The Annals of Statistics
2021, Vol. 49, No. 1, 182–202
https://doi.org/10.1214/20-AOS1952

FREQUENTIST VALIDITY OF BAYESIAN LIMITS

BY B. J. K. KLEIJN[1]

[1]*Korteweg–de Vries Institute for Mathematics, University of Amsterdam, B.Kleijn@uva.nl*

To the frequentist who computes posteriors, not all priors are useful asymptotically: in this paper, a Bayesian perspective on test sequences is proposed and Schwartz's Kullback–Leibler condition is generalised to widen the range of frequentist applications of posterior convergence. With *Bayesian tests* and a weakened form of contiguity termed *remote contiguity*, we prove simple and fully general frequentist theorems, for posterior consistency and rates of convergence, for consistency of posterior odds in model selection, and for conversion of sequences of credible sets into sequences of confidence sets with asymptotic coverage one. For frequentist uncertainty quantification, this means that a prior inducing remote contiguity allows one to enlarge credible sets of calculated, simulated or approximated posteriors to obtain asymptotically consistent confidence sets.

1. Introduction. In this paper, we examine for which model-prior pairs Bayesian asymptotic conclusions give rise to conclusions valid in the frequentist sense: how Doob's prior-almost-sure consistency is strengthened to reach Schwartz's frequentist conclusion; how a test that is consistent prior-almost-surely becomes a test that is consistent in *all* points of the model; and how sequences of Bayesian credible sets can serve as frequentist confidence sets of asymptotic coverage one.

Frequentist posterior consistency conditions focus on prior-model pairs satisfying *Schwartz's Kullback–Leibler (KL) lower bound* [34]. Before generalizing to a contiguity argument for sequential approximation, let us focus on simple circumstances in which Schwartz's condition cannot be applied [23].

EXAMPLE 1.1. Consider $X_1, X_2, \ldots$ that are *i.i.d.*-P_0 with continuous, nonzero Lebesgue density $p_0 : \mathbb{R} \to \mathbb{R}$ on an interval of known width (say, 1) but unknown location. Parametrize with a continuous density η on $[0, 1]$ with $\eta(x) > 0$ for all $x \in (0, 1)$ and $\theta \in \mathbb{R}$: $p_{\theta,\eta}(x) = \eta(x - \theta)1_{[\theta,\theta+1]}(x)$. If $\theta \neq \theta'$, then

$$-P_{\theta,\eta} \log \frac{p_{\theta',\eta'}}{p_{\theta,\eta}} = \infty$$

for all η, η', so KL neighbourhoods do not have any extent in the θ-direction and *no prior is a KL prior in this model*. Nonetheless, the posterior is consistent (see Examples 3.7 and 4.3).

Similarly, heavy tails can undermine the *Ghosal–Ghosh–van der Vaart (GGV) condition* [14]: consider an *i.i.d.* sample of integers from a distribution P_a $(a \geq 1)$, defined by $p_a(k) = P_a(X = k) = Z_a^{-1} k^{-a} (\log k)^{-3}$, for all $k \geq 2$. For $a = 1$, $b > 1$,

$$-P_a \log \frac{p_b}{p_a} < \infty, \qquad P_a \left(\log \frac{p_b}{p_a} \right)^2 = \infty.$$

No prior can satisfy the GGV condition for neighbourhoods of $a = 1$. If we change the third power of the logarithm in $p_a(k)$ to a square, Schwartz's KL-priors also cease to exist.

Received November 2017; revised January 2020.
MSC2020 subject classifications. Primary 62G05, 62G20; secondary 62G10, 62C10, 62B15.
Key words and phrases. Contiguity, frequentist consistency, posterior consistency, posterior rate of convergence, posterior odds, model selection, credible set, coverage, uncertainty quantification.

Standard frequentist conditions for suitability of priors (given the model) may therefore imply unnecessary disqualification in comparable but less-obvious ways in more complicated models. Combined with natural questions regarding generalization (e.g., what form does Schwartz's theorem take when data are non-*i.i.d.*? how are credible sets useful to the frequentist? if posteriors and tests are so close, what about frequentist model selection with posteriors? *etcetera*), the examples suggest we look for generalization of KL and GGV conditions. Below we argue that the central property to enable frequentist interpretation of posterior asymptotics is *remote contiguity* (see Section 3), a less stringent version of Le Cam's notion of contiguity [24]. We argue by example in Section 3.3 that remote contiguity has the potential to provide sequential approximations in nonparametric statistics, analogous to approximation by contiguous sequences in parametric setting [17]. This is illustrated by recent work [12, 33] that uses remote contiguity to prove consistency with respect to relatively complicated true data distributions by simpler, approximating sequences of max-stable distributions in extreme-value theory.

The second change we propose concerns weakening of Schwartz's testing condition: instead of requiring the existence of *uniform* test sequences [14, 34], we restrict type-I and type-II error probabilities (referred to together as "composite power") of tests when averaged with the prior. We show that these so-called *Bayesian tests* exist, if and only if, the posterior displays prior-almost-sure convergence [11] (rendering our understanding of Doob's consistency compatible with the occurrence of tests in Schwartz's theorem). Bayesian tests involve the prior in the testing condition, a property that is especially important in model-selection questions and is in line with *nonlocality* of priors, as in [19].

The most significant practical implication concerns frequentist uncertainty quantification: Theorem 6.4 shows that if the priors induce remote contiguity, sequences of credible sets can be enlarged to form sequences of confidence sets with asymptotic coverage one. Compare this with the main inferential conclusion of the Bernstein–von Mises theorem (asymptotic validity of credible sets as confidence sets in smooth parametric models [29]). In practice, a frequentist can calculate, simulate or approximate the posterior, construct associated credible sets and 'enlarge' them to obtain asymptotic confidence sets, provided his prior induces remote contiguity.

The rest of this paper is organized as follows: Section 2 focusses on an inequality that relates testing to posterior concentration. Section 3 introduces remote contiguity and the analogue of Le Cam's first lemma, applies remote contiguity in Bayesian context and compares contiguity with remote contiguity in the context of parametric and nonparametric regression. Section 4 applies remote contiguity to posterior consistency and convergence at a rate. In Section 5, frequentist model selection with posteriors is considered and Section 6 focusses on the conversion of sequences of credible sets into sequences of confidence sets with asymptotic coverage one. Section 7 discusses the conclusions. Although the main focus is theoretical, examples are provided throughout and Appendix B in the Supplementary Material [22] provides a larger example illustrating the main points, on goodness-of-fit testing with random walk data; more elaborate applications of remote contiguity and Bayesian limits are found in [12, 21, 33]. Definitions, notation and conventions roughly follow those of [27] and are collected in Appendix A in the Supplementary Material [22].

2. Posterior concentration and asymptotic tests. First, we consider a lemma that relates concentration of posterior mass in certain model subsets to test sequences that distinguish between those subsets: if consistent tests *exist*, the posterior concentrates its mass appropriately.

2.1. *Bayesian test sequences.* We propose to define test sequences immediately in Bayesian context by involving priors from the outset. Consider sequentially observed, (pos-

sibly non-*i.i.d.*) samples X^n, distributed according to $P_{\theta_0,n}$ for some $\theta_0 \in \Theta$, within a model $\theta \to P_{\theta,n}$. (More generally, refer to Appendix A in the Supplementary Material [22] for notation and conventions.)

DEFINITION 2.1. Given priors (Π_n) on the measurable spaces $(\Theta_n, \mathscr{G}_n)$, model subsets $(B_n), (V_n) \subset \mathscr{G}_n$ and $a_n \downarrow 0$, a sequence of $\mathscr{B}_n$-measurable maps $\phi_n : \mathscr{X}_n \to [0,1]$ is called a *Bayesian test sequence for B_n versus V_n (under Π_n) of composite power a_n*, if,

$$\int_{B_n} P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta) = o(a_n). \tag{2.1}$$

We say that (ϕ_n) is a *Bayesian test sequence for B_n versus V_n (under Π_n)* if (2.1) holds for some $a_n \downarrow 0$. If another Bayesian test sequence (ψ_n) exists of composite power $b_n = o(a_n)$, we say that (ψ_n) is *stronger* than (ϕ_n) for testing B_n versus V_n (under Π_n).

Bayesian test sequences and concentration of the posterior are related through the following lemma (in which n-dependence is suppressed for clarity).

LEMMA 2.2. *For any $B, V \in \mathscr{G}$ and any measurable $\phi : \mathscr{X} \to [0,1]$,*

$$\int_B P_\theta \Pi(V|X)\, d\Pi(\theta) \le \int_B P_\theta \phi \, d\Pi(\theta) + \int_V P_\theta(1-\phi)\, d\Pi(\theta). \tag{2.2}$$

PROOF. Due to Bayes's rule (A.2) and monotone convergence,

$$\int \big(1-\phi(X)\big)\Pi(V|X)\, dP^\Pi = \int_V P_\theta\big(1-\phi(X)\big)\, d\Pi(\theta).$$

Accordingly, $\int_B P_\theta(1-\phi)\Pi(V|X)\, d\Pi(\theta) \le \int (1-\phi)\Pi(V|X)\, dP^\Pi = \int_V P_\theta(1-\phi)\, d\Pi(\theta)$. Inequality (2.2) follows from the fact that $\Pi(V|X) \le 1$. □

So the mere existence of a test sequence is enough to guarantee posterior concentration, a fact expressed in n-dependent form through the following proposition. (*Local prior predictive distributions* $P_n^{\Pi_n|B_n}$ and $P_n^{\Pi_n|V_n}$ are defined in Definition A.2.)

PROPOSITION 2.3. *Let $(\mathscr{X}_n, \mathscr{B}_n)$, $(\Theta_n, \mathscr{G}_n)$, $(\mathscr{P}_n)$ and (Π_n) be given. Given sequences $(B_n), (V_n) \subset \mathscr{G}_n$ and $(a_n), (b_n), (c_n)$ such that $a_n = o(b_n \wedge c_n)$ and, $\Pi_n(B_n) \ge b_n > 0$, $\Pi_n(V_n) \ge c_n > 0$. If,*

(i) *there exists a Bayesian test sequence for B_n versus V_n of composite power a_n,*

then

(ii) *mutually, expected posterior weights vanish,*

$$P_n^{\Pi_n|B_n}\Pi\big(V_n|X^n\big) = o\big(a_n b_n^{-1}\big), \qquad P_n^{\Pi_n|V_n}\Pi\big(B_n|X^n\big) = o\big(a_n c_n^{-1}\big). \tag{2.3}$$

If $\Theta_n = B_n \cup V_n$ for all $n \ge 1$, then also (ii) $\Rightarrow$ (i).

