Electronic Journal of Statistics
Vol. 13 (2019) 4709–4742
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1633

Criteria for posterior consistency and convergence at a rate

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

and

Y. Y. Zhao[†]

formerly, Wenlan School of Business
Zhongnan University of Economics and Law
People's Republic of China

Abstract: Frequentist conditions for asymptotic consistency of Bayesian procedures with *i.i.d.* data focus on lower bounds for prior mass in Kullback-Leibler neighbourhoods of the data distribution. The goal of this paper is to investigate the flexibility in these criteria. We derive a versatile new posterior consistency theorem, which is used to consider Kullback-Leibler consistency and indicate when it is sufficient to have a prior that charges metric balls instead of KL-neighbourhoods. We generalize our proposal to sieved models with Barron's negligible prior mass condition and to separable models with variations on Walker's condition. Results are also applied in semi-parametric consistency: support boundary estimation is considered explicitly and consistency is proved in a model for which Kullback-Leibler priors do not exist. As a further demonstration of applicability, we consider metric consistency *at a rate*: under a mild integrability condition, the second-order Ghosal-Ghosh-van der Vaart prior mass condition can be relaxed to a lower bound for ordinary KL-neighbourhoods. The posterior rate is derived in a parametric model for heavy-tailed distributions in which the Ghosal-Ghosh-van der Vaart condition cannot be satisfied by any prior.

MSC 2010 subject classifications: Primary 62G07, 62G20.
Keywords and phrases: Asymptotic consistency, posterior consistency, Bayesian consistency, marginal consistency, posterior rate of convergence.

Received October 2019.

1. Introduction and main result

Aside from computational issues, the most restrictive aspects of non-parametric Bayesian methods result from limited availability of priors. In general, distri-

[†] Dedicated to the memory of Yanyun Zhao (1983–2018). Your dedication and enthusiasm will be sorely missed. STTL.

butions on infinite dimensional spaces are relatively hard to define and control technically, so unnecessary elimination of candidate priors is highly undesirable. Specifying to frequentist asymptotic aspects, the *conditions* that Bayesian limit theorems pose on priors play a crucial role and have received much attention, as reviewed in several excellent overview texts [14, 17, 16] over the years. It is the goal of this paper to extend the range of criteria on the prior for posterior consistency and convergence at a rate [15], showing asymptotic suitability for a wider range of priors. From the outset, we accept that this may go at the expense of additional model conditions.

1.1. Introduction

Doob [10] studied posterior limits with the help of his martingale convergence theory and gave the first general posterior consistency theorem for *i.i.d.* data. Notwithstanding the generality of its Bayesian interpretation, Doob's theorem is not quite satisfactory to the frequentist interested in non-parametric statistics, in that Doob's prior null set of possible inconsistency can be very large, as was stressed by Schwartz [33] and amplified by Freedman [11, 12, 9]. To frequentists Freedman's counterexamples discredited Bayesian methods for non-parametric statistics greatly. The resulting under-appreciation was hard to justify, given Schwartz's 1965 posterior consistency theorem [34] for *i.i.d.* data: posteriors are consistent in the frequentist sense, if consistent uniform tests exist and the prior Π is a so-called *Kullback-Leibler prior*: for all $\delta > 0$,

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \delta \Big) > 0. \tag{1.1}$$

Although there are alternatives [24], for example those based on *Le Cam's inequality* [25, 28], condition (1.1) has become the standard. Schwartz's theorem does not cover all examples, however.

Example 1.1. Consider *i.i.d.* $X_1, X_2, \ldots$ from a distribution P_0 with Lebesgue density $p_0 : \mathbb{R} \to \mathbb{R}$ that is supported on an interval of known width (say, 1) but unknown location. The model is parametrized in terms of a density η supported on $[0,1]$ with $\eta(x) > 0$ for all $x \in [0,1]$ and a location $\theta \in \mathbb{R}$:

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

Note that if θ does not equal θ',

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

for all η, η'. Therefore Kullback-Leibler neighbourhoods do not have any extent in the θ-direction and no prior can be a Kullback-Leibler prior in this model. Even in a simple model of uniform distributions $\{U[0,\theta] : \theta \in [0,1]\}$, any prior that is KL must have a non-zero point-mass at $\theta = 1$. (See example 7.2 for more.)

Schwartz's theorem for posterior consistency in metric parameter spaces requires that the model is of finite entropy with respect to the Hellinger metric. That condition is rather restrictive and can be mitigated in several ways. The (testing) problem was noted by Le Cam (see the *Le Cam-dimension* of the model [26]) and his solution can be applied in Schwartz's setting too. A more Bayesian solution partitions the model sequentially into subsets of bounded Hellinger metric entropy and subsets of negligible prior mass (see, for example, [2] and section 4.4.2 of [17]). Walker has proposed a method that does not depend on finite covers but adds a summability condition to condition (1.1) [39]. (For more, see subsection 4.2.)

In metric parameter spaces, consistency can be strengthened to posterior convergence *at a rate*: extensions of Schwartz's theorem [15, 35] apply Barron's sieve idea and a minimax test construction [4, 5], in combination with a second-order Kullback-Leibler condition on the prior:

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2,\, P_0\Big(\log \frac{dP}{dP_0}\Big)^2 < \epsilon_n^2 \Big) \geq e^{-Cn\epsilon_n^2}, \tag{1.2}$$

for some $C > 0$ and large enough n, to conclude that the posterior concentrates its mass in Hellinger balls around P_0 of radii $\epsilon_n \to 0$. But again, not all examples are covered: below, heavy-tailed distributions are found for which integrability of squared log-density ratios is violated.

Example 1.2. Consider an *i.i.d.* sample of integers $X_1, X_2, \ldots$ from a distribution P_a, $(a \geq 1)$, defined by,

$$p_a(k) = P_a(X = k) = \frac{1}{Z_a} \frac{1}{k^a (\log k)^3}, \tag{1.3}$$

for all $k \geq 2$, with $Z_a = \sum_{k \geq 2} k^{-a} (\log k)^{-3} < \infty$. As it turns out, for $a = 1$, $b > 1$,

$$-P_a \log \frac{p_b}{p_a} < \infty, \quad P_a\Big(\log \frac{p_b}{p_a}\Big)^2 = \infty.$$

Therefore, Schwartz's condition (1.1) for the prior of a can be satisfied but there exists no prior such that (1.2) is satisfied for all P_0 in the model. (See example 7.3 for more.) In fact, if we change the third power of the log-factor in the denominator of (1.3) to a square, Schwartz's KL-priors also do not exist.

Schwartz's theorem and its rate-specific version have become the standard frequentist 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.* (1.1) and (1.2) are virtually the *only* criteria frequentists apply to priors in non-parametric asymptotic analyses (notable exceptions are made in the examples of [7, 8, 18] as well as [16]; see, however, lemma 3.1 below). Since these lower bounds on prior weights of Kullback-Leibler-neighbourhoods are sufficient conditions applicable for *i.i.d.* data, it is not clear if other criteria for the prior can be formulated. The goal of this paper is to increase flexibility in criteria for prior choice, by formulating a greater variety

of suitability conditions for priors. The goal is *not* to generalize conditions of Schwartz's theorem or to sharpen its assertion; rather we want to show that stringency with regard to the prior can be relaxed at the expense of stringency with regard to conditions on the model.

1.2. Main result

The main result is summarized in the next theorem: we have in mind a fixed model subset V (*e.g.* the complement of a fixed neighbourhood of P_0) for which we want to demonstrate asymptotically vanishing posterior mass. Following the ideas of [34, 26, 4, 5] the set V is covered by a finite collection of subsets $V_1, \ldots, V_N$ to be tested against P_0 separately with the help of the minimax theorem: each V_i is matched with a model subset B_i (which can be thought of as a 'neighbourhood' of P_0 if the model is well-specified) such that $\Pi(B_i) > 0$ and inequality (1.5) below is satisfied. The B_i are often chosen as Kullback-Leibler neighbourhoods (as in Schwartz's theorem), but under a moment condition on likelihood ratios larger neighbourhoods can act as alternatives.

Throughout this paper and in the formulation below, we assume that the model is dominated and we use posterior (2.1). Let $\mathrm{co}(V)$ denote the convex hull of V and let P_n^Π $(n \geq 1)$, denote the n-fold prior predictive distributions: $P_n^\Pi(A) = \int P^n(A)\, d\Pi(P)$, for all $A \in \sigma(X_1, \ldots, X_n)$. Furthermore, for given $\alpha \in [0,1]$, model subsets B, W and a given distribution P_0, define,

$$\pi_{P_0}(W, B; \alpha) = \sup_{P \in W} \sup_{Q \in B} P_0\Big(\frac{dP}{dQ}\Big)^\alpha, \tag{1.4}$$

(and $\pi_{P_0}(W, B) = \inf_{\alpha \in [0,1]} \pi_{P_0}(W, B; \alpha)$; see appendix B).

Theorem 1.3. *Let the model $\mathscr{P}$ be given and let $X_1, X_2, \ldots$ be i.i.d.-P_0 distributed. Assume that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. For some $N \geq 1$ let $V_1, \ldots, V_N$ be measurable model subsets. If there exist measurable model subsets $B_1, \ldots, B_N$ such that for every $1 \leq i \leq N$,*

$$\pi_{P_0}(\,\mathrm{co}(V_i),\, B_i\,) < 1, \tag{1.5}$$

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

$$\Pi(\,V \,|\, X_1, \ldots, X_n\,) \xrightarrow{P_0\text{-}a.s.} 0, \tag{1.6}$$

for any $V \subset \bigcup_{1 \leq i \leq N} V_i$.

Although this angle will not be pursued further in this paper, it is noted that P_0 is *not required* to be in the model $\mathscr{P}$ so that the theorem applies both to well- and to mis-specified models [21] in the form stated. Furthermore, in subsection 3.1 it is shown that condition (1.5) is *equivalent* in quite some generality to separation of B_i and $\mathrm{co}(V_i)$ in Kullback-Leibler divergence with respect to P_0,

$$\sup_{Q \in B_i} -P_0 \log \frac{dQ}{dP_0} < \inf_{P \in \mathrm{co}(V_i)} -P_0 \log \frac{dP}{dP_0}, \tag{1.7}$$

underlining the fundamental nature of condition (1.1). But even with this equivalence in mind, the theorem is uncommitted regarding the nature of the V_i, and, more importantly, we may use any B_i that *(i)* allow uniform control of $P_0(p/q)^\alpha$, and *(ii)* allow convenient choice of a prior such that $\Pi(B_i) > 0$. The two requirements on B_i leave room for trade-offs between being 'small enough' to satisfy *(i)*, but 'large enough' to enable a choice for Π that leads to *(ii)*. The freedom to choose B's and Π lends the method the desired flexibility.

In what follows it is shown that Schwartz's theorem, Barron's sieve generalization, Walker's theorem and posterior rates of convergence can all be related to theorem 1.3. In section 2, the denominator the posterior is considered in detail and theorem 1.3 is proved. In section 3 we establish that condition (1.5) is equivalent to KL-separation. Based on that, Schwartz's theorem is re-derived with several variations, *e.g.* posterior consistency in Kullback-Leibler divergence and Hellinger consistency with priors that charge metric balls. In section 4 separable models are considered and in section 5 we consider posterior rates. To provide an example of how our proposals enhance flexibility, corollary 5.3 shows that condition (1.2) can be replaced by a Schwartz-type KL-condition: for some $K > 0$,

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2 \Big) \geq e^{-Kn\epsilon_n^2},$$

under a very simple integrability condition on the model.

Section 6 applies the results to semi-parametric estimation of support boundary points for a density on a bounded interval in $\mathbb{R}$ [23]. The last section contains a short discussion on applications, including consistency in non-parametric density estimation with various Dirichlet mixtures, and counterexamples 1.1 and 1.2. The appendices A, B and C contain two notes on supports, properties of Hellinger transforms and proofs, respectively.

2. Posterior consistency

To establish the basics, the model $(\mathscr{P}, \mathscr{B})$ is a measurable space consisting of Markov kernels P on a sample space $(\mathscr{X}, \mathscr{A})$: the map $A \mapsto P(A)$ is a probability measure for every $P \in \mathscr{P}$ and the map $P \mapsto P(A)$ is measurable for every $A \in \mathscr{A}$. Assuming the model is dominated by a σ-finite measure (with density p for $P \in \mathscr{P}$), a *prior* probability measure Π on $(\mathscr{P}, \mathscr{B})$ gives rise to the *posterior*:

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

We take the frequentist *i.i.d.* perspective, *i.e.* we assume that there exists a distribution P_0 on $(\mathscr{X}, \mathscr{A})$ such that $(X_1, \ldots, X_n) \sim P_0^n$. As a consequence expression (2.1) does not make sense automatically: for the denominator to be non-zero with P_0^n-probability one, we impose that,

$$P_0^n \ll P_n^\Pi, \tag{2.2}$$

for every $n \geq 1$, where P_n^{Π} is the prior predictive distribution. If (2.2) is not satisfied, expression (2.1) for the posterior is ill-defined for infinitely many $n \geq 1$ with P_0^{∞}-probability one. The following proposition provides sufficient condition to prevent this.

