6

Consistency: General Theory

Posterior consistency concerns the convergence of the posterior distribution to the true value of the parameter of the distribution of the data when the amount of data increases indefinitely. Schwartz's theorem concerning consistency in models with replicated observations is a core result. In this chapter we extend it in many ways, also to general observations, such as Markov processes. Other results include Doob's theorem on almost everywhere consistency, consistency of tail-free processes and examples of inconsistency. We conclude with alternative approaches to proving consistency of posterior distributions and posterior-like processes.

6.1 Consistency and Its Implications

For every $n \in \mathbb{N}$, let $X^{(n)}$ be an observation in a sample space $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)})$ with distribution $P_\theta^{(n)}$ indexed by a parameter θ belonging to a first countable topological space Θ.[1] For instance, $X^{(n)}$ may be sample of size n from a given distribution P_θ, and $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)})$ the corresponding product probability space. Given a prior Π on the Borel sets of Θ, let $\Pi_n(\cdot|\, X^{(n)})$ be a version of the posterior distribution: a given choice of a regular conditional distribution of θ given $X^{(n)}$.

Definition 6.1 (Posterior consistency) The posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is said to be *(weakly) consistent* at $\theta_0 \in \Theta$ if $\Pi_n(U^c|\, X^{(n)}) \to 0$ in $P_{\theta_0}^{(n)}$-probability, as $n \to \infty$, for every neighborhood U of θ_0. The posterior distribution is said to be *strongly consistent* at $\theta_0 \in \Theta$ if this convergence is in the almost-sure sense.

Both forms of consistency are of interest. Naturally, strong consistency is more appealing as it is stronger, but it may require more assumptions. To begin with, it presumes that the observations $X^{(n)}$ are defined on a common underlying probability space (with, for each n, the measure $P_\theta^{(n)}$ equal to the image $P_\theta^{(\infty)} \circ (X^{(n)})^{-1}$ of the probability measure $P_\theta^{(\infty)}$ on this space), or at least that their joint distribution is defined, whereas weak consistency makes perfect sense without any relation between the observations across n.

The definition of strong consistency allows the exceptional null set where convergence does not occur to depend on the neighborhood U. However, because we assume throughout that Θ is first countable, consideration of countably many neighborhoods suffices, and a single null set always works.

[1] A topological space is *first countable* if for every point there is a countable base for the neighborhoods.

In the case that the parameter set Θ is a metric space, the neighborhoods U in the definition can be restricted to balls around θ_0, and consistency is equivalent to the convergence $\Pi_n(\theta\colon d(\theta,\theta_0) > \epsilon|\, X^{(n)}) \to 0$ as $n \to \infty$ in probability or almost surely, for every $\epsilon > 0$. Thus, the posterior distribution contracts to within arbitrarily small distance ϵ to θ_0. This can also be described as weak convergence of the posterior distribution to the Dirac measure at θ_0.

Proposition 6.2 (Consistency by convergence to Dirac) *On a metric space Θ the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is consistent (or strongly consistent, respectively) at θ_0 if and only if $\Pi_n(\cdot|\, X^{(n)}) \rightsquigarrow \delta_{\theta_0}$ in $P_{\theta_0}^{(n)}$-probability (or almost surely $[P_{\theta_0}^{(\infty)}]$, respectively), as $n \to \infty$.*

Proof With the help of the Portmanteau theorem, Theorem A.2, it can be verified that a sequence of deterministic probability measures Π_n satisfies $\Pi_n \rightsquigarrow \delta_{\theta_0}$ if and only if $\Pi_n(\theta\colon d(\theta,\theta_0) > \epsilon) \to 0$ as $n \to \infty$ for every $\epsilon > 0$. The set of $\epsilon > 0$ in this statement can also be restricted to a countable set. This immediately gives the proposition for the case of strong consistency. This can be extended to weak consistency by using the fact that convergence in probability is equivalent to every subsequence having an almost surely converging subsequence (after first defining the observations on a suitable common probability space). □

Remark 6.3 Weak convergence of random measures "in probability" is best defined by metrizing the topology of weak convergence. A more precise statement of the preceding proposition is that consistency of the posterior distribution is equivalent to convergence of the random variables

$$d_W(\Pi_n(\cdot|\, X^{(n)}), \delta_{\theta_0})$$

to zero in probability (or almost surely) under the measures $P_{\theta_0}^{(n)}$ (or $P_{\theta_0}^{(\infty)}$), for d_W a metric on $\mathfrak{M}(\Theta)$ that metrizes the topology of weak convergence. For a separable metric space Θ such a metric exists, while for a general metric space it is still possible to metrize convergence in distribution to limits with separable support (in particular when the limit is a Dirac measure) (see e.g. van der Vaart and Wellner 1996, Chapter 1.12). Convergence in distribution of measures on a general topological space requires care, which makes defining consistency through convergence to a Dirac measure unattractive in that case.

Rather than through the posterior mass of abstract neighborhoods U, posterior consistency can be characterized through the posterior distributions of a set of test functions. Suppose that Ψ is a collection of real functions $\psi\colon \Theta \to \mathbb{R}$ such that, for any net $\{\theta_m\} \subset \Theta$,

$$\theta_m \to \theta_0, \qquad \text{if} \qquad \psi(\theta_m) \to \psi(\theta_0), \text{ for every } \psi \in \Psi. \tag{6.1}$$

For a countable set Ψ, it will be sufficient to limit this requirement to *sequences* θ_m.

For each ψ there is an induced posterior distribution $\Pi_n(\theta\colon \psi(\theta) \in \cdot|\, X^{(n)})$ on $\mathbb{R}$, which by definition is consistent at $\psi(\theta_0)$ if it converges in probability (or almost surely) in distribution to the Dirac measure at $\psi(\theta_0)$. Consistency of all these induced posterior distributions is equivalent to the posterior consistency for θ.

Lemma 6.4 (Consistency by functionals) *If Ψ is a set of measurable real functions on Θ so that* (6.1) *holds, then the posterior distribution is (strongly) consistent at θ_0 if for each $\psi(\theta)$ the induced posterior is (strongly) consistent at $\psi(\theta_0)$. If the functions ψ are uniformly bounded, then the latter is equivalent to the pair of conditions* $\mathrm{E}(\psi(\theta)|\, X^{(n)}) \to \psi(\theta_0)$ *and var* $(\psi(\theta)|\, X^{(n)}) \to 0$*, in probability (or almost surely).*

Proof We claim that for any open neighborhood U of θ_0, there exists a finite set $\psi_1, \ldots, \psi_k$ and $\epsilon > 0$ such that $\cap_{i=1}^k \{\theta\colon |\psi_i(\theta) - \psi_i(\theta_0)| < \epsilon\} \subset U$. In this case $\Pi_n(U^c|\, X^{(n)})$ can be bounded by the sum of the probabilities $\Pi_n(|\psi_i(\theta) - \psi_i(\theta_0)| \geq \epsilon|\, X^{(n)})$, which tends to zero for fixed k, and the theorem follows.

If the claim were false, there would exist a neighborhood U of θ_0 such that for every finite subset m of Ψ there is some $\theta_m \notin U$ such that $|\psi(\theta_m) - \psi(\theta_0)| \leq (\#m)^{-1}$, for every $\psi \in m$. Then the net θ_m with the finite subsets of Ψ partially ordered by inclusion would satisfy $\psi(\theta_m) \to \psi(\theta_0)$ as $m \to \infty$, for every fixed ψ. Hence $\theta_m \to \theta_0$ by (6.1), which contradicts that $\theta_m \notin U$ for every m.

If Ψ is countable, we may order it arbitrarily as $\psi_1, \psi_2, \ldots$ and use the special net with m the finite set $\{\psi_1, \ldots, \psi_m\}$, which is a sequence.

A sequence of deterministic distributions on a compact subset of $\mathbb{R}$ tends weakly to a Dirac measure at a point if and only if the expectations tend to this point and the variances tend to zero. This extends trivially to almost sure weak convergence of random distributions (covering strong consistency), and to weak convergence in probability by arguing along subsequences. □

The following example shows that consistency depends not only on the true parameter and the prior, but also on the version of the posterior distribution. This discrepancy can arise because the posterior distribution is defined only up to a null set under the Bayesian marginal distribution of $X^{(n)}$, and this may not dominate this variable's distribution $P_{\theta_0}^{(n)}$ under a fixed parameter (see Section 1.3).

