FREQUENTIST VALIDITY OF BAYESIAN LIMITS

By B. J. K. Kleijn,

Korteweg-de Vries institute for Mathematics

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 [21].

Example 1.1 Consider $X_1, X_2, \ldots$ that are i.i.d. $P_0$ with continuous, non-zero 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 $\eta$ on $[0,1]$ with

MSC 2010 subject classifications: Primary 62G05, 62G20; secondary 62G10, 62C10, 62B15

Keywords and phrases: contiguity, frequentist consistency, posterior consistency, posterior rate of convergence, posterior odds, model selection, credible set, coverage, uncertainty quantification
η(x) > 0 for all x ∈ (0, 1) and θ ∈ ℝ: \( p_{θ,η}(x) = η(x - θ) 1_{[θ,θ+1]}(x) \). If \( θ ≠ θ' \) then,
\[-P_{θ,η} \log \frac{p_{θ',η'}}{p_{θ,η}} = ∞,\]
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 \( P_a \), \( a ≥ 1 \), defined by \( p_a(k) = P_a(X = k) = Z_a^{-1} k^{-a} (\log k)^{-3} \), for all \( k ≥ 2 \).

For \( a = 1, b > 1 \),
\[-P_a \log \frac{p_b}{p_a} < ∞, \quad P_a \left( \log \frac{p_b}{p_a} \right)^2 = ∞.\]

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. \( □ \)

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 Subsection 3.3 that remote contiguity has the potential to provide sequential approximations in non-parametric statistics, analogous to approximation by contiguous sequences in parametric setting [18]. This is illustrated by recent work [33, 12] 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 [34, 14], 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 non-locality of priors, as in [20].

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 [30]). 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 non-parametric 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 supplement 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 [22, 12, 33]. Definitions, notation and conventions roughly follow those of [28] and are collected in appendix A in the supplement.

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, (possibly non-i.i.d.) samples $X^n$, distributed according to $P_{\theta_0,n}$ for some $\theta_0 \in \Theta$, within a model $\theta \mapsto P_{\theta,n}$. (More generally, refer to appendix A in the supplement for notation and conventions.)
**Definition 2.1** Given priors \((\Pi_n)\) on the measurable spaces \((\Theta_n, \mathcal{G}_n)\), model subsets \((B_n), (V_n) \subset \mathcal{G}_n\) and \(a_n \downarrow 0\), a sequence of \(\mathcal{B}_n\)-measurable maps \(\phi_n : \mathcal{X}_n \rightarrow [0,1]\) is called a Bayesian test sequence for \(B_n\) versus \(V_n\) (under \(\Pi_n\)) of composite power \(a_n\), if,

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

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

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 \(B,V \in \mathcal{G}\) and any measurable \(\phi : \mathcal{X} \rightarrow [0,1]\),

\[
(2.2) \quad \int_B P_{\theta}(V|X) d\Pi(\theta) \leq \int_B P_{\theta}(\phi(X)) d\Pi(\theta) + \int_V P_{\theta}(1 - \phi(X)) d\Pi(\theta).
\]

**Proof.** Due to Bayes’s Rule (A.2) and monotone convergence,

\[
\int (1 - \phi(X)) \Pi(V|X) dP_{\Pi} = \int V P_{\theta}(1 - \phi(X)) d\Pi(\theta).
\]

Accordingly, \(\int_B P_{\theta}(1-\phi) \Pi(V|X) d\Pi(\theta) \leq \int (1-\phi) \Pi(V|X) dP_{\Pi} = \int_V P_{\theta}(1-\phi) d\Pi(\theta)\). Inequality (2.2) follows from the fact that \(\Pi(V|X) \leq 1\). \(\square\)

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 \(P_{n|B_n}^{\Pi_n}\) and \(P_{n|V_n}^{\Pi_n}\) are defined in Definition A.2.)

**Proposition 2.3** Let \((\mathcal{X}_n, \mathcal{B}_n), (\Theta_n, \mathcal{G}_n), (\mathcal{P}_n)\) and \((\Pi_n)\) be given. Given sequences \((B_n), (V_n) \subset \mathcal{G}_n\) and \((a_n), (b_n), (c_n)\) such that \(a_n = o(b_n \land c_n)\) and, \(\Pi_n(B_n) \geq b_n > 0, \Pi_n(V_n) \geq c_n > 0\). 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,

\[ P_n^{\Pi_n|B_n}\Pi(V_n|X^n) = o(a_n b_n^{−1}), \quad P_n^{\Pi_n|V_n}\Pi(B_n|X^n) = o(a_n c_n^{−1}). \]

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

**Proof.** Assume (i). Then,

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

(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,

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

so \( (\phi_n) \) 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 Subsection 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,*

\[ P_{0,n}\Pi(V_n|X) \leq \|P_{0,n} - P_n^{\Pi|B_n}\| \]