Proposition 2.1. *If P_0 lies in the Hellinger support of the prior Π, then $P_0^n \ll P_n^{\Pi}$, for all $n \geq 1$. Particularly, if Π is a KL-prior, then $P_0^n \ll P_n^{\Pi}$ for all $n \geq 1$.*

So under Schwartz's prior mass condition, one does not worry about condition (2.2); it plays a role only if one is interested in priors that are not Kullback-Leibler priors (as in example 7.2, for instance).

Example 2.2. To illustrate the denominator problem by example, consider the following regression problem with one-sided error distributions: one observes pairs $(X_i, Y_i) \in \mathbb{R}^2$, $i \geq 1$, of real-valued random variables related through $Y = f(X) + e$ for some non-negative regression function f, such that for all $\delta > 0$, $P(f(X) < \delta) > 0$. Errors $e_1, e_2, \ldots$ are independent of X and *i.i.d.*, with a shared marginal distribution supported on $[\theta, \infty)$, for some unknown $\theta \in \mathbb{R}$ to be estimated. The problem occurs when the statistician believes that his errors are *positive* with probability one, while their true distribution assigns (small but) non-zero probability to *negative* outcomes. (In finance examples abound, arising when one anticipates non-negative returns (for example a hedged return, the total return on a bond or an auction price) based on an incomplete or simplified model for downside risk.)

The statistician makes a choice for the prior Π that reflects his belief, placing no mass on negative values for θ. When sequential *i.i.d.* draws are conducted, sooner or later a negative value of the error will occur in conjunction with a small value of $f(X)$, resulting in a negative value for Y. But negative $f(X) + e$ have probability zero according to all distributions in a subset of the model of prior mass one: sooner or later, the likelihood evaluates to zero Π-almost-everywhere in the model, resulting in a posterior that is ill-defined.

2.1. A sketch of the proof of theorem 1.3

Our first lemma asserts that, under the condition that certain specific test-sequences for covers of complements exist, posterior concentration in neighbourhoods follows. The proof is inspired by Le Cam's *dimensionality restrictions* and more broadly, by [34, 26, 4, 5, 28, 15, 17]. The argument is essentially an application of the minimax theorem (see section 16.4 of [28], section 45 of [36]), adapted as in [21]. The essential difference between lemma 2.3 and other Bayesian limit theorems is that posterior numerator and denominator are dealt with simultaneously rather than separately, so that the prior Π is one of the factors that determines testing power and can be balanced against model properties directly.

In the following lemma V is a fixed set (*e.g.* the complement of an open neighbourhood of P_0) for which we want to prove asymptotically vanishing posterior mass. We cover V by a finite number of model subsets $V_1, \ldots, V_N$ such

that for each V_i, a special type of test sequence exists. In the next subsection, we give conditions for the existence of such sequences.

Lemma 2.3. *Assume that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. For some $N \geq 1$, 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,*

$$P_0^n \phi_{i,n} + \sup_{P \in V_i} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{i,n}) \leq e^{-nD_i}, \tag{2.3}$$

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

$$\Pi(V|X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0. \tag{2.4}$$

The condition that covers of the model have to be of *finite* order is restrictive and problems arise already in parametric context. In such cases application of the theorem requires a bit more refinement [26] (see example C.1), or the alternatives of section 4.

Le Cam [26, 27, 28] and Birgé [4, 5] propose a seminal approach to testing that combines the minimax theorem with the Hellinger geometry of the model. Here we stay close to the methods of [21] which are inspired by the above and their application in [15]. Define $V^n = \{P^n : P \in V\}$ and denote its convex hull by $\mathrm{co}(V^n)$; elements from $\mathrm{co}(V^n)$ are denoted P_n. The following lemma says that testing power is bounded in terms of Hellinger transforms [28].

Lemma 2.4. *Let $n \geq 1$, $V \in \mathscr{B}$ be given; assume that $P_0^n(dP^n/dP_n^\Pi) < \infty$ for all $P \in V$. Then there exists a test sequence (ϕ_n) such that,*

$$P_0^n \phi_n + \sup_{P \in V} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_n) \leq \sup_{P_n \in \mathrm{co}(V^n)} \inf_{0 \leq \alpha \leq 1} P_0^n \Big(\frac{dP_n}{dP_n^\Pi}\Big)^\alpha. \tag{2.5}$$

Given Π and a measurable B such that $\Pi(B) > 0$, define the *local* prior predictive distributions $P_n^{\Pi|B}$ by conditioning the prior predictive on B:

$$P_n^{\Pi|B}(A) = \int Q^n(A)\, d\Pi(Q|B), \tag{2.6}$$

for all $n \geq 1$ and $A \in \sigma(X_1, \ldots, X_n)$. The following lemma formulates an upper bound for the right-hand side of inequality (2.5), which prescribes the (n-independent) form of the central requirement of theorem 1.3.

Lemma 2.5. *Let Π be given, fix $n \geq 1$. Let $V, B \in \mathscr{B}$ 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 $\phi_n : \mathscr{X}^n \to [0, 1]$ such that,*

$$\begin{aligned} P_0^n \phi_n + \sup_{P \in V} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_n) \\ \leq \inf_{0 \leq \alpha \leq 1} \Pi(B)^{-\alpha} \int \Big[\sup_{P \in \mathrm{co}(V)} P_0 \Big(\frac{dP}{dQ}\Big)^\alpha\Big]^n d\Pi(Q|B). \end{aligned} \tag{2.7}$$

Theorem 1.3 is the conclusion of lemma 2.3 upon substitution of lemmas 2.4 and 2.5.

3. Variations on Schwartz's theorem

In this section we apply theorem 1.3 to re-derive Schwartz's theorem, sharpen its assertion to consistency in Kullback-Leibler divergence and we consider model conditions that allow priors charging metric balls rather than Kullback-Leibler neighbourhoods.

3.1. Schwartz's theorem and Kullback-Leibler priors

The strategy to prove posterior consistency in a certain topology (or more generally, to prove posterior concentration outside 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 (1.5) for subsets B_i that are as large as possible and neighbourhoods of P_0 in an appropriate sense. Subsequently we try to find (a σ-algebra $\mathscr{B}$ on $\mathscr{P}$ and) a prior $\Pi : \mathscr{B} \to [0,1]$ such that ($B_i \in \mathscr{B}$ and) $\Pi(B_i) > 0$ for all $1 \leq i \leq N$. In this regard the following lemma offers guidance, because it relates testing power to Kullback-Leibler separation of the sets B and W in definition (1.4). It is stressed that in applications the sets W_i are *convex hulls* of model subsets V_i.

Lemma 3.1. *Let $P_0 \in B \subset \mathscr{P}$ and $W \subset \mathscr{P}$ be given and assume that there exists an $a \in (0,1)$ such that for all $Q \in B$ and $P \in W$, $P_0(dP/dQ)^a < \infty$. Then,*

$$\pi_{P_0}(W,B) < 1, \tag{3.1}$$

if and only if,

$$\sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} < \inf_{P \in W} -P_0 \log \frac{dP}{dP_0}. \tag{3.2}$$

Quite generally, lemma 3.1 shows that model subsets are consistently testable *if and only if* they can be separated from neighbourhoods of P_0 in Kullback-Leibler divergence. This illustrates the fundamental nature of Schwartz's prior mass requirement and undermines hopes for useful priors that charge different neighbourhoods of P_0 in general. However, this does not exclude the possibility of gaining freedom in the choice of the prior by strengthening requirements on the model, as we hope to demonstrate with the rest of this paper.

Due to the fact that Kullback-Leibler divergence dominates Hellinger distance, Schwartz's theorem can be proved from theorem 1.3 and lemma 3.1 (at least, for models that have $\sup_{Q \in B} P_0(dP/dQ) < \infty$ for all $P \in V$ and B a Kullback-Leibler neighbourhood of P_0). Schwartz's theorem does not fully exploit the room that (3.2) offers, because it stops short of asserting posterior consistency in Kullback-Leibler divergence. However it is well-known [39, 16], that Kullback-Leibler consistency does not require much more than Schwartz's conditions. The following theorem provides posterior Kullback-Leibler consistency without requiring more of the prior, by imposing an integrability condition on the model.

Theorem 3.2. *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 $P \in \mathscr{P}$. Let Π be a Kullback-Leibler prior. For any $\epsilon > 0$, assume that $\{P : -P_0 \log(dP/dP_0) \geq \epsilon\}$ is covered by a finite number $N \geq 1$ of model subsets $V_1, \ldots, V_N$ such that,*

$$\inf_{P\in \mathrm{co}(V_i)} -P_0 \log \frac{dP}{dP_0} > 0, \tag{3.3}$$

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

$$\Pi\big(P \in \mathscr{P} \,:\, -P_0 \log(dP/dP_0) < \epsilon \,\big|\, X_1, \ldots, X_n \big) \xrightarrow{P_0\text{-a.s.}} 1. \tag{3.4}$$

To appreciate how a finite cover of Kullback-Leibler neighbourhoods may occur in models, consider the following example that relies on relative compactness with respect to the uniform norm for log-densities.

Example 3.3. Let $\epsilon > 0$ be given and assume that the complement V of a Kullback-Leibler ball of radius $\epsilon > 0$ contains N points $P_1, \ldots, P_N$ such that the convex sets,

$$V_i = \big\{P \in \mathscr{P} : \|dP/dP_i - 1\|_\infty < \tfrac{1}{2}\epsilon\big\},$$

cover V. Finiteness of the cover can be guaranteed, for example with the Ascoli-Arzelà theorem, if the model describes data taking values in a fixed bounded interval in $\mathbb{R}$ and the associated family of log-densities is bounded and equicontinuous. (Other ways to find suitable covers refer to $\|\cdot\|_\infty$-entropy or bracketing numbers for log-likelihood ratios [38].) Then any $P \in \mathrm{co}(V_i)$ 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,

$$-P_0 \log \frac{dP}{dP_0} \geq \epsilon - P_0 \log \frac{dP}{dP_i} \geq \tfrac{1}{2}\epsilon,$$

and (3.3) holds. In such models, any prior Π satisfying (1.1) leads to a posterior that is consistent with respect to Kullback-Leibler divergence.

3.2. Priors that charge metric balls

In this subsection we discuss model conditions that allow one to relax Schwartz's condition for the prior, to the condition that the prior has full support in Hellinger (or other) metric topologies. Given some P_0 and a suitable (metric) neighbourhood B, we impose that for all $Q \in B$ and any $P \in \mathscr{P}$, $p/q \in L_2(Q)$ (with norm denoted $\|\cdot\|_{2,Q}$). Under this condition the Cauchy-Schwarz inequality leads to,

$$\begin{aligned} P_0\Big(\frac{p}{q}\Big)^{1/2} &= \int \Big(\frac{p_0}{q}\Big)^{1/2} p_0^{1/2} p^{1/2}\, d\mu \\ &= \int p_0^{1/2}\, p^{1/2}\, d\mu - \int \Big(1 - \Big(\frac{p_0}{q}\Big)^{1/2}\Big) \Big(\frac{p_0}{q}\Big)^{1/2} \Big(\frac{p}{q}\Big)^{1/2} dQ \\ &\leq 1 - \frac{1}{2} H(P_0, P)^2 + H(P_0, Q) \Big\|\frac{p_0}{q}\Big\|_{2,Q}^{1/2} \Big\|\frac{p}{q}\Big\|_{2,Q}^{1/2}. \end{aligned}$$

Combined with lemma 2.3 this gives the following theorem.

Theorem 3.4. *Assume the model $\mathscr{P}$ has finite Hellinger metric entropy numbers. Furthermore assume that there exists a constant $L > 0$ and a Hellinger ball B' centred on P_0 such that for all $P \in \mathscr{P}$ and $Q \in B'$,*

$$\left\| \frac{p}{q} \right\|_{2,Q} = \left(\int \frac{p^2}{q} \, d\mu \right)^{1/2} < L. \tag{3.5}$$

Finally assume that for any Hellinger neighbourhood B of P_0, $\Pi(B) > 0$. Then the posterior is Hellinger consistent, P_0-almost-surely.

Next choose $1 \leq r < \infty$. Analogous to the Hellinger metric ($r = 2$), define, for all P, Q probability measures, Matusita's *r-metric distance* [30],

$$d_r(P, Q) = \left(\int \left| p^{1/r} - q^{1/r} \right|^r d\mu \right)^{1/r},$$

(based on any σ-finite μ that dominates P and Q). Applying Hölder's inequality where we applied Cauchy-Schwarz before we arrive at the following theorem concerning priors that charge d_r-balls.

