Criteria for Bayesian consistency

B. J. K. Kleijn

Korteweg-de Vries Institute for Mathematics, University of Amsterdam

August 2013

Abstract

Conditions for asymptotic suitability of Bayesian procedures focus on lower bounds for prior mass in Kullback-Leibler neighbourhoods of the data distribution. The goal of this paper is to investigate whether there is more flexibility in asymptotic criteria for posterior consistency, with the ultimate goal of formulating new, alternative consistency theorems based on a wider variety of prior suitability conditions. We formulate a versatile Bayesian consistency theorem, re-derive Schwartz’ theorem (Schwartz (1965) [22]), sharpen it to Kullback-Leibler consistency and formulate several other consistency theorems in which priors charge metric (e.g. Hellinger) balls. Results also apply to marginal, semi-parametric consistency; support boundary estimation is considered explicitly and posterior consistency is proved in a model where Schwartz’ theorem fails.

1 Introduction and main result

The most restrictive aspects of nonparametric Bayesian methods result from limited availability of suitable priors. In general, distributions on infinite dimensional spaces are relatively hard to define and control technically, so unnecessary elimination of candidate priors is highly undesirable. Specifying to asymptotic aspects, the conditions that Bayesian limit theorems pose on priors play a crucial role and should be as flexible as possible. It is the goal of this paper to relax standing criteria for posterior consistency (or equivalently, to show asymptotic suitability for a wider class of priors) under stronger model conditions.

As early as the 1940’s, J. Doob [4] studied posterior limits as a part of his exploits in Martingale convergence: if the data forms an infinite i.i.d. sample from a distribution $P_{\theta_0}$ in a model $\mathcal{P} = \{P_{\theta} : \theta \in \Theta\}$ where $\Theta$ and sample space are Polish spaces and $\Theta \rightarrow \mathcal{P} : \theta \mapsto P_{\theta}$ is one-to-one, then for any prior $\Pi$ on $\Theta$ the posterior is consistent, $\Pi$-almost-surely. Notwithstanding its remarkable generality, Doob’s theorem is perhaps somewhat misleading for non-parametric models, in that the null-set of the prior on which inconsistency occurs can be very large (Freedman (1963) [5]) and the collection of $(P_0, \Pi)$ pairs that lead to a consistent posterior can be topologically very small (that is, of first Baire category in the weak-* product topology; see Freedman (1965) [6]).

Freedman’s counterexamples discredited Bayesian methods for non-parametric statistics to the point of abandonment until its revival in the 1990’s. This neglect of Bayesian methods is hard to justify given that significant improvement over Doob’s theorem existed already
since 1965: Schwartz’ consistency theorem below concerns models \( \mathcal{P} \) that are dominated by a \( \sigma \)-finite measure and departs from the standard expression for the posterior,

\[
\Pi( A \mid X_1, \ldots, X_n) = \frac{\int_A \prod_{i=1}^n p(X_i) \, d\Pi(P)}{\int \mathcal{P} \prod_{i=1}^n p(X_i) \, d\Pi(P)},
\]

(1)

(which makes sense only if \( P_0^n \) is dominated by the posterior predictive distribution \( P_\Pi^n \); c.f. lemma 2.1, prior mass condition (2) implies \( P_0^n \ll P_\Pi^n \)).

**Theorem 1.1.** (Schwartz (1965) [22])

Let \( X_1, X_2, \ldots \) be i.i.d. \( - P_0 \in \mathcal{P} \), let \( \mathcal{P} \) be totally bounded relative to the Hellinger metric and let \( \Pi \) be such that for all \( \epsilon > 0 \),

\[
\Pi\left( \{ P \in \mathcal{P} : - P_0 \log \frac{dP}{dP_0} < \epsilon \} \right) > 0.
\]

(2)

Then for every \( \epsilon > 0 \) and every i.i.d.-\( P_0 \) sample \( X_1, X_2, \ldots \),

\[
\Pi\left( \{ P \in \mathcal{P} : H(P, P_0) > \epsilon \mid X_1, \ldots, X_n \} \right) \xrightarrow{P_0-a.s.} 0.
\]

This type of formulation departs from a specific underlying distribution \( P_0 \) and thus avoids Doob’s mishaps on null-sets of the prior. The extension to a theorem on posterior rates of convergence (see Ghosh, Ghosal and van der Vaart (2000) [5]) depends on sharpened versions of condition (2). Schwartz’ theorem and its rate-specific version have become the standard tools for the asymptotic analysis of Bayesian posteriors, almost to the point of exclusivity. As a consequence, lower bounds for prior mass in Kullback-Leibler neighbourhoods c.f. (2) are the only criteria frequentists apply to priors in non-parametric asymptotic analyses. Since these lower bounds are sufficient conditions, it is not clear if other criteria can be formulated and, if so, how much extra freedom such alternatives allow.

The goal of this paper is to investigate whether there is more flexibility in asymptotic criteria for priors with the ultimate goal of formulating new consistency theorems based on a greater variety of suitability conditions for priors. The goal is not to generalize conditions for Schwartz’ theorem or to sharpen its assertion (see, however, theorem 3.1); rather we want to show that stringency with regard to the prior can be relaxed at the expense of stringency with regard to the model. The main result is summerized in the following theorem.

**Theorem 1.2.** Let \( \mathcal{P} \) be dominated by a \( \sigma \)-finite measure. Let \( X_1, X_2, \ldots \) be i.i.d.-\( P_0 \) distributed for \( P_0 \in \mathcal{P} \). Assume that the prior \( \Pi \) is such that \( P_0^n \ll P_\Pi^n \) for all \( n \geq 1 \). Let \( V_1, \ldots, V_N \) be a finite collection of measurable model subsets. If there exist model subsets \( B_1, \ldots, B_N \) such that for every \( 1 \leq i \leq N \),

\[
\inf_{0 \leq \alpha \leq 1} \sup_{P \in \mathrm{co}(V_i)} \sup_{Q \in B_i} P_0 \left( \frac{dP}{dQ} \right)^\alpha < 1,
\]

(3)

\( \Pi(B_i) > 0 \) and \( \sup_{Q \in B_i} P_0(dP/dQ) < \infty \) for all \( P \in V_i \), then any \( V \subset \bigcup_{1 \leq i \leq N} V_i \) receives posterior mass zero asymptotically,

\[
\Pi(V \mid X_1, \ldots, X_n) \xrightarrow{P_0-a.s.} 0.
\]

(4)
The novelty here lies in the fact that the theorem is uncommitted regarding the nature of the $V$, and, more importantly, that we may use any neighbourhoods $B$ of $P_0$ that (i) allow (uniform) control of $P_0(p/q)^\alpha$, and (ii) allow convenient choice of a prior such that $\Pi(B) > 0$. The two requirements on $B$ leave room for trade-offs between being ‘small enough’ to satisfy (i), but ‘large enough’ to enable a choice for $\Pi$ that leads to (ii). The freedom to choose $B$ and $\Pi$ lends the method the desired flexibility regarding Bayesian convergence questions: given $\mathcal{P}$ and $V$, can we find $B$ and $\Pi$ like above?

This paper is organised as follows: in section 2 we prove theorem 1.2 based on minimax test sequences of an unconventional kind that involve prior mass naturally. In section 3 we investigate the relationship between (3) and Kullback-Leibler divergences (see, e.g. lemma 3.1), re-derive Schwartz’ theorem and consider several variations, e.g. posterior consistency in Kullback-Leibler divergence with a prior satisfying (2) and Hellinger consistency with a prior that charges Hellinger balls. In section 4 we consider semi-parametric estimation of domain boundary points for a density on $\mathbb{R}$ (see also Knapik and Kleijn (2013) [14], including a problem in which Schwartz’ theorem does not apply. We conclude with a short discussion.

A note on density supports

Below the focus is on expectations of the form $P_0(p/q)^\alpha$, where $p$ and $q$ are probability densities and $P_0$ is the (true) marginal for the i.i.d. sample. Because the proof of theorem 2.1 is in $P_0$-expectation, an indicator $1_{\{p_0 > 0\}}(x)$ is implicit in all calculations that follow. Because of (1) and because we look at moments of $p/q$, an indicator $1_{\{p > 0\}}(x)$ can also be thought of as a factor in the integrand. Because we require finiteness of $P_0(p/q)$, $q > 0$ is implicit whenever $p_0 > 0$ and $p > 0$, so in expressions of this form an indicator $1_{\{q > 0\}}(x)$ may also be thought of as implicit.