Example 6.5 (Dependence of consistency on version) Consider the model with observations $X_1, \ldots, X_n|\, P \overset{\text{iid}}{\sim} P$ and a Dirichlet prior $P \sim \mathrm{DP}(\alpha)$, for some base measure α. By Theorem 4.6 the $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$-distribution is a version of the posterior distribution, and this is consistent at any distribution P_0, by Corollary 4.17. Now, given a measurable set B with $\alpha(B) = 0$ and some probability measure Q, consider the random measure

$$\mathcal{P}^*_{X_1,\ldots,X_n} = \begin{cases} Q, & \text{if } X_i \in B \ \forall i = 1, \ldots, n, \\ \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i}), & \text{otherwise.} \end{cases}$$

The Pólya urn scheme given in Section 4.1.4 describes the marginal distribution of the observation $(X_1, \ldots, X_n)$ as a mixture of distributions of vectors of the type $(Y_1, \ldots, Y_n)$, where $Y_i = Y_j$ if i and j belong to the same set in a given partition of $\{1, 2, \ldots, n\}$ and the tied values are generated i.i.d. from $\bar{\alpha}$. This shows that it gives probability zero to the set $B^n = B \times \cdots \times B$. As a posterior distribution is unique only up to null sets in the marginal distribution of the data, it follows that $\mathcal{P}^*_{X_1,\ldots,X_n}$ is another version of the posterior distribution.

If $P_0^n(B^n) \to 0$, then $\mathcal{P}^*_{X_1,\ldots,X_n}$ differs from the consistent sequence $\mathrm{DP}(\Phi + \sum_{i=1}^n \delta_{X_i})$ only on events with probability tending to zero, and hence it is (also) consistent at P_0. However, in the other case the new version is inconsistent whenever $Q \neq P_0$. For instance, this situation arises if P_0 and α are orthogonal and the set B is chosen so that $P_0(B) = 1$ and $\alpha(B) = 0$. As another example, if $B = \{0\}$, then $P_0^n(B^n) \to 0$ for every $P_0 \neq \delta_0$ and hence consistency pertains, but the posterior is inconsistent at $P_0 = \delta_0$ unless $Q = \delta_0$.

Consistency is clearly dependent on the topology on Θ under consideration. If Θ is a class of distributions, common choices are the weak topology, the topology induced by the Kolmogorov-Smirnov distance (if the sample space is Euclidean) and the topology induced by the total variation distance. The weak topology gives the weakest form of consistency of these three, but can be appropriate if only the probability measure needs to be estimated. The total variation distance is appropriate in density estimation, as it is one-half times the $\mathbb{L}_1$-distance on the densities, and is also equivalent to the Hellinger distance (in that they generate the same topology).

Consistency is a frequentist concept, but should be of interest also to a subjective Bayesian, by the following reasoning. A prior or posterior distribution expresses our knowledge about the unknown parameter, and a degenerate prior models perfect knowledge. Thus, by its definition, consistency entails convergence toward perfect knowledge. This is desirable, for instance, if n is the sample size and increases indefinitely, and also in every other setting where information accumulates without bound as $n \to \infty$.

From the point of view of an objective Bayesian (who believes in an unknown true model), consistency can be motivated by a "what if" argument. Suppose that an experimenter generates observations from a given distribution and presents the resulting data to a Bayesian without revealing the source distribution. It would be embarrassing if, even with large sample size, the Bayesian failed to come close to finding the mechanism used by the experimenter. Thus, consistency can be thought of as a validation of the Bayesian method: it ensures that an infinite amount of data overrides the prior opinion.

Consistency can also be linked to robustness with respect to the choice of a prior distribution: a Bayesian procedure should change little when the prior is only slightly modified. Now for any prior distribution Π_0, the posterior distribution $\Pi_{n,1}(\cdot|\, X^{(n)})$ based on a random sample $X^{(n)}$ of observations and the "contamination" prior $\Pi_1 = (1-\epsilon)\Pi_0 + \epsilon\delta_{p_0}$ is consistent at p_0 (see Problem 6.7). Therefore, if the posterior $\Pi_{n,0}(\cdot|\, X^{(n)})$ based on the prior Π_0 were *in*consistent at p_0, then for a metric d_W for the weak convergence the distance $d_W(\Pi_0(\cdot|\, X^{(n)}), \Pi_1(\cdot|\, X^{(n)}))$ between the two sequences of posterior distributions would not go to zero. Because this would be true even for small $\epsilon > 0$, robustness would be violated.

From a subjective Bayesian's point of view no true parameter exists, whence consistency may appear a useless concept at first. However, the true parameter may be eliminated and consistency be equivalently defined in terms of *merging* of predictive distributions of future observations, fully within the Bayesian paradigm. Roughly speaking, the calculations of different Bayesians using different priors will tend to agree (in the sense of the weak topology) if and only if their posteriors are consistent. In other words, two Bayesians with different priors will have their opinions "merging" if and only if consistency holds.

This is formalized in the following theorem, whose proof we omit (see Diaconis and Freedman 1986b, Appendix A). Consider i.i.d. observations $X_1, X_2, \ldots$ from a distribution P_θ on a sample space $\mathfrak{X}$, and denote the (posterior) predictive distribution of the future observations $(X_{n+1}, X_{n+2}, \ldots)$, given $X^{(n)} = (X_1, \ldots, X_n)$ when using prior Π by $Q_{\Pi,n} = \int P_\theta^\infty \, d\Pi_n(\theta|\, X^{(n)})$.

Theorem 6.6 (Merging of opinions) *Let Θ be a Borel subset of a Polish space and $\mathfrak{X}$ a standard Borel space, and assume that the map $\theta \mapsto P_\theta$ is one-to-one and measurable from Θ into $\mathfrak{M}(\Theta)$, equipped with the weak topology. Then, for any prior probability distribution Π on Θ equivalent are*

(i) *The posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ relative to Π is consistent at every $\theta \in \Theta$.*
(ii) *The posterior distributions $\Pi_n(\cdot|\, X^{(n)})$ and $\Gamma_n(\cdot|\, X^{(n)})$ relative to Π and Γ merge in the sense that $d_W(\Pi_n(\cdot|\, X^{(n)}), \Gamma_n(\cdot|\, X^{(n)})) \to 0$ as $n \to \infty$ a.s. $[\int P_\theta^\infty \, d\Gamma(\theta)]$, for any prior probability distribution Γ.*

If, furthermore, the map $\theta \mapsto P_\theta$ is continuous and has a continuous inverse, then these statements are also equivalent to

(iii) *The predictive distributions relative to Π and Γ merge in that $d_W(Q_{\Pi,n}, Q_{\Gamma,n}) \to 0$ as $n \to \infty$ a.s. $[\int P_\theta^\infty \, d\Gamma(\theta)]$, for any prior probability distribution Γ.*

Posterior consistency entails the contraction of the full posterior distribution to the true value of the parameter. Naturally, an appropriate summary of its location should provide a point estimator that is consistent, in the usual sense of consistency of estimators. The following proposition gives a summary that works without further conditions. (The value $1/2$ can be replaced by any other number strictly between 0 and 1; possible problems with selecting a measurable version can be alleviated by restricting the estimator to a countable dense subset of Θ.)

Proposition 6.7 (Point estimator) *Suppose that the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is consistent (or strongly consistent) at θ_0 relative to the metric d on Θ. Then $\hat\theta_n$, defined as the center of a (nearly) smallest ball that contains posterior mass at least $1/2$ satisfies $d(\hat\theta_n, \theta_0) \to 0$ in $P_{\theta_0}^{(n)}$-probability (or almost surely $[P_{\theta_0}^{(\infty)}]$, respectively).*

Proof For $B(\theta, r) = \{s \in \Theta\colon d(s, \theta) \leq r\}$ the closed ball of radius r around $\theta \in \Theta$, let $\hat r_n(\theta) = \inf\{r\colon \Pi_n(B(\theta, r)|\, X^{(n)}) \geq 1/2\}$, where the infimum over the empty set is infinity. Taking the balls closed ensures that $\Pi_n(B(\theta, \hat r_n(\theta))|\, X^{(n)}) \geq 1/2$, for every θ. Let $\hat\theta_n$ be a near minimizer of $\theta \mapsto \hat r_n(\theta)$ in the sense that $\hat r_n(\hat\theta_n) \leq \inf_\theta \hat r_n(\theta) + 1/n$.