Theorem 3.5. *Let $1 \leq r < \infty$ be given and let the model $\mathscr{P}$ be has finite d_r-metric entropy numbers. Let $X_1, X_2, \ldots$ be i.i.d.-P_0 distributed for some $P_0 \in \mathscr{P}$. Assume that the prior is such that $P_0^n \ll P_n^\Pi$, for all $n \geq 1$ and satisfies,*

$$\Pi\big(P \in \mathscr{P} \,:\, d_r(P_0, P) < \delta \big) > 0, \tag{3.6}$$

for all $\delta > 0$. In addition, assume that there is an $L > 0$ and a d_r-ball B such that for all $P \in \mathscr{P}$ and $Q \in B$, $P_0(p/q)^{s/r \vee 1} \leq L^s$, where $1/r + 1/s = 1$. Then the posterior is consistent for the d_r-metric, P_0-almost-surely.

Remark 3.6. For the models under discussion, we note the following general construction of so-called *net priors* [25, 28, 13, 15, 20]: denote the metric on $\mathscr{P}$ by d. Initially, assume that $\mathscr{P}$ has finite d-metric entropy numbers. Let (η_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 an η_m-net for $\mathscr{P}$ and define Π_m to be the measure that places mass $1/M_m$ at every P_i, $(1 \leq i \leq M_m)$. Choose a sequence (λ_m) such that $\lambda_m > 0$ for all $m \geq 1$ and $\sum_{m \geq 1} \lambda_m = 1$, to define the net prior $\Pi = \sum_{m \geq 1} \lambda_m \, \Pi_m$. A net prior assigns non-zero mass to every open set. In addition, lower-bounds for prior mass in metric balls are proportional to inverses of upper bounds for metric entropy numbers, provided we choose (λ_m) appropriately. In case $\mathscr{P}$ is not totally bounded, one may generalize the above construction by choosing an sieve (K_m) of relatively compact submodels.

Net priors, or more generally, Borel priors of full support are helpful if one is interested in the construction of Kullback-Leibler priors, at least, if the corresponding topology is fine enough.

Lemma 3.7. *If for every $P \in \mathscr{P}$, the Kullback-Leibler divergence $\mathscr{P} \to \mathbb{R} : Q \mapsto -P \log(dQ/dP)$ is continuous, then a Borel prior of full support is a Kullback-Leibler prior.*

When discussing consistency, requirements on the model like (3.5) are present to guarantee continuity of the Kullback-Leibler-divergence. For example, the perceptive reader may have recognized in (3.5) sufficiency to invoke theorem 5 of [41] which provides an upper bound for the Kullback-Leibler divergence in terms of the Hellinger distance. The latter is a stronger, Lipschitz-like variation on the continuity condition of the above lemma.

4. Posterior consistency on separable models

Requiring *finiteness* of the order of the cover in theorem 1.3 and lemma 2.3 is somewhat crude. There are several ways out: firstly, in subsection 4.1 we explore the possibility of letting a sieve of totally bounded submodels approximate the full model analogous to Barron's theorem. Secondly, Hellinger consistency of the posterior on separable models formed the assertion of a remarkable theorem of Walker for a Kullback-Leibler prior that also satisfies a summability condition [39]. In subsection 4.2 we show that variations on Walker's theorem can be derived with the methods of section 2.

4.1. Generalization to sieves

When the model is (a measurable subset of) a Polish space, inner regularity guarantees that the model is approximated in prior measure by (relatively) compact submodels. Since the latter are of finite metric entropy, a proof is conceivable based on a sieve of compact submodels with complements of 'negligible' prior mass.

Theorem 4.1. *Let $X_1, X_2, \ldots$ be i.i.d. $- P_0$ for some $P_0 \in \mathscr{P}$ and let V be given. Assume that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$ and that there exist constants $K, L > 0$ and a sequence of submodels $(\mathscr{P}_n)$ such that for large enough $n \geq 1$,*

(i.) there is a cover $V_1, \ldots, V_{N_n}$ for $V \cap \mathscr{P}_n$ of order $N_n \leq \exp(\frac{1}{2}Ln)$ with tests $\phi_{1,n}, \ldots, \phi_{N_n,n}$ such that,

$$P_0^n \phi_{i,n} + \sup_{P \in V_i} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{i,n}) \leq e^{-nL},$$

for all $1 \leq i \leq N_n$;

(ii.) the prior mass $\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp(-nK)$ and,

$$\sup_{P \in V \setminus \mathscr{P}_n} \sup_{Q \in B} P_0\Big(\frac{dP}{dQ}\Big) \leq e^{\frac{K}{2}}, \tag{4.1}$$

for some model subset B such that $\Pi(B) > 0$.

Then $\Pi(\, V \mid X_1, \ldots, X_n \,) \xrightarrow{P_0\text{-a.s.}} 0$.

Condition *(i.)* of theorem 4.1 corresponds with condition (2.3) and upper bounds for testing power of the preceding subsections remain applicable. More particularly, condition *(i.)* has the following alternative.

(i'.) there exist a model subset B *with* $\Pi(B) > 0$ *and a cover* $V_1, \ldots, V_{N_n}$ *for* $V \cap \mathscr{P}_n$ *of order* $N_n \leq \exp(\frac{1}{2}Ln)$*, such that for every* $1 \leq i \leq N_n$*,*

$$\pi_{P_0}\big(\operatorname{co}(V_i), B\,\big) \leq e^{-L},$$

and $\sup_{Q\in B} P_0(dP/dQ) < \infty$ *for all* $P \in V_i$*.*

Condition *(ii.)* of theorem 4.1 defines what 'negligibility' of prior mass outside the sieve means. If we think of B as a small neighbourhood around P_0, it appears that the freedom to choose B small enables upper bounds for the *l.h.s.* of (4.1) arbitrarily close to one. In such cases, condition *(ii.)* reduces to the requirement that $\Pi(\mathscr{P} \setminus \mathscr{P}_n)$ decreases exponentially [1]. The following example illustrates this point.

Example 4.2. Assume that $X_1, X_2, \ldots$ are *i.i.d.*-P_0 for some P_0 in a model $\mathscr{P}$ that is dominated by a σ-finite measure μ. Consider a prior Π that charges all $L_\infty(\mu)$-balls around $\log p_0$ (where p_0, p denote the μ-densities for P_0, P respectively):

$$\Pi\big(\,P \in \mathscr{P} \,:\, \|\log p - \log p_0\|_\infty < \epsilon\,\big) > 0,$$

for all $\epsilon > 0$. Note that, for all $P \in \mathscr{P}$,

$$P_0\Big(\frac{dP}{dQ}\Big) = \int \frac{p_0\, p}{q}\, d\mu = \int \frac{p_0}{q}\, dP \leq e^{\epsilon},$$

whenever $\|\log q - \log p_0\|_\infty \leq \epsilon$. Hence, a sieve $(\mathscr{P}_n)$ satisfying condition *(i.)* such that $\Pi(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp(-nK')$ for *some* small $K' > 0$ would suffice in this case and similar ones.

A generalization of condition *(ii.)* of theorem 4.1 involving n-dependent choices for B can be found in appendix C. Theorem 4.1 is applied in the support boundary problem of section 6, see remark 6.4.

4.2. Variations on Walker's theorem

In this subsection we abandon constructions based on finite covers altogether and require only that the cover is *countable*. Le Cam's *dimensionality restrictions* [26, 27] are related to in example C.1. More generally, a natural setting arises when we consider models that are *separable* in some metric topology, in which case countable covers by balls of any radius exist.

Theorem 4.3. *Let* $\mathscr{P}$ *and* Π *be given and assume that* $P_0^n \ll P_n^\Pi$ *for all* $n \geq 1$*. Let* V *be a model subset, with a countable cover* $V_1, V_2, \ldots$ *and* $B_1, B_2, \ldots$ *such that for all* $i \geq 1$*, we have* $\Pi(B_i) > 0$ *and for all* $P \in V_i$*,* $\sup_{Q\in B_i} P_0(dP/dQ) < \infty$*. Then,*

$$P_0^n \Pi(V|X_1, \ldots, X_n) \leq \sum_{i\geq 1} \inf_{0\leq\alpha\leq 1} \frac{\Pi(V_i)^\alpha}{\Pi(B_i)^\alpha}\, \pi_{P_0}\big(\operatorname{co}(V_i), B_i; \alpha\,\big)^n. \tag{4.2}$$

Compare the upper bound in (4.2) to that of example C.1. Two corollaries show how theorem 4.3 is related to Walker's condition [39]. Note that in the first, the prior is not required to be a KL-prior.

Corollary 4.4. *Let $\mathscr{P}$ and Π be given and assume that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. Let V be a model subset, with a countable cover $V_1, V_2, \ldots$, and a $B \subset \mathscr{P}$ such that $\Pi(B) > 0$ and for all $i \geq 1$, $P \in V_i$, $\sup_{Q \in B} P_0(dP/dQ) < \infty$. Furthermore, assume that,*

$$\sup_{i \geq 1} \sup_{P \in \mathrm{co}(V_i)} \sup_{Q \in B} P_0 \Big(\frac{dP}{dQ} \Big)^{1/2} < 1. \tag{4.3}$$

If the prior satisfies the summability condition,

$$\sum_{i \geq 1} \Pi(V_i)^{1/2} < \infty, \tag{4.4}$$

then the posterior satisfies, $\Pi(V | X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0$.

For another perspective on Walker's condition, see exercises 8.11–8.12 in [16]. The second corollary does not impose model conditions like (4.3), and, instead, requires a Kullback-Leibler prior that satisfies a slightly different summability condition.

Corollary 4.5. *Let $\mathscr{P}$ be separable in the Hellinger topology. Assume that there is Kullback-Leibler neighbourhood B of P_0 such that for all $P \in \mathscr{P}$, $\sup_{Q \in B} P_0(dP/dQ) < \infty$. Let Π be a Kullback-Leibler prior such that for all $\beta > 0$,*

$$\sum_{i \geq 1} \Pi(V_i)^{\beta} < \infty, \tag{4.5}$$

where the V_i, ($i \geq 1$) are any cover of $\mathscr{P}$ by Hellinger balls of a fixed radius. Then the posterior is P_0-almost-surely Hellinger consistent.

5. Posterior rates of convergence

Minimax rates of convergence for (estimators based on) posterior distributions were considered more or less simultaneously in Ghosal, Ghosh and van der Vaart [15] and Shen and Wasserman [35], with conditions that display very close resemblance. Both pose (1.2) as the condition on the prior and both appear to be inspired by Wong and Shen [41], as well as Ghosal *et al.* [13] and/or Barron *et al.* [2], which concern posterior consistency based on controlled bracketing entropy for a sieve, up to subsets of negligible prior mass, following ideas that were first laid down in [1]. In [40] Walker, Lijoi and Prünster extend the results of [39] to posterior rates of convergence.

Note that methods proposed in the preceding sections hold at finite values of $n \geq 1$: the hypothesis B, V as well as the constant α can be made n-dependent without changing the basic building blocks. As such, not much needs to be adapted to preceding results to extend also to rates of posterior convergence.

Below we follow Barron's ideas again and sharpen theorem 4.1 to accomodate rates of posterior convergence. For the theorem below, we endow the model with a metric d and assume that the prior is Borel with respect to the associated metric topology.

Theorem 5.1. *Let $X_1, X_2, \ldots$ be i.i.d. $- P_0$ for some $P_0 \in \mathscr{P}$. Assume that the prior Π is such that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. Let (ϵ_n) be a sequence with $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Define $V_n = \{P \in \mathscr{P} : d(P, P_0) > \epsilon_n\}$, a sequence of measurable submodels $\mathscr{P}_n \subset \mathscr{P}$ and measurable model subsets B_n such that $\sup_{Q \in B_n} P_0(dP/dQ) < \infty$ for all $P \in V_n$. Assume that, for sufficiently large $n \geq 1$,*

(i) there is an $L > 0$ such that $V_n \cap \mathscr{P}_n$ has a cover $V_{n,1}, V_{n,2}, \ldots, V_{n,N_n} \subset \mathscr{P}_n$ of order $N_n \leq \exp(\frac{1}{2}Ln\epsilon_n^2)$, such that for all $1 \leq i \leq N_n$,

$$\pi_{P_0}\big(\,\mathrm{co}(V_{n,i}), B_n\,\big) \leq e^{-L\epsilon_n^2}, \tag{5.1}$$

(ii) there is a $K > 0$ such that $\Pi(\mathscr{P} \backslash \mathscr{P}_n) \leq e^{-Kn\epsilon_n^2}$ and $\Pi(B_n) \geq e^{-\frac{K}{2}n\epsilon_n^2}$, while also,

$$\sup_{P \in \mathscr{P} \backslash \mathscr{P}_n} \sup_{Q \in B_n} P_0\Big(\frac{dP}{dQ}\Big) < e^{\frac{K}{4}\epsilon_n^2}. \tag{5.2}$$

Then, $\Pi(\,P \in \mathscr{P} :\, d(P, P_0) > \epsilon_n \,|\, X_1, \ldots, X_n\,) \xrightarrow{P_0} 0$.