2 Posterior consistency

To establish the basic framework, we assume that the data $X_1, X_2, \ldots$ form an i.i.d. sample from a distribution $P_0$. The model $(\mathcal{P}, \mathcal{B})$ is represented by a Markov kernel $P$ into a sample space $(\mathcal{X}, \mathcal{A})$, i.e. $A \mapsto P(A)$ is a probability measure for every $P$ and $P \mapsto P(A)$ is measurable for every $A \in \mathcal{A}$. Assuming that the model is dominated by a $\sigma$-finite measure with densities $p$, a given prior probability measure $\Pi$ on $(\mathcal{P}, \mathcal{B})$ gives rise to the posterior (1), which is a Markov kernel from $(\mathcal{P}^n, \mathcal{A}^n)$ into $(\mathcal{P}, \mathcal{B})$. In order for the denominator of (1) to be non-zero $P_0^n$-almost-surely, it is required for every $n \geq 1$ that,

$$P_0^n \ll P_n^{\Pi},$$

where $P_n^{\Pi}$ is the prior predictive distribution,

$$P_n^{\Pi}(A) = \int_{\mathcal{P}} P^n(A) d\Pi(P),$$

for all $n \geq 1$ and $A \in \sigma(X_1, \ldots, X_n)$. 

\begin{align*}
\text{A note on density supports} \\
\text{Below the focus is on expectations of the form } P_0(p/q)^\alpha, \text{ where } p \text{ and } q \text{ are probability densities and } P_0 \text{ is the (true) marginal for the i.i.d. sample. Because the proof of theorem 2.1 is in } P_0-\text{expectation, an indicator } 1_{\{p_0 > 0\}}(x) \text{ is implicit in all calculations that follow. Because of (1) and because we look at moments of } p/q, \text{ an indicator } 1_{\{p > 0\}}(x) \text{ can also be thought of as a factor in the integrand. Because we require finiteness of } P_0(p/q), q > 0 \text{ is implicit whenever } p_0 > 0 \text{ and } p > 0, \text{ so in expressions of this form an indicator } 1_{\{q > 0\}}(x) \text{ may also be thought of as implicit.} \\
\text{2 Posterior consistency} \\
\text{To establish the basic framework, we assume that the data } X_1, X_2, \ldots \text{ form an i.i.d. sample from a distribution } P_0. \text{ The model } (\mathcal{P}, \mathcal{B}) \text{ is represented by a Markov kernel } P \text{ into a sample space } (\mathcal{X}, \mathcal{A}), \text{ i.e. } A \mapsto P(A) \text{ is a probability measure for every } P \text{ and } P \mapsto P(A) \text{ is measurable for every } A \in \mathcal{A}. \text{ Assuming that the model is dominated by a } \sigma-\text{finite measure with densities } p, \text{ a given prior probability measure } \Pi \text{ on } (\mathcal{P}, \mathcal{B}) \text{ gives rise to the posterior (1), which is a Markov kernel from } (\mathcal{P}^n, \mathcal{A}^n) \text{ into } (\mathcal{P}, \mathcal{B}). \text{ In order for the denominator of (1) to be non-zero } P_0^n-\text{almost-surely, it is required for every } n \geq 1 \text{ that,} \\
P_0^n \ll P_n^{\Pi}, \\
\text{where } P_n^{\Pi} \text{ is the prior predictive distribution,} \\
P_n^{\Pi}(A) = \int_{\mathcal{P}} P^n(A) d\Pi(P), \\
\text{for all } n \geq 1 \text{ and } A \in \sigma(X_1, \ldots, X_n).
Lemma 2.1. If $\Pi(U) > 0$ for all total-variational/Hellinger neighbourhoods $U$ of $P_0$, (5) is satisfied.

**proof** For any $A \in \sigma(X_1, \ldots, X_n)$ and any model subset $U'$ such that $\Pi(U') > 0$, $P^n_0(A) \leq \int P^n(A) \, d\Pi(P) + \sup_{P \in U'} |P^n(A) - P^n_0(A)|$. Let $A$ be a null-set of $P^n_0$; since $\Pi(U') > 0$, $\int P^n(A) \, d\Pi(P) = 0$. For some $\epsilon > 0$, take $U' = \{P : |P^n(A) - P^n_0(A)| < \epsilon\}$, note that $U'$ contains a total-variational neighbourhood and conclude $P^n_0(A) < \epsilon$ for all $\epsilon > 0$, so (5) holds. (See remark 3.6 (2) in Strasser (1985) [23]: in fact, using the notation of section 3 of [23], requiring $\Pi(U) > 0$ for all $U \in \mathcal{F}_n$ is enough for (5).) \hfill \square

### 2.1 Posterior concentration

To prove consistency given a prior one tries to show that the posterior concentrates all its mass in neighbourhoods of $P_0$ asymptotically, often metric balls centred on $P_0$. The following theorem asserts that, under the condition that specific test-sequences for covers of the complement exist, posterior concentration obtains. The proof is inspired by Schwartz’ proof [22] in that its essence is an application of the minimax theorem (see, e.g., Strasser (1985) [23]), but it takes a short-cut that will pay off later. The essential difference between theorem 2.1 and existing Bayesian limit theorems is that posterior numerator and denominator are dealt with simultaneously rather than separately. As a result the prior $\Pi$ is one of the factors that determines testing power.

**Theorem 2.1.** Let $\Pi$ and $1 \leq N < \infty$ be given. Assume that $P^n_0 \ll P^n_\Pi$ for all $n \geq 1$. Let $V_1, \ldots, V_N$ be a finite collection of measurable model subsets. If there exist constants $D_i > 0$ and test sequences $(\phi_{i,n})$ for all $1 \leq i \leq N$ such that,

$$P^n_0 \phi_{i,n} + \sup_{P \in V_i} P^n_0 \frac{dP^n}{dP^n_\Pi}(1 - \phi_{i,n}) \leq e^{-nD_i},$$

(7)

for large enough $n$, then any $V \subset \bigcup_{1 \leq i \leq N} V_i$ receives posterior mass zero asymptotically,

$$\Pi(V \mid X_1, \ldots, X_n) \underset{P^n_0}{\longrightarrow} 0.$$  

(8)

**proof** For a set $V$ covered by $V_1, \ldots, V_N$ like above, almost-sure convergence per individual $V_i$ implies the assertion. Fix some $1 \leq i \leq N$ and note that,

$$P^n_0 \Pi(V_i \mid X_1, \ldots, X_n) \leq P^n_0 \Pi(V_i \mid X_1, \ldots, X_n)(1 - \phi_{i,n}) + P^n_0 \phi_{i,n},$$

for large enough $n$. By Fubini’s theorem,

$$P^n_0 \Pi(V_i \mid X_1, \ldots, X_n)(1 - \phi_{n}) = P^n_0 \int_{V_i} \frac{dP^n}{dP^n_\Pi}(1 - \phi_{i,n}) \, d\Pi(P) \leq \sup_{P \in V_i} P^n_0 \frac{dP^n}{dP^n_\Pi}(1 - \phi_{i,n}).$$

From (7) we conclude that $P^n_0 \Pi(V_i \mid X_1, \ldots, X_n) \leq e^{-nD_i}$, for large enough $n$. Apply Markov’s inequality to find that,

$$P^n_0 \{ \Pi(V_i \mid X_1, \ldots, X_n) \geq e^{-nD_i} \} \leq e^{-\frac{n}{2}D_i},$$

4
so that the first Borel-Cantelli lemma guarantees,

\[ P_0^\infty \left( \limsup_{n \to \infty} (\Pi(V_1, \ldots, X_n) - e^{-\frac{n}{2} D_1}) > 0 \right) = 0. \]

Replicating this argument for all \( 1 \leq i \leq N \), assertion (3) follows.

Requiring finiteness of covering numbers is somewhat crude: for example, if we use the above to prove consistency relative to some metric, we would let the sets \( V_i \) be open balls of arbitrarily small radius. Thus requiring total boundedness, we are only a completeness statement away from restricting applicability to compact models. There are several ways out: first of all, Ulam’s theorem says that if the model is a Polish space, any prior is Radon; inner regularity of the prior implies that the model can be approximated in prior measure by compact submodels. Since the latter are totally bounded, a proof is conceivable based on an approximating sequence of compact submodels. Alternatively, proofs based on the so-called Le Cam dimension \([16]\) rather than covering numbers of the model could also solve the totally-boundedness issue. These constructions are not the main focus of this paper and are not addressed here.

2.2 Existence and power of test sequences

Le Cam (1973, 1975, 1986) \([15, 19, 20]\) and Birgé (1983, 1984) \([2, 3]\) put forth a versatile approach to testing that combines the minimax theorem with the Hellinger geometry of the model, in particular its Helinger metric entropy numbers. Below, we make a carefully chosen variation on this theme that is technically reminiscent of Kleijn and van der Vaart (2006) \([12]\).

Define \( V^n = \{ P^n : P \in \mathcal{P} \} \) and denote its convex hull by \( \text{co}(V^n) \).