PROOF. Assume (i). Then

$$P_n^{\Pi_n|B_n}\Pi\big(V_n|X^n\big) = b_n^{-1}\int_{B_n} P_{\theta,n}\Pi\big(V_n|X^n\big)\, d\Pi_n(\theta) = o\big(a_n b_n^{-1}\big)$$

(and analogously for V_n). Assume (ii) and $B_n \cup V_n = \Theta_n$. Define maps $\phi_n(X^n) = \Pi(V_n|X^n)$, then

$$b_n P_n^{\Pi_n|B_n}\Pi\big(V_n|X^n\big) + c_n P_n^{\Pi_n|V_n}\Pi\big(B_n|X^n\big) = o(a_n),$$

so (ϕ_n) defines a Bayesian test sequence for B_n versus V_n of composite power a_n. □

We come back to the equivalence of Bayesian test existence and posterior concentration in Section 2.2, as well as in Section 4. To illustrate how Proposition 2.3 relates to frequentist posterior concentration and how this involves remote contiguity, consider model subsets $V_n = V$ that are all equal to the complement of a neighbourhood U of P_0. The subsets $B_n = B$ are thought of as being even closer to the $P_{0,n}$, in such a way that the expectations of the random variables $X^n \mapsto \Pi(V|X^n)$ under $P_n^{\Pi_n|B_n}$ "dominate" their expectations under $P_{0,n}$ in a suitable way. Then sufficiency of prior mass b_n given composite power a_n, is enough to assert that $P_{0,n}\Pi(V|X^n) \to 0$. Remote contiguity makes this notion of domination precise.

REMARK 2.4. To conclude this section, take inequality (2.2) one step further, to obtain *Le Cam's inequality*,

$$P_{0,n}\Pi(V_n|X) \le \bigl\| P_{0,n} - P_n^{\Pi|B_n} \bigr\| + \int P_{\theta,n}\phi_n \, d\Pi_n(\theta|B_n) + \frac{\Pi_n(V_n)}{\Pi_n(B_n)} \int P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta|V_n) \tag{2.4}$$

for B_n and V_n such that $\Pi_n(B_n) > 0$ and $\Pi_n(V_n) > 0$. Inequality (2.4) is used in the proof of the Bernstein–von Mises theorem; see Section 8.4 of [29]. A less successful application pertains to nonparametric posterior rates of convergence for *i.i.d.* data, in an unpublished paper [26].

2.2. *Existence of Bayesian test sequences.* Lemma 2.2 and Proposition 2.3 require the existence of test sequences of the Bayesian type. That question is unfamiliar, frequentists are used to test sequences for uniform testing, like the minimax Hellinger tests of Section 16.4 in [27], or uniform tests for *weak* neighbourhoods [34] based on Hoeffding's inequality. Requiring the existence of a Bayesian test sequence *c.f.* (2.1) is quite different: first of all the existence of a Bayesian test sequence is linked directly to behaviour of the posterior itself.

THEOREM 2.5. *Let $(\Theta, \mathscr{G}, \Pi)$ be given and assume that there is a coupling $X \in \mathscr{X}^\infty$ with distribution P_θ and marginals $X^n \sim P_{\theta,n}$ for every $\theta \in \Theta$ and $n \ge 1$. For any $B, V \in \mathscr{G}$ with $\Pi(B) > 0$, $\Pi(V) > 0$, the following are equivalent*:

(i) *there are $\mathscr{B}_n$-msb. $\phi_n : \mathscr{X}_n \to [0,1]$ such that for Π-almost-all $\theta \in B, \theta' \in V$,*

$$\phi_n(X^n) \xrightarrow{P_\theta\text{-a.s.}} 0, \qquad \phi_n(X^n) \xrightarrow{P_{\theta'}\text{-a.s.}} 1,$$

(ii) *there are $\mathscr{B}_n$-msb. $\phi_n : \mathscr{X}_n \to [0,1]$ such that for Π-almost-all $\theta \in B, \theta' \in V$,*

$$P_{\theta,n}\phi_n \to 0, \qquad P_{\theta',n}(1-\phi_n) \to 0,$$

(iii) *there are $\mathscr{B}_n$-msb. $\phi_n : \mathscr{X}_n \to [0,1]$ such that*

$$\int_B P_{\theta,n}\phi_n \, d\Pi(\theta) + \int_V P_{\theta,n}(1-\phi_n)\, d\Pi(\theta) \to 0,$$

(iv) *for Π-almost-all $\theta \in B$, $\theta' \in V$,*

$$\Pi(V|X^n) \xrightarrow{P_{\theta,n}\text{-a.s.}} 0, \qquad \Pi(B|X^n) \xrightarrow{P_{\theta',n}\text{-a.s.}} 0.$$

PROOF. (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) by dominated convergence. Assume (iii) and note that by Lemma 2.2,

$$\int P_{\theta,n}\Pi(V|X^n)\, d\Pi(\theta|B) \to 0.$$

With the coupling X of the observations X^n, martingale convergence in $L^1(\mathcal{X}^\infty \times \Theta)$ (relative to the probability measure Π^* defined by $\Pi^*(A \times B) = \int_B P_\theta(A)\,d\Pi(\theta)$ for measurable $A \subset \mathcal{X}^\infty$ and $B \subset \Theta$), shows there is a measurable $g : \mathcal{X}^\infty \to [0,1]$ such that

$$\int P_\theta \big|\Pi\big(V|X^n\big) - g(X)\big|\,d\Pi(\theta|B) \to 0.$$

So $\int P_\theta g(X)\,d\Pi(\theta|B) = 0$, implying that $g = 0$, P_θ-almost-surely for Π-almost-all $\theta \in B$. Using martingale convergence again (now in $L^\infty(\mathcal{X}^\infty \times \Theta)$), conclude $\Pi(V|X^n) \to 0$, P_θ-almost-surely for Π-almost-all $\theta \in B$, from which (iv) follows. Choose $\phi(X^n) = \Pi(V|X^n, \theta \in B \cup V)$ to conclude that (iv) $\Rightarrow$ (i). □

The interpretation of this theorem is gratifying to supporters of the likelihood principle and pure Bayesians: distinctions between model subsets are Bayesian testable, *if and only if*, they are picked up by the posterior asymptotically, *if and only if*, there exists a pointwise test for B versus V that is Π-almost-surely consistent. There is also a constructivist interpretation: where the mathematical existence of test sequences to separate model subsets is fully *abstract*, posteriors can in principle be calculated and actually perform said separation *concretely*.

A second perspective on the existence of Bayesian tests arises from Doob's argument (see [11], as well as Section 17.7, Proposition 2 in [27]): if Θ is Polish (more precisely, a Borel subset of a complete metric spaces), there exists a Borel measurable $\vartheta : \mathcal{X}^\infty \to \Theta$ such that $P_\theta(\vartheta(X) = \theta) = 1$, for Π-almost-all $\theta \in \Theta$. (Note: here and elsewhere in *i.i.d.* setting, the parameter space Θ is the single-observation model $\mathcal{P}$, θ is the single-observation distribution P and $\theta \mapsto P_{\theta,n}$ is $P \mapsto P^n$.)

PROPOSITION 2.6. *Consider a model $\mathcal{P}$ of single-observation distributions P for i.i.d. data $(X_1, X_2, \ldots, X_n) \sim P^n$ $(n \geq 1)$. Assume that $\mathcal{P}$ is a Polish space with Borel prior Π. For any Borel set V there is a Bayesian test sequence for V versus $\mathcal{P} \setminus V$ under Π.*

PROOF. (See [11] and [27], Section 17.7, Proposition 1 with the indicator for V; see also [8].) Note that if $\vartheta : \mathcal{X}^\infty \to \Theta$ exists, then by martingale convergence in $L^\infty(\mathcal{X}^\infty \times \Theta)$, $\Pi(V|X^n) \to \int 1_V(\theta)\,d\Pi(\theta|X) = 1_V(\vartheta(X))$, Π^*-almost-surely, implying posterior convergence. To conclude, use that (iv) $\Rightarrow$ (i) in Theorem 2.5. □

Theorem 2.5 is seen to be related to Doob's consistency theorem, if we let V be the complement of any open neighbourhood of P_0 in Proposition 2.6.

Compared to uniform tests, Bayesian tests are quite abundant, because Bayesian testing really only amounts to testing of *barycentres*: to see this, let priors (Π_n) and $\mathcal{G}$-measurable model subsets B_n, V_n be given. For given tests (ϕ_n) and composite power a_n, write (2.1) as follows:

$$\Pi_n(B_n) P_n^{\Pi_n|B_n} \phi_n\big(X^n\big) + \Pi_n(V_n) P_n^{\Pi_n|V_n} \big(1 - \phi_n\big(X^n\big)\big) = o(a_n),$$

and note that what is required here, is a (weighted) test sequence for $(P_n^{\Pi_n|B_n})$ versus $(P_n^{\Pi_n|V_n})$. The likelihood-ratio test (denote densities for $P_n^{\Pi_n|B_n}$ and $P_n^{\Pi_n|V_n}$ by $p_{B_n,n}$ and $p_{V_n,n}$),

$$\phi_n\big(X^n\big) = 1_{\{\Pi_n(V_n) p_{V_n,n}(X^n) > \Pi_n(B_n) p_{B_n,n}(X^n)\}},$$

is optimal and has composite power $\|\Pi_n(B_n) P_n^{\Pi_n|B_n} \wedge \Pi_n(V_n) P_n^{\Pi_n|B_n}\|$. (Here, $P \wedge Q$ denotes the *minimum* of P and Q [27], the largest (sub-probability) measure λ that satisfies $\lambda \leq P$ and $\lambda \leq Q$. Explicitly, if $\mu = P + Q$ and $p = dP/d\mu$, $q = dQ/d\mu$, the minimum is given by $(P \wedge Q)(A) = \int_A (p(x) \wedge q(x))\,d\mu(x)$.) This leads to the following lemma based on the so-called Hellinger transform (see Section 16.4, Remark 1 in [27]).