This theorem has been formulated generally and this generality obscures the interpretation of conditions somewhat: the first condition plays the same role as the entropy condition in the Ghosal-Ghosh-van der Vaart theorem; it enables construction of a suitable minimax test. Sufficiency of prior mass around P_0 forms part of the second condition, which also assures that the sieve approximates the model closely enough, by upper-bounding prior mass outside the sieve. Under an integrability condition, condition (5.1) for the sets $\mathrm{co}(V_{n,i})$ and B_n follows from a minimal amount of separation of $\mathrm{co}(V_{n,i})$ and B_n in Kullback-Leibler divergence.

Lemma 5.2. *Consider two model subsets B, W such that $P_0 \in B$. Suppose that for some $a \in (0,1)$, $P_0(dP/dQ)^a$ is finite for all $P \in W$, $Q \in B$. If, for some $\Delta > 0$,*

$$\sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} \leq \inf_{P \in W} -P_0 \log \frac{dP}{dP_0} - \Delta, \tag{5.3}$$

then there exists an $\alpha \in (0,1)$ such that,

$$\pi_{P_0}(B, W) \leq e^{-\alpha\Delta}.$$

Conversely, if for some $\Delta > 0$,

$$\sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} > \inf_{P \in W} -P_0 \log \frac{dP}{dP_0} - \Delta,$$

then $\pi_{P_0}(B, W; \alpha) > e^{-\alpha\Delta}$ for all $\alpha \in (0,1)$.

Lemma 5.2 says that if B and W are separated in Kullback-Leibler divergence by some small difference Δ, then the logarithm of the Hellinger transform $\log \pi_{P_0}(B, W)$ is upper-bounded by a multiple of $-\Delta$. This emphasizes the role played by the Kullback-Leibler divergence and illustrates the associated limitation: not all models have integrable likelihood ratios, and Kullback-Leibler divergences that are infinite make inequality (5.3) void.

With lemma 5.2 in hand, we can simplify and specify theorem 5.1 considerably, to bring us closer to the Ghosal-Ghosh-van der Vaart theorem. For simplicity of presentation, we do not incorporate Barron's negligible prior mass argument (although one could trivially).

Corollary 5.3. *Let $X_1, X_2, \ldots$ be i.i.d.-P_0 for some $P_0 \in \mathscr{P}$. Specify that the metric on $\mathscr{P}$ is the Hellinger metric H; define (ϵ_n) with $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$, and take $V_n = \{P \in \mathscr{P} : H(P_0, P) > M\epsilon_n\}$, for $M > 0$, and $B_n = \{Q \in \mathscr{P} : -P_0 \log(dQ/dP_0) < \epsilon_n^2\}$. Assume that for n large enough and all $P \in V_n$, $\sup\{P_0(dP/dQ) : Q \in B_n\} < \infty$. If, for large enough $n \geq 1$,*

(i) there is an $L > 0$, such that $N(\epsilon_n, \mathscr{P}, H) \leq e^{Ln\epsilon_n^2}$;
(ii) there is a $K > 0$, such that

$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2 \Big) \geq e^{-Kn\epsilon_n^2}, \tag{5.4}$$

then $\Pi(\, P \in \mathscr{P} :\, H(P, P_0) > M\epsilon_n \,|\, X_1, \ldots, X_n \,) \xrightarrow{P_0} 0$, for M large enough.

Comparison with inequality (1.2) shows that the requirement on the prior is in terms of *Schwartz's KL-neighbourhoods* rather than the second-order KL-neighbourhoods of (1.2), a convenience that comes at the expense of an integrability condition for likelihood ratios. (It is noted that Ghosal and van der Vaart [16] provide a refinement of [15] that also does not involve second-order KL-neighbourhoods. The simplicity of the integrability condition of corollary 5.3 seems preferable to the technical intricacy of their theorem 8.11, however.)

For an analysis of example 1.2 using corollary 5.3, see example 7.3.

6. Marginal consistency

In this section, we consider a problem of the following basic, semi-parametric type [3]: let Θ be an open subset of $\mathbb{R}^k$ parametrizing the *parameter of interest* θ and let H be a measurable (and typically infinite-dimensional) parameter space for the *nuisance parameter* η. The model is $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ where $\Theta \times H \to \mathscr{P} : (\theta, \eta) \mapsto P_{\theta,\eta}$ is a Markov kernel on the sample space $(\mathscr{X}, \mathscr{A})$ describing the distributions of individual points from an infinite *i.i.d.* sample $X_1, X_2, \ldots \in \mathscr{X}$. Given a metric $g : \Theta \times \Theta \to [0, \infty)$ and a prior measure Π on $\Theta \times H$ we say that the posterior is *marginally consistent* for the parameter of interest, if for all $\epsilon > 0$,

$$\Pi\big(P_{\theta,\eta} \in \mathscr{P} : g(\theta, \theta_0) > \epsilon, \eta \in H \;\big|\; X_1, \ldots, X_n \big) \xrightarrow{P_{\theta_0,\eta_0}\text{-a.s.}} 0, \tag{6.1}$$

for all $\theta_0 \in \Theta$ and $\eta_0 \in H$. Marginal consistency amounts to consistency with respect to the pseudo-metric $d : \mathscr{P} \times \mathscr{P} \to [0,\infty)$, $d\big(P_{\theta,\eta}, P_{\theta',\eta'}\big) = g(\theta,\theta')$, for all $\theta,\theta' \in \Theta$ and $\eta,\eta' \in H$. The following theorem is a formulation of theorem 1.3 specific to marginal consistency.

Theorem 6.1. *Let $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ be a model for data $X_1, X_2, \ldots$ assumed distributed i.i.d.-P_0 for some $P_0 \in \mathscr{P}$ in the Hellinger support of Π. Let $\epsilon > 0$ be given, define $V = \{P_{\theta,\eta} \in \mathscr{P} : g(\theta,\theta_0) > \epsilon, \eta \in H\}$ 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 $1 \leq i \leq N$,*

$$\pi_{P_0}(\,\mathrm{co}(V_i),\, B_i\,) < 1,$$

$\Pi(B_i) > 0$ and $\sup_{Q \in B_i} P_0(dP/dQ) < \infty$ for all $P \in V_i$, then the posterior is marginally consistent, P_0-almost-surely.

6.1. Density support boundaries

Consistent support boundary estimation (see [19], or [31] for a more recent, Bayesian reference), 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 [32]. The Bernstein-Von Mises phenomenon for this type of problem is studied in Kleijn and Knapik [23] and leads to exponential rather than normal limiting form for the posterior. Below, we prove consistency using theorem 1.3 and note that the result generalizes relatively straightforwardly to rates (see below).

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

$$\inf_{\eta \in H} \min\Big\{ \int_0^\epsilon \eta \, d\mu, \int_{1-\epsilon}^1 \eta \, d\mu \Big\} \geq f(\epsilon), \quad (0 < \epsilon < a). \tag{6.2}$$

The model $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is defined in terms of Lebesgue densities of the following semi-parametric form,

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

for some $(\theta_1,\theta_2) \in \Theta$ and $\eta \in H$. A condition like (6.2) is necessarily part of the analysis, because questions concerning support boundary points make sense *only* if the distributions under consideration put mass in every neighbourhood of θ_1 and θ_2. (Let $\|\cdot\|_{s,Q}$ denote the $L_s(Q)$-norm, for $s \geq 1$.)

Theorem 6.2. *For some $\sigma > 0$, let Θ be $\{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ and let the space H with associated function f as in (6.2) be given. Assume that there exists an $s \geq 1$ such that the sets B,*

$$B = \Big\{ Q \in \mathscr{P} : \Big\| \frac{dP_0}{dQ} - 1 \Big\|_{s,Q} < \delta \Big\},$$

satisfy $\Pi(B) > 0$ for all $\delta > 0$. Also assume there exists a constant $K > 0$ such that for all $P \in \mathscr{P}$ and $Q \in B$, $\|dP/dQ\|_{r,Q} \leq K$, where $1/r + 1/s = 1$. If $X_1, X_2, \ldots$ form an i.i.d.-P_0 sample for $P_0 = P_{\theta_0,\eta_0} \in \mathscr{P}$ then,

$$\Pi\big(\theta \in \Theta \,:\, \|\theta - \theta_0\| < \epsilon \,\big|\, X_1, \ldots, X_n \big) \xrightarrow{P_0\text{-a.s.}} 1, \tag{6.3}$$

for all $\epsilon > 0$.

Example 6.3. To apply theorem 6.2, let $P_0 = P_{\theta_0,\eta_0}$ 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 $g : [0,1] \to [0, \infty)$ be a known Lebesgue probability density such that $g(x) > 0$ for all $x \in (0,1)$. For some constant $M > 0$ consider the subset C_M of $C[0,1]$ of all continuous $h : [0,1] \to [0, \infty)$ such that $e^{-M} \leq h \leq e^M$. To define the model's dependence on the nuisance parameter h, let H contain all $\eta : [0,1] \to [0, \infty)$ that are Esscher transforms [29] 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 η under the prior Π_H by taking $h = e^Z$. On Θ let Π_Θ denote a prior with a Lebesgue density that is continuous and strictly positive on Θ. One verifies easily that the model satisfies (6.2) with f defined by,

$$f(\epsilon) = e^{-2M} \min\Big\{ \int_0^\epsilon g(x)\, dx, \int_{1-\epsilon}^1 g(x)\, dx \Big\},$$

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.

Remark 6.4. If the assumed bound $\sigma > 0$ is set to infinity, testing power is lost (see the proof of theorem 6.2, or note that if one pictures distributions P of wider and wider support, the minimal mass bound (6.2) implies less and less mass remains to lower-bound $P(p_0 = 0)$ and $P_0(p = 0)$). To see that the bound is of a technical rather than essential nature, note that if a model of bounded-support

distributions satisfies (6.2) and is uniformly tight, such a constant $\sigma > 0$ exists. Consequently, a sequence of models with growing σ's can be used: for given $P_0 = P_{\theta_0,\eta_0}$, there is a lower bound $\bar{\sigma} > 0$ such that the model of theorem 6.2 is well-specified for all $\sigma > \bar{\sigma}$. So if $\sigma_m \to \infty$, the corresponding models $\mathscr{P}_m$ are well-specified for large enough m and the posteriors on those $\mathscr{P}_m$ are consistent, *c.f.* theorem 6.2. By diagonalization there exists a sequence $(\sigma_{m(n)})_{n\geq 1}$ that traverses (σ_m) slowly enough in order to guarantee that consistency obtains while we increase $m(n)$ with the sample size n.

To know exactly how slowly we should let σ go to infinity, we use theorem 4.1: let σ_n increase with n and define $\mathscr{P}_n = \{P_{\theta,\eta} \in \mathscr{P} : |\theta_1 - \theta_2| < \sigma_n, \eta \in H\}$. Since $N_n = 4$ for all $n \geq 1$ (namely the sets $V_{+,1}$, $V_{-,1}$, $V_{+,2}$ and $V_{-,2}$ in the proof of theorem 6.2) any constant $L > 0$ will do, as long as,

$$n\, f(\epsilon/\sigma_n) \to \infty.$$

(Similarly, rates of convergence can be studied with the choice $\epsilon = \epsilon_n$: the modulus of continuity f then determines how ϵ_n, σ_n and other n-dependencies must be fine-tuned.) A glance at inequality (C.8) suggests that condition (4.1) applies, if we choose Π such that,

$$\Pi\big(P_{\theta,\eta} \in \mathscr{P} : |\theta_1 - \theta_2| \geq \sigma_n, \eta \in H\big) \leq e^{-n\,K},$$

for some $K > 0$. For example, if the family H consists of densities that display jumps at both θ_1 and θ_2 of some minimal size $\delta > 0$, then $f(x) \geq \frac{1}{2}\delta\, x$ for values of $x > 0$ that are close enough to $x = 0$. Consequently, for a model in which support boundaries represent discontinuous jumps, marginal posterior consistency obtains if we let $\sigma_n = o(n)$. If H consists of densities that are continuous ($k = 0$) or $k \geq 1$ times continuously differentiable at the boundary points, then $f(x)$ is lower-bounded by a multiple of x^{k+2}, which implies that σ_n must be of order $o(n^{1/k+2})$.

7. Conclusion and examples

Schwartz's theorem is central to the frequentist perspective on Bayesian non-parametric statistics and it has been in place for more 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. Since choices for non-parametric priors are usually not abundant, overly stringent criteria form a problem. 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.

7.1. Some easy examples

Because Hellinger consistent density estimation using mixtures is a well-studied subject, especially with Dirichlet priors, we discuss that example below in quite some generality, to illustrate practicality of the proposed methods.