(2.4)

\[ + \int P_{\theta,n}\phi_n d\Pi_n(\theta|B_n) + \frac{\Pi_n(V_n)}{\Pi_n(B_n)} \int P_{\theta,n}(1 - \phi_n) d\Pi_n(\theta|V_n), \]

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 [30]. A less successful application pertains to non-parametric posterior rates of convergence for i.i.d. data, in an unpublished paper [27].
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 [28], 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 \((\Theta, \mathcal{G}, \Pi)\) be given and assume that there is a coupling \(X \in \mathcal{X}^\infty\) with distribution \(P_\theta\) and marginals \(X^n \sim P_{\theta,n}\) for every \(\theta \in \Theta\) and \(n \geq 1\). For any \(B, V \in \mathcal{G}\) with \(\Pi(B) > 0, \Pi(V) > 0\), the following are equivalent:

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

\[
\phi_n(X^n) \overset{P_{\theta,n}-a.s.}{\rightarrow} 0, \quad \phi_n(X^n) \overset{P_{\theta'-n}-a.s.}{\rightarrow} 1,
\]

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

\[
P_{\theta,n}\phi_n \to 0, \quad P_{\theta',n}(1 - \phi_n) \to 0,
\]

(iii) there are \(\mathcal{B}_n\)-msb. \(\phi_n : \mathcal{X}_n \to [0,1]\) 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 \(\Pi\)-almost-all \(\theta \in B, \theta' \in V\),

\[
\Pi(V|X^n) \overset{P_{\theta,n} - a.s.}{\rightarrow} 0, \quad \Pi(B|X^n) \overset{P_{\theta',n} - a.s.}{\rightarrow} 0.
\]

**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 \(\Pi^*\) 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,

\[
\int P_{\theta}|\Pi(V|X^n) - g(X)| d\Pi(\theta|B) \to 0.
\]
So $\int P_\theta g(X) \, d\Pi(\theta|B) = 0$, implying that $g = 0$, $P_\theta$-almost-surely for $\Pi$-almost-all $\theta \in B$. Using martingale convergence again (now in $L^\infty(\mathcal{F}^\infty \times \Theta)$), conclude $\Pi(V|X^n) \to 0$, $P_\theta$-almost-surely for $\Pi$-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 $\Pi$-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 [28]): if $\Theta$ is Polish (more precisely, a Borel subset of a complete metric spaces), there exists a Borel measurable $\vartheta : \mathcal{F}^\infty \to \Theta$ such that $P_\theta(\vartheta(X) = \theta) = 1$, for $\Pi$-almost-all $\theta \in \Theta$. (Note: here and elsewhere in i.i.d. setting, the parameter space $\Theta$ is the single-observation model $\mathcal{P}$, $\theta$ 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 $\Pi$. For any Borel set $V$ there is a Bayesian test sequence for $V$ versus $\mathcal{P} \setminus V$ under $\Pi$.

**Proof.** (See [11] and [28], Section 17.7, Proposition 1 with the indicator for $V$; see also [8].) Note that if $\vartheta : \mathcal{F}^\infty \to \Theta$ exists, then by martingale convergence in $L^\infty(\mathcal{F}^\infty \times \Theta)$, $\Pi(V|X^n) \to \int 1_V(\theta) \, d\Pi(\theta|X) = 1_V(\vartheta(X))$, $\Pi^*$-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 $(\Pi_n)$ and $\mathcal{G}$-measurable model subsets $B_n, V_n$ be given. For given
tests \((\phi_n)\) and composite power \(a_n\), write (2.1) as follows:

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

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 \(p_{V_n,n}\)),

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

is optimal and has composite power \(|\Pi_n(B_n) P_n^{\Pi_n[B_n]} \land \Pi_n(V_n) P_n^{\Pi_n[V_n]}||\). (Here, \(P \land Q\) denotes the minimum of \(P\) and \(Q\) [28], the largest (sub-probability) measure \(\lambda\) 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 \((P \land Q)(A) = \int_A (p(x) \land q(x)) d\mu(x)\).) This leads to the following lemma based on the so-called Hellinger transform (see Section 16.4, Remark 1 in [28]).

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

\[
(2.5) \int_{B_n} P_{\theta,n} \phi_n d\Pi_n(\theta) + \int_{V_n} P_{\theta,n} (1 - \phi_n) d\Pi_n(\theta) \\
\leq \int \left( \Pi_n(B_n) p_{B_n,n}(x) \right)^\alpha \left( \Pi_n(V_n) p_{V_n,n}(x) \right)^{1-\alpha} d\mu_n(x),
\]

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,

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

Combination with (2.5) for \(\alpha = 1/2\), implies that the minimax upper bound in \(i.i.d.\) cases [28] remains valid:

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

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 supplement). This falls in line with the argument that underpins non-locality of priors for variable selection, as in [20].