**Lemma 2.2.** Let \( n \geq 1 \) and \( V \in \mathcal{B} \) be given and assume that \( P_0^n (dP^n / dP \Pi_n) < \infty \) for all \( P \in V \). Then there exists a test sequence \( (\phi_n) \) such that,

\[ P_0^n \phi_n + \sup_{P \in V} P_0^n \frac{dP^n}{dP \Pi_n} (1 - \phi_n) \leq \sup_{P^{(n)} \in \text{co}(V^n)} \inf_{0 \leq \alpha \leq 1} P_0^n \left( \frac{dP^{(n)}(\cdot)}{dP \Pi_n} \right)^\alpha, \tag{9} \]

i.e. testing power is bounded in terms of so-called Hellinger transforms.

**proof** We replicate an argument given in Kleijn and van der Vaart (2006) \([12]\), based on the minimax theorem (see, e.g., theorem 45.8 in Strasser (1985) \([23]\)). Use lemma 6.1 of \([12]\) to conclude that there exists a test \( (\phi_n) \) that minimizes the l.h.s. of \((9)\) and,

\[ \sup_{P \in V} \left( P_0^n \phi_n + P_0^n \frac{dP^n}{dP \Pi_n} (1 - \phi_n) \right) \leq \sup_{P^{(n)} \in \text{co}(V^n)} \inf_{\phi} \left( P_0^n \phi + P_0^n \frac{dP^{(n)}(\cdot)}{dP \Pi_n} (1 - \phi) \right). \]

The infimal \( \phi \) on the right-hand side may now be chosen specifically tuned to \( P^{(n)} \), and equals the indicator \( \phi = 1_{\{dP^{(n)}(\cdot)/dP \Pi_n > 1\}} \). For any \( \alpha \in [0, 1] \),

\[ \int 1_{\{dP^{(n)}(\cdot)/dP \Pi_n > 1\}} dP^n_0 + \int \frac{dP^{(n)}}{dP \Pi_n} 1_{\{dP^{(n)}(\cdot)/dP \Pi_n \leq 1\}} dP^n_0 \leq \int \left( \frac{dP^{(n)}(\cdot)}{dP \Pi_n} \right)^\alpha dP_0^n, \]

5
which enables an upper-bound of familiar form (see inequality (6.39) in [12] and remark 2 in section 16.4 of Le Cam (1986) [20]; see also section 3.6 of [20]),

\[
\sup_{P^{(n)} \in \text{co}(V^n)} \left( P_n^{(n)} 1_{\{dP^{(n)}/dP_n^{(n)} > 1\}} + P_0^{n} \frac{dP^{(n)}}{dP_n^{(n)}} 1_{\{dP^{(n)}/dP_n^{(n)} \leq 1\}} \right) \leq \sup_{P^{(n)} \in \text{co}(V^n)} \inf_{0 \leq \alpha \leq 1} P_0^n \left( \frac{dP^{(n)}}{dP_n^{(n)}} \right)^\alpha,
\]

in terms of the Hellinger transform.

With the next definition, we localize the prior in a flexible sense and cast the discussion into a frame that also feature centrally in Wong ans Shen (1995) [21]; where their approximation of \( P_0 \) pertains to a sieve, here it is required that the set \( B \) approximate \( P_0 \) in the same technical sense. Given \( \Pi \) and a measurable \( B \) such that \( \Pi(B) > 0 \), define the local prior predictive distributions \( Q_n^{\Pi} \) by conditioning the prior predictive on \( B \):

\[
Q_n^{\Pi}(A) = \int Q^n(A) \, d\Pi(Q|B),
\]

for all \( n \geq 1 \) and \( A \in \sigma(X_1, \ldots, X_n) \). The following lemma gives rise to more concrete conditions on model and prior.

**Lemma 2.3.** Let \( \Pi \) be given, fix \( n \geq 1 \). Let \( V, B \in \mathcal{B} \) be such that \( \Pi(B) > 0 \) and for all \( P \in V \), \( \sup_{Q \in B} P_0(dP/dQ) < \infty \). Then there exists a test function \( \phi_n \) such that,

\[
P_n^n \phi_n + \sup_{P \in V} P_n^n \frac{dP^n}{dP_n^n}(1 - \phi_n) \leq \inf_{0 \leq \alpha \leq 1} \Pi(\Pi(B)^{-\alpha}) \left\{ \sup_{P \in \text{co}(V)} P_0^n \left( \frac{dP^{(n)}}{dQ} \right)^\alpha \right\}^n\, d\Pi(Q|B).
\]

**proof** Let \( 0 \leq \alpha \leq 1 \) and a \( B \in \mathcal{B} \) like above be given. Note that for all \( n \geq 1 \), \( P_n^{\Pi}(A) \geq \Pi(B) Q_n^{\Pi}(A) \) for all \( A \in \sigma(X_1, \ldots, X_n) \). Combining that with the convexity of \( x \mapsto x^{-\alpha} \) on \((0, \infty)\), we see that,

\[
P_n^n \left( \frac{dP^{(n)}}{dP_n^{(n)}} \right)^\alpha \leq \Pi(B)^{-\alpha} P_0^n \left( \frac{dP^{(n)}}{dQ_n^{(n)}} \right)^\alpha \leq \Pi(B)^{-\alpha} P_0^n \int \left( \frac{dP^{(n)}}{dQ^n} \right)^\alpha \, d\Pi(Q|B).
\]

With the use of Fubini’s theorem and lemma 6.2 in Kleijn and van der Vaart (2006) [12] which says that Hellinger transforms factorize when taken over convex hulls of products, we find:

\[
\sup_{P^{(n)} \in \text{co}(V^n)} \inf_{0 \leq \alpha \leq 1} \Pi(B)^{-\alpha} \int P_0^n \left( \frac{dP^{(n)}}{dQ^n} \right)^\alpha \, d\Pi(Q|B)
\]

\[
\leq \inf_{0 \leq \alpha \leq 1} \Pi(B)^{-\alpha} \int \sup_{P^{(n)} \in \text{co}(V^n)} P_0^n \left( \frac{dP^{(n)}}{dQ^n} \right)^\alpha \, d\Pi(Q|B)
\]

\[
\leq \inf_{0 \leq \alpha \leq 1} \Pi(B)^{-\alpha} \int \left[ \sup_{P \in \text{co}(V)} P_0^n \left( \frac{dP^{(n)}}{dQ} \right)^\alpha \right]^n \, d\Pi(Q|B).
\]

Applying [12] with \( \alpha = 1 \) and using that for all \( P \in V \), \( \sup_{Q \in B} P_0(dP/dQ) < \infty \), we see that also \( P_0^n(dP^n/dP_n^n) < \infty \). By [9], we obtain [11].

Theorem [12] is the conclusion of theorem 2.1 upon substitution of lemmas 2.2 and 2.3.

### 3 Variations on Schwartz’ theorem

In this section we apply theorem 1.2 to re-derive Schwartz’ theorem, sharpen its assertion to consistency in Kullback-Leibler divergence and we consider a model condition that allows priors charging Hellinger rather than Kullback-Leibler neighbourhoods.

---

6
3.1 Schwartz’ theorem and Kullback-Leibler priors

The strategy to prove posterior consistency in a certain topology, or more generally, to prove posterior concentration on a set $V$ now runs as follows: one looks for a finite cover of $V$ by model subsets $V_i$, $(1 \leq i \leq N)$ satisfying the inequalities (3) for subsets $B_i$ that are as large as possible (usually neighbourhoods of $P_0$ in a suitable sense). Subsequently, we try to find (a $\sigma$-algebra $\mathcal{A}$ on $\mathcal{P}$ and) a prior $\Pi$ : $\mathcal{A} \rightarrow [0,1]$ such that $(B_i \in \mathcal{A})$ and $\Pi(B_i) > 0$ for all $1 \leq i \leq N$. In this light, the following lemma offers a great deal of guidance.

**Lemma 3.1.** Let $P_0 \in B \subset \mathcal{P}$ and $V \subset \mathcal{P}$ be given. Then,

$$\inf_{0 \leq \alpha \leq 1} \sup_{Q \in B} \sup_{P \in V} P_0 \left( \frac{dP}{dQ} \right)^\alpha < 1,$$

if and only if,

$$\sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} < \inf_{P \in V} -P_0 \log \frac{dP}{dP_0},$$

that is, if and only if $B$ and $V$ are strictly separated in Kullback-Leibler divergence.