LEMMA 2.7. *Fix $n \geq 1$ and let a prior (Π_n) and measurable model subsets B_n, V_n be given. There exists a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that*

$$\begin{aligned}\int_{B_n} P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta) \\ \leq \int \big(\Pi_n(B_n)\, p_{B_n,n}(x)\big)^{\alpha} \big(\Pi_n(V_n)\, p_{V_n,n}(x)\big)^{1-\alpha}\, d\mu_n(x)\end{aligned} \tag{2.5}$$

for any $0 \leq \alpha \leq 1$.

Lemma 2.7 generalises Proposition 2.6 and makes Bayesian tests available with a sharp bound on composite power. This bound can be related to more familiar minimax upper bounds as follows. If $\{P_{\theta,n} : \theta \in B_n\}$ and $\{P_{\theta,n} : \theta \in V_n\}$ are *convex* sets, then

$$H\big(P_n^{\Pi_n|B_n}, P_n^{\Pi_n|V_n}\big) \geq \inf\big\{H(P_{\theta,n}, P_{\theta',n}) : \theta \in B_n, \theta' \in V_n\big\}.$$

Combination with (2.5) for $\alpha = 1/2$, implies that the minimax upper bound in *i.i.d.* cases [27] remains valid:

$$\int_{B_n} P^n \phi_n \, d\Pi_n(P) + \int_{V_n} Q^n(1-\phi_n)\, d\Pi_n(Q) \leq \sqrt{\Pi_n(B_n)\Pi_n(V_n)}\, e^{-n\epsilon_n^2}, \tag{2.6}$$

where $\epsilon_n = \inf\{H(P,Q) : P \in B_n, Q \in V_n\}$.

Note that Bayesian tests enhance the role of the prior in the frequentist discussion of the asymptotic behaviour of posteriors: the prior must not only assign enough mass to KL- or GGV-neighbourhoods of the truth, but is also of influence in the testing condition: *where the test is least powerful, prior mass should be scarce to compensate and where the test is more powerful, prior mass can be plentiful*. To optimize composite power, one imposes *upper bounds on prior mass* in hard-to-test subsets of the model (see Appendix B in the Supplementary Material [22]). This falls in line with the argument that underpins *nonlocality* of priors for variable selection, as in [19].

3. Remote contiguity. In this section, we weaken the notion of contiguity (see [24], Chapter 6 in [27] and [17, 29]) in a way that is suitable to promote Π-almost-everywhere Bayesian limits to frequentist limits that hold everywhere in the model.

3.1. *Definition and criteria for remote contiguity.* The notion of "domination" left undefined in the argument following Proposition 2.3 is made rigorous here.

DEFINITION 3.1. Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$, $n \geq 1$ with two sequences (P_n) and (Q_n) of probability measures and a sequence $\rho_n \downarrow 0$, we say that Q_n is ρ_n-remotely contiguous with respect to P_n, notation $Q_n \lhd \rho_n^{-1} P_n$, if

$$P_n\phi_n\big(X^n\big) = o(\rho_n) \quad \Rightarrow \quad Q_n\phi_n\big(X^n\big) = o(1) \tag{3.1}$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0,1]$.

Note that for a sequence (Q_n) that is a_n-remotely contiguous with respect to (P_n), there exists no test sequence that distinguishes between P_n and Q_n with composite power of order $o(a_n)$. Note also that given two sequences (P_n) and (Q_n), contiguity $P_n \lhd Q_n$ is equivalent to remote contiguity $P_n \lhd a_n^{-1} Q_n$ for all $a_n \downarrow 0$.

EXAMPLE 3.2. Let $\mathscr{P}$ be a model for the distribution of a single observation in *i.i.d.* samples $X^n = (X_1, \ldots, X_n)$. Let P_0, P and $\epsilon > 0$ be such that $-P_0 \log(dP/dP_0) < \epsilon^2$. The law of large numbers implies that for large enough n,

$$\frac{dP^n}{dP_0^n}(X^n) \geq e^{-\frac{n}{2}\epsilon^2}, \tag{3.2}$$

with P_0^n-probability one. Consequently, for large enough n and for any $\mathscr{B}_n$-measurable sequence $\psi_n : \mathscr{X}_n \to [0,1]$, $P^n\psi_n \geq e^{-\frac{1}{2}n\epsilon^2} P_0^n\psi_n$. Therefore, if $P^n\phi_n = o(\exp(-\frac{1}{2}n\epsilon^2))$ then $P_0^n\phi_n = o(1)$. Conclude that for every $\epsilon > 0$, the Kullback–Leibler neighbourhood $\{P : -P_0 \log(dP/dP_0) < \epsilon^2\}$ consists of model distributions for which the sequence (P_0^n) of product distributions are $\exp(-\frac{1}{2}n\epsilon^2)$-remotely contiguous with respect to (P^n).

Criteria for remote contiguity are given in the lemma below; note that, here, we give sufficient conditions, rather than necessary and sufficient, as in Le Cam's first lemma. (For the Q_n-almost-sure definition of $(dP_n/dQ_n)^{-1}$, see Appendix A in the Supplementary Material [22].)

LEMMA 3.3. *Let probability measures (P_n), (Q_n) on measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$ and $a_n \downarrow 0$ be given, then $Q_n \lhd a_n^{-1}P_n$ if any of the following hold*:

(i) *for any bounded, $\mathscr{B}_n$-msb. $T_n : \mathscr{X}_n \to [0,1]$, $a_n^{-1}T_n \xrightarrow{P_n} 0 \Rightarrow T_n \xrightarrow{Q_n} 0$,*

(ii) *for any $\epsilon > 0$, there is a $\delta > 0$ such that $Q_n(dP_n/dQ_n < \delta a_n) < \epsilon$, for large enough n,*

(iii) *there is a $b > 0$ such that $\liminf_n ba_n^{-1}P_n(dQ_n/dP_n > ba_n^{-1}) = 1$,*

(iv) *for any $\epsilon > 0$, there is a constant $c > 0$ such that $\|Q_n - Q_n \wedge ca_n^{-1}P_n\| < \epsilon$, for large enough n,*

(v) *under Q_n every subsequence of $(a_n(dP_n/dQ_n)^{-1})$ has a weakly convergent subsequence.*

PROOF. (The proof of this lemma actually shows that ((i) or (iv)) implies remote contiguity; that ((ii) or (iii)) $\Rightarrow$ (iv) and that (v) $\Leftrightarrow$ (ii).) Assume (i). Let $\phi_n : \mathscr{X}_n \to [0,1]$ be given and assume that $P_n\phi_n = o(a_n)$. By Markov's inequality, for every $\epsilon > 0$, $P_n(a_n^{-1}\phi_n > \epsilon) = o(1)$. Then $\phi_n \xrightarrow{Q_n} 0$ and since ϕ_n is bounded, that implies $Q_n\phi_n = o(1)$, so that $Q_n \lhd a_n^{-1}P_n$. Next, assume (iv). Let $\epsilon > 0$ and $\phi_n : \mathscr{X}_n \to [0,1]$ be given. By assumption, there exist $c > 0$ and $N \geq 1$ such that for all $n \geq N$,

$$Q_n\phi_n < ca_n^{-1}P_n\phi_n + \frac{\epsilon}{2}.$$

Assuming $P_n\phi_n = o(a_n)$, $Q_n\phi_n < \epsilon$ for large enough n. Conclude that $Q_n \lhd a_n^{-1}P_n$. To show that (ii) $\Rightarrow$ (iv), let $\mu_n = P_n + Q_n$ and denote μ_n-densities for P_n, Q_n by $p_n, q_n : \mathscr{X}_n \to \mathbb{R}$. Then, for any $n \geq 1$, $c > 0$,

$$\begin{aligned}
\|Q_n - Q_n \wedge ca_n^{-1}P_n\| &= \sup_{A \in \mathscr{B}_n} \left(\int_A q_n \, d\mu_n - \int_A q_n \, d\mu_n \wedge \int_A ca_n^{-1}p_n \, d\mu_n \right) \\
&\leq \sup_{A \in \mathscr{B}_n} \int_A (q_n - q_n \wedge ca_n^{-1}p_n) \, d\mu_n \\
&= \int 1\{q_n > ca_n^{-1}p_n\}(q_n - ca_n^{-1}p_n) \, d\mu_n.
\end{aligned} \tag{3.3}$$

Note that the right-hand side of (3.3) is bounded above by $Q_n(dP_n/dQ_n < c^{-1}a_n)$. To show that (iii) $\Rightarrow$ (iv), it is noted that, for all $c > 0$ and $n \geq 1$,

$$0 \leq \int ca_n^{-1} P_n(q_n > ca_n^{-1} p_n) \leq Q_n(q_n > ca_n^{-1} p_n) \leq 1,$$

so (3.3) goes to zero if $\liminf_{n\to\infty} ca_n^{-1} P_n(dQ_n/dP_n > ca_n^{-1}) = 1$. To prove that (v) $\Leftrightarrow$ (ii), note that Prohorov's theorem says that (v) is equivalent to the uniform tightness of $(a_n(dP_n/dQ_n)^{-1} : n \geq 1)$ under Q_n, which is equivalent to (ii). $\square$

To conclude this subsection, we specify the definition of remote contiguity slightly further.

DEFINITION 3.4. Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$ $(n \geq 1)$ with two sequences (P_n) and (Q_n) of probability measures and sequences $\rho_n, \sigma_n > 0$, $\rho_n, \sigma_n \to 0$, we say that Q_n is ρ_n-to-σ_n remotely contiguous with respect to P_n, notation $\sigma_n^{-1} Q_n \lhd \rho_n^{-1} P_n$, if

$$P_n\phi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad Q_n\phi_n(X^n) = o(\sigma_n)$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0, 1]$.

Like Definition 3.1, Definition 3.4 allows for reformulation similar to Lemma 3.3, for example, if for some sequences ρ_n, σ_n like in Definition 3.4,

$$\| Q_n - Q_n \wedge \sigma_n \rho_n^{-1} P_n \| = o(\sigma_n),$$

then $\sigma_n^{-1} Q_n \lhd \rho_n^{-1} P_n$. We leave the formulation of other sufficient conditions to the reader.

EXAMPLE 3.5. The inequality of Example 3.2 implies that $b_n^{-1} P_0^n \lhd a_n^{-1} P^n$, for any $a_n \leq \exp(-n\alpha^2)$ with $\alpha^2 > \frac{1}{2}\epsilon^2$ and $b_n = \exp(-n(\alpha^2 - \frac{1}{2}\epsilon^2))$. It is noted that this implies that $\phi_n(X^n) \xrightarrow{P_0\text{-a.s.}} 0$ for any $\phi_n : \mathscr{X}_n \to [0, 1]$ such that $P^n\phi_n(X^n) = o(a_n)$ (more generally, this holds whenever $\sum_n \sigma_n < \infty$, as a consequence of the first Borel–Cantelli lemma).