By consistency $\Pi_n(B(\theta_0, \epsilon)|\, X^{(n)}) \to 1$ in probability or almost surely, for every $\epsilon > 0$. As a first consequence $\hat r_n(\theta_0) \leq \epsilon$ with probability tending to one, or eventually almost surely, and hence $\hat r_n(\hat\theta_n) \leq \hat r_n(\theta_0) + 1/n$ is bounded by $\epsilon + 1/n$ with probability tending to one, or eventually almost surely. As a second consequence, the balls $B(\theta_0, \epsilon)$ and $B(\hat\theta_n, \hat r_n(\hat\theta_n))$ cannot be disjoint, as their union would contain mass nearly $1 + 1/2$. This shows that $d(\theta_0, \hat\theta_n) \leq \epsilon + \hat r_n(\hat\theta_n)$ with probability tending to one, or eventually almost surely, which is further bounded by $2\epsilon + 1/n$. □

An alternative point estimator is the *posterior mean* $\int \theta\, d\Pi_n(\theta|\, X^{(n)})$. This is attractive for computational reasons, as it can be approximated by the average of the output of a simulation run. Usually the posterior mean is also consistent, but in general this may require additional assumptions. For instance, for the parameter set equal to a Euclidean space, convergence of the posterior distributions relative to the weak topology does not imply convergence of its moments.

For the posterior mean to make sense, the parameter set Θ must be a subset of a vector space and some other conditions must be fulfilled. If the parameter θ is a probability measure, e.g. describing the law of a single observation, then the posterior mean is well defined as the *mean measure* of the posterior distribution by Fubini's theorem. If the parameter is an element of a separable normed space, then the posterior mean can be defined as a Pettis integral (see Section I.4), provided that $\int \|\theta\|\, d\Pi_n(\theta|\, X^{(n)}) < \infty$. The geometric version of *Jensen's inequality* says that the mean $\int \theta\, d\Pi(\theta)$ of a probability measure Π that concentrates on a closed convex set is contained in this set. Even when the metric on Θ is not induced by a norm, it is often true that the balls around points are convex. In the following theorem, only assume that Θ is a metric space where notions of expectation and convex combinations are defined.

Theorem 6.8 (Posterior mean) *Assume that the balls of the metric space (Θ, d) are convex and the geometric version of Jensen's inequality holds on closed balls. Suppose that $d(\theta_{n,1}, (1-\lambda_n)\theta_{n,1}+\lambda_n\theta_{n,2}) \to 0$ as $\lambda_n \to 0$, for any sequences $\theta_{n,1}$ and $\theta_{n,2}$. Then (strong) consistency of the posterior distributions $\Pi_n(\cdot|\, X^{(n)})$ at θ_0 implies (strong) consistency of the sequence of posterior means $\int \theta\, d\Pi_n(\theta|\, X^{(n)})$ at θ_0.*

Proof For fixed $\epsilon > 0$ we can decompose the posterior mean $\bar\theta_n = \int \theta\, d\Pi_n(\theta|\, X^{(n)})$ as $\bar\theta_n = (1-\lambda_n)\bar\theta_{n,1} + \lambda_n\bar\theta_{n,2}$, where $1-\lambda_n = \Pi_n(\theta\colon d(\theta,\theta_0) \le \epsilon|\, X^{(n)})$, and $\bar\theta_{n,1}$ and $\bar\theta_{n,2}$ are the means of the probability measures obtained by restricting and renormalizing the posterior distribution to the sets $\{\theta\colon d(\theta,\theta_0) \le \epsilon\}$ and $\{\theta\colon d(\theta,\theta_0) > \epsilon\}$, respectively. By the geometric form of Jensen's inequality and the assumed convexity of balls, $d(\bar\theta_{n,1},\theta_0) \le \epsilon$. By the consistency of the posterior distribution, the constants λ_n tend to zero. The result now follows with the help of the triangle inequality $d(\theta_0,\bar\theta_n) \le d(\theta_0,\bar\theta_{n,1}) + d(\bar\theta_{n,1}, (1-\lambda_n)\bar\theta_{n,1} + \lambda_n\bar\theta_{n,2})$ and the condition on d. □

If the metric d is convex in its arguments and uniformly bounded, then its balls are convex, and the condition on the metric d holds, as $d(\theta_{n,1}, (1-\lambda_n)\theta_{n,1} + \lambda_n\theta_{n,2}) \le 0 + \lambda_n d(\theta_{n,1},\theta_{n,2}) \to 0$, as $\lambda_n \to 0$. This shows that the theorem applies to the total variation norm, the Kolmogorov-Smirnov distance, the $\mathbb{L}_r$-norms for $r \ge 1$ and also to the weak topology on probability measures, metrized by the bounded Lipshitz distance (see Section A.2).

The Hellinger distance is not convex, but its square $P \mapsto d_H^2(P_0, P)$ is convex, which allows us to reach the same conclusion.

Provided that consistency is defined also for more general "divergence measures" like the Kullback-Leibler divergence, these could also be included if convex.

6.2 Doob's Theorem

Doob's theorem basically says that for any fixed prior, the posterior distribution is consistent at every θ except those in a "bad set" that is "small" when seen from the prior point of view. We first present the theorem and next argue that the message is not as positive as it may first seem.

The result requires that the experiments $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}: \theta \in \Theta)$ are suitably linked and that Θ is a separable metric space with Borel σ-field $\mathscr{B}$. We assume that the observations $X^{(n)}$ can be defined as measurable functions $X^{(n)}: \mathfrak{X}^{(\infty)} \to \mathfrak{X}^{(n)}$ in a single experiment $(\mathfrak{X}^{(\infty)}, \mathscr{X}^{(\infty)}, P_\theta^{(\infty)}: \theta \in \Theta)$, with corresponding induced laws $P_\theta^{(n)} = P_\theta^{(\infty)} \circ (X^{(n)})^{-1}$. Furthermore, we assume that the informativeness increases with n in the sense that the induced σ-fields $\sigma\langle X^{(n)}\rangle$ are increasing in n (i.e. form a filtration on $(\mathfrak{X}^{(\infty)}, \mathscr{X}^{(\infty)})$. Given a prior Π, we next define a distribution Q on the product space $(\mathfrak{X}^{(\infty)} \times \Theta, \mathscr{X}^{(\infty)} \otimes \mathscr{B})$ by

$$Q(A \times B) = \int_B P_\theta^{(\infty)}(A)\, d\Pi(\theta).$$

The variable $\vartheta: \mathfrak{X}^{(\infty)} \times \Theta \to \Theta$ defined by $\vartheta(x, \theta) = \theta$ possesses the prior distribution Π as its law under Q. With abuse of notation we view $X^{(n)}$ also as maps defined on the product space (depending only on the first coordinate), with the induced σ-fields $\sigma\langle X^{(n)}\rangle$ contained in the product σ-field $\mathscr{X}^{(\infty)} \otimes \mathscr{B}$ (but being of product form). In this setting, the conditional distribution of $X^{(n)}$ given $\vartheta = \theta$ is $P_\theta^{(n)}$ and the posterior law $\Pi_n(\cdot|\, X^{(n)})$ is the conditional law of ϑ given $X^{(n)}$.

Theorem 6.9 (Doob) *Let Π be an arbitrary prior on the Borel sets of a separable metric space Θ. If ϑ is measurable relative to the Q-completion of $\sigma\langle X^{(1)}, X^{(2)}, \ldots\rangle$, then the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is strongly consistent at θ, for Π-almost every θ. In fact $\int f(\theta')\, d\Pi_n(\theta'|\, X^{(n)}) \to f(\theta)$, almost surely $[P_\theta^{(\infty)}]$, for Π-almost every θ and every $f \in \mathbb{L}_1(\Pi)$.*

Proof For a given bounded measurable real-valued function f on Θ, the martingale convergence theorem gives that $\mu_n(X^{(n)}) := \mathrm{E}(f(\vartheta)|\, X^{(n)}) \to \mathrm{E}(f(\vartheta)|\, \sigma\langle X^{(1)}, X^{(2)}, \ldots\rangle)$, almost surely under Q. By assumption, the limit is equal to $f(\vartheta)$, almost surely. It follows that

$$1 = Q\Big(\lim_{n\to\infty} \mu_n(X^{(n)}) = f(\vartheta)\Big) = \int P_\theta^{(\infty)}\Big(\lim_{n\to\infty} \mu_n(X^{(n)}) = f(\theta)\Big)\, d\Pi(\theta).$$

The integrand must be equal to 1 for all θ except those in a Π-null set N. For every $\theta \notin N$ we have $\int f(\theta)\, d\Pi_n(\theta|\, X^{(n)}) = \mu_n(X^{(n)}) \to f(\theta)$.