**Proof** Assume that (14) holds. Lemma A.1 says that $\alpha \mapsto P_0 \left( \frac{p}{q} \right)^\alpha$ is smaller than or equal to one at $\alpha = 0$ and continuously differentiable on $[0,1]$. A Taylor expansion in $\alpha = 0$ shows that for all $\alpha \in [0,1]$,

$$\sup_{Q \in B} \sup_{P \in V} P_0 \left( \frac{dP}{dQ} \right)^\alpha \leq 1 + \alpha \sup_{\alpha' \in [0,\alpha]} \sup_{Q \in B} \sup_{P \in V} P_0 \left( \frac{dP}{dQ} \right)^\alpha' \log \frac{dP}{dQ}.$$  

(15)

The function $\alpha \mapsto \sup_{Q \in B} \sup_{P \in V} P_0 \left( \frac{dP}{dQ} \right)^\alpha \log \frac{dP}{dQ}$ is convex (hence continuous on $(0,1)$ and upper-semicontinuous at 0) and, due to (14), strictly negative at $\alpha = 0$. As a consequence, there exists an interval $[0,\alpha_0]$ on which the function in the above display is strictly negative. Based on (15) there exists an $\alpha_0 \in [0,1]$ such that $\sup_{P,Q} P_0 \left( \frac{dP}{dQ} \right)^{\alpha_0} < 1$ and we conclude that (13) holds. Conversely, assume that (14) does not hold. Let $P \in V$, $Q \in B$ and $\alpha \in [0,1]$ be given; by Jensen’s inequality,

$$P_0 \left( \frac{dP}{dQ} \right)^\alpha \geq \exp \left( \alpha P_0 \log \frac{dP}{dQ} \right) = \exp \left( \alpha \left( P_0 \log \frac{dP}{dP_0} - P_0 \log \frac{dQ}{dP_0} \right) \right).$$

Therefore,

$$\sup_{Q \in B} \sup_{P \in V} P_0 \left( \frac{dP}{dQ} \right)^\alpha \geq \exp \left( \alpha \sup_{Q \in B} -P_0 \log \frac{dP}{dP_0} \right) \exp \left( -\alpha \inf_{P \in V} -P_0 \log \frac{dP}{dP_0} \right),$$

which is greater than or equal to one for all $\alpha \in [0,1]$ if (14) fails. \hfill $\Box$

Due to the fact that Kullback-Leibler divergence dominates Hellinger distance, the proof of Schwartz’ theorem is now immediate, at least, for models that have $P_0 \left( \frac{dP}{dQ} \right) < \infty$ for all $P \in \mathcal{P}$ and all $Q$ in some Kullback-Leibler neighbourhood of $P_0$.

**Proof** (of Schwartz’ theorem 1.1)

Let $\epsilon > 0$ be given. If one covers the complement of a Hellinger ball $V$ of radius $2\epsilon$ centred on
P by a finite, convex cover of Hellinger balls \( V_i \) \( (1 \leq i \leq N) \) of radii \( \epsilon \), then for every \( P \in V_i \), 
\[-P_0 \log \frac{dP}{dP_0} \geq H(P, P_0) > \epsilon, \] and (2) is such that (14) is satisfied. Under the assumption that there exists a Kullback-Leibler neighbourhood \( B \) such \( \sup_{Q \in B} P_0(dP/dQ) < \infty \) for all \( P \in \mathcal{P} \) and (2), lemma 2.3 guarantees the existence of tests that satisfy the testing condition of theorem 2.1 Because every total-variational ball contains a Kullback-Leibler neighbourhood, lemma 2.1 asserts that (5) is satisfied and we conclude that the posterior satisfies (4).

It is clear that the above proof does not fully exploit the room that (14) offers because it does not prove posterior consistency in Kullback-Leibler divergence. The following theorem provides such an assertion without requiring more of the prior.

**Theorem 3.1.** Let \( P_0 \) and the model be such that for some Kullback-Leibler neighbourhood \( B \) of \( P_0 \), \( \sup_{Q \in B} P_0(dP/dQ) < \infty \) for all \( P \in \mathcal{P} \). Let \( \Pi \) satisfy (2) for all \( \epsilon > 0 \). For any \( \epsilon > 0 \), assume that \( \{ P \in \mathcal{P} : -P_0 \log(dP/dP_0) \geq \epsilon \} \) is covered by a finite number of model subsets \( V_1, \ldots, V_N \) such that,

\[
\inf_{P \in \text{co}(V_i)} -P_0 \log \frac{dP}{dP_0} > 0, \tag{16}
\]

for all \( 1 \leq i \leq N \). Then for i.i.d. \( P_0 \)-distributed \( X_1, X_2, \ldots, \)

\[
\Pi \left( \left\{ P \in \mathcal{P} : -P_0 \log(dP/dP_0) < \epsilon \right\} \bigg| X_1, \ldots, X_n \right) \xrightarrow{P_0-a.s.} 1.
\]

**proof** For every \( 1 \leq i \leq N \), there exists a constant \( b_i > 0 \) such that for every \( P \in \text{co}(V_i) \),

\[-P_0 \log(dP/dP_0) > b_i.\]

Denoting the Kullback-Leibler radius of \( B \) by \( b > 0 \), we define \( B_i = \{ P \in \mathcal{P} : -P_0 \log(dP/dP_0) < b_i \wedge b \} \) to satisfy (14). Note that, by assumption, \( \Pi(B_i) > 0 \)

and \( \sup_{Q \in B_i} P_0(dP/dQ) < \infty \) for all \( P \in \mathcal{P} \). Every total-variational neighbourhood of \( P_0 \) contains a Kullback-Leibler neighbourhood, so combination of lemma 3.1, lemma 2.3 and theorem 1.2 then proves posterior consistency. □

A straightforward way to satisfy condition (16) depends on uniform bounds for likelihood ratios: let \( \epsilon > 0 \) be given and assume that the model is such that \( V \) contains \( N \) points \( P_1, \ldots, P_N \) such that the sets,

\[ V_i = \{ P \in \mathcal{P} : \|dP/dP_i - 1\|_{\infty} < \frac{1}{2}\epsilon \}, \]

cover \( V \). Then any \( P \in \text{co}(V_i) \) satisfies \( \|dP/dP_i - 1\|_{\infty} < \frac{1}{2}\epsilon \) as well, and hence, \( \log(dP/dP_i) \leq \log(1 + \frac{1}{2}\epsilon) \leq \frac{1}{2}\epsilon. \) As a result,

\[-P_0 \log \frac{dP}{dP_0} \geq \epsilon + P_0 \log \frac{dP}{dP_i} \geq \frac{1}{2}\epsilon, \]

and (10) holds. In such models, any prior \( \Pi \) satisfying (2) leads to a posterior that is consistent in Kullback-Leibler divergence.

### 3.2 Hellinger priors

In quite some generality, lemma 3.1 shows that model subsets are consistently testable if and only if they are uniformly separable from ‘neighbourhoods’ of \( P_0 \) in Kullback-Leibler divergence. This illustrates the fundamental nature of Schwartz’ prior mass requirement and
underscores hopes for useful priors that charge different neighbourhoods of \( P_0 \) (like, \textit{e.g.}, Hellinger balls) in general. But that does not exclude the possibility of gaining freedom in the choice of the prior by strengthening requirements on the model. In this subsection, we give an example that is quite general; in the next section, we consider a specific model in which Schwartz’ prior mass condition cannot be satisfied.

Fix \( \alpha = 1/2 \) and assume that the model is dominated by a \( \sigma \)-finite measure \( \mu \). In fact, given \( P_0 \) and a suitable neighbourhood \( B \), we impose that for all \( Q \in B \) and any \( P \in \mathcal{P}, \)
\[
\frac{p}{q} \in L_2(Q). \tag{17}
\]
(The perceptive reader notes immediately that this requirement is a strong version of a classical condition of (theorem 5 in) Wong and Shen (1995) \cite{24} that guarantees that Kullback-Leibler divergences are dominated by Hellinger distances. Therefore, the conclusion of theorem 3.2 is not as novel as its proof.) By the Cauchy-Schwartz inequality,
\[
P_0\left(\frac{p}{q}\right)^{1/2} = \int \left(\frac{p_0}{q}\right)^{1/2} p_0^{1/2} p^{1/2} d\mu = \int p_0^{1/2} p^{1/2} d\mu + \int \left(1 - \sqrt{\frac{p_0}{q}}\right) \left(\frac{p_0}{q}\right)^{1/2} \left(\frac{p}{q}\right)^{1/2} dQ \leq 1 - \frac{1}{2} H(P_0, P)^2 + H(P_0, Q) \left\| \frac{p_0}{q} \right\|_{2, Q} \left\| \frac{p}{q} \right\|_{2, Q}.
\]
To enable the use of priors that charge Hellinger balls, we strengthen (17) to a uniform bound over the model, making it possible to separate \( B \) from \( V \) in Hellinger distance while maintaining testing power. Combined with theorem 2.1 this leads to the following theorem.