3.2. *Remote contiguity for Bayesian limits.* Applications in the context of Bayesian limit theorems concern remote contiguity of the sequence of true distributions $P_{\theta_0,n}$ with respect to local prior predictive distributions $P_n^{\Pi_n|B_n}$, where the sets $B_n \subset \Theta$ are such that

$$P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi_n|B_n} \tag{3.4}$$

for some rate $a_n \downarrow 0$. Let us first demonstrate how Schwartz's KL-priors induce remote contiguity.

EXAMPLE 3.6. Let $\mathscr{P}$ be a model for *i.i.d.* samples X^n as in Example 3.2. Fix P_0 and $\epsilon > 0$, define $K(\epsilon) = \{P \in \mathscr{P} : -P_0 \log(dP/dP_0) < \epsilon^2\}$ and recall that a KL-prior Π satisfies, $\Pi(K(\epsilon)) > 0$ for every $\epsilon > 0$. The exponential lower bound (3.2) implies that $\liminf_n \exp(\frac{1}{2}n\epsilon^2)(dP^n/dP_0^n)(X^n) \geq 1$ with P_0^∞-probability one for every $P \in K(\epsilon)$. With Fatou's lemma,

$$\liminf_{n\to\infty} \frac{e^{\frac{1}{2}n\epsilon^2}}{\Pi(K(\epsilon))} \int_{K(\epsilon)} \frac{dP_\theta^n}{dP_{\theta_0}^n}(X^n)\, d\Pi(\theta) \geq 1,$$

with $P_{\theta_0}^\infty$-probability one, showing that sufficient condition (ii) of Lemma 3.3 holds. Conclude that

$$P_0^n \lhd e^{\frac{1}{2}n\epsilon^2} P_n^{\Pi|K(\epsilon)}.$$

A version of the form $b_n^{-1} P_0^n \lhd a_n^{-1} P^n$ based on Example 3.5 is also possible.

Remote contiguity also applies in more irregular situations: Example 1.1 does not admit KL priors, but satisfies the requirement of remote contiguity.

EXAMPLE 3.7. Consider again Example 1.1 in the case of an *i.i.d.* sample from a uniform distribution on $[\theta, \theta+1]$, for unknown $\theta \in \mathbb{R}$. Model distributions P_θ have Lebesgue densities $p_\theta(x) = 1_{[\theta,\theta+1]}(x)$, for $\theta \in \Theta = \mathbb{R}$. Pick a prior Π on Θ with a continuous and strictly positive Lebesgue density $\pi : \mathbb{R} \to \mathbb{R}$ and, for some rate $\delta_n \downarrow 0$, choose $B_n = (\theta_0, \theta_0 + \delta_n)$. For any $\alpha > 0$, $(1-\alpha)\pi(\theta_0)\delta_n \leq \Pi(B_n) \leq (1+\alpha)\pi(\theta_0)\delta_n$ for large enough n. Note that for any $\theta \in B_n$ and $X^n \sim P_{\theta_0}^n$, $dP_\theta^n / dP_{\theta_0}^n(X^n) = 1\{X_{(1)} \geq \theta\}$, and correspondingly,

$$\frac{dP_n^{\Pi|B_n}}{dP_{\theta_0}^n}(X^n) = \Pi_n(B_n)^{-1} \int_{\theta_0}^{\theta_0+\delta_n} 1\{X_{(1)} \geq \theta\}\, d\Pi(\theta)$$
$$\geq \frac{1-\alpha}{1+\alpha} \frac{\delta_n \wedge (X_{(1)} - \theta_0)}{\delta_n}$$

for large enough n. As a consequence, for every $\delta > 0$ and all $a_n \downarrow 0$,

$$P_{\theta_0}^n\left(\frac{dP_n^{\Pi|B_n}}{dP_{\theta_0}^n}(X^n) < \delta a_n \right) \leq P_{\theta_0}^n\big(\delta_n^{-1}(X_{(1)} - \theta_0) < (1+\alpha)\delta a_n\big)$$

for large enough $n \geq 1$. Since $n(X_{(1)} - \theta_0)$ has an exponential weak limit under $P_{\theta_0}^n$, we choose $\delta_n = n^{-1}$, so that the right-hand side in the above display goes to zero. So $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi_n|B_n}$, for any $a_n \downarrow 0$. Conclude that with these choices for Π and B_n, (3.4) holds, for any a_n.

Example 3.7 emphasizes the role of *weak convergence of likelihood ratios*, similar to *limits of experiments* [25, 27, 39]. To emphasize this relation further, consider the following proposition. Proposition 3.8 should be viewed in light of [28], which considers contiguity under statistical information loss. To make the present case compatible, think of (remote) contiguity for probability measures that arise as marginals for the data X^n when information concerning the (Bayesian random) parameter θ is unavailable.

PROPOSITION 3.8. *Let $\theta_0 \in \Theta$ and priors $\Pi_n : \mathscr{G} \to [0,1]$, $n \geq 1$ be given. Let (B_n) be a sequence of measurable subsets of Θ_n such that $\Pi_n(B_n) > 0$ for all $n \geq 1$. Assume that for some $a_n \downarrow 0$, the family,*

$$\left\{ a_n \left(\frac{dP_{\theta,n}}{dP_{\theta_0,n}} \right)^{-1}(X^n) : n \geq 1, \theta \in B_n \right\},$$

is uniformly tight under $P_{\theta_0,n}$. Then $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi_n|B_n}$.

PROOF. For every $\epsilon > 0$, there exists a constant $\delta > 0$ such that

$$P_{\theta_0,n}\left(a_n \left(\frac{dP_{\theta,n}}{dP_{\theta_0,n}} \right)^{-1}(X^n) > \frac{1}{\delta} \right) < \epsilon$$

for all $n \geq 1$, $\theta \in B_n$. For this choice of δ, condition (ii) of Lemma 3.3 is satisfied for all $\theta \in B_n$ simultaneously, and according to the proof of said lemma, for given $\epsilon > 0$, there exists a $c > 0$ such that

$$\left\| P_{\theta_0,n} - P_{\theta_0,n} \wedge c a_n^{-1} P_{\theta,n} \right\| < \epsilon \tag{3.5}$$

for all $n \geq 1$, $\theta \in B_n$. Now note that for any $A \in \mathscr{B}_n$,

$$\begin{aligned} 0 &\leq P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge ca_n^{-1} P_n^{\Pi_n|B_n}(A) \\ &\leq \int \bigl(P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge ca_n^{-1} P_{\theta,n}(A)\bigr)\, d\Pi_n(\theta|B_n). \end{aligned}$$

Taking the supremum with respect to A, we find the following inequality in terms of total variational norms,

$$\bigl\| P_{\theta_0,n} - P_{\theta_0,n} \wedge ca_n^{-1} P_n^{\Pi_n|B_n} \bigr\| \leq \int \bigl\| P_{\theta_0,n} - P_{\theta_0,n} \wedge ca_n^{-1} P_{\theta,n} \bigr\| d\Pi_n(\theta|B_n).$$

Based on (3.5), condition (iv) of Lemma 3.3 is satisfied. □

If we think of Proposition 3.8 in the context of density estimation, one sees that remote contiguity benefits from *model distributions that have heavier tails than the true distribution of the data*. This rhymes with experience in example models (see, e.g., Theorem 3.1 in [38]) and holds true more generally: if model distributions are 'not concentrated enough' in regions of sample spaces where the true data-generating mechanism assigns 'too much probability mass', then posteriors may display instances of inconsistency. Remote contiguity makes precise what heuristic notions like 'not concentrated' and 'too much mass' mean.

3.3. *Comparison of contiguity and remote contiguity.* To compare contiguity and its remote analogue in parametric and nonparametric context, consider the following standard example.

Let $\mathscr{F}$ denote a class of functions $\mathscr{X} \to \mathbb{R}$, where $\mathscr{X}$ is a compact, convex subset of $\mathbb{R}^d$. We consider samples $X^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ $(n \geq 1)$ of points in $\mathscr{X} \times \mathbb{R}$, assumed to be related through

$$Y_i = f_0(X_i) + e_i$$

for some unknown $f_0 \in \mathscr{F}$, where the errors are *i.i.d.* standard normal $e_1, \ldots, e_n \sim N(0,1)^n$ and independent of the *i.i.d.* covariates $X_1, \ldots, X_n \sim P^n$, for some ancillary distribution P on $\mathbb{R}$. Assume that $\mathscr{F} \subset L^2(P)$ and that $Pf(X) = 0$ for all $f \in \mathscr{F}$. We distinguish two cases: (a) the case of linear regression, $\mathscr{F} = \{f_\theta : \mathscr{X} \subset \mathbb{R} \to \mathbb{R} : \theta \in \Theta\}$, where $\theta = (a, b) \in \Theta = \mathbb{R}^2$ and $f_\theta(x) = ax + b$; (b) the case of nonparametric regression (to maintain concreteness, we keep in mind the special case $\mathscr{F} = C_1^\alpha(\mathscr{X})$, the collection of all α-smooth functions on $\mathscr{X}$ with Hölder-α-norm $\|\cdot\|_\alpha$ bounded by 1).

For (ρ_n) to be fixed later, define $a_n = \exp(-\frac{1}{2}n\rho_n^2)$. A bit of manipulation casts the a_n-rescaled likelihood ratio for $f_0, f \in \mathscr{F}$ in the following form:

$$a_n^{-1} \frac{dP_{f,n}}{dP_{f_0,n}}(X^n) = e^{-\frac{1}{2}\sum_{i=1}^n (2e_i(f-f_0)(X_i) + (f-f_0)^2(X_i) - n\rho_n^2)} \tag{3.6}$$

for $X^n \sim P_{f_0,n}$.