3. Remote contiguity. In this section we weaken the notion of contiguity (see [24], chapter 6 in [28] and [18, 30]) in a way that is suitable to promote II-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 \((\mathcal{X}_n, \mathcal{B}_n)\), \(n \geq 1\) with two sequences \((P_n)\) and \((Q_n)\) of probability measures and a sequence \(\rho_n \downarrow 0\), we say that \(Q_n\) is \(\rho_n\)-remotely contiguous with respect to \(P_n\), notation \(Q_n \triangleright \rho_n^{-1}P_n\), if,

\[
P_n \phi_n(X^n) = o(\rho_n) \implies Q_n \phi_n(X^n) = o(1),
\]

for every sequence of \(\mathcal{B}_n\)-measurable \(\phi_n : X_n \to [0, 1]\).

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 \(P_n \triangleright Q_n\) is equivalent to remote contiguity \(P_n \triangleright a_n^{-1}Q_n\) for all \(a_n \downarrow 0\).

**Example 3.2** Let \(\mathcal{P}\) 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 \(n\),

\[
\frac{dP^n}{dP_0^n}(X^n) \geq e^{-\frac{n}{2}\epsilon^2},
\]

with \(P_0^n\)-probability one. Consequently, for large enough \(n\) and for any \(\mathcal{B}_n\)-measurable sequence \(\psi_n : X_n \to [0, 1]\), \(P^n \psi_n \geq e^{-\frac{1}{2}n\epsilon^2} P_0^n \psi_n\). Therefore, if
\( P^n \phi_n = o(\exp(-\frac{1}{2}n\epsilon^2)) \) then \( P^n_0 \phi_n = o(1) \). 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 \( (P^n_0) \) 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 supplement.)

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

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

(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 \( \liminf_n b a_n^{-1} P_n(dQ_n/dP_n > b a_n^{-1}) = 1 \),

(iv) for any \( \epsilon > 0 \), there is a constant \( c > 0 \) such that \( \| Q_n - Q_n \wedge c a_n^{-1} P_n \| < \epsilon \), 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 \((i) \text{ or } (iv)\) implies remote contiguity; that \((ii) \text{ or } (iii)\) \(\Rightarrow\) \((iv)\) and that \((v) \Leftrightarrow (ii)\).) Assume \((i)\). Let \( \phi_n : \mathcal{X}_n \to [0, 1] \) 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 \( \phi_n Q_n \xrightarrow{P} 0 \) and since \( \phi_n \) is bounded, that implies \( Q_n \phi_n = o(1) \), so that \( Q_n \prec a_n^{-1} P_n \). Next, assume \((iv)\). Let \( \epsilon > 0 \) and \( \phi_n : \mathcal{X}_n \to [0, 1] \) be given. By assumption, there exist \( c > 0 \) and \( N \geq 1 \) such that for all \( n \geq N \),

\[
Q_n \phi_n < c a_n^{-1} P_n \phi_n + \frac{\epsilon}{2}.
\]

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

\[ (3.9) \]

\[
\left\| Q_n - Q_n \wedge c a_n^{-1} P_n \right\| = \sup_{A \in \mathcal{A}_n} \left( \int_A q_n \, d\mu_n - \int_A q_n \wedge c a_n^{-1} p_n \, d\mu_n \right)
\]

\[
\leq \sup_{A \in \mathcal{A}_n} \int_A (q_n - q_n \wedge c a_n^{-1} p_n) \, d\mu_n
\]

\[
= \int 1\{q_n > c a_n^{-1} p_n\} (q_n - c a_n^{-1} p_n) \, d\mu_n.
\]

Note that the right-hand side of (3.9) 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 \( n \geq 1 \),

\[
0 \leq \int c a_n^{-1} P_n(q_n > c a_n^{-1} p_n) \leq Q_n(q_n > c a_n^{-1} p_n) \leq 1,
\]

so (3.9) goes to zero if \( \lim \inf_{n \to \infty} c a_n^{-1} P_n(dQ_n/dP_n > c a_n^{-1}) = 1 \). 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 \( (\mathcal{X}_n, \mathcal{B}_n), (n \geq 1) \) 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 \( \rho_n \)-to-\( \sigma_n \) remotely contiguous with respect to \( P_n \), notation \( \sigma_n^{-1} Q_n \ll \rho_n^{-1} P_n \), if,

\[
P_n \phi_n(X^n) = o(\rho_n) \Rightarrow Q_n \phi_n(X^n) = o(\sigma_n),
\]

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

Like Definition 3.1, Definition 3.4 allows for reformulation similar to Lemma 3.3, e.g. if for some sequences \( \rho_n, \sigma_n \) like in Definition 3.4,

\[
\left\| Q_n - Q_n \wedge \sigma_n \rho_n^{-1} P_n \right\| = o(\sigma_n),
\]