**Theorem 3.2.** Let \( \mathcal{P} \) be totally bounded with respect to the Hellinger metric. Assume also that there exists a constant \( L > 0 \) and a Hellinger ball \( B' \) centred on \( P_0 \) such that for all \( P \in \mathcal{P} \) and \( Q \in B' \),
\[
\left\| \frac{p}{q} \right\|_{2, Q} = \left(\int p^2_q d\mu\right)^{1/2} < L. \tag{18}
\]
Finally assume that for any Hellinger neighbourhood \( B \) of \( P_0 \), \( \Pi(B) > 0 \). Then the posterior is almost-surely Hellinger consistent.

**proof** For given \( \epsilon > 0 \), let \( V \) denote \( \{ P \in \mathcal{P} : H(P, P_0) > 2\epsilon \} \). Since \( \mathcal{P} \) is totally bounded in the Hellinger metric, there exist \( P_1, \ldots, P_N \) such that the model subsets \( V_i = \{ P \in \mathcal{P} : H(P, P_i) < \epsilon \} \) form a cover of \( V \). On the basis of the constant \( L \) of (17), define \( B = \{ P \in \mathcal{P} : H(Q, P_0) < \epsilon^2/(4L^2) \wedge \epsilon' \} \), where \( \epsilon' \) is the Hellinger radius of \( B' \). Since Hellinger balls are convex, we have for all \( 1 \leq i \leq N \),
\[
\sup_{P \in \text{col}(V_i)} \sup_{Q \in B} P_0 \left(\frac{dP}{dQ}\right)^{1/2} \leq 1 - \frac{1}{4} \epsilon^2 \leq \epsilon - \frac{1}{4} \epsilon^2.
\]
By the Cauchy-Schwarz inequality, for every \( P \in V \),
\[
\sup_{Q \in B} P_0(dP/dQ) \leq \sup_{Q \in B} \left\| \frac{p_0}{q} \right\|_{2, Q} \left\| \frac{p}{q} \right\|_{2, Q} < L^2 < \infty.
\]
According to lemma 2.3 and theorem 2.1, the posterior is consistent. \( \square \)

As a side-remark, note that it is possible that (13) is not satisfied without extra conditions on \( Q \). In that case impose that \( B \) is \textit{included} in a Hellinger ball, while satisfying other
conditions as well; the theorem remains valid as long as we also change the prior, i.e. as long as \( \Pi(B) > 0 \) is maintained.

Since the models under consideration have finite Hellinger covering numbers, we note the following construction of so-called net priors \([17, 7, 8, 11]\): let \((\eta_m)\) be any sequence such that \(\eta_m > 0\) for all \(m \geq 1\) and \(\eta_m \downarrow 0\). For fixed \(m \geq 1\), let \(P_1, \ldots, P_{M_m}\) denote a (finite) \(\eta_m\)-net for \(\mathcal{P}\) and define \(\Pi_m\) to be the measure that places mass \(1/M_m\) at every \(P_i\), \((1 \leq m \leq M_m)\). Choose a sequence \((\lambda_m)\) such that \(\lambda_m > 0\) for all \(m \geq 1\) and \(\sum_{m \geq 1} \lambda_m = 1\), to define the net prior \(\Pi = \sum_{m \geq 1} \lambda_m \Pi_m\). By construction, such a net prior satisfies (5) and theorem 3.2 applies if (18) holds.

4 Marginal consistency

Semiparametric estimation problems form an important sub-class of questions in nonparametric statistics. The field presents a well-developed frequentist theory of finite-dimensional parameter estimation in infinite-dimensional models, including notions of optimality for parameters that are smooth functionals of model distributions. By comparison, Bayesian semiparametric methods are still in the early stages of development. In the approach of Bickel and Kleijn (2012) \([1]\), Kleijn and Knapik (2013) \([14]\) and Kleijn (2013) \([15]\), so-called Bernstein-von Mises limits for LAN and LAE model parameters are derived and it turns out that the central requirement is marginal consistency (at optimal rate) of the posterior for the parameter of interest. However, methods to establish marginal consistency leave a lot to be desired: for example, lemma 6.1 in \([1]\) requires likelihood ratios to be bounded appropriately, uniformly over the non-parametric component of the model. In this section an alternative method of demonstrating marginal consistency is formulated, based on the material in preceding sections.

The basic problem (see section 5 regarding rate optimality) can be formulated as follows: let \(\Theta\) be an open subset of \(\mathbb{R}^k\) parametrizing the parameter of interest \(\theta\) and let \(H\) be a (typically infinite-dimensional) metric parameter space for the nuisance parameter \(\eta\). Both \(\Theta\) and \(H\) are endowed with their respective Borel \(\sigma\)-algebras. The model is \(\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}\) where \(\Theta \times H : (\theta, \eta) \mapsto P_{\theta,\eta}\) is a Markov kernel on the sample space \((\mathcal{X}, \mathcal{A})\) describing the distributions of individual points from an i.i.d. sample \(X_1, X_2, \ldots \in \mathcal{X}\). Given a metric \(g : \Theta \times \Theta \rightarrow [0, \infty)\) and a prior measure \(\Pi\) on \(\Theta \times H\) we say that the posterior is marginally consistent for the parameter of interest, if for all \(\epsilon > 0\),

\[
\Pi\left( \{P_{\theta,\eta} \in \mathcal{P} : g(\theta, \theta_0) > \epsilon, \eta \in H\} \mid X_1, \ldots, X_n \right) \xrightarrow{P_{\theta_0,\eta_0} - a.s.} 0, \quad (19)
\]

for all \(\theta_0 \in \Theta\) and all \(\eta_0 \in H\). Assuming that \(\theta\) is identifiable by means of a functional \(\theta(P_{\theta,\eta}) = \theta\) on the model, marginal consistency amounts to consistency with respect to the pseudo-metric \(d : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty)\),

\[
d(P_{\theta,\eta}, P_{\theta',\eta'}) = g(\theta, \theta'),
\]

for all \(\theta, \theta' \in \Theta\) and \(\eta, \eta' \in H\). The following theorem is theorem 1.2 specified to marginal consistency.
Theorem 4.1. Let $\mathcal{P} = \{P_{\theta, \eta} : \theta \in \Theta, \eta \in H\}$ be a model with semiparametric parametrization. Assume that $X_1, X_2, \ldots$ are distributed i.i.d. $P_0$ and that $P_0$ lies in the total-variational support of $\Pi$. Let $\epsilon > 0$ be given, define $V = \{P_{\theta, \eta} \in \mathcal{P} : \|\theta - \theta_0\| > \epsilon\}$ and assume that $V_1, \ldots, V_N$ form a finite cover of $V$. If there exist model subsets $B_1, \ldots, B_N$ such that for every $1 \leq i \leq N$, 

$$\inf_{0 < \alpha < 1} \sup_{P \in \text{co}(V_i)} \sup_{Q \in B_i} P_0\left(\frac{dP}{dQ}\right)^\alpha < 1,$$

$\Pi(B_i) > 0$ and $\sup_{Q \in B_i} P_0(dP/dQ) < \infty$ for all $P \in V_i$, then the posterior satisfies (19).

An important special case concerns semiparametric models for which $H$ is a convex subset of a linear space and the maps $\eta \mapsto P_{\theta, \eta}$ is affine, like mixture models in which the mixing distribution is the nuisance and the example of domain boundary estimation of the next subsection.

4.1 Density support boundaries

Consistent domain boundary estimation (see, e.g. Ibragimov and Has’minskii (1981) [10]), though easy from the perspective of point-estimation, is not a triviality when using Bayesian methods because one is required to specify a nuisance space with a non-parametric prior and show that certain prior mass requirements are met. The Bernstein-Von Mises phenomenon for this type of problem is studied in Kleijn and Knapik (2013) [14]. Below, we prove consistency using theorem 1.2.

Consider the following simple model: for some constant $\sigma > 0$ define the parameter of interest to lie in the space $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ equipped with the Euclidean norm $\| \cdot \|$. Let $H$ be a collection of Lebesgue probability densities $\eta : [0, 1] \to [0, \infty)$ with a continuous, monotone increasing $f : (0, a) \to (0, \infty)$, such that,

$$\inf_{\eta \in H} \min \left\{ \int_0^\epsilon \eta \, d\mu, \int_1^{1-\epsilon} \eta \, d\mu \right\} \geq f(\epsilon), \quad (0 < \epsilon < a).$$

(20)

A condition like (20) is necessarily part of the analysis, because questions concerning domain boundary points make sense only if the distributions under consideration put mass in every small neighbourhood of $\theta_1$ and $\theta_2$. The model $\mathcal{P} = \{P_{\theta, \eta} : \theta \in \Theta, \eta \in H\}$ is defined in terms of Lebesgue densities of the following semiparametric form,

$$p_{\theta, \eta}(x) = \frac{1}{\theta_2 - \theta_1} \eta\left(\frac{x - \theta_1}{\theta_2 - \theta_1}\right) 1_{\{\theta_1 \leq x \leq \theta_2\}},$$

for some $(\theta_1, \theta_2) \in \Theta$ and $\eta \in H$. The semiparametric problem concerns the estimation of $\theta$; more specifically we are interested in formulation of (conditions for) priors that lead to marginal consistency for the parameter of interest. Let $\| \cdot \|_{s,Q}$ denote the $L_s(Q)$-norm (for $s \geq 1$).