Example 7.1. Consider a model $\mathscr{P}$ for observation of one of two real-valued, dependent random variables X, Z, assuming that *if* we would observe Z, the distribution for X would be known: $X|Z = z$ is assumed to have a Lebesgue density $p(\cdot|z) : \mathbb{R} \to \mathbb{R}$ such that $z \mapsto p(x|z)$ is bounded and continuous for every x. We observe only an *i.i.d.* sample $X_1, X_2, \ldots$ from $P_0 \in \mathscr{P}$ and the corresponding $Z_1, Z_2, \ldots$ remain hidden. The model $\mathscr{P}$ then consists of distributions P_F for X with Lebesgue densities of the form,

$$p_F(x) = \int_{\mathbb{R}} p(x|z)\, dF(z),$$

where the parameter F represents the unknown distribution of Z. For reasons explained below, assume that $Z \in [0,1]$, so that the space $\mathscr{D}$ of all distributions on $[0,1]$ is compact in Prokhorov's weak topology. Note that for any fixed $x \in \mathbb{R}$, $F \mapsto p_F(x)$ is weakly continuous. By Scheffé's lemma this pointwise continuity implies weak-to-total-variational continuity of the map $F \mapsto P_F$, which is equivalent to weak-to-Hellinger continuity. Since $\mathscr{D}$ is weakly compact, this implies that the model $\mathscr{P}$ is Hellinger compact (and consequently, Hellinger entropy numbers are all finite). Additionally we make the assumption that the L_2-condition (3.5) is satisfied; for example in the well-known normal location mixture model, where $X|Z = z$ is distributed normally with mean z [14], the family $\mathscr{P} = \{p_F : F \in \mathscr{D}\}$ is contained in an envelope that allows straightforward verification of (3.5) (for details, see the proof of theorem 3.2 in [21]).

With finite entropy numbers and (3.5) established, note that any prior Π on $\mathscr{D}$ that is Borel for the weak topology induces a prior that is Borel for the Hellinger topology on the model $\mathscr{P}$. If the weak support of Π equals $\mathscr{D}$ then the induced Hellinger support includes $\mathscr{P}$. For instance, a Dirichlet prior for F with base measure of full support on $[0,1]$ will suffice to conclude from theorem 3.4 that the posterior is Hellinger consistent. Other priors on $\mathscr{D}$, like Gibbs-type measures of full weak support [6] would also suffice. In fact, consistency applies for any bounded, continuous (and some semi-continuous) kernel(s) $x \mapsto p(x|z)$ such that mixture densities satisfy (3.5).

To conclude we demonstrate that the approach advocated in this paper applies in counterexamples 1.1 and 1.2.

Example 7.2. Assume that the width of the support of p_0 is equal to one. The model consists of densities η supported on $[0,1]$ shifted over θ in $\mathbb{R}$,

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

Consider H with some prior Π_H and a prior Π_Θ 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 $\eta, \eta' \in H$. As noted, KL-neighbourhoods do not have any extent in the θ-direction, however, the construction of example 6.3 remains applicable. 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 supports implies that,

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

for all $P \in \mathrm{co}(V_\pm)$, based on (6.2). If H is chosen such that for all $P \in V_\pm$, $\sup_{Q \in B_\pm} P_0(p/q) < \infty$, then (6.3) follows (even *regardless of the prior on* H, which is remarkable). Larger spaces H can be considered if the sets $B_\pm$ are restricted appropriately while maintaining $\Pi(B_\pm) > 0$. Conclude that for the estimation of an unknown $\theta_0 \in \mathbb{R}$, Schwartz's theorem does not apply, while example 6.3 remains in effect.

Example 7.3. Recall example 1.2: the sample $X_1, X_2, \ldots$ consists of *i.i.d.* integers from a distribution P_a, $(a \geq 1)$, defined by,

$$p_a(k) = P_a(X = k) = \frac{1}{Z_a} \frac{1}{k^a (\log k)^3},$$

for all $k \geq 2$ (with normalization Z_a). The parameter a is smooth and the Fisher information is non-singular, so a can be estimated at parametric rate, but as noted, there exists no prior for the parameter a such that condition (1.2) can be satisfied for all P_0 in the model. Corollary 5.3 remains valid, however, and demonstrates that the posterior converges at $\sqrt{n}$-rate. Because corollary 5.3 is formulated for totally-bounded parameter spaces only, without a negligiblility condition like (5.2), we restrict the parameter a to a bounded interval $I = [1, L]$, for some $L > 1$. (However the result below is expected to hold also without this restriction.)

For any rate ϵ_n that is slower than $n^{-1/2}$, write $\epsilon_n = n^{-1/2} M_n$, with $M_n \to \infty$ and note that we only have to consider M_n that diverge very slowly, *i.e.* ϵ_n that are arbitrarily close to the parametric rate. Also note that there exist constants $M_1, M_2 > 0$ such that,

$$M_1^2 (b - a)^2 \leq -P_a \log(p_b/p_a) \leq M_2^2 (b - a)^2, \tag{7.1}$$

Define $V_n = \{P : H(P, P_0) \geq M\epsilon_n\}$ for some $M > 0$. We cover V_n with Hellinger balls $V_{n,i}$ $(1 \leq i \leq N_n)$ of radius $\frac{1}{2} M\epsilon_n$. Note that $H(P_b, P_c) \leq M_2 |c - b|$ for all $b, c \in I$, so $N_n = N(\frac{1}{2} M\epsilon_n, \mathscr{P}, H) \leq N((M/2M_2)\epsilon_n, I, |\cdot|) \leq 2M_2 |I|/(M\epsilon_n)$.

Defining also $B_n = \{Q : -P_0 \log(dQ/dP_0) < \epsilon_n^2\}$, we note that $B_n \subset \{P_b : |b - a| < \epsilon_n/M_1\}$. Hence, for any $a \geq 1$, any $P_c \in V_n$ and any P_b with $|b - a| < \epsilon_n/M_2$, we have,

$$P_a\Big(\frac{p_c}{p_b}\Big) = \frac{Z_b}{Z_c} \sum_{k \geq 2} \frac{1}{Z_a} \frac{1}{k^a (\log k)^3} \frac{k^b}{k^c} \leq \frac{Z_b}{Z_c},$$

because $b < c$ if M is chosen large enough. Since I is compact and $I \to \mathbb{R} : b \mapsto Z_b$ is continuous, $b \mapsto Z_b$ is bounded, so that for every $P_c \in V_{n,i}$, the integrability condition $\sup\{P_a(dP_c/dQ) : Q \in B_n\} < \infty$ holds. Due to the second inequality

of (7.1), any Borel prior on I of full support is a KL-prior. More specifically, if we choose the uniform prior on I, $\Pi(B_n) \geq \Pi(b \in I : |b-a| \leq \epsilon_n/M_2) \geq (|I|M_2)^{-1}\epsilon_n$. Conclude that the conditions of corollary 5.3 are met for any rate above $n^{-1/2}$, so the posterior for a converges at parametric rate.

Appendix A: Two notes on supports

Remark A.1. Throughout the main text, 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 marginal for the *i.i.d.* sample. Because the central point of lemma 2.3 concerns only P_0-expectations, an indicator $1_{\{p_0>0\}}(x)$ may be thought of as implicitly present in all calculations; because we look at moments of p/q, an indicator $1_{\{p>0\}}(x)$ can also be thought of as implicit; 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 $1_{\{q>0\}}(x)$ is also implicit.

Remark A.2. It is easy to misinterpret KL-divergence in cases where distributions have mismatching domains. Le Cam (1986) and Le Cam-Yang (1990) make it explicit that where ever a RN-derivative dP_θ/dP_{θ_0} is used, it concerns *only the P_{θ_0}-dominated part of P_θ*. With that in mind we define the KL-divergence as,

$$KL(\theta_0;\theta) = -P_{\theta_0}\log\frac{dP_\theta}{dP_{\theta_0}} = \int_{\{x:p_{\theta_0}(x)>0\}} \log\frac{p_\theta}{p_{\theta_0}}(x)\,dx, \tag{A.1}$$

which connects properly with Schwartz's proof for sufficiency of prior mass (since with a P_{θ_0}-distributed sample, log-likelihood ratios converge to KL-divergences of the form (A.1)).

Appendix B: Some properties of Hellinger transforms

Given two finite measures μ and ν, the Hellinger transform is defined as follows for all $0 \leq \alpha \leq 1$:

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

where σ is a σ-finite measure that dominates both μ and ν (*e.g.* $\sigma = \mu+\nu$).

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

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

Properties of the Hellinger transform that are used in the main text are listed in the following lemma, which extends lemma 6.3 in [21].

Lemma B.1. *For a probability measure P and a finite measure ν (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:*

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

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

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

(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 $y > 0$, is decreasing for $y < 1$, increasing for $y > 1$ and constant for $y = 1$. By monotone convergence, as $\alpha \downarrow 0$,

$$\nu\Big(\frac{p}{r}\Big)^\alpha 1_{\{0<p<r\}} \uparrow \nu\Big(\frac{p}{r}\Big)^0 1_{\{0<p<r\}} = \nu(0<p<r).$$

By the dominated convergence theorem (note that $(p/r)^{1/2}1_{\{p\geq r\}}$ upper-bounds $(p/r)^\alpha 1_{\{p\geq r\}}$ for $\alpha \leq 1/2$) we have,

$$\nu\Big(\frac{p}{r}\Big)^\alpha 1_{\{p\geq r\}} \to \nu\Big(\frac{p}{r}\Big)^0 1_{\{p\geq r\}} = \nu(p\geq r),$$

as $\alpha \downarrow 0$. Combining the two preceding displays, we have $\rho_\alpha(\nu,P) = P(p/r)^\alpha \to P(r>0)$ as $\alpha \downarrow 0$. Upon substitution of α 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 $y \geq 0$,

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

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

$$\Big(\frac{r}{p}\Big)^{\alpha_0} \frac{e^{(\alpha-\alpha_0)\log(r/p)} - 1}{\alpha - \alpha_0} \downarrow \Big(\frac{r}{p}\Big)^{\alpha_0} \log\Big(\frac{r}{p}\Big), \quad \text{as } \alpha \downarrow \alpha_0,$$

and is bounded above by $0 \vee \epsilon^{-1}(r/p)^{\alpha_0+2\epsilon} 1_{r\geq p}$ for small $\epsilon > \alpha - \alpha_0 > 0$, which is P-integrable for small enough ϵ. We conclude that,

$$\frac{1}{\alpha-\alpha_0}(\rho_\alpha(\nu,P) - \rho_{\alpha_0}(\nu,P)) \downarrow P\,1_{r>0}\Big(\frac{r}{p}\Big)^{\alpha_0} \log\Big(\frac{r}{p}\Big), \quad \text{as } \alpha \downarrow \alpha_0,$$

by monotone convergence. For $\alpha < \alpha_0$ a similar argument can be given. Convexity of $\alpha \mapsto P\,1_{r>0}\,(r/p)^\alpha \log(r/p)$ implies continuity of the derivative. □

Appendix C: Proofs

This section contains all proofs of theorems and lemmas in the main text.

C.1. Proofs for section 2

Proof of proposition 2.1. For any $A \in \sigma_n := \sigma(X_1, \ldots, X_n)$ and any model subset U' such that $\Pi(U') > 0$,

$$P_0^n(A) \leq \int P^n(A)\, d\Pi(P|U') + \sup_{P \in U'} |P^n(A) - P_0^n(A)|.$$

Now assume that A is a null-set of P_n^Π; since $\Pi(U') > 0$, $\int P^n(A)\, d\Pi(P|U') = 0$. For some $\epsilon > 0$, take $U' = \{P : |P^n(A) - P_0^n(A)| < \epsilon\}$, note that U' contains a total-variational neighbourhood and therefore a Hellinger neighbourhood, to conclude that $P_0^n(A) < \epsilon$ for all $\epsilon > 0$. Since every Hellinger ball contains a Kullback-Leibler-ball, (1.1) implies that $\Pi(U) > 0$ for every Hellinger ball U. □

Proof of lemma 2.3. For a set V covered by measurable $V_1, \ldots, V_N$, almost-sure convergence per individual V_i implies the assertion. So we fix some $1 \leq i \leq N$ and note that,

$$P_0^n \Pi(V_i|X_1, \ldots, X_n) \leq P_0^n \phi_{i,n} + P_0^n \Pi(V_i|X_1, \ldots, X_n)(1 - \phi_{i,n}).$$

By Fubini's theorem,

$$\begin{aligned} P_0^n \Pi(V_i|X_1, \ldots, X_n)(1 - \phi_{i,n}) &= P_0^n \int_{V_i} \frac{dP^n}{dP_n^\Pi}(1 - \phi_{i,n})\, d\Pi(P) \\ &\leq \Pi(V_i) \sup_{P \in V_i} P_0^n \frac{dP^n}{dP_n^\Pi}(1 - \phi_{i,n}). \end{aligned} \tag{C.1}$$

From (2.3) we conclude that $P_0^n \Pi(V_i|X_1, \ldots, X_n) \leq e^{-nD_i}$, for large enough n. Apply Markov's inequality to find that,

$$P_0^n \left(\Pi(V_i|X_1, \ldots, X_n) \geq e^{-\frac{n}{2} D_i} \right) \leq e^{-\frac{n}{2} D_i},$$

so that the first Borel-Cantelli lemma guarantees,

$$P_0^\infty \left(\limsup_{n \to \infty} \left(\Pi(V_i|X_1, \ldots, X_n) - e^{-\frac{n}{2} D_i} \right) > 0 \right) = 0.$$

Replicating this argument for all $1 \leq i \leq N$, assertion (2.4) follows. □

Example C.1. Suppose that we wish to prove consistency relative to some metric d on $\mathscr{P}$ but coverings of the model by d-balls are not finite. Then we may try the following construction: for $\epsilon > 0$, we define $W = \{P \in \mathscr{P} : d(P, P_0) > \epsilon\}$ and $W_k = \{P \in \mathscr{P} : 2^{k-1}\epsilon \leq d(P, P_0) < 2^k \epsilon\}$, $(k \geq 1)$. Assume that the covering numbers $N_k := N_k(2^{k-2}\epsilon, W_k, d)$ of the model subsets W_k (related to the so-called *Le Cam dimension* of the model [26]) are finite. Let $V_{k,1}, \ldots, V_{k,N_k}$ be d-balls of radius $2^{k-2}\epsilon$ covering W_k. Assume that for every d-ball $V_{k,i}$, $(1 \leq$

$i \leq N_k$), there exists a test sequence $(\phi_{k,i,n})_{n\geq 1}$ such that (2.3) is satisfied with $D_{k,i} \geq d^2(V_{k,i}, P_0)$. Then, for every $n \geq 1$,

$$P_0^n \Pi(W|X_1,\ldots,X_n) \leq \sum_{k\geq 1}\sum_{1\leq i\leq N_k} P_0^n \Pi(V_{k,i}|X_1,\ldots,X_n) \leq \sum_{k\geq 1} N_k\, e^{-2^{2k-4}\, n\, \epsilon^2}.$$

If we show that the right-hand side goes to zero as $n \to \infty$, the posterior is d-consistent.