EXAMPLE 3.9. In the parametric case, expression (3.6) can be written in terms of a local parameter $h \in \mathbb{R}^2$ which, for given θ_0 and $n \geq 1$, is related to θ by $\theta = \theta_0 + n^{-1/2}h$. For $h \in \mathbb{R}^2$, we write $P_{h,n} = P_{\theta_0 + n^{-1/2}h,n}$, $P_{0,n} = P_{\theta_0,n}$ and write

$$\frac{dP_{h,n}}{dP_{0,n}}(X^n) = e^{\frac{1}{\sqrt{n}}\sum_{i=1}^n h \cdot \ell_{\theta_0}(X_i,Y_i) - \frac{1}{2}h \cdot I_{\theta_0} \cdot h + o_{P_{\theta_0,n}}(1)}, \tag{3.7}$$

where $\ell_{\theta_0}: \mathbb{R}^2 \to \mathbb{R}^2 : (x, y) \mapsto (y - a_0 x - b_0)(x, 1)$ is the score function for θ at θ_0, $I_{\theta_0} = P_{\theta_0,1}\ell_{\theta_0}\ell_{\theta_0}^T$ is the Fisher information matrix. Assume I_{θ_0} is nonsingular and note the central limit,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_{\theta_0}(X_i, Y_i) \xrightarrow{P_{\theta_0,n}\text{-w.}} N_2(0, I_{\theta_0}),$$

which expresses local asymptotic normality of the model and implies that for any fixed $h \in \mathbb{R}^2$, $P_{h,n} \lhd P_{0,n}$. It is well known that contiguity extends to $n^{-1/2}$-localized prior averages (see Lemma 3, Section 8.4 in [29]):

$$P_{\theta_0,n} \lhd P_n^{\Pi|B_n} \tag{3.8}$$

(where $B_n = \{\theta \in \Theta : \|\theta - \theta_0\| \leq Mn^{-1/2}\}$, for any $M > 0$) provided $\Pi(B_n) > 0$ for all n.

EXAMPLE 3.10. In the nonparametric case, define $B(\rho) = \{f \in \mathscr{F} : \|f - f_0\| < \rho\}$ (where $\|\cdot\|$ denotes the $L_2(\mathbb{P}_n)$-norm, with $\mathbb{P}_n$ the empirical distribution of observed design points [40]). Theorem 3.4.1 and, more specifically, Section 3.4.3 of [40] prove that the (outer) expectation of the supremum of the empirical process for scores satisfies the maximal inequality,

$$P_{f_0,n}\sup_{f\in B(\rho)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i(f - f_0)(X_i)\right| \leq \phi_n(\rho)$$

for all $\rho > 0$, where $\phi_n(\rho)$ is a bracketing integral. If we choose $\rho_n > 0$ such that $n\rho_n^2 \to \infty$ and $\rho_n^{-2}\phi_n(\rho_n) = \beta_n$ with $\beta_n = o(\sqrt{n})$, then Markov's inequality shows that, for any $\epsilon > 0$,

$$P_{f_0,n}\left(\sup_{f\in B(\rho_n)}\left|\sum_{i=1}^{n}e_i(f - f_0)(X_i)\right| > \frac{n\rho_n^2\beta_n}{\epsilon}\right) \leq \epsilon. \tag{3.9}$$

If the ancillary distribution P is such that $\{(f - f_0)^2 : f \in B(\rho_n)\}$ satisfy the Glivenko–Cantelli-like requirement that

$$\sup_{f\in B(\rho_n)}\left|\frac{1}{n}\sum_{i=1}^{n}(f - f_0)^2(X_i) - \|f - f_0\|_{P,2}^2\right| \xrightarrow{P^\infty\text{-a.s.}} 0,$$

then for any $\delta, \delta' > 0$, using (3.9) and assuming that $\Pi_n(B(\rho_n)) > 0$,

$$\begin{aligned}
&P_{f_0,n}\left(\frac{1}{\Pi_n(B(\rho_n))}\int_{B(\rho_n)}\frac{dP_{f,n}}{dP_{f_0,n}}(X^n)\,d\Pi_n(f) < \delta a_n\right)\\
&\quad\leq P_{f_0,n}\left(\inf_{f\in B(\rho_n)}a_n^{-1}\frac{dP_{f,n}}{dP_{f_0,n}}(X^n) < \delta\right)\\
&\quad\leq P_{f_0,n}\left(\inf_{f\in B(\rho_n)} -\sum_{i=1}^{n}e_i(f - f_0)(X_i) + \frac{1}{2}n\rho_n^2 < \log\delta - \delta'\right) \leq \epsilon
\end{aligned}$$

for large enough n. Conclude that

$$P_{f_0,n} \lhd e^{\frac{1}{2}n\rho_n^2}P_n^{\Pi|B(\rho_n)}. \tag{3.10}$$

A similar proof based on Proposition 3.8 is also possible. For a smoothness class $\mathscr{F} = C_1^\alpha(\mathscr{X})$ (and provided certain technical conditions are met, see Section 3.4.3.2 in [40]), rates ρ_n that solve $\rho_n^{-2}\phi_n(\rho_n) = o(n^{1/2})$ exist arbitrarily close to $n^{-\alpha/(2\alpha+2d)}$, the minimax $L^2(P)$-rate of estimation of f. Note that the argument extends to other sequences (Q_n) that

approximate $(P_{f_0,n})$ well enough. (For example, if we define (Q_n) by substitution of estimators $\hat{f}_n$ that are $L^2(P)$-consistent at rate ρ_n, and we can show that $P_{\hat{f}_n,n}(A_n) = o(e^{\frac{1}{2}n\rho_n^2})$, then also $P_{f_0,n}(A_n) = o(1)$.)

The analogy between (3.8) and (3.10) establishes in this regression example (and many others that allow the same empirical-process argument), that *remote contiguity has the potential to provide sequential approximations in nonparametric statistics, analogous to approximation by contiguous sequences in parametric setting* [17]. More examples of sequential approximation by remote contiguity are provided in [12, 33] and [21].

4. Posterior concentration for frequentists. From the perspective of the Bayesian, asymptotic concentration of the posterior is covered by Lemma 2.2, particularly as in Proposition 2.3. To existence of Bayesian tests, we add the requirement of remote contiguity to arrive at the frequentist conclusion that the posterior concentrates.

THEOREM 4.1. *Let* $(\mathscr{X}_n, \mathscr{B}_n)$, $(\Theta_n, \mathscr{G}_n)$, $(\mathscr{P}_n)$ *and* (Π_n) *be given. Assume that for all* $n \geq 1$, *the data* $X^n \sim P_{0,n}$ *and that, for given* $B_n, V_n \in \mathscr{G}_n$ *and* $a_n, b_n \downarrow 0$ *with* $a_n = o(b_n)$:

(i) *there are Bayesian tests* $\phi_n : \mathscr{X}_n \to [0,1]$ *such that*

$$\int_{B_n} P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta) = o(a_n), \tag{4.1}$$

(ii) *the prior mass of* B_n *is lower-bounded,* $\Pi_n(B_n) \geq b_n$,

(iii) *the sequence* $P_{0,n}$ *satisfies* $P_{0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n}$.

Then $\Pi(V_n|X^n) \xrightarrow{P_{\theta_0,n}} 0$.

PROOF. Proposition 2.3 says that $P_n^{\Pi_n|B_n}\Pi(V_n|X^n)$ is of order $o(b_n^{-1}a_n)$. Condition (iii) then implies that $P_{\theta_0,n}\Pi(V_n|X^n) = o(1)$, or equivalently, $\Pi(V_n|X^n)$ goes to zero in $P_{\theta_0,n}0$-probability. □

This theorem requires very little of $P_{0,n}$: it is not required that $P_{0,n}$ describes *i.i.d.* data, nor does $P_{0,n}$ need to correspond to an element of B_n (or even lie in $\mathscr{P}_n$): the true data-distributions need to relate to the rest of the problem *only* through remote contiguity.

4.1. *Posterior consistency.* The most basic interpretation is that in which $\Theta_n = \Theta$, $\Pi_n = \Pi$, $B_n = B$, $V_n = V$ and $P_{0,n} = P_{\theta_0,n}$ for some $\theta_0 \in B$, with V the complement of a neighbourhood U of θ_0 in Θ and $B \subset U$. If, moreover, we have data X^n that is *i.i.d.*, we arrive at Schwartz's consistency in $\mathscr{P}$. In that case, require that $b_n = \Pi_n(B_n) = \Pi(B) = b > 0$, to restate Schwartz's theorem.

THEOREM 4.2. *Assume that for all* $n \geq 1$, *the data* $X^n \sim P_0^n$ *for some* $P_0 \in \Theta$. *Fix a prior* $\Pi : \mathscr{G} \to [0,1]$ *and assume that for given* $B, V \in \mathscr{G}$ *with* $\Pi(B) > 0$ *and* $a_n \downarrow 0$:

(i) *there exist Bayesian tests* ϕ_n *for* B *versus* V,

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n(1-\phi_n)\, d\Pi(Q) = o(a_n), \tag{4.2}$$

(ii) *the sequence* $P_{\theta_0}^n$ *satisfies* $P_{\theta_0}^n \lhd a_n^{-1} P_n^{\Pi|B}$.

Then $\Pi(V|X^n) \xrightarrow{P_{\theta_0,n}} 0$.

Theorem 4.2 relates to Schwartz's conditions as follows: Schwartz requires that uniform tests exist; a well-known argument based on Hoeffding's inequality then guarantees the existence of a uniform test sequence *of exponential composite power*. According to Example 3.6, KL-priors induce remote contiguity of P_0^n with respect to KL-localized prior predictive distributions based on $B = K(\epsilon)$ at exponential rate.

Next, observe that B is contained in a Hellinger ball in $\mathscr{P}$ centred on P_0. So if we let U be a Hellinger ball centred on P_0 of some larger radius, B and V are separated by nonzero Hellinger distance. Assuming that $\mathscr{P}$ is dominated, any Π that is Borel for the Hellinger topology on $\mathscr{P}$ is Radon in the completion, so for every $\delta > 0$, there exists a Hellinger pre-compact (that is, totally-bounded) $K \subset \mathscr{P}$, such that $\Pi(K) > 1 - \delta$. Totally-boundedness is the entropy argument needed in a well-known construction [5, 14, 27, 30] of a finite cover of $V \cap K$ by Hellinger balls and combination of the corresponding uniform minimax tests versus B (Section 16.4 in [27]) into uniform test ϕ_n of B versus $V \cap K$ of exponential composite power:

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n) \, d\Pi(Q) \leq N(\epsilon, V \cap K, H) e^{-n\epsilon^2} + \delta \tag{4.3}$$

for some $\epsilon > 0$. Diagonalization with respect to exponentially decreasing δ's and an upper bound on the Hellinger covering numbers of the corresponding pre-compact K's then formulates *Barron's negligible prior mass condition* [2, 3].