Theorem 4.2. Let $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ (for some $\sigma > 0$) and the space $H$ with associated lower bound $f$ be given. Assume that there exists an $s \geq 1$ such that the sets
for all \( \alpha \) cases that for all \( P \in \mathcal{P} \) and \( Q \in B \),
\[
\left\| \frac{dP}{dQ} \right\|_{r,Q} \leq K,
\]
where \( 1/r + 1/s = 1 \). If \( X_1, X_2, \ldots \) form an i.i.d. \( P_0 \) sample, where \( P_0 = P_{\theta_0, \eta_0} \in \mathcal{P} \), then,
\[
\Pi(\|\theta - \theta_0\| < \epsilon \mid X_1, \ldots, X_n) \xrightarrow{P-a.s.} 1,
\]
for all \( \epsilon > 0 \).

**proof** Let \( \epsilon > 0 \) be given and consider the (equivalent) metric \( g : \Theta \times \Theta \to [0, \infty) \) by \( g(\theta, \theta') = \max\{|\theta_1 - \theta'_1|, |\theta_2 - \theta'_2|\} \). Define \( V = \{P_{\theta, \eta} \in \mathcal{P} : g(\theta, \theta') > \epsilon\} \). Concentrate on the cases \( \alpha = 0 \) and \( \alpha = 1 \); pick \( 0 < \delta < f(\epsilon/\sigma)/(2K) \) and define \( B \) as above. Lemma A.1 says that for all \( P \in V \) and \( Q \in B \),
\[
P_0 \left( \frac{dP}{dQ} \right) \bigg|_{\alpha=0} = P_0(p > 0),
\]
\[
P_0 \left( \frac{dP}{dQ} \right) \bigg|_{\alpha=1} = \int \frac{dP_0}{dQ} 1_{\{p_0>0, p>0, q>0\}} dP \leq P(p_0 > 0) + \int \left| \frac{dP_0}{dQ} - 1 \right| dP
\]
\[
\leq P(p_0 > 0) + \left| \frac{dP_0}{dQ} - 1 \right|_{s,Q} \left\| \frac{dP}{dQ} \right\|_{r,Q} < P(p_0 > 0) + \frac{1}{2} f\left(\frac{\epsilon}{\sigma}\right),
\]
by Hölder’s inequality. Note that every total-variational neighbourhood of \( P_0 \) contains a model subset of the form \( B \) and, by assumption, \( \Pi(B) > 0 \), so that lemma 2.1 guarantees that \( P_0^n \ll P^n \) for all \( n \geq 1 \). For all \( P \in V \), \( \sup_{Q \in B} P_0(dP/dQ) \leq 1 + (1/2)f(\epsilon/\sigma) < \infty \) and for all \( Q \in B \), we have,
\[
\inf_{0 \leq \alpha \leq 1} P_0 \left( \frac{dP}{dQ} \right) \bigg|_{\alpha} \leq \min\{P_0(p > 0), P(p_0 > 0)\} + \frac{1}{2} f\left(\frac{\epsilon}{\sigma}\right).
\]
Identify \( P_0 \) and \( P \) with parameters \((\theta_0, \eta_0)\) and \((\theta, \eta)\), writing \( P_0 = P_{(\theta_0,1,\theta_0,2),\eta_0} \) and \( P = P_{(\theta,1,\theta,2),\eta} \). By definition of \( V \), the support intervals for \( p \) and \( p_0 \) are disjoint by an interval of length greater than or equal to \( \epsilon \). Cover \( V \) by four sets, \( V_{+,1} = \{P_{\theta, \eta} : \theta_1 \geq \theta_{0,1} + \epsilon, \eta, \eta \in H\} \), \( V_{-,1} = \{P_{\theta, \eta} : \theta_1 \leq \theta_{0,1} - \epsilon, \eta, \eta \in H\} \), \( V_{+,2} = \{P_{\theta, \eta} : \theta_2 \geq \theta_{0,2} + \epsilon, \eta, \eta \in H\} \) and \( V_{-,2} = \{P_{\theta, \eta} : \theta_2 \leq \theta_{0,2} - \epsilon, \eta, \eta \in H\} \). For \( P \in \text{co}(V_{+,1}) \), we have,
\[
P_0(p = 0) \geq \int_{\theta_{0,1}}^{\theta_{0,1} + \epsilon} p_0(x) dx = \int_{\theta_{0,1}}^{\theta_{0,1} + \epsilon} \frac{1}{\theta_{0,2} - \theta_{0,1}} \eta_0\left(\frac{x - \theta_{0,1}}{\theta_{0,2} - \theta_{0,1}}\right) dx
\]
\[
= \int_0^{\epsilon/(\theta_{0,2} - \theta_{0,1})} \eta_0(z) dz \geq \int_0^{\epsilon/\sigma} \eta_0(z) dz \geq f\left(\frac{\epsilon}{\sigma}\right),
\]
using (20). Similarly for \( P \in \text{co}(V_{-,1}) \), with some \( I \geq 1 \) write \( P = \sum_{i=1}^I \lambda_i P_i \) where \( \sum_{i=1}^I \lambda_i = 1 \) and \( \lambda_i \geq 0 \), \( P_i = P_{\theta_i, \eta_i} \) for \( \theta_i = (\theta_{i,1}, \theta_{i,2}) \) with \( \theta_{i,1} \leq \theta_{0,1} - \epsilon \) and \( \eta_i \in H_i \) for all \( 1 \leq i \leq I \).
Note that,
\[
P(p_0 = 0) = \sum_{i=1}^{I} \lambda_i P_i(p_0 = 0) \geq \sum_{i=1}^{I} \lambda_i \int_{\theta_i,1}^{\theta_{i,1} + \epsilon} p_i(x) \, dx
\]
\[
= \sum_{i=1}^{I} \lambda_i \int_{\theta_i,1}^{\theta_{i,1} + \epsilon} \frac{1}{\theta_{i,2} - \theta_{i,1}} \eta_i \left( \frac{x - \theta_{i,1}}{\theta_{i,2} - \theta_{i,1}} \right) \, dx
\]
\[
= \sum_{i=1}^{I} \lambda_i \int_{0}^{\epsilon/(\theta_{i,2} - \theta_{i,1})} \eta_i(z) \, dz \geq \sum_{i=1}^{I} \lambda_i \int_{0}^{\epsilon} \frac{z}{\sigma} \eta_i(z) \, dz \geq f \left( \frac{\epsilon}{\sigma} \right),
\]
using (20). Analogously we obtain bounds for \( P \in \operatorname{co}(V_{+2}) \) and \( P \in \operatorname{co}(V_{-2}) \), giving rise to the inequalities
\[
\sup_{P \in \operatorname{co}(V)} \min \{ P(p > 0), P(p_0 > 0) \} \leq 1 - f \left( \frac{\epsilon}{\sigma} \right),
\]
for \( V \) equal to \( V_{+1} \), \( V_{-1} \), \( V_{+2} \) and \( V_{-2} \). Combination of lemma 2.3 and theorem 1.2 now shows that,
\[
\Pi \left( \left\{ g(\theta, \theta_0) < \epsilon \mid X_1, \ldots, X_n \right\} \right) \xrightarrow{P_{\theta_0=\sigma a.s.}} 1.
\]
The topology associated with the metric \( g \) on \( \Theta \) is equivalent to the restriction to \( \Theta \) of the usual norm topology on \( \mathbb{R}^2 \), so that consistency with respect to the pseudo-metric \( g \) is equivalent to (21).