Proof of lemma 2.4. According to lemma 6.1 of [21] (see the minimax theorem 45.8 in Strasser (1985) [36] and [28], p. 478) there exists a test (ϕ_n) that minimizes the *l.h.s.* of (2.5) and,

$$\sup_{P\in V}\Big(P_0^n\phi_n + P_0^n\frac{dP^n}{dP_n^\Pi}(1-\phi_n)\Big) \leq \sup_{P_n\in\mathrm{co}(V^n)}\ \inf_{\phi}\ \Big(P_0^n\phi + P_0^n\frac{dP_n}{dP_n^\Pi}(1-\phi)\Big).$$

The infimal ϕ equals $1_{\{dP_n/dP_n^\Pi>1\}}$. For any $\alpha\in[0,1]$,

$$\int 1_{\{dP_n/dP_n^\Pi>1\}}\, dP_0^n + \int \frac{dP_n}{dP_n^\Pi}\, 1_{\{dP_n/dP_n^\Pi\leq 1\}}\, dP_0^n \leq \int\Big(\frac{dP_n}{dP_n^\Pi}\Big)^\alpha dP_0^n,$$

which enables an upper-bound for testing power,

$$\sup_{P_n\in\mathrm{co}(V^n)}\Big(P_0^n\,\phi + P_0^n\frac{dP_n}{dP_n^\Pi}\,(1-\phi)\Big) \leq \sup_{P_n\in\mathrm{co}(V^n)}\ \inf_{0\leq\alpha\leq 1}\ P_0^n\Big(\frac{dP_n}{dP_n^\Pi}\Big)^\alpha,$$

in terms of the Hellinger transform. □

The following lemma reduces the testing criterion to an n-independent condition, *c.f.* (1.5).

Proof of lemma 2.5. Let $0\leq\alpha\leq 1$ be given. Note that for all $n\geq 1$, $P_n^\Pi(A) \geq \Pi(B)\,P_n^{\Pi|B}(A)$ 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,

$$P_0^n\Big(\frac{dP_n}{dP_n^\Pi}\Big)^\alpha \leq \Pi(B)^{-\alpha}\,P_0^n\Big(\frac{dP_n}{dP_n^{\Pi|B}}\Big)^\alpha \leq \Pi(B)^{-\alpha}\,P_0^n\int\Big(\frac{dP_n}{dQ^n}\Big)^\alpha d\Pi(Q|B). \tag{C.2}$$

With the use of Fubini's theorem and lemma 6.2 in Kleijn and van der Vaart (2006) [21] which says that Hellinger transforms factorize when taken over convex hulls of products, we find:

$$\begin{aligned}\sup_{P_n\in\mathrm{co}(V^n)}\ \inf_{0\leq\alpha\leq 1}\ \Pi(B)^{-\alpha}\int P_0^n\Big(\frac{dP_n}{dQ^n}\Big)^\alpha d\Pi(Q|B)&\\ \leq \inf_{0\leq\alpha\leq 1}\Pi(B)^{-\alpha}\int\sup_{P_n\in\mathrm{co}(V^n)}P_0^n\Big(\frac{dP_n}{dQ^n}\Big)^\alpha d\Pi(Q|B)&\\ \leq \inf_{0\leq\alpha\leq 1}\Pi(B)^{-\alpha}\int\Big[\sup_{P\in\mathrm{co}(V)}P_0\Big(\frac{dP}{dQ}\Big)^\alpha\Big]^n d\Pi(Q|B).&\end{aligned}$$

Applying (C.2) with $\alpha = 1$, $P_n = P^n$, and using that for all $P \in V$, $P_0(dP/dQ)$ is bounded uniformly over B, we see that also $P_0^n(dP^n/dP_n^{\Pi}) < \infty$. By (2.5), we obtain (2.7). □

C.2. Proofs for section 3

Proof of lemma 3.1. Assume that (3.2) holds. Lemma B.1 says that $\alpha \mapsto P_0(dP/dQ)^\alpha$ is convex and continuously differentiable on $(0, a)$. So for all $\alpha \in (0, a)$,

$$\sup_{Q\in B}\sup_{P\in W} P_0\Big(\frac{dP}{dQ}\Big)^\alpha \le 1 + \alpha \sup_{Q\in B}\sup_{P\in W} P_0\Big(\frac{dP}{dQ}\Big)^\alpha \log\frac{dP}{dQ}. \tag{C.3}$$

The function

$$\alpha \mapsto \sup_{Q\in B}\sup_{P\in W} P_0\Big(\frac{dP}{dQ}\Big)^\alpha \log\frac{dP}{dQ},$$

is convex (hence continuous on $(0, a)$ and upper-semi-continuous at 0) and, due to (3.2), 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 (C.3) there exists an $\alpha_0 \in [0, 1]$ such that $\sup_{P,Q} P_0(dP/dQ)^{\alpha_0} < 1$ and we conclude that (3.1) holds. Conversely, assume that (3.2) does not hold. Let $P \in W$, $Q \in B$ and $\alpha \in [0, 1]$ be given; by Jensen's inequality,

$$P_0\Big(\frac{dP}{dQ}\Big)^\alpha \ge \exp\Big(\alpha P_0 \log\frac{dP}{dQ}\Big) = \exp\Big(\alpha\Big(P_0\log\frac{dP}{dP_0} - P_0\log\frac{dQ}{dP_0}\Big)\Big).$$

Therefore,

$$\sup_{Q\in B}\sup_{P\in W} P_0\Big(\frac{dP}{dQ}\Big)^\alpha \ge \exp\Big(\alpha \sup_{Q\in B} -P_0\log\frac{dQ}{dP_0}\Big)\exp\Big(-\alpha \inf_{P\in W} -P_0\log\frac{dP}{dP_0}\Big),$$

which is greater than or equal to one for all $\alpha \in [0, 1]$. □

Proof of theorem 3.2. For every $1 \le i \le N$, there exists a constant $b_i > 0$ such that for every $P \in W_i := \mathrm{co}(V_i)$, $-P_0\log(dP/dP_0) > b_i$. Denoting the Kullback-Leibler radius of B by $b > 0$, we define $B_i = \{P \in \mathscr{P} : -P_0\log(dP/dP_0) < b_i \wedge b\}$ to satisfy (3.2). Note that, by assumption, $\Pi(B_i) > 0$ and $\sup_{Q\in B_i} P_0(dP/dQ) < \infty$ for all $P \in \mathscr{P}$. Every total-variational neighbourhood of P_0 contains a Kullback-Leibler neighbourhood, so combination of lemma 3.1, lemma 2.5 and theorem 1.3 proves posterior consistency in Kullback-Leibler divergence *c.f.* (3.4). □

Proof of theorem 3.4. Proposition 2.1 guarantees that $P_0^n \ll P_n^{\Pi}$, for all $n \ge 1$. For given $\epsilon > 0$, let V denote $\{P \in \mathscr{P} : H(P, P_0) > 2\epsilon\}$. Since $\mathscr{P}$ is totally bounded in the Hellinger metric, there exist $P_1, \ldots, P_N$ such that the model subsets $V_i = \{P \in \mathscr{P} : H(P, P_i) < \epsilon\}$ form a cover of V. On the basis of the constant L of (3.5), define $B = \{Q \in \mathscr{P} : H(Q, P_0) < \epsilon^2/(4L) \wedge \epsilon'\}$, where ϵ'

is the Hellinger radius of B'. Since Hellinger balls are convex, we have for all $1 \leq i \leq N$,

$$\sup_{P \in \mathrm{co}(V_i)} \sup_{Q \in B} P_0 \Big(\frac{p}{q}\Big)^{1/2} \leq 1 - \tfrac{1}{4}\epsilon^2 \leq e^{-\frac{1}{4}\epsilon^2}.$$

By the Cauchy-Schwarz inequality, for every $P \in V$,

$$\sup_{Q \in B} P_0 \Big(\frac{p}{q}\Big) \leq \sup_{Q \in B} \Big\| \frac{p_0}{q} \Big\|_{2,Q} \Big\| \frac{p}{q} \Big\|_{2,Q} < L^2 < \infty.$$

According to lemmas 2.5 and 2.3, the posterior is consistent. □

Proof of theorem 3.5. Reasoning like in the introduction of subsection 3.2, but now with Hölder's inequality, one finds,

$$P_0 \Big(\frac{p}{q}\Big)^{1/r} \leq \rho_{1/r}(P, P_0) + d_r(P_0, Q) \left(P_0 (p/q)^{s/r}\right)^{1/s}$$

Let $\epsilon > 0$ be given and let V be the complement of a d_r-ball of radius 2ϵ. Cover V by N d_r-balls $V_1, \ldots, V_N$ of radii ϵ (which are convex) and note that for all $1 \leq i \leq N$ and $P \in V_i$, $d_r(P, P_0) \geq \epsilon$. It is shown in the corollary of theorem 1 of [37] that,

$$\rho_{1/r}(P_0, P) \leq \Big(\frac{2(r-1)}{r}\Big)^{1/2} \big(1 - \tfrac{1}{4}\, d_r(P, P_0)^{2r}\big)^{1/2},$$

and with $K = (2(r-1)/r)^{1/2}$, it follows that,

$$P_0 \Big(\frac{p}{q}\Big)^{1/r} \leq K(1 - \tfrac{1}{4}\, \epsilon^{2r})^{1/2} + L^{s/r \wedge 1}\, d_r(P_0, Q).$$

Since $(1-x)^{1/2} \leq 1 - \frac{1}{2}x$ for $x \in (0,1)$, the choice $\delta = (K/16) L^{-(s/r \wedge 1)} \epsilon^{2r}$ in (3.6) guarantees that $P_0(p/q)^{1/r} \leq K(1 - (1/16)\epsilon^{2r})$ for all $1 \leq i \leq N$, $P \in V_i$ and $Q \in B$. If $s \geq r$, Jensen's inequality implies that $\sup_{Q \in B} P_0(p/q) < \infty$; if $s < r$, $\sup_{Q \in B} P_0(p/q) < \infty$ by assumption. According to lemma 2.5 and lemma 2.3, the posterior is consistent. □

Proof of lemma 3.7. Continuity implies that every Kullback-Leibler ball around P_0 contains an open neighbourhood of P_0. □

C.3. Proofs for section 4

Proof of theorem 4.1. For given V and $n \geq 1$, denote the cover of condition *(i.)* by $V_1, \ldots, V_{N_n}$ with tests $\phi_{i,n}$, $1 \leq i \leq N_n$. Define $\psi_n = \max_i \phi_{i,n}$ and decompose the n-th posterior for V as follows,

$$\begin{aligned} P_0^n \Pi(V | X_1, \ldots, X_n) \leq\ & P_0^n \psi_n + \\ & P_0^n \Pi(V \cap \mathscr{P}_n | X_1, \ldots, X_n)(1 - \psi_n) + P_0^n \Pi(V \setminus \mathscr{P}_n | X_1, \ldots, X_n). \end{aligned}$$