4.2. *Rates of posterior concentration.* A significant extension to the theory on posterior convergence is formed by results concerning posterior convergence in metric spaces *at a rate* [3, 14, 23, 26, 35, 42]. To establish the exceptional case first, we start with application of Theorem 4.1 to the rate of posterior convergence in Examples 1.1 and 3.7, where no KL- or GGV-priors exist.

EXAMPLE 4.3. Consider again the situation of a uniform distribution with an unknown location, as in Examples 1.1 and 3.7. Take V_n equal to $\{\theta : \theta - \theta_0 > \epsilon_n\}$ with $\epsilon_n = M_n/n$ for some $M_n \to \infty$. It is noted that, for every $0 < c < 1$, the likelihood ratio test,

$$\phi_n(X^n) = 1\{dP_{\theta_0+\epsilon_n,n}/dP_{\theta_0,n}(X^n) > c\} = 1\{X_{(1)} > \theta_0 + \epsilon_n\},$$

satisfies $P_\theta^n(1 - \phi_n)(X^n) = 0$ for all $\theta \in V_n$, and if we choose $\delta_n = 1/2$ and $\epsilon_n = M_n/n$ for some $M_n \to \infty$, $P_\theta^n \phi_n \leq e^{-M_n+1}$ for all $\theta \in B_n$, so that

$$\int_{B_n} P_\theta^n \phi_n \, d\Pi(\theta) + \int_{V_n} P_\theta^n (1 - \phi_n) \, d\Pi(\theta) \leq \Pi(B_n) e^{-M_n+1}.$$

Using Lemma 2.2, we see that $P_n^{\Pi|B_n} \Pi(V_n|X^n) \leq e^{-M_n+1}$. Based on the conclusion of Example 3.7, contiguity implies that $P_{\theta_0}^n \Pi(V_n|X^n) \to 0$. Treating the case $\theta < \theta_0 - \epsilon_n$ similarly, we conclude that the posterior is consistent at any rate $\epsilon_n = M_n/n$, with $M_n \to \infty$.

Let us also review the conditions of [3, 14, 35] in light of Theorem 4.1.

EXAMPLE 4.4. Let $\epsilon_n \downarrow 0$ such that $n\epsilon_n^2 \to \infty$ denote a Hellinger rate of convergence, let $M > 1$ be some constant and define

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq M\epsilon_n\},$$

$$B_n = \{P \in \mathscr{P} : -P_0 \log dP/dP_0 < \epsilon_n^2, P_0 \log^2 dP/dP_0 < \epsilon_n^2\}.$$

We repeat the argument leading to (4.3) for every n, with $\epsilon = \epsilon_n$, $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. If we require Barron's δ-contribution in (4.3) to be of $n\epsilon_n^2$-exponentially small order,

$$\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp(-nM\epsilon_n^2),$$

and the sieve of pre-compact $\mathscr{P}_n$ has Hellinger entropies that are upper-bounded (see [5, 30])

$$N(\epsilon_n, \mathscr{P}_n, H) \le e^{Kn\epsilon_n^2}$$

for some $K > 0$, then the minimax construction extends to tests that separate $V_n = \{P \in \mathscr{P} : H(P_0, P) \ge 4\epsilon_n\}$ from $B_n = \{P \in \mathscr{P} : H(P_0, P) < \epsilon_n\}$ asymptotically, with composite power $\exp(-nL\epsilon_n^2)$ for some $L > 0$.

Note that B_n is contained in the Hellinger ball of radius ϵ_n around P_0, so (4.1) holds. Remote contiguity therefore requires that for some $C > 0$,

$$\Pi_n(B_n) \ge e^{-Cn\epsilon_n^2}. \tag{4.4}$$

We note Lemma 8.1 in [14], which says that if (4.4) is satisfied then Lemma 3.3(ii) holds, so that

$$P_0^n \lhd e^{cn\epsilon_n^2} P_n^{\Pi|B_n} \tag{4.5}$$

for any $c > 1$. For large enough M, Theorem 4.1 then reproduces the GGV-result, that is, the posterior is Hellinger consistent at rate ϵ_n. Due to relations that exist between metrics for model parameters and the Hellinger metric in many examples and applications, the material covered here is widely applicable in (nonparametric) models for *i.i.d.* data. (For much more on this and many similar constructions, see [15].)

Experience teaches that the sharpest results on posterior concentration are achieved when the alternatives V_n are split into pieces, each according to the strength of the optimal test versus B_n. Combination of the tests per piece and re-summation weighted by prior masses can often be employed to arrive at sharp results.

EXAMPLE 4.5. Consider a model $\mathscr{P}$ of distributions P for *i.i.d.* data $X^n \sim P^n$ $(n \ge 1)$ and suppose that $\mathscr{P}$ is *Hellinger-separable*. Let $P_0 \in \mathscr{P}$ and $\epsilon_n \to 0$ be given, denote $V(\epsilon) = \{P \in \mathscr{P} : H(P_0, P) \ge 4\epsilon\}$, $B_H(\epsilon) = \{P \in \mathscr{P} : H(P_0, P) < \epsilon\}$ for all $\epsilon > 0$. There exist $N(\epsilon_n) \ge 1$ (possibly infinite) and a cover of $V(\epsilon_n)$ by $N(\epsilon_n)$ Hellinger balls $V_{n,1}, V_{n,2}, \ldots$ of radius ϵ_n and for any point Q in any $V_{n,i}$ and any $P \in B_H(\epsilon_n)$, $H(Q, P) > \epsilon_{i,n}$. According to Lemma 2.7 with $\alpha = 1/2$ and (2.6), for each $1 \le i \le N(\epsilon_n)$ there exists a Bayesian test sequence $(\phi_{n,i})$ for $B_H(\epsilon_n)$ versus $V_{n,i}$ of composite power $\exp(-\frac{1}{2}n\epsilon_{i,n}^2)$. Then, for any subsets $B_n' \subset B_H(\epsilon_n)$,

$$\begin{aligned} P_n^{\Pi|B_n'}\Pi\big(V(\epsilon_n)|X^n\big) &\le \sum_{i=1}^{N(\epsilon_n)} P_n^{\Pi|B_n'}\Pi\big(V_{n,i}|X^n\big) \\ &\le \frac{1}{\Pi(B_n')} \sum_{i=1}^{N(\epsilon_n)} \left(\int_{B_n'} P^n \phi_{n,i}\, d\Pi(P) + \int_{V_{n,i}} P^n(1-\phi_{n,i})\, d\Pi(P) \right) \\ &\le \sum_{i=1}^{N(\epsilon_n)} \sqrt{\frac{\Pi(V_{n,i})}{\Pi(B_n')}} \exp\Big(-\frac{1}{2}n\epsilon_{i,n}^2\Big). \end{aligned} \tag{4.6}$$

The requirement that the above upper bound converges to zero leads directly to the summability requirements for square-root prior masses of Hellinger covers of separable models posed by [41, 42].

Summability of this type leads [30] to define the so-called *Le Cam dimension* of the model, as well as to various subtle results on posterior behaviour in nonparametric applications, and also explains the sharpness of the posterior concentration results of [21]. We emphasize that (4.6) makes explicit the balancing of prior masses and composite power, as intended by the remark that closes Section 2.2.

5. Consistent hypothesis testing with posterior odds. *Model selection* describes all statistical methods that attempt to determine from the data which model to use for further inferential statistical analysis (for an overview, see [37]). For example, consider projection of a high-dimensional vector of co-variates onto a sparse subset for subsequent regression analysis, or the selection of a directed a-cyclical graph to formulate a graphical model. Model selection also makes an appearance in very high-dimensional models, which often leave room for over-fitting, requiring regularization [6, 7, 9].

Frequentist methods for model selection vary widely, ranging from very simple rules-of-thumb, to cross-validation and penalization of the likelihood function. Here, we propose to conduct the frequentist analysis with the help of the posterior [4]: when faced with a (dichotomous) model choice, we let posterior odds determine our preference. An (objective) Bayesian perspective on model selection is provided in [43].

For hypotheses $B, V \subset \Theta$ and any $n \geq 1$, define *posterior odds* G_n,

$$G_n = \frac{\Pi(B|X^n)}{\Pi(V|X^n)}$$

for B versus V. Analysing the question first from a purely Bayesian perspective, we see that for a fixed prior Π, Theorem 2.5 says that the posterior gives rise to consistent posterior odds G_n for B versus V in a Bayesian (i.e., Π-almost-sure) way, if and only if a Bayesian test sequence for B versus V exists. Proposition 2.6 says that in Polish models, any Borel set V is Bayesian testable versus its complement. So basically, for the Bayesian, measurable distinctions are consistently testable with posterior odds. In fact, posterior odds are *optimal* [20], in the sense that $\phi_n(X^n) = 1\{X^n \in \mathcal{X}_n : \Pi(B|X^n) > \Pi(V|X^n)\}$ satisfies

$$\begin{aligned}&\int_B P_{\theta,n}\phi_n(X^n)\,d\Pi(\theta) + \int_V P_{\theta,n}(1-\phi_n(X^n))\,d\Pi(\theta)\\ &\quad = \inf_\psi \int_B P_{\theta,n}\psi(X^n)\,d\Pi(\theta) + \int_V P_{\theta,n}(1-\psi(X^n))\,d\Pi(\theta),\end{aligned}$$

where the infimum runs over all measurable $\psi_n : \mathcal{X}_n \to [0,1]$.

However, the frequentist requires convergence in *all points* of the model.

DEFINITION 5.1. For all $n \geq 1$, let the model be parametrized by maps $\theta \mapsto P_{\theta,n}$ on a parameter space $(\Theta, \mathcal{G})$ with priors $\Pi_n : \mathcal{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$. Posterior odds G_n are frequentist consistent for testing B versus V, if

$$G_n \xrightarrow{P_{\theta,n}} 0, \qquad G_n \xrightarrow{P_{\theta',n}} \infty,$$

for all $\theta \in V$, and all $\theta' \in B$.

We employ remote contiguity again to bridge the gap between Bayesian and frequentist formulations.

THEOREM 5.2. *For all $n \geq 1$, let the model be parametrized by maps $\theta \mapsto P_{\theta,n}$ on a parameter space with $(\Theta, \mathcal{G})$ with priors $\Pi_n : \mathcal{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$ with $\Pi_n(B), \Pi_n(V) > 0$ such that*:

(i) *there exist Bayesian tests for B versus V of composite power $a_n \downarrow 0$,*

$$\int_B P^n\phi_n\,d\Pi_n(P) + \int_V Q^n(1-\phi_n)\,d\Pi_n(Q) = o(a_n),$$

(ii) *for every $\theta \in B$ and every $\theta' \in V$,*

$$P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|B}, \qquad P_{\theta',n} \lhd a_n^{-1} P_n^{\Pi_n|V}.$$

Then posterior odds are frequentist consistent for B versus V.