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

**Example 3.5** The inequality of Example 3.2 implies that \( b_n^{-1} P_0^n \ll a_n^{-1} P^n \), 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
B. J. K. KLEIJN

noted that this implies that $\phi_n(X^n) \xrightarrow{P_{\theta_n}} 0$ for any $\phi_n : \mathcal{X}_n \to [0, 1]$ 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 $P_{\Pi_n|B_n}$, where the sets $B_n \subset \Theta$ are such that,

$$P_{\theta_0,n} \ll a_n^{-1} P_{\Pi_n|B_n},$$

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

Example 3.6 Let $\mathcal{P}$ be a model for i.i.d. samples $X^n$ as in Example 3.2. Fix $P_0$ and $\epsilon > 0$, define $K(\epsilon) = \{ P \in \mathcal{P} : -P_0 \log(dP/dP_0) < \epsilon^2 \}$ and recall that a KL-prior $\Pi$ satisfies, $\Pi(K(\epsilon)) > 0$ for every $\epsilon > 0$. The exponential lower bound (3.8) implies that $\liminf_n \exp(\frac{1}{2} n \epsilon^2) (dP_n/\Pi_n)(X^n) \geq 1$ with $P_0$ probability one for every $P \in K(\epsilon)$. With Fatou’s lemma,

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

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

$$P_0 \ll e^{\frac{1}{2} n \epsilon^2} P_{\Pi_n|K(\epsilon)}.$$

A version of the form $b_n^{-1} P_0 \ll a_n^{-1} P^n$ 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 $P_{\theta}$ have Lebesgue densities $p_{\theta}(x) = 1_{[\theta, \theta+1]}(x)$, for $\theta \in \Theta = \mathbb{R}$. Pick a prior $\Pi$ on $\Theta$ 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 $X^n \sim P_{\theta_0,n}$, $dP_n/dP_{\theta_0}(X^n) = 1\{X(1) \geq \theta\}$, and
correspondingly,
\[
\frac{dP_n^{\Pi|B_n}}{dP_{\theta_0}^n}(X^n) = \Pi_n(B_n)^{-1} \int_{\theta_0}^{\theta_0 + \delta_n} 1\{X(1) \geq \theta\} \, d\Pi(\theta)
\geq \frac{1 - \alpha \delta_n (X(1) - \theta_0)}{1 + \alpha \delta_n},
\]
for large enough \( n \). As a consequence, for every \( \delta > 0 \) and all \( a_n \downarrow 0 \),
\[
P_{\theta_0}^n \left( \frac{dP_n^{\Pi|B_n}}{dP_{\theta_0}^n}(X^n) < \delta a_n \right) \leq P_{\theta_0}^n \left( \frac{\delta_n^{-1}(X(1) - \theta_0)}{(1 + \alpha)\delta a_n} \right),
\]
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 r.h.s. in the above display goes to zero. So \( P_{\theta_0,n} \ll a_n^{-1} P_n^{\Pi_n|B_n} \), for any \( a_n \downarrow 0 \). Conclude that with these choices for \( \Pi \) and \( B_n \), (3.10) holds, for any \( a_n \).

Example 3.7 emphasizes the role of weak convergence of likelihood ratios, similar to limits of experiments \([25, 28, 39]\). To emphasize this relation further, consider the following proposition. Proposition 3.8 should be viewed in light of \([29]\), 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 \( \theta \) is unavailable.

**Proposition 3.8** Let \( \theta_0 \in \Theta \) and priors \( \Pi_n : \mathcal{G} \to [0,1], n \geq 1 \) be given. Let \( (B_n) \) be a sequence of measurable subsets of \( \Theta_n \) 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 \( P_{\theta_0,n} \ll a^{-1}_n P_n^{\Pi_n|B_n} \).

**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 \( \delta \), 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 \( c \geq 0 \) such that,
\[
\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c a^{-1}_n P_{\theta_0,n}\| < \epsilon,
\]
(3.11)
for all $n \geq 1$, $\theta \in B_n$. Now note that for any $A \in \mathcal{B}_n$,

$$0 \leq P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \land c a_n^{-1} P_{\theta_0,n}^{B_n}(A)$$

$$\leq \int (P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \land c a_n^{-1} P_{\theta_0,n}(A)) d\Pi_n(\theta | B_n).$$

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

$$\|P_{\theta_0,n} - P_{\theta_0,n} \land c a_n^{-1} P_{\theta_0,n}^{B_n}\| \leq \int \|P_{\theta_0,n} - P_{\theta_0,n} \land c a_n^{-1} P_{\theta_0,n}\| d\Pi_n(\theta | B_n).$$

Based on (3.11), condition (iv) of Lemma 3.3 is satisfied. $\square$

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, for example, Theorem 3.1 in [16]) 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 non-parametric context, consider the following standard example.