Note $P_0^n\psi_n \le \sum_{i=1}^{N_n} P_0^n\phi_{i,n} \le N_n \exp(-nL) \le \exp(-\frac{1}{2}nL)$, and,

$$\begin{aligned}
&P_0^n\Pi(V\cap\mathscr{P}_n|X_1,\dots,X_n)(1-\psi_n)\\
&\qquad\le \sum_{i=1}^{N_n} P_0^n\Pi(V_i\cap\mathscr{P}_n|X_1,\dots,X_n)(1-\psi_n)\\
&\qquad\le \sum_{i=1}^{N_n} P_0^n\Pi(V_i\cap\mathscr{P}_n|X_1,\dots,X_n)(1-\phi_{i,n})\\
&\qquad\le \sum_{i=1}^{N_n} \sup_{P\in V_i} P_0^n\frac{dP^n}{dP_n^{\Pi}}(1-\phi_{i,n}) \le N_n\, e^{-nL} \le e^{-\frac{1}{2}nL},
\end{aligned}$$

where we have followed the steps in the proof of theorem 1.3. Using again the local prior predictive distribution $P_n^{\Pi|B}$ of (2.6), the third term satisfies,

$$\begin{aligned}
&P_0^n\Pi(V\setminus\mathscr{P}_n|X_1,\dots,X_n) \le P_0^n\Pi(\mathscr{P}\setminus\mathscr{P}_n|X_1,\dots,X_n)\\
&\quad= \int_{\mathscr{P}\setminus\mathscr{P}_n} P_0^n\Big(\frac{dP^n}{dP_n^{\Pi}}\Big)d\Pi(P) \le \frac{1}{\Pi(B)}\int_{\mathscr{P}\setminus\mathscr{P}_n} P_0^n\Big(\frac{dP^n}{dP_n^{\Pi|B}}\Big)d\Pi(P)\\
&\quad\le \frac{\Pi(\mathscr{P}\setminus\mathscr{P}_n)}{\Pi(B)} \sup_{P\in\mathscr{P}\setminus\mathscr{P}_n}\sup_{Q\in B}\big[P_0(p/q)\big]^n \le \Pi(B)^{-1}e^{-\frac{1}{2}Kn}.
\end{aligned}$$

Like at the end of the proof of lemma 2.3, an application of the Borel-Cantelli proves the assertion. □

Proof of theorem 4.3. By monotone convergence,

$$P_0^n\Pi(V|X_1,\dots,X_n) \le P_0^n\Pi\big(\cup_{i\ge1}V_i \;\big|\; X_1,\dots,X_n\big) \le \sum_{i\ge1} P_0^n\Pi(V_i|X_1,\dots,X_n).$$

We treat the terms in the sum separately with the help of test sequences $(\phi_{i,n})$, for all $i\ge1$, following the proof of lemma 2.3:

$$P_0^n\Pi(V|X_1,\dots,X_n) \le \sum_{i\ge1}\Big(P_0^n\phi_{i,n} + \Pi(V_i)\sup_{P\in V_i} P_0^n\frac{dP^n}{dP_n^{\Pi}}(1-\phi_{i,n})\Big). \tag{C.4}$$

Like in the proof of lemma 2.5, the assumptions that $\sup_{Q\in B_i} P_0(dP/dQ)<\infty$ and $\Pi(B_i)>0$, imply that $P_0^n(dP^n/dP_n^{\Pi})<\infty$, for all $P\in V_i$. So $\phi_{i,n}$ can be chosen in such a way that,

$$\begin{aligned}
&P_0^n\phi_{i,n} + \Pi(V_i)\sup_{P\in V_i} P_0^n\frac{dP^n}{dP_n^{\Pi}}(1-\phi_{i,n})\\
&\qquad= \sup_{P_n\in\mathrm{co}(V_i^n)}\ \inf_{\phi}\ \Big(P_0^n\phi + \Pi(V_i)\,P_0^n\frac{dP_n}{dP_n^{\Pi}}(1-\phi)\Big)
\end{aligned} \tag{C.5}$$

by the minimax theorem. To minimize the *r.h.s.*, choose ϕ as follows,

$$\phi(X_1,\dots,X_n) = 1\Big\{(X_1,\dots,X_n)\in\mathscr{X}^n \,:\, \Pi(V_i)\frac{dP^n}{dP_n^{\Pi}}(X_1,\dots,X_n) > 1\Big\},$$

and follow the proof of lemma 2.5 to conclude that the *r.h.s.* of (C.5) is upper bounded by,

$$\inf_{0\leq\alpha\leq 1}\frac{\Pi(V_i)^\alpha}{\Pi(B_i)^\alpha}\Big[\sup_{P\in\mathrm{co}(V_i)}\sup_{Q\in B_i}P_0\Big(\frac{dP}{dQ}\Big)^\alpha\Big]^n.$$

Combine with (C.4) to arrive at the assertion. □

Proof of corollary 4.4. Fix $\alpha = 1/2$ and $B_i = B$ in (4.2) and use (4.3) to arrive at,

$$P_0^n\Pi(V|X_1,\ldots,X_n)\leq\Pi(B)^{-1/2}(1-\gamma)^n\sum_{i\geq 1}\Pi(V_i)^{1/2},$$

(for some constant $0<\gamma<1$) which goes to zero at geometric rate, if (4.4) holds, so that $\Pi(V|X_1,\ldots,X_n)\xrightarrow{P_0\text{-a.s.}}0$. □

Proof of corollary 4.5. Given $\epsilon>0$, define $V=\{P: H(P,P_0)\geq\epsilon\}$ and let $\{V_i: i\geq 1\}$ denote a countable collection of Hellinger balls of radius $\frac{1}{4}\epsilon$ with centre points in V that cover V, so that,

$$\inf_{i\geq 1}\inf_{P\in\mathrm{co}(V_i)}H(P,P_0)\geq\tfrac{3}{4}\epsilon. \tag{C.6}$$

Inspection of the proof of lemma 3.1 reveals that it generalizes to the statement that:

$$\inf_{0\leq\alpha\leq 1}\sup_{i\geq 1}\sup_{Q\in B}\sup_{P\in\mathrm{co}(V_i)}P_0\Big(\frac{dP}{dQ}\Big)^\alpha<1,$$

if and only if, B and the $\mathrm{co}(V_i)$ are KL-separated:

$$\sup_{Q\in B}-P_0\log\frac{dQ}{dP_0}<\inf_{i\geq 1}\inf_{P\in\mathrm{co}(V_i)}-P_0\log\frac{dP}{dP_0}.$$

Note that (C.6) serves as a lower bound for the *r.h.s.* of the previous display, which enables the choice $B=\{P\in\mathscr{P}: -P_0\log(p/p_0)<\epsilon/4\}$ to guarantee that there exist constants $0<\alpha',\gamma<1$ such that,

$$P_0^n\Pi(V|X_1,\ldots,X_n)\leq\Pi(B)^{-\alpha'}(1-\gamma)^{n\alpha'}\sum_{i\geq 1}\Pi(V_i)^{\alpha'},$$

which goes to zero since $\Pi(B)>0$ and the sum is finite by assumption. □

C.4. Proofs for section 5

Proof of theorem 5.1. Fix $n\geq 1$ large enough to satisfy conditions *(i)* and *(ii)*. According to lemma 2.5, there exist test functions $\phi_{n,i}:\mathcal{X}^n\to[0,1]$ for all $1\leq i\leq N_n$, such that, for all $\alpha\in[0,1]$,

$$P_0^n\phi_{n,i}+\sup_{P\in V_{n,i}}P_0^n\frac{dP^n}{d\Pi_n^\Pi}(1-\phi_{n,i})\leq\Pi(B_n)^{-\alpha}\pi_{P_0}(\mathrm{co}(V_{n,i}),B_n;\alpha)^n.$$

Define $\psi_n = \max_i \phi_{n,i}$ and decompose the n-th posterior for $V_n = \{P \in \mathscr{P} : d(P, P_0) \geq \epsilon_n\}$, as follows,

$$P_0^n \Pi(V_n | X_1, \ldots, X_n) \leq P_0^n \psi_n + P_0^n \Pi(V_n \cap \mathscr{P}_n | X_1, \ldots, X_n)(1 - \psi_n) + P_0^n \Pi(\mathscr{P} \setminus \mathscr{P}_n | X_1, \ldots, X_n).$$

The first term is upper-bounded as follows,

$$P_0^n \psi_n \leq \sum_{i=1}^{N_n} P_0^n \phi_{n,i} \leq N_n \Pi(B_n)^{-\alpha} \pi_{P_0}(\mathrm{co}(V_{n,i}), B_n; \alpha)^n \leq e^{(\frac{\alpha K}{2} - L) n \epsilon^2},$$

for all $\alpha \in [0, 1]$. The second term is bounded by,

$$\begin{aligned} P_0^n \Pi(V_n \cap \mathscr{P}_n | X_1, \ldots, X_n)(1 - \psi_n) &\leq \sum_{i=1}^{N_n} P_0^n \Pi(V_{n,i} | X_1, \ldots, X_n)(1 - \psi_n) \\ &\leq \sum_{i=1}^{N_n} P_0^n \Pi(V_{n,i} | X_1, \ldots, X_n)(1 - \phi_{n,i}) \leq \sum_{i=1}^{N_n} \sup_{P \in V_{n,i}} P_0^n \frac{dP^n}{d\Pi_n^{\Pi}}(1 - \phi_{n,i}) \\ &\leq N_n \Pi(B_n)^{-\alpha} \pi_{P_0}(\mathrm{co}(V_{n,i}), B_n; \alpha)^n \leq e^{(\frac{\alpha K}{2} - L) n \epsilon^2} \end{aligned}$$

for all $\alpha \in [0, 1]$. The third term requires condition *(ii)*,

$$\begin{aligned} P_0^n \Pi(\mathscr{P} \setminus \mathscr{P}_n | X_1, \ldots, X_n) &= \int_{\mathscr{P} \setminus \mathscr{P}_n} P_0^n \Big(\frac{dP^n}{dP_n^{\Pi}} \Big) d\Pi(P) \leq \frac{1}{\Pi(B_n)} \int_{\mathscr{P} \setminus \mathscr{P}_n} P_0^n \Big(\frac{dP^n}{dP_n^{\Pi | B_n}} \Big) d\Pi(P) \\ &\leq \frac{\Pi(\mathscr{P} \setminus \mathscr{P}_n)}{\Pi(B_n)} \sup_{P \in \mathscr{P} \setminus \mathscr{P}_n} \sup_{Q \in B_n} \big[P_0(dP/dQ) \big]^n \leq e^{-\frac{K}{2} n \epsilon_n^2}. \end{aligned}$$

Choosing $0 < \alpha < 2L/K$, all three contributions go to zero as $n \to \infty$. □

Proof of lemma 5.2. Assume that (5.3) holds. Like in the proof of lemma 3.1, we have for all $\alpha \in (0, a)$,

$$\sup_{Q \in B} \sup_{P \in W} P_0 \Big(\frac{dP}{dQ} \Big)^{\alpha} \leq 1 + \alpha\, z(\alpha), \tag{C.7}$$

where the function $z : [0, a) \to \mathbb{R}$ is given by,

$$z(\alpha) = \sup_{Q \in B} \sup_{P \in W} P_0 \Big(\frac{dP}{dQ} \Big)^{\alpha} \log \frac{dP}{dQ}.$$

The function z is convex and increasing, hence continuous on $(0, a)$ and upper-semicontinuous at $a = 0$ and maximal at $\alpha = a$. Clearly, we have,

$$\lim_{\alpha \downarrow 0} z(\alpha) \leq \sup_{Q \in B} \sup_{P \in W} P_0 \log \frac{dP}{dQ} = \sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} - \inf_{P \in W} -P_0 \log \frac{dP}{dP_0},$$

and the right-hand side is less than or equal to $-\Delta$. By the continuity of z, there exists an $a' \in (0,a)$ such that, $z(\alpha) \leq -\frac{1}{2}\Delta$ for all $\alpha \in (0,a')$. Combining (C.7) with the latter conclusion, we see that, for all $\alpha \in (0,a')$,

$$\pi_{P_0}(B,W) = \inf_{0\leq\alpha\leq 1} \sup_{Q\in B} \sup_{P\in W} P_0\Big(\frac{dP}{dQ}\Big)^{\alpha} \leq 1 - \tfrac{1}{2}\alpha\Delta \leq e^{-\alpha''\Delta},$$

where $\alpha'' = \frac{1}{2}\alpha$. Conversely, assume that (5.3) does not hold. Let $P \in W$, $Q \in B$ and $\alpha \in [0,1]$ be given; by Jensen's inequality,

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

Therefore, for all $\alpha \in (0,1)$,

$$\sup_{Q\in B} \sup_{P\in W} P_0\Big(\frac{dP}{dQ}\Big)^{\alpha} > e^{-\alpha\Delta}. \qquad \square$$