For every bounded measurable function f there exists such a null set. Because Θ is a separable metric space, there exists a *countable* collection of such functions that determines convergence in distribution (see e.g. Theorem 1.12.2 in van der Vaart and Wellner 1996). For every θ not in the union of these countably many null sets, we have that $\Pi_n(\cdot|\, X^{(n)}) \rightsquigarrow \delta_\theta$. □

The assumption that $\vartheta \in \sigma\langle X^{(1)}, X^{(2)}, \ldots\rangle$ can be paraphrased as saying that, in the Bayesian model, with unlimited observations, it is possible to find the parameter exactly. The condition holds, for instance, if θ, or even a bimeasurable transformation $h(\theta)$ of the parameter, is consistently estimable from the observations $X^{(n)}$. Indeed, if $T_n(X^{(n)}) \to h(\theta)$ almost surely $[P_\theta^{(\infty)}]$ for every $\theta \in \Theta$, then

$$Q\Bigl(\lim_{n\to\infty} T_n(X^{(n)}) = h(\vartheta)\Bigr) = \int P_\theta^{(\infty)}\Bigl(\lim_{n\to\infty} T_n(X^{(n)}) = h(\theta)\Bigr)\, d\Pi(\theta) = 1,$$

by the dominated convergence theorem. Thus $h(\vartheta)$ inherits its measurability (up to a Q-null set) from $\lim_{n\to\infty} T_n(X^{(n)})$. Various conditions ensuring existence of consistent estimators are available in the literature (e.g. Ibragimov and Has'minskiĭ 1981, Bickel et al. 1998, or van der Vaart (1998), Chapters 5, 10). Alternatively, the assumption may be verified directly. In the i.i.d. setup, with $\mathfrak{X}^{(\infty)} = \mathfrak{X}^\infty$ and $X^{(n)}$ the projection on the first n coordinates, some basic regularity conditions on the model suffice.

Proposition 6.10 *Let $(\mathfrak{X}, \mathscr{X}, P_\theta\colon \theta \in \Theta)$ be experiments with $(\mathfrak{X}, \mathscr{X})$ a standard Borel space and Θ a Borel subset of a Polish space. If $\theta \mapsto P_\theta(A)$ is Borel measurable for every $A \in \mathscr{X}$ and the model $\{P_\theta\colon \theta \in \Theta\}$ is identifiable, then there exists a Borel measurable function $f\colon \mathfrak{X}^\infty \to \Theta$ such that $f(x_1, x_2, \ldots) = \theta$ a.s. $[P_\theta^\infty]$, for every $\theta \in \Theta$.*

Proof By Proposition A.5 and the measurability of $\theta \mapsto P_\theta(A)$ for every Borel set A, the map $P\colon \theta \mapsto P_\theta$ can be viewed as a Borel measurable map from Θ into the space $\mathfrak{M}(\mathfrak{X})$ equipped with the weak topology (with metric d_W). Likewise, the empirical distribution $\mathbb{P}_n$ of n i.i.d. observations $X_1, \ldots, X_n$ from P_θ can be viewed as a measurable map from $\mathfrak{X}^\infty$ into $\mathfrak{M}(\mathfrak{X})$. By Proposition F.3, $d_W(\mathbb{P}_n, P_\theta) \to 0$ almost surely $[P_\theta^\infty]$. We can also view both $\mathbb{P}_n$ and P as maps on the product space $\mathfrak{X}^\infty \times \Theta$ (depending only on the first and second coordinates, respectively), and then by Fubini's theorem have that $d_W(\mathbb{P}_n, P) \to 0$ almost surely under the measure Q on $\mathfrak{X}^\infty \times \Theta$, defined as previously by $Q(A \times B) = \int_B P_\theta^\infty(A)\, d\Pi(\theta)$. Hence the random variable P taking values in the Polish space $(\mathfrak{M}(\mathfrak{X}), d_W)$ is the a.s. limit of the sequence of the random variables $\mathbb{P}_n$, taking values in the same space. As $\lim \mathbb{P}_n$ is measurable relative to the σ-field $\sigma\langle X_1, X_2, \ldots\rangle$ on $\mathfrak{X}^\infty \times \Theta$ and $\mathfrak{X}^\infty$ is Polish by assumption, it follows that $P_\theta = \tilde{f}(X_1, X_2, \ldots)$, Q-almost surely for some measurable function $\tilde{f}\colon \mathfrak{X}^\infty \to \mathfrak{M}(\mathfrak{X})$ (see Dudley 2002, Theorem 4.2.8).

Since $\theta \mapsto P_\theta$ is one-to-one, its range is a measurable subset of $\mathfrak{M}(\mathfrak{X})$ and its inverse $\jmath\colon P_\theta \mapsto \theta$ is measurable by Kuratowski's theorem (Parthasarathy 2005; or Srivastava 1998, Theorems 4.5.1 and 4.5.4). We define f as the composition $f = \jmath \circ \tilde{f}$. □

Doob's theorem is remarkable in many respects. Virtually no condition is imposed on the model regarding dependence or distribution, or about the parameter space. As posterior consistency implies existence of a consistent estimator (see Proposition 6.7), the only condition of the theorem is also necessary. Also of interest is that the result applies to any version of the posterior distribution.

The implication of Doob's theorem goes far beyond consistency because the posterior measure converges to the degenerate measure at the true θ not just weakly, but in a much stronger sense where the posterior expectation of every integrable function f approaches the function value $f(\theta)$. Weak convergence requires that the latter hold only for bounded, continuous functions. In particular, this means that the posterior probability of every set A

converges to the indicator $\mathbb{1}\{\theta \in A\}$ under $P_\theta^{(\infty)}$ a.s. θ $[\Pi]$, and hence if $\Pi(\{\theta_0\}) > 0$, then $\Pi_n(\{\theta_0\}|X^{(n)}) \to 1$ a.s. under $P_{\theta_0}^{\infty}$.

The theorem implies that a Bayesian will "almost always" have consistency, as long as she is certain of her prior. Since null sets are "negligibly small," a troublesome value of the parameter "will not obtain."

However, such a view is very dogmatic. No one in practice can be certain of the prior, and troublesome values of the parameter may really obtain. In fact, the Π-null set could be very large if *not* judged from the point of view of the prior. To see an extreme example, consider a prior that assigns all its mass to some fixed point θ_0. The posterior then also assigns mass one to θ_0 and hence is *in*consistent at every $\theta \neq \theta_0$. Doob's theorem is still true, of course; the point is that the set $\{\theta: \theta \neq \theta_0\}$ is a null set under the present prior. Thus Doob's theorem should not create a false sense of satisfaction about Bayesian procedures in general. It is important to know, for a given "reasonable" prior, at which parameter values consistency holds. Consistency at every parameter in a set of prior probability one is not enough. This explains that in Section 6.3 we reach a very different conclusion when measuring "smallness" of exceptional sets in a different way.

An exception is the case that the parameter set Θ is countable. Then Doob's theorem shows that consistency holds at θ as long as Π assigns positive mass to it. More generally, consistency holds at any atom of a prior. In fact, if θ_0 is an atom of Π, then $\Pi(\theta = \theta_0|X^{(n)}) \to 1$ a.s. $[P_{\theta_0}^{(\infty)}]$. If the parameter space is Euclidean and the null sets of the prior distribution are also Lebesgue null sets, then the conclusion from Doob's theorem may be still attractive even though this does not give consistency at all points. In a semiparametric model, if the parametric part is only of interest, the method may give rise to posterior consistency except on a Lebesgue null set for the parametric part. Even in these cases the theorem is of "asymptopia" type only, in that at best it gives convergence without quantification of the approximation error, or uniformity in the parameter.

6.3 Inconsistency

If $\Pi(U) = 0$ for some set U, then $\Pi_n(U|\, X^{(n)}) = 0$ almost surely under the Bayesian marginal distribution $\int P_\theta^{(n)}\, d\Pi(\theta)$ of the observation $X^{(n)}$, and hence also under its "true" law $P_{\theta_0}^{(n)}$ if this is absolutely continuous with respect to its marginal law. (The latter is necessary to ensure that the posterior distribution is almost surely uniquely defined under the true law.) Therefore, in this situation it is necessary for consistency that the prior charge every neighborhood of the true parameter θ_0. In most finite-dimensional problems, this necessary condition is also sufficient, but not so in infinite dimensions.

Consider the simplest infinite-dimensional problem, that of estimating a discrete distribution on $\mathbb{N}$. The parameter set $\Theta = \mathfrak{M}(\mathbb{N})$ can be identified with the infinite-dimensional unit simplex $\mathbb{S}_\infty$, and the weak and total variation topologies are equivalent. One might expect that the posterior is consistent at every point if the prior assigns positive probability to every of its neighborhoods, but this is not the case.