Let $\mathcal{F}$ denote a class of functions $\mathcal{X} \rightarrow \mathbb{R}$, where $\mathcal{X}$ 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 $\mathcal{X} \times \mathbb{R}$, assumed to be related through,

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

for some unknown $f_0 \in \mathcal{F}$, 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 $X_1,\ldots,X_n \sim \mathcal{P}^n$, for some ancillary distribution $\mathcal{P}$ on $\mathbb{R}$. Assume that $\mathcal{F} \subset L^2(\mathcal{P})$ and that $Pf(X) = 0$ for all $f \in \mathcal{F}$. We distinguish two cases: (a) the case of linear regression, $\mathcal{F} = \{f_\theta : \mathcal{X} \subset \mathbb{R} \rightarrow \mathbb{R} : \theta \in \Theta\}$, where $\theta = (a,b) \in \Theta = \mathbb{R}^2$ and $f_\theta(x) = ax+b$; (b) the case of non-parametric regression (to maintain concreteness, we keep in mind the special case $\mathcal{F} = C^\alpha_1(\mathcal{X})$, the collection of all $\alpha$-smooth functions on $\mathcal{X}$ with Hölder-$\alpha$-norm $\|\cdot\|_\alpha$ bounded by 1).
For \((\rho_n)\) 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 \(f_0, f \in \mathcal{F}\) in the following form,

\[
(3.12) \quad a_n^{-1} \frac{dP_{f_0, n}}{dP_{f_0, n}}(X^n) = e^{-\frac{1}{2} \sum_{i=1}^{n}(2\varepsilon_i(f-f_0)(X_i)+(f-f_0)^2(X_i)-n\rho_n^2)}
\]

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

**Example 3.9** In the parametric case, expression (3.12) can be written in terms of a local parameter \(h \in \mathbb{R}^2\) which, for given \(\theta_0\) and \(n \geq 1\), is related to \(\theta\) 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,

\[
(3.13) \quad \frac{dP_{h, n}}{dP_{0, n}}(X^n) = e^{\frac{1}{2} \sum_{i=1}^{n}h \cdot \ell_{\theta_0}(X_i,Y_i)-\frac{1}{2}h \cdot \ell_{\theta_0}h+o_{P_{\theta_0, n}}(1)},
\]

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 \(\theta\) at \(\theta_0\), \(I_{\theta_0} = P_{\theta_0,1} \ell_{\theta_0} \ell_{\theta_0}^T\) is the Fisher information matrix. Assume \(I_{\theta_0}\) is non-singular and note the central limit,

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_{\theta_0}(X_i,Y_i) \xrightarrow{P_{\theta_0, n \to w}} N_2(0, I_{\theta_0}),
\]

which expresses local asymptotic normality of the model and implies that for any fixed \(h \in \mathbb{R}^2\), \(P_{h, n} \ll P_{0, n}\). It is well-known that contiguity extends to \(n^{-1/2}\)-localized prior averages (see Lemma 3, Section 8.4 in [30]):

\[
(3.14) \quad P_{\theta_0, n} \ll P_{\Pi|B_n},
\]

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

**Example 3.10** In the non-parametric case, define \(B(\rho) = \{f \in \mathcal{F} : \|f-f_0\| < \rho\}\) (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, Subsection 3.4.3 of [40] prove that the (outer) expectation of the supremum of the empirical process for scores satisfies the maximal inequality,

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

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 \( \epsilon > 0 \),

\[
P_{f_0,n} \left( \sup_{f \in B(\rho_n)} \left| \sum_{i=1}^{n} e_i(f - f_0)(X_i) \right| > \frac{n\rho_n^2 \beta_n}{\epsilon} \right) \leq \epsilon. \tag{3.15}
\]

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

\[
\sup_{f \in B(\rho_n)} \left| \frac{1}{n} \sum_{i=1}^{n} (f - f_0)^2(X_i) - \|f - f_0\|_{L^2(P)}^2 \right|_{P^{\infty}-a.s.} \rightarrow 0.
\]

then for any \( \delta, \delta' > 0 \), using (3.15) and assuming that \( \Pi_n(B(\rho_n)) > 0 \),

\[
P_{f_0,n} \left( \frac{1}{\Pi_n(B(\rho_n))} \int_{B(\rho_n)} \frac{dP_{f,n}}{dP_{f_0,n}}(X^n) d\Pi_n(f) < \delta a_n \right)
\]

\[
\leq P_{f_0,n} \left( \inf_{f \in B(\rho_n)} a_n^{-1} \frac{dP_{f,n}}{dP_{f_0,n}}(X^n) < \delta \right)
\]

\[
\leq P_{f_0,n} \left( \inf_{f \in B(\rho_n)} - \sum_{i=1}^{n} e_i(f - f_0)(X_i) + \frac{1}{2} n\rho_n^2 < \log \delta - \delta' \right) \leq \epsilon,
\]

for large enough \( n \). Conclude that,

\[
P_{f_0,n} \leq e^{\frac{1}{2} n\rho_n^2} P_{n\Pi \mid B(\rho_n)}. \tag{3.16}
\]

A similar proof based on Proposition 3.8 is also possible. For a smoothness class \( \mathcal{F} = C_1^\alpha(\mathcal{X}) \) (and provided certain technical conditions are met, see Subsection 3.4.3.2 in [40]), rates \( \rho_n \) 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 \( \rho_n \), and we can show that \( P_{f_0,n}(A_n) = o(e^{\frac{1}{2} n\rho_n^2}) \), then also \( P_{f_0,n}(A_n) = o(1) \).)