Proof of corollary 5.3. Take $\mathscr{P}_n = \mathscr{P}$ for all $n \geq 1$. Note that (5.4) implies that Π is a Kullback-Leibler prior, which implies that $P_0^n \ll P_n^{\Pi}$. Let $V_n = \{P \in \mathscr{P} : H(P,P_0) \geq \epsilon_n\}$ and $B_n = \{P \in \mathscr{P} : -P_0 \log(dP/dP_0) < \epsilon_n^2/8\}$. By condition *(i)* there is a cover of V_n consisting of Hellinger balls of radii $\epsilon_n/2$ of order $N_n = N(\epsilon_n, \mathscr{P}, H) \leq \exp(Ln\epsilon_n^2)$. Note that for every $1 \leq i \leq N_n$ and all $P \in \mathrm{co}(V_{n,i})$, we have $-P_0 \log(dP/dP_0) \geq H^2(P,P_0) \geq (H(V_n,P_0) - \epsilon_n/2)^2 = \epsilon_n^2/4$, while $-P_0 \log(dQ/dP_0) \leq \epsilon_n^2/8$ for all $Q \in B_n$. According to lemma 5.2, the separation in Kullback-Leibler divergence between B_n and V_n implies that $\pi_{P_0}\big(\mathrm{co}(V_{n,i}), B_n\big) \leq e^{-\alpha\epsilon_n^2}$ for some $\alpha > 0$. Possibly after rescaling of ϵ_n by an n-independent constant (which leads to larger α, effectively), π_{P_0} satisfies condition (5.1). The assertion then follows from theorem 5.1. $\square$

C.5. Proofs for section 6

Proof of theorem 6.2. Let $\epsilon > 0$ be given and consider the (equivalent) metric $g : \Theta \times \Theta \to [0,\infty)$ defined by $g(\theta,\theta') = \max\{|\theta_1 - \theta_1'|, |\theta_2 - \theta_2'|\}$. Define $V = \{P_{\theta,\eta} \in \mathscr{P} : g(\theta,\theta') > \epsilon\}$. Concentrate on the cases $\alpha = 0+$ and $\alpha = 1-$; pick $0 < \delta < f(\epsilon/\sigma)/(2K)$ and define B as in theorem 6.2. Lemma B.1 says that for all $P \in V$ and $Q \in B$,

$$\begin{aligned}
P_0\Big(\frac{dP}{dQ}\Big)^{0+} &= P_0(p>0),\\
P_0\Big(\frac{dP}{dQ}\Big)^{1-} &= \int \frac{dP_0}{dQ}\, 1_{\{p_0>0,p>0,q>0\}}\, dP\\
&\leq P(p_0>0) + \int \Big|\frac{dP_0}{dQ} - 1\Big| 1_{\{q>0\}}\, dP\\
&\leq P(p_0>0) + \Big\|\frac{dP_0}{dQ} - 1\Big\|_{s,Q} \Big\|\frac{dP}{dQ}\Big\|_{r,Q} < P(p_0>0) + \tfrac{1}{2} f\big(\tfrac{\epsilon}{\sigma}\big),
\end{aligned}$$

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 proposition 2.1 guarantees that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. For all $P \in V$, $\sup_{Q\in B} P_0(dP/dQ) \leq 1 + \frac{1}{2}f(\epsilon/\sigma) < \infty$ and for all $Q \in B$, we have,

$$\inf_{0\leq\alpha\leq 1} P_0\Big(\frac{dP}{dQ}\Big)^\alpha \leq \min\big\{P_0(p>0), P(p_0>0)\big\} + \tfrac{1}{2}f\big(\tfrac{\epsilon}{\sigma}\big),$$

as an upper bound for testing power.

Choose P_0, P with parameters (θ_0, η_0) and (θ, η), 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 ϵ. Cover V by four sets, $V_{+,1} = \{P_{\theta,\eta} : \theta_1 \geq \theta_{0,1} + \epsilon, \eta \in H\}$, $V_{-,1} = \{P_{\theta,\eta} : \theta_1 \leq \theta_{0,1} - \epsilon, \eta \in H\}$, $V_{+,2} = \{P_{\theta,\eta} : \theta_2 \geq \theta_{0,2} + \epsilon, \eta \in H\}$ and $V_{-,2} = \{P_{\theta,\eta} : \theta_2 \leq \theta_{0,2} - \epsilon, \eta \in H\}$. For $P \in \mathrm{co}(V_{+,1})$, we have,

$$\begin{aligned} P_0(p=0) &\geq \int_{\theta_{0,1}}^{\theta_{0,1}+\epsilon} p_0(x)\,dx = \int_{\theta_{0,1}}^{\theta_{0,1}+\epsilon} \frac{1}{\theta_{0,2}-\theta_{0,1}}\eta_0\Big(\frac{x-\theta_{0,1}}{\theta_{0,2}-\theta_{0,1}}\Big)dx \\ &= \int_0^{\epsilon/(\theta_{0,2}-\theta_{0,1})} \eta_0(z)\,dz \geq \int_0^{\frac{\epsilon}{\sigma}} \eta_0(z)\,dz \geq f\big(\tfrac{\epsilon}{\sigma}\big), \end{aligned}$$

using (6.2). For $P \in \mathrm{co}(V_{-,1})$, with some $I \geq 1$ write $P = \sum_{i=1}^I \lambda_i P_i$ with $\sum_{i=1}^I \lambda_i = 1$ 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 $\eta_i \in H$, for all $1 \leq i \leq I$. Note that,

$$\begin{aligned} P(p_0=0) &= \sum_{i=1}^I \lambda_i P_i(p_0=0) \geq \sum_{i=1}^I \lambda_i \int_{\theta_{i,1}}^{\theta_{i,1}+\epsilon} p_i(x)\,dx \\ &= \sum_{i=1}^I \lambda_i \int_{\theta_{i,1}}^{\theta_{i,1}+\epsilon} \frac{1}{\theta_{i,2}-\theta_{i,1}}\eta_i\Big(\frac{x-\theta_{i,1}}{\theta_{i,2}-\theta_{i,1}}\Big)dx \\ &= \sum_{i=1}^I \lambda_i \int_0^{\epsilon/(\theta_{i,2}-\theta_{i,1})} \eta_i(z)\,dz \geq \sum_{i=1}^I \lambda_i \int_0^{\frac{\epsilon}{\sigma}} \eta_i(z)\,dz \geq f\big(\tfrac{\epsilon}{\sigma}\big), \end{aligned}$$

using (6.2). Analogously we obtain bounds for $P \in \mathrm{co}(V_{+,2})$ and $P \in \mathrm{co}(V_{-,2})$, giving rise to the inequalities

$$\sup_{P\in\mathrm{co}(V_\cdot)} \min\big\{P_0(p>0), P(p_0>0)\big\} \leq 1 - f\big(\tfrac{\epsilon}{\sigma}\big), \tag{C.8}$$

for $V_\cdot$ equal to $V_{+,1}$, $V_{-,1}$, $V_{+,2}$ and $V_{-,2}$. Combination of lemma 2.5 and theorem 1.3 now shows that,

$$\Pi\big(\, g(\theta,\theta_0) < \epsilon \;\big|\; X_1, \ldots, X_n \,\big) \xrightarrow{P_0\text{-a.s.}} 1.$$

The topology associated with the metric g on Θ is equivalent to the restriction to Θ of the usual norm topology on $\mathbb{R}^2$, so that consistency with respect to the pseudo-metric g is equivalent to (6.3). □

Acknowledgements

The authors would like to thank P. Grünwald for a very useful suggestion and the anonymous referees for this paper for many helpful comments. BK also thanks Y. D. Kim and S. Petrone and the *Statistics Department of Seoul National University, South Korea*, and the *Dipartimento di Scienze delle Decisioni, Univesita Bocconi, Milano, Italy* for their kind hospitality. YYZ thanks the *Korteweg-de Vries Institute* of the *University of Amsterdam* for its kind hospitality.

References

[1] A. Barron, *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions,* Technical report nr. 7 (1988), Dept. of Statistics, Univ. of Illinois.

[2] A. Barron, M. Schervish and L. Wasserman, *The Consistency of posterior distributions in nonparametric problems,* Ann. Statist. **27** (1999), 536–561. MR1714718

[3] P. Bickel and B. Kleijn, *The semiparametric Bernstein-Von Mises theorem,* Ann. Statist. **40** (2012), 206–237. MR3013185

[4] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), 181–238. MR0722129

[5] L. Birgé, *Sur un théorème de minimax et son application aux tests,* Probability and Mathematical Statistics **3** (1984), 259–282. MR0764150

[6] P. De Blasi, A. Lijoi and I. Prünster, *An asymptotic analysis of a class of discrete nonparametric priors,* Statist. Sinica **23** (2013), 1299–1321. MR3114715

[7] I. Castillo, *On Bayesian supremum norm contraction rates,* Ann. Statist. **42** (2014), 2058–2091. MR3262477

[8] I. Castillo, J. Schmidt-Hieber, and A. van der Vaart, *Bayesian linear regression with sparse priors,* Ann. Statist. **43** (2015), 1986–2018. MR3375874

[9] P. Diaconis and D. Freedman, *On the consistency of Bayes estimates,* Ann. Statist. **14** (1986), 1–26. MR0829555

[10] J. Doob, *Applications of the theory of martingales,* Le calcul des Probabilités et ses Applications, Colloques Internationales du CNRS, Paris (1948), 22–28. MR0033460

[11] D. Freedman, *On the asymptotic behavior of Bayes estimates in the discrete case I,* Ann. Math. Statist. **34** (1963), 1386–1403. MR0158483

[12] D. Freedman, *On the asymptotic behavior of Bayes estimates in the discrete case II,* Ann. Math. Statist. **36** (1965), 454–456. MR0174146

[13] 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). MR1479180

[14] S. Ghosal, J. Ghosh and R. Ramamoorthi, *Posterior consistency of Dirichlet mixtures in density estimation,* Ann. Statist. **27** (1999), 143–158. MR1701105
[15] S. Ghosal, J. Ghosh and A. van der Vaart, *Convergence rates of posterior distributions,* Ann. Statist. **28** (2000), 500–531. MR1790007
[16] S. Ghosal, and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference,* Cambridge University Press, Cambridge (2017). MR3587782
[17] J. Ghosh and R. Ramamoorthi, *Bayesian Nonparametrics,* Springer, New York (2003). MR1992245
[18] M. Hoffmann, J. Schmidt-Hieber, *On adaptive posterior concentration rates,* Ann. Statist. **43** (2015), 2259—2295. MR3396985
[19] I. Ibragimov, R. Has'minskii, *Statistical Estimation: Asymptotic Theory,* Springer, New York (1981).
[20] B. Kleijn, *Bayesian asymptotics under misspecification,* PhD. Thesis, Free University Amsterdam (2004).
[21] B. Kleijn and A. van der Vaart, *Misspecification in infinite-dimensional Bayesian statistics,* Ann. Statist. **34** (2006), 837–877. MR2283395
[22] B. Kleijn and A. van der Vaart, *The Bernstein-Von-Mises theorem under misspecification,* Electron. J. Statist. **6** (2012), 354–381. MR2988412
[23] B. Kleijn and B. Knapik, *Semiparametric posterior limits under local asymptotic exponentiality,* 1210.6204.
[24] B. Kleijn, *On the frequentist validity of Bayesian limits,* 1611.08444.
[25] L. Le Cam, *On the speed of convergence of posterior distributions,* (unpublished preprint) University of California, Berkeley (197?).
[26] L. Le Cam, *Convergence of estimates under dimensionality restrictions,* Ann. Statist. **1** (1973), 38–55. MR0334381
[27] 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. MR0395005
[28] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory,* Springer, New York (1986). MR0856411
[29] T. Leonard, *Density estimation, stochastic processes and prior information,* J. Roy. Statist. Soc. **B40** (1978), 113–146. MR0517434
[30] K. Matusita, *Some propoerties of affinity and applications,* Ann. Inst. Statist. Math. **23** (1971), 137–155. MR0348925
[31] M. Reiss, J. Schmidt-Hieber, *Nonparametric Bayesian analysis for support boundary recovery,* 1703.08358.
[32] Y. Ritov, P. Bickel, A. Gamst and B. Kleijn, *The Bayesian analysis of complex, high-dimensional models: can it be CODA?* Statist. Sci. **29** (2014), 619–639. MR3300362
[33] L. Schwartz, *Consistency of Bayes procedures,* PhD. thesis, UC Berkeley Statistics Department (1961). MR2939195
[34] L. Schwartz, *On Bayes procedures,* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26. MR0184378

[35] X. Shen and L. Wasserman, *Rates of convergence of posterior distributions,* Ann. Statist. **29** (2001), 687–714. MR1865337
[36] H. Strasser, *Mathematical Theory of Statistics,* de Gruyter, Berlin (1985). MR0812467
[37] G. Toussiant, *Some properties of Matusita's measure of affinity of several distributions,* Ann. Inst. Statist. Math. **26** (1974), 389–394. MR0362730
[38] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes,* Springer, New York (1996). MR1385671
[39] S. Walker, *New approaches to Bayesian consistency,* Ann. Statist. **32** (2004), 2028–2043. MR2102501
[40] S. Walker, A. Lijoi and I. Prünster, *On rates of convergence for posterior distributions in infinite-dimensional models,* Ann. Statist. **35** (2007), 738–746. MR2336866
[41] W.H. Wong and X. Shen, *Probability inequalities for likelihood ratios and convergence rates of sieve MLEs,* Ann. Statist. **23** (1995), 339–362. MR1332570