\begin{example}
To apply theorem 1.2, let \( P_0 = P_{\theta_0, \sigma} \) be a distribution on \( \mathbb{R} \) with Lebesgue density \( p_0 : \mathbb{R} \mapsto [0, \infty) \) supported on an interval \([\theta_{0,1}, \theta_{0,2}]\) of a width smaller than or equal to a (known) constant \( \sigma > 0 \). Furthermore, let \( g : [0, 1] \mapsto [0, \infty) \) be a known probability density such that \( g(x) > 0 \) for all \( x \in (0, 1) \). For a known constant \( M > 0 \) consider the (convex) subset \( C_M \) of \( C[0, 1] \) of all continuous \( h : [0, 1] \mapsto [0, \infty) \) such that \( e^{-M} \leq h \leq e^M \). Define \( H \) to contain all \( \eta : [0, 1] \mapsto [0, \infty) \) that are Esscher transforms of the form,
\[
\eta(x) = \frac{g(x) h(x)}{\int_0^1 g(y) h(y) \, dy},
\]
for some \( h \in C_M \) and all \( x \in [0, 1] \). To define a prior on \( H \), let \( U \sim U[-M, M] \) be uniformly distributed on \([-M, M]\) and let \( W = \{ W(x) : x \in [0, 1] \} \) be Brownian motion on \([0, 1]\), independent of \( U \). Note that it is possible to condition the process \( Z(x) = U + W(x) \) on \(-M \leq Z(x) \leq M \) for all \( x \in [0, 1] \) (or reflect \( Z \) in \( z = -M \) and \( z = M \)). Define the distribution of \( \eta \) under the prior \( \Pi_H \) by taking \( h = e^Z \). On \( \Theta \) let \( \Pi_\Theta \) denote a prior with a Lebesgue density that is continuous and strictly positive on \( \Theta \). One verifies easily that the model satisfies (20) with \( f \) defined by,
\[
f(\epsilon) = e^{-2M} \min \left\{ \int_0^\epsilon g(x) \, dx, \int_{1-\epsilon}^1 g(x) \, dx \right\},
\]
for all \( \epsilon > 0 \) small enough. The prior mass requirement is satisfied because the distribution of the process \( Z \) has full support relative to the uniform norm in the collection of all continuous functions on \([0, 1]\) bounded by \( M \).
\end{example}
To demonstrate that the approach followed in this paper applies in cases where Schwartz’ theorem fails, we look at the following simplified version of the domain boundary problem sketched above.

Example 4.2. Instead of leaving both boundary points free, assume that the width of the domain of \( p_0 \) is known (say, equal to one). The model consists of densities supported on intervals of the form \([\theta, \theta + 1]\) for some \( \theta \in \mathbb{R} \),

\[
p_{\theta, \eta}(x) = \eta(x - \theta) 1_{[\theta, \theta+1]}(x).
\]

Consider \( H \) with a prior \( \Pi_H \) and a prior \( \Pi_\Theta \) on \( \Theta = \mathbb{R} \) with a Lebesgue density that is continuous and strictly positive on all of \( \mathbb{R} \).

Note that if \( \theta \neq \theta' \) the Kullback-Leibler divergence of \( P_{\theta, \eta} \) with respect to \( P_{\theta', \eta'} \) is infinite (for all \( \eta, \eta' \in H \)) because the supports of the corresponding densities suffer from a mismatch. Hence, for given \( P_0 = P_{\theta_0, \eta_0} \in \mathcal{P} \), Kullback-Leibler neighbourhoods do not have any extent in the \( \theta \)-direction:

\[
\{ P_{\theta, \eta} \in \mathcal{P} : -P_0 \log \frac{dP_{\theta, \eta}}{dP_0} < \epsilon \} \subset \{ P_{\theta_0, \eta} : \eta \in H \}.
\]

In order for Schwartz’ theorem to apply, a prior satisfying (2) is required: in this example, that requirement implies that,

\[
\Pi(P_{\theta_0, \eta} : \eta \in H) > 0.
\]

Conclude that for the estimation of an unknown \( \theta_0 \in \mathbb{R} \), Schwartz’ theorem does not work.

The construction of example 4.1 continues to work however. In fact, in the present, fixed-width simplification the situation is more transparent: if we write \( P_0 = P_{\theta_0, \eta_0} \) and \( V = V_+ \cup V_- \) with \( V_+ = \{ P_{\theta, \eta} : \theta > \theta_0 + \epsilon, \eta \in H \} \) and \( V_- = \{ P_{\theta, \eta} : \theta < \theta_0 - \epsilon, \eta \in H \} \) for some \( \epsilon > 0 \), then we choose \( B_+ = \{ P_{\theta, \eta} : \theta_0 + \frac{1}{2} \epsilon < \theta < \theta_0 + \epsilon, \eta \in H \} \) and \( B_- = \{ P_{\theta, \eta} : \theta_0 - \epsilon < \theta < \theta_0 - \frac{1}{2} \epsilon, \eta \in H \} \), so that \( \Pi(B_{\pm}) > 0 \). Consider only \( \alpha = 0 \) and notice that the mismatch in extent of domains implies that,

\[
P_0(p > 0) \leq 1 - f(\epsilon) < 1,
\]

for all \( P \in \text{co}(V_\pm) \), based on (20). If \( H \) is chosen such that for all \( P \in V_\pm \), \( \sup_{Q \in B_{\pm}} P_0(p/q) < \infty \), then (21) follows (regardless of the prior on \( H \)). Larger spaces \( H \) can be considered if the sets \( B_{\pm} \) are restricted appropriately while maintaining \( \Pi(B_{\pm}) > 0 \).

5 Discussion

Schwartz’ theorem is the basic cornerstone of (the frequentist perspective on) Bayesian non-parametric statistics and it has been in place for no less than fifty years: it is beautiful and powerful, in that it applies to a very wide class of models. However, its generality with respect to the model implies that it is rather stringent with respect to the prior.

In this paper, an attempt has been made to demonstrate that there is more flexibility in the criteria for the prior, if one is willing to accept more strict model conditions. The proposed
method implies Schwartz’ theorem and gives rise to a consistency theorem in Kullback-Leibler divergence, as well as a Hellinger consistency theorem for Hellinger priors. Additionally, the approach has been applied to a semiparametric domain boundary model in which Schwartz’ prior mass condition fails. It appears safe to conclude that the approach we propose is more versatile than Schwartz’ theorem: its flexibility allows that we ‘tailor’ the prior to the problem using model properties, rather than being forced to deal with Kullback-Leibler neighbourhoods.

The technical background of the approach can be traced back to Wong and Shen (1995) [24] and Kleijn and van der Vaart (2006) [12]. The former makes extensive use of Hellinger transforms to control sieve approximations. Regarding the latter, note that there is a form of misspecification [11, 12, 13] that applies: $P_0$ is not equal to the prior predictive distribution. So the Bayesian methodology implies a marginal distribution for the data that does not coincide with $P_0$. It appears that the asymptotic manifestation of this mismatch depends on the local prior predictive distributions $Q_\Pi$: if those match $P_n$ closely enough (see lemma 2.3), testability is maintained and consistency obtains.

To conclude, we note that proposed methods do not just concern the limit $n \to \infty$, but hold at finite values of $n \geq 1$ (see, e.g. lemma 2.3). Therefore, the alternative hypothesis $V$ can be made $n$-dependent without changing the basic building blocks of the proofs. As such, not much needs to be adapted to these results to extend also to rates of posterior convergence. Ghosal, Ghosh and van der Vaart (2000) [8] build on Schwartz’ theorem to arrive at a theorem for posterior convergence at rate $(\varepsilon_2 n)$, which, besides a more stringent test-condition requires sufficiency of prior mass in specialized Kullback-Leibler neighbourhoods:

$$\Pi\left( P \in \mathcal{P} : -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2, P_0 \left( \log \frac{dP}{dP_0} \right)^2 < \epsilon_n^2 \right) \geq e^{-n\epsilon_2^2}.$$  

Usually in proofs that a prior satisfies the above display, one appeals to theorem 5 of Wong and Shen (1995) [24] to show that these specialized Kullback-Leibler neighbourhoods contain Hellinger balls. The material presented here suggests that we apply the minimax theorem for every $n \geq 1$, with $V$, $B$ and $\alpha$ that display $n$-dependence: for convex ($V_n$) and ($B_n$) such that $\Pi(B_n)$ is greater than some $n$-dependent lower bound, there exists a test sequence ($\phi_n$) such that, for every sequence $(\alpha_n) \subset [0,1]$:

$$P_0^n \phi_n + \sup_{P \in V_n} P_0^n \frac{dP_n}{dP_0^n} (1 - \phi_n) \leq \Pi(B_n)^{-\alpha_n} \int \left( \sup_{P \in V_n} P_0 \left( \frac{dP}{dQ} \right)^\alpha \right)^n d\Pi(Q|B_n).$$

The freedom to choose $(\alpha_n)$ makes it possible to balance the lower bound on prior masses $\Pi(B_n)$ versus upper bounds on Hellinger transforms. This possibility is not part of the analysis of [8] and one wonders if this extra freedom translates into concrete differences and, perhaps, more flexibility in the criteria for prior selection. For example, it would be interesting to see whether the resulting rates-of-convergence theorem assigns the correct rate of convergence in smoothness classes relative to the uniform norm (see Giné and Nickl (2012) [9], Rousseau (2013) [21]).
A Some properties of Hellinger transforms

Given two finite measures $\mu$ and $\nu$, the Hellinger transform is defined as follows:

$$\rho_\alpha(\mu, \nu) = \int d\mu^\alpha d\nu^{1-\alpha}.$$  

For $P$ and $Q$ such that $P_0(dP/dQ) < \infty$ define $d\nu_{P,Q} = (dP_0/dQ)dP$ and note that,

$$P_0\left(\frac{dP}{dQ}\right)^\alpha = \rho_\alpha(\nu_{P,Q}, P_0) = \rho_{1-\alpha}(P_0, \nu_{P,Q}).$$  

Properties of the Hellinger transform that are used in the main text are listed in the following lemma, which extends lemma 6.3 in Kleijn and van der Vaart (2006).