The analogy between (3.14) and (3.16) 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 non-parametric statistics, analogous to approximation by contiguous sequences in parametric setting [18]. More examples of sequential approximation by remote contiguity are provided in [12, 33] and [22].
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 \((\mathcal{X}_n, \mathcal{B}_n), (\Theta_n, \mathcal{G}_n), (\mathcal{P}_n)\) and \((\Pi_n)\) be given. Assume that for all \(n \geq 1\), the data \(X^n \sim P_{0,n}\) and that, for given \(B_n, V_n \in \mathcal{G}_n\) and \(a_n, b_n \downarrow 0\) with \(a_n = o(b_n)\),

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

\[
(4.17) \quad \int_{B_n} P_{\theta,n} \phi_n d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1 - \phi_n) d\Pi_n(\theta) = o(a_n),
\]

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

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

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

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

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 \(B_0\) (or even lie in \(\mathcal{P}_n\)): 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 \(\theta_0\) in \(\Theta\) and \(B \subset U\). If, moreover, we have data \(X^n\) that is \(i.i.d.\), we arrive at Schwartz’s consistency in \(\mathcal{P}\). 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 \(X^n \sim P^n_0\) for some \(P_0 \in \Theta\). Fix a prior \(\Pi : \mathcal{G} \to [0, 1]\) and assume that for given \(B, V \in \mathcal{G}\) with \(\Pi(B) > 0\) and \(a_n \downarrow 0\),
(i) there exist Bayesian tests $\phi_n$ for $B$ versus $V$,

\[(4.18) \quad \int_B P^n \phi_n d\Pi(P) + \int_V Q^n (1 - \phi_n) d\Pi(Q) = o(a_n),\]

(ii) the sequence $P^n_{\theta_0}$ satisfies $P^n_{\theta_0} \prec a_n^{-1} P^{|B|}_{\Pi}$. 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 $P^n_0$ 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 $\mathcal{P}$ 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 non-zero Hellinger distance. Assuming that $\mathcal{P}$ is dominated, any $\Pi$ that is Borel for the Hellinger topology on $\mathcal{P}$ is Radon in the completion, so for every $\delta > 0$, there exists a Hellinger pre-compact (that is, totally-bounded) $K \subset \mathcal{P}$, such that $\Pi(K) > 1 - \delta$. Totally-boundedness is the entropy argument needed in a well-known construction [26, 5, 28, 14] 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 [28]) into uniform test $\phi_n$ of $B$ versus $V \cap K$ of exponential composite power:

\[(4.19) \quad \int_B P^n \phi_n d\Pi(P) + \int_V Q^n (1 - \phi_n) d\Pi(Q) \leq N(\epsilon, V \cap K, H) e^{-n\epsilon^2} + \delta,\]

for some $\epsilon > 0$. Diagonalization with respect to exponentially decreasing $\delta$’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 [27, 3, 14, 35, 42, 21]. 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 \( \Pi_{|B_n} \Pi(V_n|X^n) \leq e^{-M_n + 1} \). 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,

\[
V_n = \{ P \in \mathcal{P} : H(P, P_0) \geq M \epsilon_n \},
\]

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

We repeat the argument leading to (4.19) for every \( n \), with \( \epsilon = \epsilon_n, \epsilon_n \downarrow 0 \) and \( n\epsilon_n^2 \to \infty \). If we require Barron’s \( \delta \)-contribution in (4.19) to be of \( n\epsilon_n^2 \)-exponentially small order,

\[
\Pi(\mathcal{P} \setminus \mathcal{P}_n) \leq \exp(-nM\epsilon_n^2),
\]

and the sieve of pre-compact \( \mathcal{P}_n \) has Hellinger entropies that are upper-bounded (see [26, 5]),

\[
N(\epsilon_n, \mathcal{P}_n, H) \leq e^{Kn\epsilon_n^2},
\]

for some \( K > 0 \), then the minimax construction extends to tests that separate \( V_n = \{ P \in \mathcal{P} : H(P_0,P) \geq 4\epsilon_n \} \) from \( B_n = \{ P \in \mathcal{P} : H(P_0,P) < \epsilon_n \} \) 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 \( \epsilon_n \) around \( P_0 \), so (4.17) holds. Remote contiguity therefore requires that for some \( C > 0 \),

\[
\Pi_n(B_n) \geq e^{-Cn\epsilon_n^2},
\]
We note Lemma 8.1 in [14], which says that if (4.20) is satisfied then Lemma 3.3-(ii) holds, so that,

\[(4.21)\quad P_0^n \ll e^{c n \epsilon^2_n} P_n^|B_n|,\]