Note that the second condition of Theorem 5.2 can be replaced by a local condition: if, for every $\theta \in B$, there exists a sequence $B_n(\theta) \subset B$ such that $\Pi_n(B_n(\theta)) \geq b_n$ and $P_{\theta,n} \lhd a_n^{-1} b_n P_n^{\Pi_n|B_n}$, then $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|B}$.

This device for model selection is used in the application of Appendix B in the Supplementary Material [22]: it is shown that for stationary Markov chains, the transition kernel for a random walk X^n can be subjected to a goodness-of-fit test inspired by Pearson's χ^2-test, based on a finite partition of the state-space. Proposition B.2 emphasizes the enhancement of the role of the prior, as intended by the remark that closes Section 2.2: *where the test is less powerful, prior mass should be scarce to compensate and where the test is more powerful, prior mass can be plentiful*. In model selection, alternative hypotheses often 'touch' and a continuous power function leads to problems with testing power in the vicinity of the boundary separating them: in such cases, prior mass is upper-bounded in model subsets near that boundary, in line with *nonlocality* of priors as in [19].

6. Confidence sets from credible sets. The assertion of the Bernstein–von Mises theorem [29] has the methodological implication that Bayesian credible sets can be interpreted as asymptotically efficient confidence sets, at least, in the setting of smooth parametric models. Extension to nonparametric models is highly desirable and has been explored in many examples and counterexamples [10, 13]. In recent years, much effort has gone into calculations that balance posterior expectation and variance so that credible metric balls have asymptotic frequentist coverage, mostly in Gaussian models with conjugate posteriors, often with empirically chosen prior to control posterior bias [36]. Below we formulate a general theorem that asserts that certain enlargements of credible sets have an interpretation as asymptotic confidence sets, based on remote contiguity.

DEFINITION 6.1. Given $(\Theta, \mathscr{G})$ with priors Π_n, denote the sequence of posteriors by $\Pi(\cdot|\cdot) : \mathscr{G} \times \mathscr{X}_n \to [0,1]$. Let $\mathscr{D}$ denote a collection of measurable subsets of Θ. A *sequence of credible sets* (D_n) of *credible levels* $1 - a_n$ (where $0 \leq a_n \leq 1$, $a_n \downarrow 0$) is a sequence of set-valued maps $D_n : \mathscr{X}_n \to \mathscr{D}$ such that $\Pi(\Theta \setminus D_n(x)|x) = o(a_n)$ for $P_n^{\Pi_n}$-almost-all $x \in \mathscr{X}_n$.

DEFINITION 6.2. For $0 \leq a \leq 1$, a set-valued map $x \mapsto C(x)$ defined on $\mathscr{X}$ such that, for all $\theta \in \Theta$, $P_\theta(\theta \notin C(X)) \leq a$, is called a confidence set of level $1 - a$. If the levels $1 - a_n$ of a sequence of confidence sets $C_n(X^n)$ go to 1 as $n \to \infty$, the $C_n(X^n)$ are said to be asymptotically consistent.

DEFINITION 6.3. Let D be a (credible) set in Θ and let $B = \{B(\theta) : \theta \in \Theta\}$ denote a collection of model subsets such that $\theta \in B(\theta)$ for all $\theta \in \Theta$. A model subset C' is said to be (a confidence set) associated with D under B, if for all $\theta \in \Theta \setminus C'$, $B(\theta) \cap D = \varnothing$. The intersection C of all C' like above equals $\{\theta \in \Theta : B(\theta) \cap D \neq \varnothing\}$ and is called the minimal (confidence) set associated with D under B (see Figure 1).

Example 6.6 makes this construction explicit in uniform spaces and specializes to metric context.

THEOREM 6.4. *Let $\theta_0 \in \Theta$ and $0 \leq a_n \leq 1$, $b_n > 0$ such that $a_n = o(b_n)$ be given. Choose priors Π_n and let D_n denote level-$(1 - a_n)$ credible sets. Furthermore, for all $\theta \in \Theta$, let $B_n = \{B_n(\theta) \in \mathscr{G} : \theta \in \Theta\}$ denote a sequence such that*:

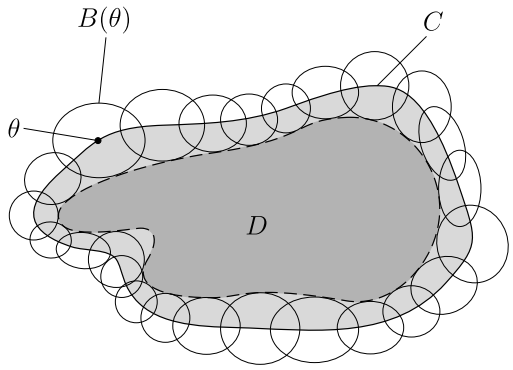

FIG. 1. *The relation between a credible set D and its associated (minimal) confidence set C under B in Venn diagrams: the extra points θ in the associated confidence set C not included in the credible set D are characterized by nonempty intersection $B(\theta) \cap D \neq \varnothing$.*

(i) $\Pi_n(B_n(\theta_0)) \geq b_n$,
(ii) $P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n(\theta_0)}$.

Then any confidence sets C_n associated with the credible sets D_n under B_n are asymptotically consistent,

$$P_{\theta_0,n}\big(\theta_0 \in C_n\big(X^n\big)\big) \to 1. \tag{6.1}$$

PROOF. Fix $n \geq 1$ and let D_n denote a credible set of level $1 - o(a_n)$, defined for all $x \in F_n \subset \mathcal{X}_n$ such that $P_n^{\Pi_n}(F_n) = 1$. For any $x \in F_n$, let $C_n(x)$ denote a confidence set associated with $D_n(x)$ under B. Due to Definition 6.3, $\theta_0 \in \Theta \setminus C_n(x)$ implies that $B_n(\theta_0) \cap D_n(x) = \varnothing$. Hence the posterior mass of $B(\theta_0)$ satisfies $\Pi(B_n(\theta_0)|x) = o(a_n)$. Consequently, the functions $x \mapsto 1\{\theta_0 \in \Theta \setminus C_n(x)\}\Pi(B(\theta_0)|x)$ are $o(a_n)$ for all $x \in F_n$. Integrating with respect to the nth prior predictive distribution and dividing by the prior mass of $B_n(\theta_0)$, one obtains

$$\frac{1}{\Pi_n(B_n(\theta_0))} \int 1\{\theta_0 \in \Theta \setminus C_n\}\Pi\big(B_n(\theta_0)|X^n\big)\, dP_n^{\Pi_n} \leq \frac{a_n}{b_n}.$$

Applying Bayes's rule in the form (A.2), we see that

$$P_n^{\Pi_n|B_n(\theta_0)}\big(\theta_0 \in \Theta \setminus C_n\big(X^n\big)\big) = \int P_{\theta,n}\big(\theta_0 \in \Theta \setminus C_n\big(X^n\big)\big)\, d\Pi_n(\theta|B_n) \leq \frac{a_n}{b_n}.$$

By the definition of remote contiguity, this implies asymptotic coverage; cf. (6.1). □

Theorem 6.4 can be interpreted as follows: the credible sets D_n at its heart are 'statistically informative', according to the Bayesian notion of what 'statistically informative' means. To render that compatible with the frequentist notion asymptotically, Theorem 6.4 employs enlargement by sets B_n and remote contiguity to carry one into the other. This entails a trade-off: the larger the sets B_n are chosen, the greater the enlargements; but also, the larger the sets B_n, the higher the lower bounds b_n, and thence, the more slowly the credible levels a_n can go to zero (allowing for smaller choices of D_n). The fact that Theorem 6.4 holds generally implies practical ways to obtain confidence sets from posteriors: to illustrate, [21] uses Theorem 6.4 to derive confidence sets for the community assignment in a sparse stochastic block model.

In order for the assertion of Theorem 6.4 to be specific regarding the confidence level (rather than just resulting in asymptotic coverage), we re-write the last condition of Theorem 6.4 as follows:

(ii′) $c_n^{-1} P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n(\theta_0)}$,

so that the last step in the proof of Theorem 6.4 is more specific; particularly, assertion (6.1) becomes

$$P_{\theta_0,n}\big(\theta \notin C_n(X^n)\big) = o(c_n),$$

controlling asymptotic confidence levels.

6.1. *Credible/confidence sets in metric spaces*. Next, we specialize to parameter spaces that are metric. First we note a theorem proved in [21], showing that posterior convergence at a rate ensures coverage of enlarged minimal-radius credible balls.

THEOREM 6.5. *Suppose that* (Θ, d) *with Borel priors* (Π_n) *parametrizes models* $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$ *for data* X^n *distributed according to* $P_{\theta_0,n}$ *for some* $\theta_0 \in \Theta$. *Assume that posteriors concentrate in metric balls of radii* r_n:

$$\Pi\big(d(\theta,\theta_0) \le r_n | X^n\big) \xrightarrow{P_{\theta_0,n}} 1.$$

Given X^n *and some* $0 < \epsilon < 1$, *let* $\hat{D}_n = B_n(\hat{\theta}_n, \hat{r}_n)$ *be level-*$1-\epsilon$ *credible balls of minimal radii. With high* $P_{\theta_0,n}$*-probability,* $\hat{r}_n \le r_n$ *and the sequence* $C_n(X^n) = B(\hat{\theta}_n, \hat{r}_n + r_n) \subset B(\hat{\theta}_n, 2r_n)$ *is asymptotically consistent*,

$$P_{\theta_0,n}\big(\theta_{0,n} \in C_n(X^n)\big) \to 1.$$

However, posterior convergence at a known rate is a relatively strong condition and, in practice, one may not be able to guarantee it. For that reason, we also explore the direct method of Theorem 6.4 in metric spaces.

When enlarging credible sets to confidence sets using a collection of subsets B as in Definition 6.3, measurability of confidence sets is guaranteed if $B(\theta)$ is *open* in Θ for all $\theta \in \Theta$. It is worth recalling that KL-divergence is *not automatically continuous* with respect to Hellinger distance (for specifics, see Theorem 5 of [44]).

