[31] LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics*: *Some Basic Concepts*. Springer, New York. MR1066869

[32] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

[33] MAMMEN, E. and VAN DE GEER, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035. MR1447739

[34] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168

[35] RIVOIRARD, V. and ROUSSEAU, J. (2009). Bernstein–von Mises theorem for linear functionals of the density. Preprint. Available at arXiv:0908.4167v1.

[36] ROBERT, C. P. (2001). *The Bayesian Choice*: *From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. Springer, New York. MR1835885

[37] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378

[38] SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1802. MR1193312

[39] SHEN, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* **97** 222–235. MR1947282

[40] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[41] SHIVELY, T. S., KOHN, R. and WOOD, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *J. Amer. Statist. Assoc.* **94** 777–804.

[42] STEIN, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 187–195. Univ. California Press, Berkeley. MR0084921

[43] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[44] 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

[45] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[46] WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372. MR0522220

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94710-3860
USA
E-MAIL: bickel@stat.berkeley.edu

KORTEWEG-DE VRIES INSTITUTE
UNIVERSITY OF AMSTERDAM
P.O. BOX 94248
1090 GE, AMSTERDAM
THE NETHERLANDS
E-MAIL: b.kleijn@uva.nl