for any $c > 1$. For large enough $M$, Theorem 4.1 then reproduces the GGV-result, i.e. the posterior is Hellinger consistent at rate $\epsilon_n$. 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 (non-parametric) 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 $\mathcal{P}$ of distributions $P$ for i.i.d. data $X^n \sim P^n$, $(n \geq 1)$ and suppose that $\mathcal{P}$ is Hellinger-separable. Let $P_0 \in \mathcal{P}$ and $\epsilon_n \rightarrow 0$ be given, denote $V(\epsilon) = \{P \in \mathcal{P} : H(P_0, P) \geq 4 \epsilon\}$, $B_H(\epsilon) = \{P \in \mathcal{P} : H(P_0, P) < \epsilon\}$ for all $\epsilon > 0$. There exist $N(\epsilon_n) \geq 1$ (possibly infinite) and a cover of $V(\epsilon_n)$ by $N(\epsilon_n)$ Hellinger balls $V_{n,1}, V_{n,2}, \ldots$ of radius $\epsilon_n$ 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 $1 \leq i \leq N(\epsilon_n)$ 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 $B_{n,i}' \subset B_H(\epsilon_n)$,

\[(4.22)\quad P_n^{|B_n'| \Pi(V(\epsilon_n)|X^n)} \leq \sum_{i=1}^{N(\epsilon_n)} P_n^{|B_n'| \Pi(V_{n,i}|X^n)} \]

\[\leq \frac{1}{\Pi(B_n')} \sum_{i=1}^{N(\epsilon_n)} \left( \int_{B_n'} P^n \phi_{n,i} d\Pi(P) + \int_{V_{n,i}} P^n (1 - \phi_{n,i}) d\Pi(P) \right) \]

\[\leq \sum_{i=1}^{N(\epsilon_n)} \sqrt{\frac{\Pi(V_{n,i})}{\Pi(B_n')}} \exp(-\frac{1}{2} n \epsilon_{i,n}^2).\]

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 [26] to define the so-called Le Cam dimension of the model, as well as to various subtle results on posterior behaviour in non-parametric applications, and also explains the sharpness of the posterior concentration results of [22]. We emphasize that (4.22) makes explicit the balancing of prior masses and composite power, as intended by the remark that closes Subsection 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 [38]). 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 $\Pi$, Theorem 2.5 says that the posterior gives rise to consistent posterior odds $G_n$ for $B$ versus $V$ in a Bayesian (that is, $\Pi$-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 [23], in the sense that $\phi_n(X^n) = 1\{X^n \in \mathcal{X}_n : \Pi(B|X^n) > \Pi(V|X^n)\}$ satisfies,

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

$$= \inf_\psi \int_B P_{\theta,n}\psi(X^n) d\Pi(\theta) + \int_V P_{\theta,n}(1 - \psi(X^n)) d\Pi(\theta),$$

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

imsart-aos ver. 2014/10/16 file: rc-aos-main-3.tex date: January 21, 2020
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,

\[
G_n \xrightarrow{P_{\theta,n}} 0, \quad G_n \xrightarrow{P_{\theta',n}} \infty,
\]

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 \( \theta' \in V \),

\[
P_{\theta,n} \searrow a_n^{-1} P_n^{\Pi_n|B}, \quad P_{\theta',n} \nearrow a_n^{-1} P_n^{\Pi_n|V}.
\]

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 \( P_{\theta,n} \nearrow a_n^{-1} P_n^{\Pi_n|B_n} \), then \( P_{\theta,n} \searrow a_n^{-1} P_n^{\Pi_n|B} \).

This device for model selection is used in the application of appendix B in the supplement: 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 \( \chi^2 \)-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 Subsection 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 non-locality of priors as in [20].
6. Confidence sets from credible sets. The assertion of the Bernstein-von Mises theorem [30] 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 non-parametric 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 [37]. 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 $\left(\Theta, \mathcal{G}\right)$ with priors $\Pi_n$, denote the sequence of posteriors by $\Pi(\cdot|\cdot): \mathcal{G} \times \mathcal{X}_n \rightarrow [0, 1]$. Let $\mathcal{D}$ denote a collection of measurable subsets of $\Theta$. 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 $D_n: \mathcal{X}_n \rightarrow \mathcal{D}$ such that $\Pi(\Theta \setminus D_n(x)|x) = o(a_n)$ for $P\Pi_n$-almost-all $x \in \mathcal{X}_n$.

Definition 6.2 For $0 \leq a \leq 1$, a set-valued map $x \mapsto C(x)$ defined on $\mathcal{X}$ 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 $n \rightarrow \infty$, the $C_n(X^n)$ are said to be asymptotically consistent.