EXAMPLE 6.6. Let $\mathscr{G}$ be the Borel σ-algebra for a uniform topology on Θ. Let W denote a symmetric entourage and, for every $\theta \in \Theta$, define $B(\theta) = \{\theta' \in \Theta : (\theta, \theta') \in W\}$, a neighbourhood of θ. Let D denote any credible set. A confidence set associated with D under B is any set C' such that the complement of D contains the W-enlargement of the complement of C'. Equivalently (by the symmetry of W), the W-enlargement of D does not meet the complement of C'. Then the minimal confidence set C associated with D is the W-enlargement of D. If the $B(\theta)$ are all open neighbourhoods (e.g., whenever W is a symmetric entourage from a fundamental system for the uniformity on Θ), the minimal confidence set associated with D is open.

The most common examples include the Hellinger or total-variational metric uniformities, but weak topologies and polar topologies are uniform, too.

EXAMPLE 6.7. To illustrate Example 6.6 with a customary situation, consider a parameter space Θ with parametrization $\theta \mapsto P_\theta^n$, to define a model for *i.i.d.* data $X^n = (X_1, \ldots, X_n) \sim P_{\theta_0}^n$, for some $\theta_0 \in \Theta$. Let $\mathscr{D}$ be the class of all pre-images of Hellinger balls, that is, sets $D(\theta, \epsilon) \subset \Theta$ of the form,

$$D(\theta, \epsilon) = \big\{\theta' \in \Theta : H(P_\theta, P_{\theta'}) < \epsilon\big\}$$

for any $\theta \in \Theta$ and $\epsilon > 0$. After choice of a Kullback–Leibler prior Π for θ and calculation of the posteriors, choose D_n equal to the pre-image $D(\hat{\theta}_n, \hat{\epsilon}_n)$ of a minimal-radius Hellinger ball

with credible level $1 - o(a_n)$, $a_n = \exp(-n\alpha^2)$ for some $\alpha > 0$. Assume, now, that for some $0 < \epsilon < \alpha$, the W of Example 6.6 is the Hellinger entourage $W = \{(\theta, \theta') : H(P_\theta, P_{\theta'}) < \epsilon\}$. Since Kullback–Leibler neighbourhoods are contained in Hellinger balls, the sets $D(\hat{\theta}_n, \hat{\epsilon}_n + \epsilon)$ (associated with D_n under the entourage W), is a sequence of asymptotically consistent confidence sets, provided the prior satisfies Schwartz's KL condition. If we make ϵ vary with n, like before, $C_n(X^n) = D(\hat{\theta}_n, \hat{\epsilon}_n + \epsilon_n)$ are asymptotic confidence sets, provided that the prior satisfies (4.4).

In the case ϵ_n is the minimax rate of convergence for the problem, the confidence sets $C_n(X^n)$ attain rate-optimality [31]. Rate-adaptivity [16, 18, 36] is not possible with Theorem 6.4 because a definite, nondata-dependent choice for the B_n is required. An interesting option concerns the exploration of data-driven choices for priors Π_n and B_n, as in [36].

7. Conclusions. We list and discuss the main conclusions below.

Frequentist validity of Bayesian limits. There exists a systematic way of taking Bayesian limits into frequentist ones, if priors satisfy an extra condition relating true data distributions to localized prior predictive distributions. This extra condition generalises Schwartz's Kullback–Leibler condition and amounts to a weakened form of contiguity, termed *remote contiguity*. Remote contiguity has the potential to provide sequential approximations in nonparametric statistics, analogous to approximation by contiguous sequences in parametric statistics (e.g., see [12, 33]).

Given steadily growing interest in the analysis of large datasets gathered from networks (e.g., by *webcrawlers* that perform branching random walks across linked webpages), or from time-series/stochastic processes (e.g., in statistical physics or financial markets), or in the form of high-dimensional, functional or random-graph data (e.g., from biological, financial, medical and meteorological fields), the development of new Bayesian methods benefits from a simple asymptotic perspective to guide the search for suitable priors. Theory presented here is general enough to enable new frequentist applications of Bayesian methodology in models from applied probability, machine learning and statistical physics that involve (large and often dependent) data X^n of nonstandard types. An example with random-walk data concerns the goodness-of-fit tests of Appendix B in the Supplementary Material [22]. An example with random-graph data concerns recovery of the community structure in the planted bisection model, which is known to be possible *if and only if* the sparsity levels for edges within and between communities satisfy certain limits [1, 32]. In [21], these *necessary* conditions are found to be *sufficient* for (almost-)exact recovery with posteriors, showing that theory presented here does not impose overly stringent conditions (at least in this random graph model).

The nature of Bayesian test sequences. The existence of a Bayesian test sequence is equivalent to consistent posterior convergence in the Bayesian, prior-almost-sure sense. Bayesian test sequences are more abundant than uniform or pointwise test sequences. To optimize the composite power of a Bayesian test *the prior should assign little mass where the test is less powerful, and much where the test is more powerful*, ideally.

This point appears to be especially relevant in *model selection* with posterior odds, which requires careful construction of Bayesian tests with little prior mass near the boundaries between hypotheses, leading to *upper bounds for prior mass*, as in [19]. Appendix B in the Supplementary Material [22] illustrates the influence of the prior on frequentist hypothesis testing with posterior odds.

Frequentist uncertainty quantification. Use of a prior that induces remote contiguity allows one to convert credible sets of calculated, simulated or approximated posteriors into asymptotically consistent confidence sets.

The latter conclusion forms the most important and practically useful aspect of this paper. For example, in the planted bisection model, the devices of Section 6 give rise to frequentist *uncertainty quantification for community structure*: if exact recovery is possible, credible sets are asymptotic confidence sets; if recovery is almost-exact, enlarged credible sets are asymptotic confidence sets [21].

Acknowledgements. The author thanks J. van Waaij and S. Rizzelli for interesting discussions and perspectives.

SUPPLEMENTARY MATERIAL

Appendices (DOI: 10.1214/20-AOS1952SUPP; .pdf). A. Definitions and conventions; B. Goodness-of-fit for random walks.

REFERENCES

[1] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[2] BARRON, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. Statistics, Univ. Illinois.

[3] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718 https://doi.org/10.1214/aos/1018031206

[4] BAYARRI, M. J. and BERGER, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.* **19** 58–80. MR2082147 https://doi.org/10.1214/088342304000000116

[5] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129 https://doi.org/10.1007/BF00532480

[6] BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* 55–87. Springer, New York. MR1462939

[7] BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031

[8] BREIMAN, L., LE CAM, L. and SCHWARTZ, L. (1964). Consistent estimates and zero–one sets. *Ann. Math. Stat.* **35** 157–161. MR0161413 https://doi.org/10.1214/aoms/1177703737

[9] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[10] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525 https://doi.org/10.1214/aos/1176349157

[11] DOOB, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et Ses Applications*. *Colloques Internationaux du Centre National de la Recherche Scientifique* **13** 23–27. Centre National de la Recherche Scientifique, Paris. MR0033460

[12] FALK, M., PADOAN, S. and RIZZELLI, S. (2019). Strong convergence of multivariate maxima. Preprint. Available at arXiv:1903.10596.

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

[14] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

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

[16] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707 https://doi.org/10.1214/09-AOS738

[17] GREENWOOD, P. E. and SHIRYAYEV, A. N. (1985). *Contiguity and the Statistical Invariance Principle*. *Stochastics Monographs* **1**. Gordon & Breach, New York. MR0822226

[18] HENGARTNER, N. W. and STARK, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550. MR1332580 https://doi.org/10.1214/aos/1176324534
[19] JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. MR2830762 https://doi.org/10.1111/j.1467-9868.2009.00730.x
[20] KLEIJN, B. The frequentist theory of Bayesian statistics, Ch. 8, Springer, Berlin (in preparation).
[21] KLEIJN, B. and VAN WAAIJ, J. (2018). Recovery and confidence sets of communities in a sparse stochastic block model. Preprint. Available at arXiv:1810.09533.
[22] KLEIJN, B. J. K. (2021). Supplement to "Frequentist validity of Bayesian limits." https://doi.org/10.1214/20-AOS1952SUPP
[23] KLEIJN, B. J. K. and ZHAO, Y. Y. (2019). Criteria for posterior consistency and convergence at a rate. *Electron. J. Stat.* **13** 4709–4742. MR4033683 https://doi.org/10.1214/19-EJS1633
[24] LE CAM, L. (1960). Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. Calif. Publ. Statist.* **3** 37–98. MR0126903
[25] 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
[26] LE CAM, L. (1979). An inequality concerning Bayes estimates. Preprint, Univ. California, Berkeley, CA.
[27] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. *Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7
[28] LE CAM, L. and YANG, G. L. (1988). On the preservation of local asymptotic normality under information loss. *Ann. Statist.* **16** 483–520. MR0947559 https://doi.org/10.1214/aos/1176350817
[29] LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics*: *Some Basic Concepts*. *Springer Series in Statistics*. Springer, New York. MR1066869 https://doi.org/10.1007/978-1-4684-0377-0
[30] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381
[31] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 https://doi.org/10.1214/aos/1030741084
[32] MOSSEL, E., NEEMAN, J. and SLY, A. (2016). Consistency thresholds for the planted bisection model. *Electron. J. Probab.* **21** Art. ID 21. MR3485363 https://doi.org/10.1214/16-EJP4185
[33] PADOAN, S. and RIZZELLI, S. (2019). Strong consistency of nonparametric Bayesian inferential methods for multivariate max-stable distributions. Preprint. Available at arXiv:1904.00245.
[34] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378 https://doi.org/10.1007/BF00535479
[35] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337 https://doi.org/10.1214/aos/1009210686
[36] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861 https://doi.org/10.1214/14-AOS1270
[37] TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112
[38] TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. MR2719655 https://doi.org/10.1214/10-BA605
[39] TORGERSEN, E. (1991). *Comparison of Statistical Experiments*. *Encyclopedia of Mathematics and Its Applications* **36**. Cambridge Univ. Press, Cambridge. MR1104437 https://doi.org/10.1017/CBO9780511666353
[40] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
[41] WALKER, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043. MR2102501 https://doi.org/10.1214/009053604000000409
[42] WALKER, S. G., LIJOI, A. and PRÜNSTER, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866 https://doi.org/10.1214/009053606000001361
[43] WASSERMAN, L. (2000). Bayesian model selection and model averaging. *J. Math. Psych.* **44** 92–107. MR1770003 https://doi.org/10.1006/jmps.1999.1278
[44] WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. MR1332570 https://doi.org/10.1214/aos/1176324524