Example 6.11 (Freedman 1963) Let θ_0 be the geometric distribution with parameter $1/4$. There exists a prior Π on $\mathbb{S}_\infty$ that gives positive mass to every neighborhood of

θ_0, but the posterior concentrates in the neighborhoods of the geometric distribution with parameter $3/4$.

This counter-example is generic in a topological sense. A subset A of a topological space is considered to be topologically small, and said to be *meager* (or of the *first category*), if it can be written as a countable union of *nowhere dense* sets (sets whose closures have no interior points). The following result shows that the misbehavior of the posterior described in the preceding example is not just a pathological case, but a common phenomenon.

Theorem 6.12 *Under the product of the usual topology on $\Theta = \mathfrak{M}(\mathbb{N})$ and the weak topology on $\mathfrak{M}(\Theta)$, the set of all pairs $(\theta, \Pi) \in \Theta \times \mathfrak{M}(\Theta)$ such that the posterior is consistent at θ is meager in $\Theta \times \mathfrak{M}(\Theta)$. Furthermore, the set of $(\theta, \Pi) \in \Theta \times \mathfrak{M}(\Theta)$ such that*

$$\limsup_{n\to\infty} \mathrm{E}_\theta \Pi_n(U | X_1, \ldots, X_n) = 1, \text{ for all open } U \neq \varnothing, \tag{6.2}$$

is the complement of a meager set.

Proof The first assertion is a consequence of the second. For the proof of the second let $\Theta^+ = \{\theta = (\theta_1, \theta_2, \ldots) : \theta_i > 0, \forall i\}$ be the interior of the infinite-dimensional simplex, and let $\mathfrak{M}^+(\Theta^+) = \{\Pi : \Pi(\Theta^+) > 0\}$ be the set of priors that assign some mass to this, i.e. do not concentrate all their mass on the "faces" $\{\theta : \theta_i = 0, \text{ some } i\}$ of Θ.

The set $\mathfrak{M}^+(\Theta^+)$ is *residual*, i.e. its complement $\{\Pi : \Pi(\Theta^+) = 0\}$ is meager within $\mathfrak{M}(\Theta)$. Indeed, as Θ^+ is open, this complement is closed by the Portmanteau theorem; it also has empty interior, since $(1 - 1/m)\Pi + (1/m)\delta_\theta \rightsquigarrow \Pi$ as $m \to \infty$ for any θ, and is clearly not in the set if $\theta \in \Theta^+$. As a consequence the set $\Theta \times \mathfrak{M}^+(\Theta^+)$ is residual within $\Theta \times \mathfrak{M}(\Theta)$. The remainder of the proof may therefore focus on showing that the set of $(\theta, \Pi) \in \Theta \times \mathfrak{M}^+(\Theta^+)$ that satisfy (6.2) is residual within $\Theta \times \mathfrak{M}^+(\Theta^+)$.

The likelihood for an i.i.d. sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \theta$ relative to counting measure on $\mathfrak{X}^n$ is the function $\theta \mapsto \prod_{i=1}^n \theta_{X_i}$, and the posterior distribution satisfies $d\Pi_n(\theta | X_1, \ldots, X_n) \propto \prod_{i=1}^n \theta_{X_i}\, d\Pi(\theta)$. For $\Pi \in \mathfrak{M}^+(\Theta^+)$, the norming constant is positive and the posterior is uniquely defined. Furthermore, when restricted to the domain $\Theta \times \mathfrak{M}^+(\Theta^+)$, the map

$$(\theta_0, \Pi) \mapsto \mathrm{E}_{\theta_0} \int \chi(\theta)\, d\Pi_n(\theta | X_1, \ldots, X_n) = \mathrm{E}_{\theta_0} \frac{\int \chi(\theta) \prod_{i=1}^n \theta_{X_i}\, d\Pi(\theta)}{\int \prod_{i=1}^n \theta_{X_i}\, d\Pi(\theta)}$$

is continuous for any fixed, bounded, continuous function $\chi : \Theta \to \mathbb{R}$.

For $\theta^+ \in \Theta^+$ let $\mathfrak{M}^f(\Theta, \theta^+)$ be the set of priors Π that are finitely supported with exactly one support point θ^+ belonging to Θ^+ (and hence the finitely many other support points belonging to one of the faces $\{\theta : \theta_i = 0\}$ of the simplex). From the facts that any prior Π can be approximated arbitrarily closely by a finitely supported measure, the Dirac measure δ_θ can be approximated by Dirac measures $\delta_{(\theta_1, \ldots, \theta_N, 0, 0, \ldots)}$ and θ^+ can be given arbitrarily small weight, it follows that $\mathfrak{M}^f(\Theta, \theta^+)$ is dense in $\mathfrak{M}^+(\Theta^+)$.

For $\Pi \in \mathfrak{M}^f(\Theta, \theta^+)$ and every $\theta_0 \in \Theta^+$,

$$\Pi_n(\cdot | X_1, \ldots, X_n) \rightsquigarrow \delta_{\theta^+}, \quad \text{a.s. } [P_{\theta_0}^\infty].$$

In other words, the posterior distribution for data resulting from a fully supported θ_0 eventually chooses for the (unique) fully supported point θ^+ in the support of the prior, also if $\theta^+ \neq \theta_0$. The reason is that for every $\theta \neq \theta^+$ in the support of Π eventually some observation will be equal to a value $x \in \mathfrak{X}$ such that $\theta_x = 0$, whence the posterior distribution $\Pi_n(\{\theta\}|\, X_1 \ldots, X_n) \propto \prod_{i=1}^n \theta_{X_i}\, \Pi(\{\theta\})$ attaches zero mass to that point.

For $\theta^+ \in \Theta^+$, let $\theta \mapsto \chi_k(\theta; \theta^+)$ be a sequence of bounded continuous functions with $1 \geq \chi_k(\cdot; \theta^+) \downarrow \mathbb{1}_{\{\theta^+\}}$, as $k \to \infty$. By the preceding display, for every k,

$$\limsup_{n\to\infty} \mathrm{E}_{\theta_0} \int \chi_k(\theta; \theta^+)\, d\Pi_n(\theta|\, X_1, \ldots, X_n) = 1. \tag{6.3}$$

This is true for any prior $\Pi \in \mathfrak{M}^f(\Theta, \theta^+)$, and any $\theta_0 \in \Theta^+$.

For $\theta^+ \in \Theta^+$ and $k, j, m \in \mathbb{N}$, define sets $F_{\theta^+,k,j,m}$ as the intersections

$$\bigcap_{n\geq m} \Big\{(\theta_0, \Pi) \in \Theta \times \mathfrak{M}^+(\Theta^+) \colon \mathrm{E}_{\theta_0}\!\int \chi_k(\theta; \theta^+)\, d\Pi_n(\theta|\, X_1, \ldots, X_n) \leq 1 - j^{-1}\Big\}.$$

These sets are closed in $\Theta \times \mathfrak{M}^+(\Theta^+)$ by the continuity of the maps in their definition. They have empty interior, because any pair (θ_0, Π) can be arbitrarily closely approximated by an element of $\Theta^+ \times \mathfrak{M}^f(\Theta, \theta^+)$, but no such pair is contained in $F_{\theta^+,k,j,m}$ by (6.3). Thus $F_{\theta^+,k,j,m}$ is nowhere dense.

Finally the set of $(\theta, \Pi) \in \Theta \times \mathfrak{M}^+(\Theta^+)$ such that (6.2) fails is the union of all $F_{\theta^+,k,j,m}$ over θ^+ in a countable, dense subset of Θ^+, $k \in \mathbb{N}$, $j \in \mathbb{N}$, $m \in \mathbb{N}$. This follows since every open U contains θ^+ with $\mathbb{1}_{\{\theta^+\}} \leq \chi_k(\cdot; \theta^+) \leq \mathbb{1}_U$. □

Thus, except for a relatively small collection of pairs of (θ, Π), the posterior distribution is inconsistent. The second assertion of the theorem even shows that the posterior distribution visits every open set in the parameter set infinitely often, thus "wanders aimlessly around the parameter space." While this result cautions about the naive use of Bayesian methods, it does not mean that Bayesian methods are useless. Indeed, a pragmatic Bayesian's only aim may be to find some prior that is consistent at various parameter values (and complies with his subjective belief, if available). Plenty of such priors may exist, even though "many more" are not appropriate. This phenomenon may be compared with the role of differentiable functions within the class of all functions. Functions that are differentiable at some point are a small minority in the topological sense, and nowhere differentiable functions are abundant. Nevertheless, the functions which we generally work with are differentiable at nearly all places.