Definition 6.3 Let $D$ be a (credible) set in $\Theta$ 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 $\theta \in \Theta \setminus C'$, $B(\theta) \cap D = \emptyset$. The intersection $C$ of all $C'$ like above equals $\{\theta \in \Theta: B(\theta) \cap D \neq \emptyset\}$ and is called the minimal (confidence) set associated with $D$ under $B$ (see Fig 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 $\Pi_n$ and let $D_n$ denote level $(1 - a_n)$ credible sets. Furthermore, for all $\theta \in \Theta$, let $B_n = \{B_n(\theta) \in \mathcal{G}: \theta \in \Theta\}$ denote a sequence such that,

\begin{enumerate}
\item $\Pi_n(B_n(\theta_0)) \geq b_n$,
\item $P_{\theta_0,n} < b_n a_n^{-1} P\Pi_n|B_n(\theta_0)$.
\end{enumerate}
Then any confidence sets $C_n$ associated with the credible sets $D_n$ under $B_n$ are asymptotically consistent,

\begin{align}
P_{\theta_0,n}(\theta_0 \in C_n(X^n)) \to 1.
\end{align}

**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 $P^\Pi_n(F_n) = 1$. 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 $B_n(\theta_0) \cap D_n(x) = \emptyset$. 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 $n$-th prior predictive distribution and dividing by the prior mass of $B_n(\theta_0)$, one obtains,

\begin{align*}
\frac{1}{\Pi_n(B_n(\theta_0))} \int 1\{\theta_0 \in \Theta \setminus C_n\} \Pi(B_n(\theta_0)|X^n) dP^\Pi_n \leq \frac{a_n}{b_n}.
\end{align*}

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

\begin{align*}
P^\Pi_n|B_n(\theta_0)(\theta_0 \in \Theta \setminus C_n(X^n)) = \int P_{\theta,n}(\theta_0 \in \Theta \setminus C_n(X^n)) d\Pi_n(\theta|B_n) \leq \frac{a_n}{b_n}.
\end{align*}
By the definition of remote contiguity, this implies asymptotic coverage c.f. (6.23).

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, [22] 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,

$$(ii') \quad c_n^{-1} P_{\theta_0,n} < b_n a_n^{-1} F_{\Pi_n|B_n(\theta_0)},$$

so that the last step in the proof of Theorem 6.4 is more specific; particularly, assertion (6.23) becomes,

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

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 [22], showing that posterior convergence at a rate ensures coverage of enlarged minimal-radius credible balls.

**Theorem 6.5** Suppose that $(\Theta, d)$ with Borel priors $(\Pi_n)$ parametrizes models $\Theta \to \mathcal{P}_n : \theta \mapsto P_{\theta,n}$ 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 $r_n$:

$$\Pi(d(\theta, \theta_0) \leq r_n \mid X^n) \xrightarrow{P_{\theta_0,n}} 1.$$ 

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, $\hat{r}_n \leq r_n$ 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, 
\[ P_{\theta_0,n}(\theta_{0,n} \in C_n(X^n)) \to 1, \]

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 \( \Theta \) 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 \( G \) be the Borel \( \sigma \)-algebra for a uniform topology on \( \Theta \). 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 \( \theta \). Let \( D \) denote any credible set. A confidence set associated with \( D \) under \( B(\theta) \) 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 \( \Theta \)), 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 \( \Theta \) with parametrization \( \theta \mapsto P^n_\theta \), to define a model for i.i.d. data \( X^n = (X_1, \ldots, X_n) \sim P^n_{\theta_0} \), for some \( \theta_0 \in \Theta \). Let \( D \) be the class of all pre-images of Hellinger balls, i.e. sets \( D(\theta, \epsilon) \subset \Theta \) of the form,
\[ D(\theta, \epsilon) = \{ \theta' \in \Theta : H(P_\theta, P_{\theta'}) < \epsilon \} \]
for any \( \theta \in \Theta \) and \( \epsilon > 0 \). After choice of a Kullback-Leibler prior \( \Pi \) for \( \theta \) 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 \( 1 - o(a_n), a_n = \exp(-na^n) \) for some \( a > 0 \). Assume, now, that for some \( 0 < \epsilon < a \), 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 $\epsilon$ 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 (4.20).

In the case $\epsilon_n$ is the minimax rate of convergence for the problem, the confidence sets $C_n(X^n)$ attain rate-optimality [31]. Rate-adaptivity [19, 17, 37] is not possible with Theorem 6.4 because a definite, non-data-dependent choice for the $B_n$ is required. An interesting option concerns the exploration of data-driven choices for priors $\Pi_n$ and $B_n$, as in [37].

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 non-parametric statistics, analogous to approximation by contiguous sequences in parametric statistics (for examples, see [33, 12]).

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 non-standard types. An example with random-walk data concerns the goodness-of-fit tests of appendix B in the supplement. An example with random-graph data concerns recovery of the community structure in the planted bi-section 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 [22], 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 [20]. Appendix B in the supplement 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 bi-section 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 [22].

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

SUPPLEMENTARY MATERIAL

Supplement: Appendices
(doi: 10.1214/00-AOSXXXXSUPP; .pdf).

REFERENCES


[27] L. Le Cam, An inequality concerning Bayes estimates, University of California, Berkeley (197X), unpublished.

ADDRESS OF THE AUTHOR
B. J. K. KLEIJN
KORTEWEG-DE VRIES INSTITUTE FOR MATHEMATICS
UNIVERSITY OF AMSTERDAM
P.O. BOX 94248
1090 GE AMSTERDAM
THE NETHERLANDS
E-MAIL: B.Kleijn@uva.nl
URL: https://staff.fnwi.uva.nl/b.j.k.kleijn