**Lemma A.1.** For a probability measure $P$ and a finite measure $\nu$ (with densities $p$ and $r$ respectively), the function $\rho : [0, 1] \to \mathbb{R} : \alpha \mapsto \rho_\alpha(\nu, P)$ is convex on $[0, 1]$ with:

$$\rho_\alpha(\nu, P) \to P(r > 0), \quad \text{as} \quad \alpha \downarrow 0, \quad \rho_\alpha(\nu, P) \to \nu(p > 0), \quad \text{as} \quad \alpha \uparrow 1.$$

Furthermore $\alpha \mapsto \rho_\alpha(\nu, P)$ is continuously differentiable on $[0, 1]$ with derivative,

$$\frac{d\rho_\alpha(\nu, P)}{d\alpha} = P 1_{r>0} \left(\frac{p}{r}\right)^\alpha \log(r/p),$$

(which may be equal to $-\infty$).

**proof** The function $\alpha \mapsto e^{\alpha y}$ is convex on $(0, 1)$ for all $y \in [-\infty, \infty)$, implying the convexity of $\alpha \mapsto \rho_\alpha(\nu, P) = P(r/p)^\alpha$ on $(0, 1)$. The function $\alpha \mapsto y^\alpha = e^{\alpha \log y}$ is continuous on $[0, 1]$ for any $y > 0$, decreasing for $y < 1$, increasing for $y > 1$ and constant for $y = 1$. By monotone convergence, as $\alpha \downarrow 0$,

$$\nu\left(\frac{p}{r}\right)^\alpha 1_{[0<p<r]} \uparrow \nu\left(\frac{p}{r}\right)^0 1_{[0<p<r]} = \nu(0 < p < r).$$

By the dominated convergence theorem, with dominating function $(p/r)^{1/2}1_{[p \geq r]}$ (which lies above $(p/r)^\alpha 1_{[p \geq r]}$ for $\alpha \leq 1/2$), we have,

$$\nu\left(\frac{p}{r}\right)^\alpha 1_{[p \geq r]} \to \nu\left(\frac{p}{r}\right)^0 1_{[p \geq r]} = \nu(p \geq r),$$

as $\alpha \downarrow 0$. Combining the two preceding displays, we see that $\rho_\alpha(\nu, P) = P(p/r)^\alpha \to P(r > 0)$ as $\alpha \downarrow 0$. Upon substitution of $\alpha$ by $1 - \alpha$, one finds that $\rho_\alpha(\nu, P) \to \nu(p > 0)$ as $\alpha \uparrow 1$.

Let $\alpha_0 \in [0, 1]$ be given. By the convexity of $\alpha \mapsto e^{\alpha y}$ for all $y \in \mathbb{R}$, the map $\alpha \mapsto f_\alpha(y) = (e^{\alpha y} - e^{\alpha_0 y})/(\alpha - \alpha_0)$ decreases to $y e^{\alpha_0 y}$ as $\alpha \downarrow \alpha_0$, and it increases to $y e^{\alpha_0 y}$ as $\alpha \uparrow \alpha_0$. First consider the case that $\alpha \geq \alpha_0$: for $y \leq 0$ we have $f_\alpha(y) \leq 0$, while for $y \geq 0$,

$$f_\alpha(y) \leq \sup_{\alpha_0 < \alpha' \leq \alpha} ye^{\alpha'y} \leq ye^{\alpha y} \leq \frac{1}{e}e^{(\alpha + \epsilon)y},$$

so that $f_\alpha(y) \leq 0 \lor e^{-1}e^{(\alpha + \epsilon)y}1_{y \geq 0}$. Consequently, we have:

$$\left(\frac{r}{p}\right)^{\alpha_0} \frac{e^{(\alpha - \alpha_0)\log(r/p)} - 1}{\alpha - \alpha_0} \downarrow \left(\frac{r}{p}\right)^{\alpha_0} \log\left(\frac{r}{p}\right), \quad \text{as} \quad \alpha \downarrow 0,$$
and is bounded above by \(0 \lor \epsilon^{-1}(r/p)^{\alpha_0 + 2\epsilon}1_{r \geq p}\) for small \(\alpha > 0\), which is \(P\)-integrable for small enough \(\epsilon\). We conclude that,

\[
\frac{1}{\alpha - \alpha_0}(\rho_\alpha(v, P) - \rho_{\alpha_0}(v, P)) \downarrow P1_{r > 0}\left(\frac{r}{p}\right)^{\alpha_0} \log\left(\frac{r}{p}\right), \quad \text{as } \alpha \downarrow 0,
\]

by monotone convergence. For \(\alpha < \alpha_0\) a similar argument can be given. Convexity of \(\alpha \mapsto P1_{r > 0}(r/p)^{\alpha} \log(r/p)\) implies continuity of the derivative.

\[
\square
\]

Acknowledgements

The author would like to thank Y. D. Kim and S. Petrone for valuable discussions and suggestions, and the Statistics Department of Seoul National University, South Korea, and the Dipartimento di Scienze delle Decisioni, Universita Bocconi, Milano for their kind hospitality.

References

[1] P. Bickel and B. Kleijn, The semiparametric Bernstein-Von Mises theorem, Ann. Statist. 40 (2012), 206–237.
[2] L. Birgé, Approximation dans les espaces métriques et théorie de l’estimation, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65 (1983), 181–238.
[3] L. Birgé, Sur un théorème de minimax et son application aux tests, Probability and Mathematical Statistics 3 (1984), 259–282.
[4] J. Doob, Applications of the theory of martingales, Le calcul des Probabilités et ses Applications, Colloques Internationaux du CNRS, Paris (1948), 22–28.
[5] D. Freedman, On the asymptotic behavior of Bayes estimates in the discrete case I, Ann. Math. Statist. 34 (1963), 1386–1403.
[6] D. Freedman, On the asymptotic behavior of Bayes estimates in the discrete case II, Ann. Math. Statist. 36 (1965), 454–456.
[7] S. Ghosal, J. Ghosh and R. Ramamoorthi, Non-informative priors via sieves and packing numbers, in “Advances in Statistical Decision Theory and Applications” (S. Panchapakesan and N. Balakrishnan, eds.), pp. 119–132, Birkhauser, Boston, 1997.
[8] S. Ghosal, J. Ghosh and A. van der Vaart, Convergence rates of posterior distributions, Ann. Statist. 28 (2000), 500–531.
[9] E. Giné and R. Nickl, Rates of contraction for posterior distributions in \(L_r\)-metrics, \(1 \leq r \leq \infty\), Ann. Statist. 39 (2011), 2883–2911.
[10] I. Ibragimov and R. Has’minskii, Statistical estimation: asymptotic theory, Springer, New York (1981).
[11] B. Kleijn, Bayesian asymptotics under misspecification, PhD. Thesis, Free University Amsterdam (2003).
[12] B. Kleijn and A. van der Vaart, Misspecification in Infinite-Dimensional Bayesian Statistics. Ann. Statist. 34 (2006), 837–877.
[13] B. Kleijn and A. van der Vaart, The Bernstein-Von-Mises theorem under misspecification. Electron. J. Statist. 6 (2012), 354–381.
[14] B. Kleijn and B. Knäpik, Semiparametric posterior limits under local asymptotic exponentiality, (submitted for publication).
[15] B. Kleijn, Semiparametric posterior limits, (submitted for publication).
[16] L. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 1 (1972), 38–53.
[17] L. Le Cam, On posterior rates of convergence, (unpublished).
[18] L. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 22 (1973), 38–55.
[19] L. Le Cam, *On local and global properties in the theory of asymptotic normality of experiments*, Stochastic Process. and Related Topics 1 (1975), 13–53. (ed. M.L. Puri), Academic Press, New York.

[20] L. Le Cam, *Asymptotic methods in statistical decision theory*, Springer, New York (1986).

[21] J. Rousseau, *Some recent advances in the asymptotic properties of Bayesian nonparametric approaches*, Invited talk at BNP9 conference, Amsterdam (2013).

[22] L. Schwartz, *On Bayes procedures*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 4 (1965), 10–26.

[23] H. Strasser, *Mathematical theory of statistics*, de Gruyter, Berlin, 1985.

[24] W.H. Wong and X. Shen, *Probability inequalities for likelihood ratios and convergence rates of sieve MLEs*, Ann. Statist. 23 (1995), 339–362.