Dirichlet process priors can be seen to be consistent by inspection of the (Dirichlet) posterior (see Chapter 4), and this extends to mixtures of Dirichlet processes, provided the prior precision parameters of the Dirichlet components are uniformly bounded (see (4.29) and the subsequent discussion in Section 4.5). The following example shows that boundedness of the prior precision cannot be missed. It also exemplifies that a convex mixture of consistent priors may itself be inconsistent.

Example 6.13 (Mixtures of Dirichlet processes) Consider again estimating a discrete distribution on $\mathbb{N}$ based on a random sample $X_1, \ldots, X_n$ from a distribution $\theta \in \Theta = \mathfrak{M}(\mathbb{N}) =$

$\mathbb{S}_\infty$. Consider a prior of the form $\Pi = \frac{1}{2}\mathrm{DP}(\alpha) + \frac{1}{2}\delta_\phi$, for α and ϕ the elements of Θ satisfying

$$\alpha_i = 2^{-i}, \qquad\qquad \phi_i \propto \frac{1}{i(\log i)^2}.$$

Then the posterior distribution in the model with observations $X_1, \ldots, X_n|\,\theta \sim \theta$ and prior $\theta \sim \Pi$ satisfies $\Pi_n(\cdot|\, X_1, \ldots, X_n) \rightsquigarrow \delta_\phi$, θ_0^∞-almost surely, for every $\theta_0 \in \Theta$ such that $\theta_{0i} = \phi_i$ for some $i \geq 10$ and the Kullback-Leibler divergence $K(\theta_0; \phi) < \infty$. In other words, the posterior distribution is strongly *in*consistent whenever $\theta_0 \neq \phi$. This is true in spite of the fact that the prior has full support, through its $\mathrm{DP}(\alpha)$-component.

If the Dirac measure δ_ϕ is viewed as a Dirichlet process with center measure ϕ and infinite precision, then the prior Π is a mixture of two Dirichlet processes. Alternatively, the same inconsistency can be created with infinite mixtures of Dirichlet processes of finite precision. The idea is to let the components have precision tending to infinity, so that in the limit they approximate the degenerate Dirichlet δ_ϕ: as alternate prior, consider $\Pi = \sum_{j=1}^\infty 2^{-j}\mathrm{DP}(\alpha_j)$, for $\alpha_j \in \Theta$ is defined by

$$\bar{\alpha}_{j,i} \propto \begin{cases} \phi_i, & \text{if } i \leq j, \\ 2^{-i}, & \text{otherwise}\,, \end{cases} \qquad\qquad |\alpha_j| \to \infty.$$

In this setting the posterior distribution behaves the same as before.

The intuition is that "thick-tailed" distributions θ (with $\theta_i \to 0$ slowly) are unexpected according to every Dirichlet prior $\mathrm{DP}(\alpha_j)$, but less so as $j \to \infty$. A thick-tailed true distribution will produce large observations, which makes the posterior distribution choose for the heavy-tailed components of the prior.

The proofs of these results can be based on the general Theorem 6.17 ahead. For the two-component mixture this is particularly straightforward. The one-point set $\{\phi\}$ has Kullback-Leibler divergence $K(\theta_0; \phi)$ equal to a finite constant c, and its prior mass is bigger than $\frac{1}{2}$, both by construction. Thus the posterior distributions contract to $\{\phi\}$ (and hence are inconsistent) if for every $\epsilon > 0$ there exist tests such that $P_{\theta_0}^n \psi_n \to 0$ and $\int_{d(\theta,\phi)>\epsilon} P_\theta^n(1 - \psi_n) \leq e^{-nC}$ for some constant $C > 0$. Now for any $c_n > 0$ and DP_α the Dirichlet prior,

$$P_{\theta_0}^n(X_{(n)} \leq c_n) = \Bigl(1 - \sum_{i>c_n} \theta_{0i}\Bigr)^n \leq \Bigl(1 - \frac{1}{\log c_n}\Bigr)^n,$$

$$\int_{\theta\neq\phi} P_\theta^n(X_{(n)} > c_n)\, d\Pi(\theta) \leq \tfrac{1}{2}n \int P_\theta(X_1 > c_n)\, d\mathrm{DP}_\alpha(\theta) \leq \tfrac{1}{2}n \sum_{i>c_n} \alpha_i \leq 2^{-c_n}.$$

Therefore, the tests $\psi_n = \mathbb{1}\{X_{(n)} \leq c_n\}$ for $c_n = e^{\sqrt{n}}$ satisfy the requirements, and contraction to ϕ follows.

For the mixture prior we first note that $\mathrm{DP}(\alpha_j) \rightsquigarrow \delta_\phi$ as $j \to \infty$, by Theorem 4.16 as $\bar{\alpha}_j \rightsquigarrow \phi$ and $|\alpha_j| \to \infty$. Therefore $\liminf_{j\to\infty} \mathrm{DP}_{\alpha_j}(\theta\colon K(\theta_0; \theta) < c)$ is bounded below by $\delta_\phi(\theta\colon K(\theta_0; \theta) < c) = 1$, by the Portmanteau theorem, and hence $\Pi(\theta\colon K(\theta_0; \theta) < c) > 0$. Next we use the same tests ψ_n, but replace the estimate of the errors of the second kind by

$$\int_{\theta\neq\phi} P_\theta^n(X_{(n)} > c_n)\, d\Pi(\theta) \le \sum_{j<c_n} 2^{-j}\tfrac{1}{2}n \sum_{i>c_n} \alpha_{j,i} + \sum_{j\ge c_n} 2^{-j}.$$

This is readily seen to be exponentially small.

The inconsistency in the preceding example is caused by having too little prior mass near the true value of the parameter. More drastic examples of inconsistency occur in semiparametric problems. In the following example we consider the problem of estimating a location parameter θ based on observations $X_i = \theta + e_i$, with a (symmetrized) Dirichlet process prior on the distribution of the errors. Consistency of the posterior distribution for θ then depends strongly on the combination of shapes of the base measure and the error distribution. This is a bit surprising, as the parameter of interest is only one-dimensional. In the example, the location parameter is identified by the symmetry of the error distribution, but similar results hold if the median of the errors is zero and the Dirichlet prior is conditioned to have median zero.

Example 6.14 (Symmetric location) Consider a sample of observations $X_i = \theta + e_i$, where θ is an unknown constant and the errors $e_1, \ldots, e_n$ are an i.i.d. sample from a distribution H that is symmetric about 0. Thus θ is identified as the point of symmetry of the distribution of the observations.

We equip the error distribution H with a Dirichlet prior, and independently the parameter θ with a Nor(0, 1)-prior. As H is symmetric, a symmetrized Dirichlet is natural, but for simplicity consider first an ordinary (true) Dirichlet with a symmetric absolutely continuous base measure $\alpha := MC$, $M > 0$, C with density c. The resulting model can be written hierarchically as

$$X_i = Y_i + \theta, \qquad Y_1, \ldots, Y_n \mid H \overset{\text{iid}}{\sim} H, \qquad H \sim \mathrm{DP}(MC), \qquad \theta \sim \mathrm{Nor}(0, 1).$$

The marginal distribution of the variables $Y_1, \ldots, Y_n$ is given by the Pólya urn with base measure MC, described in Section 4.1.4. Given θ, the observations $X_1, \ldots, X_n$ are simple shifts of these variables and hence follow a Pólya urn with base measure $MC(\cdot - \theta)$. Conditioned on the event F that the outputs of the urn are all distinct, the outputs are simply a random sample from the base measure; in other words:

$$p(X_1, \ldots, X_n \mid \theta, F) = \prod_{i=1}^{n} c(X_i - \theta). \tag{6.4}$$

If the true (non-Bayesian) distribution of the observations is continuous, then the event F *will* occur with probability one, and it is unnecessary to consider its complement. By Bayes's rule the posterior density for θ is then proportional to the expression in the preceding display times the prior density of θ. We may analyze its performance by methods for parametric models: the infinite-dimensional parameter has been removed. However, although the expression in the display is a parametric likelihood, it is not the likelihood of the data, and the analysis must proceed by treating the expression as a misspecified likelihood. Under regularity conditions the posterior will asymptotically concentrate on the set of θ that minimizes the Kullback-Leibler divergence between the true distribution of $(X_1, \ldots, X_n)$ and the misspecified likelihood (see Theorem 8.36 ahead).

We do not pursue the full details of this analysis, but point out some striking results (found by Diaconis and Freedman 1986a). If the base measure C is standard Cauchy and the true error distribution H_0 has two point masses of size $1/2$ at the points $\pm a$, where $a > 1$, then there are two points of minimum Kullback-Leibler divergence: $\gamma_{+,-} := \theta_0 \pm \sqrt{a^2 - 1}$, for θ_0 the true location. If the true error distribution is continuous, but with a density that puts mass nearly $1/2$ close to $\pm a$, then this situation is approximated, and can be matched exactly by fine-tuning the distribution. Then the posterior distribution contracts to the points $\{\gamma_+, \gamma_-\}$ rather than to θ_0. It is even true that the posterior distribution will choose between the two points based on the fractions of observations near $\pm a$ and alternate between the γ_+ and γ_-, with the result that, for any $\epsilon > 0$, for *both* $\gamma = \gamma_+$ and $\gamma = \gamma_-$,

$$\limsup_{n\to\infty} \Pi_n(|\theta - \gamma| < \epsilon |\, X_1, \ldots, X_n) = 1, \qquad \text{a.s.}$$

Similar results are true if the base measure is a general t-distribution.

On the positive side, if the true error distribution is symmetric and strongly unimodal, then the true value θ_0 is the unique point of minimum of the Kullback-Leibler divergence, and consistency will obtain. Furthermore, if the base measure is log-concave (like the normal or double-exponential distribution), then consistency obtains at any symmetric density.

Regrettably, the negative performance of the posterior distribution does not disappear if the prior for the error distribution H is symmetrized. The observations for a symmetrized Dirichlet process prior (see Section 4.6.1) on H, and an independent Nor(0, 1)-prior on θ can be written $X_i = S_i Y_i + \theta$, for the hierarchy

$$\theta \sim \text{Nor}(0, 1), \quad S_1, \ldots, S_n \overset{\text{iid}}{\sim} \text{Rad}, \quad Y_1, \ldots, Y_n |\, H \overset{\text{iid}}{\sim} H, \quad H \sim \text{DP}(C).$$

(The S_i are *Rademacher variables*: $\text{P}(S_i = 1) = \text{P}(S_i = -1) = 1/2$.) As before, the marginal distribution of $(Y_1, \ldots, Y_n)$ follows the Pólya urn scheme (4.13). Presently we need to consider this distribution, both given the event F that there are no ties among the Y_i and given the events $G_{k,l}$ that $Y_k = Y_l$ is the only tie (for $1 \leq k < l \leq n$). Given F the variables $Y_1, \ldots, Y_n$ are a random sample from $C(\cdot - \theta)$, as before, and taking account of the symmetry of C, we again find formula (6.4) for the likelihood (given F). The posterior distribution of θ, given F, is continuous with density proportional to (6.4) times $\phi(\theta)$.

In the present situation there is no one-to-one relationship between ties among the X_i and the Y_i, whence F is *not* the event that the X_i are distinct. In fact, on the event $\{S_k = S_l\}$ we have $Y_k = Y_l$ if and only if $X_k = X_l$, but on the event $\{S_k = -S_l\}$ we have $Y_k = Y_l$ if and only if $X_k + X_l = 2\theta$. Because the second possibility will arise in the data (for some θ), we also consider the likelihood given $G_{k,l}$. Given $G_{k,l}$, the Pólya scheme produces $n - 1$ i.i.d. variables Y_i for $i \neq l$ from the base measure C and sets $Y_l = Y_k$. The resulting X_i consist of a random sample $(X_i : i \neq l)$ from C and have either $X_l = X_k$ (when $S_k = S_l$) or $X_k + X_l = 2\theta$ (when $S_k = -S_l$), both with probability $1/2$. The conditional likelihood given the intersection $G'_{k,l} = G_{k,l} \cap \{S_k = -S_l\}$ can be written symbolically in the form

$$p(X_1, \ldots, X_n |\, \theta, G'_{k,l}) = \prod_{\substack{i=1 \\ i \neq l}}^{n} c(X_i - \theta)\delta_{2\theta}(X_k + X_l).$$

This implies that the posterior distribution of θ given $X_1, \ldots, X_n, G'_{k,l}$ is a Dirac measure at the point $\theta_{k,l} := (X_k + X_l)/2$.

The posterior distribution, given the union $E := F \cup \cup_{k<l} G'_{k,l}$, is the convex combination of the posterior distributions given the events in this union, with weights equal to $\mathrm{P}(F|\, X_1, \ldots, X_n, E)$ and $\mathrm{P}(G'_{k,l}|\, X_1, \ldots, X_n, E)$. From the Pólya urn scheme it is easy to see that the unconditional probabilities of F and a single $G_{k,l}$ are equal to $M^n/M^{[n]}$ and $M^{n-1}/M^{[n]}$, respectively, whence the prior weights of these events satisfy $\mathrm{P}(F|\, E) = M\mathrm{P}(G'_{k,l}|\, E)$. For the posterior weights we compute

$$p(X_1, \ldots, X_n|\, F) = \int \prod_{i=1}^{n} c(X_i - \theta)\, \phi(\theta)\, d\theta,$$

$$p(X_1, \ldots, X_n|\, G'_{k,l}) = \int \prod_{\substack{i=1 \\ i \neq l}}^{n} c(X_i - \theta) \delta_{2\theta}(X_k + X_l)\, \phi(\theta)\, d\theta.$$

The second marginal conditional distribution shifts the $(n-1)$-dimensional, singular distribution of $X_1, \ldots, X_n|\, \theta, G'_{k,l}$ over the full space $\mathbb{R}^n$, and is equivalent to Lebesgue measure (if C has full support). Therefore, the posterior weights of the events F and $G'_{k,l}$ satisfy

$$\mathrm{P}(F|\, X_1, \ldots, X_n, E) \propto \int \prod_{i=1}^{n} c(X_i - \theta)\, \phi(\theta)\, d\theta M,$$

$$\mathrm{P}(G'_{k,l}|\, X_1, \ldots, X_n, E) \propto \int \prod_{\substack{i=1 \\ i \neq l}}^{n} c(X_i - \theta) \delta_{2\theta}(X_k + X_l)\, \phi(\theta)\, d\theta,$$

where the inverse proportionality factor is the sum of the right sides of the first equation and of $n(n-1)/2$ second equations (for $1 \leq k < l \leq n$). Thus we find that the posterior distribution of θ given E (and the observations) is the convex combination of a continuous distribution and a discrete distribution on the points $\theta_{k,l}$.

If the true distribution of the observations is continuous, then all observations X_i will be distinct with probability one, and so will all pairs $X_k + X_l$. Since this sure event can be seen to be contained in $E := F \cup \cup_{k<l} G'_{k,l}$, it is unnecessary to consider the posterior distribution outside E.

The remainder of the analysis is again to study the posterior of a misspecified likelihood. The analysis is more involved, as both the absolutely continuous and the discrete parts of the posterior distribution may dominate, depending on the true density of the observations (in particular the induced density of the pairs $(X_k + X_l)/2$). However, the final result turns out to be the same (see Diaconis and Freedman 1986a for details).

6.4 Schwartz's Theorem

In this section we take the parameter equal to a probability density p, belonging to a parameter set $\mathcal{P}$, a class of probability densities relative to a given dominating measure ν on a sample space $(\mathfrak{X}, \mathscr{X})$. The parameter set $\mathcal{P}$ is equipped with an appropriate topology and σ-field, and the prior is a probability measure Π on $\mathcal{P}$. We consider estimating p based on a random sample $X_1, \dots, X_n$ of observations, with p_0 denoting the true density. As notational convention we use the corresponding uppercase letter P to denote the probability measure specified by a density p.

Throughout the section the prior is considered to be fixed. Priors that change with n are considered in Section 6.7.

A key condition for posterior consistency is that the prior assigns positive probability to any Kullback-Leibler (or KL) neighborhood of the true density. Write $K(p_0; p) = \int p_0 \log(p_0/p)\, d\nu$ for the Kullback-Leibler divergence, and for a set $\mathcal{P}_0$ of densities let $K(p_0; \mathcal{P}_0) = \inf_{p \in \mathcal{P}_0} K(p_0; p)$ be the minimal divergence of p_0 to an element of $\mathcal{P}_0$.

Definition 6.15 (Kullback-Leibler property) A density p_0 is said possess the *Kullback-Leibler property* relative to a prior Π if $\Pi(p: K(p_0; p) < \epsilon) > 0$ for every $\epsilon > 0$. This is denoted $p_0 \in \mathrm{KL}(\Pi)$, and we also say that p_0 belongs to the *Kullback-Leibler support* of Π.

Restricted to a class of densities on which pointwise evaluation is continuous, the Kullback-Leibler divergence is a measurable function of its second argument relative to the Borel σ-field of the total variation metric, and hence Kullback-Leibler neighborhoods are measurable in the space of densities equipped with this Borel σ-field (see Problem B.5). However, for our purpose it is sufficient to interpret the Kullback-Leibler property in the sense of inner probability: it suffices that there exist measurable sets $\mathcal{B} \subset \{p: K(p_0; p) < \epsilon\}$ with $\Pi(\mathcal{B}) > 0$.

Schwartz's theorem is the basic result on posterior consistency for dominated models. It has two conditions: the true density p_0 should belong to the Kullback-Leibler support of the prior, and the hypothesis $p = p_0$ should be testable against complements of neighborhoods of p_0. The first is clearly a Bayesian condition, but the second may be considered a condition to enable recovery of p_0 by any statistical method. Although in its original form the theorem has limited applicability, extensions go deeper, and lead to a rich theory of posterior consistency. Also the theory of contraction rates, developed in Chapter 8, uses similar ideas.

In the present context *tests* ϕ_n are understood to refer both to measurable mappings $\phi_n: \mathfrak{X}^n \to [0, 1]$ and to the corresponding statistics $\phi_n(X_1, \dots, X_n)$. We write $P^n \phi_n$ for the power $\mathrm{E}_p \phi_n(X_1, \dots, X_n) = \int \phi_n \, dP^n$ of the tests. The topology in the following theorem is assumed to have a countable local basis (e.g. metrizable), but is otherwise arbitrary.

Theorem 6.16 (Schwartz) *If $p_0 \in \mathrm{KL}(\Pi)$ and for every neighborhood $\mathcal{U}$ of p_0 there exist tests ϕ_n such that $P_0^n \phi_n \to 0$ and $\sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \to 0$, then the posterior distribution $\Pi_n(\cdot \mid X_1, \dots, X_n)$ in the model $X_1, \dots, X_n \mid p \overset{iid}{\sim} p$ and $p \sim \Pi$ is strongly consistent at p_0.*

Proof It must be shown that $\Pi_n(\mathcal{U}^c \mid X_1, \dots, X_n) \to 0$ almost surely, for every given neighborhood $\mathcal{U}$ of p_0. By Lemma D.11 it is not a loss of generality to assume that the tests ϕ_n as in the theorem have exponentially small error probabilities in the sense that $P_0^n \phi_n \le e^{-Cn}$ and $\sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \le e^{-Cn}$, for some positive constant C. By the Borel-Cantelli lemma $\phi_n \to 0$ almost surely under P_0^∞. Then the theorem follows from an application of Theorem 6.17(a), with $\mathcal{P}_0 = \{p: K(p_0; p) < \epsilon\}$ a Kullback-Leibler neighborhood of size $c = \epsilon < C$ and $\mathcal{P}_n = \mathcal{U}^c$. □

Part (a) of the following theorem contains the crux of Schwartz's theorem. It has a much wider applicability. For instance, it was used in Example 6.13 to prove the *in*consistency of certain priors. Part (b) of the result shows that the posterior does not concentrate on sets of exponentially small mass. Although it is a special case of (a) it is often applied explicitly to extend the applicability of Schwartz's theorem.

Theorem 6.17 (Extended Schwartz) *If there exist a set $\mathcal{P}_0 \subset \mathcal{P}$ and number c with $\Pi(\mathcal{P}_0) > 0$ and $K(p_0; \mathcal{P}_0) \le c$, then $\Pi_n(\mathcal{P}_n|\, X_1, \dots, X_n) \to 0$ a.s. $[P_0^\infty]$ for any sets $\mathcal{P}_n \subset \mathcal{P}$ such that either* (a) *or* (b) *holds for some constant $C > c$:*

(a) *there exist tests ϕ_n such that $\phi_n \to 0$ a.s. $[P_0^\infty]$ and $\int_{\mathcal{P}_n} P^n(1-\phi_n)\, d\Pi(p) \le e^{-Cn}$.*
(b) $\Pi(\mathcal{P}_n) \le e^{-Cn}$.

In particular, the posterior distribution $\Pi_n(\cdot|\, X_1, \dots, X_n)$ in the model $X_1 \dots, X_n|\, p \overset{iid}{\sim} p$ and $p \sim \Pi$ is strongly consistent at p_0 if $p_0 \in \mathrm{KL}(\Pi)$ and for every neighborhood $\mathcal{U}$ of p_0 there exists a constant $C > 0$, measurable partitions $\mathcal{P}_n = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$, and tests ϕ_n such that

(i) $P_0^n \phi_n < e^{-Cn}$ *and* $\sup_{p \in \mathcal{P}_{n,1} \cap \mathcal{U}^c} P^n(1-\phi_n) < e^{-Cn}$,
(ii) $\Pi(\mathcal{P}_{n,2}) < e^{-Cn}$.

Proof The second part of the theorem follows by applying, for every given neighborhood $\mathcal{U}$ of p_0 of square radius smaller than C, the first part (a) with the choices $\mathcal{P}_0 = \mathcal{U}$ and $\mathcal{P}_n = \mathcal{U}^c \cap \mathcal{P}_{n,1}$, and the first part (b) with the choices $\mathcal{P}_0 = \mathcal{U}$ and $\mathcal{P}_n = \mathcal{P}_{n,2}$.

For the proof of the first part we first show that, for any $c' > c$, eventually a.s. $[P_0^\infty]$:

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \ge \Pi(\mathcal{P}_0) e^{-c'n}. \tag{6.5}$$

The integral in the left side is bounded below by $\Pi(\mathcal{P}_0) \int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_0(p)$, for Π_0 the renormalized restriction of Π to $\mathcal{P}_0$. By Jensen's inequality its logarithm is bounded below by $\log \Pi(\mathcal{P}_0) + \int \log \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_0(p)$. The second term times n^{-1} coincides with the average $n^{-1} \sum_{i=1}^n \int \log(p/p_0)(X_i)\, d\Pi_0(p)$, and tends almost surely to its expectation, by the strong law of large numbers. By Fubini's theorem the expectation is $-\int K(p_0; p)\, d\Pi_0(p) \ge -c$, by the assumption and the definition of Π_0. This concludes the proof of (6.5).

For the proof of (a) we note that, in view of Bayes's rule (1.1), with probability one under P_0^∞,

$$\Pi_n(\mathcal{P}_n|\, X_1, \dots, X_n) \le \phi_n + \frac{(1-\phi_n) \int_{\mathcal{P}_n} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}{\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}.$$

The first term on the right tends to zero by assumption. By (6.5), the denominator of the second term is bounded below by $\Pi(\mathcal{P}_0) e^{-c'n}$ eventually a.s., for every $c' > c$. The proof is complete once it is shown that $e^{c'n}$ times the numerator tends to zero almost surely, for some $c' > c$. By Fubini's theorem,

$$P_0^n\Big[(1-\phi_n)\int_{\mathcal{P}_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(p)\Big] = \int_{\mathcal{P}_n} P_0^n\Big[(1-\phi_n)\prod_{i=1}^n \frac{p}{p_0}(X_i)\Big]\,d\Pi(p)$$
$$\leq \int_{\mathcal{P}_n} P^n(1-\phi_n)\,d\Pi(p) \leq e^{-Cn}.$$

Since $\sum_n e^{c'n}e^{-Cn} < \infty$ if $c < c' < C$, it follows by Markov's inequality that $e^{c'n}$ times the numerator tends to zero almost surely.

Assertion (b) is the special case of (a) with the tests chosen equal to $\phi_n = 0$, □

Remark 6.18 If $p_0 \in \mathrm{KL}(\Pi)$, then the integral $I_n = \int \prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)$ satisfies $n^{-1}\log I_n \to 0$, almost surely $[P_0^\infty]$. Indeed, (6.5) applied with $\mathcal{P}_0$ a Kullback-Leibler neighborhood of radius ϵ gives $\liminf n^{-1}\log I_n \geq -\epsilon'$, for every $\epsilon' > \epsilon > 0$. Conversely, by Markov's inequality $P_0^n(n^{-1}\log I_n > \epsilon) \leq e^{-n\epsilon}P_0^n I_n \leq e^{-n\epsilon}$, for every n, and therefore we have $\limsup n^{-1}\log I_n \leq 0$, by the Borel-Cantelli lemma.

The convergence $n^{-1}\log I_n \to 0$ shows that the lower bound $e^{-c'n}$ in (6.5) is pessimistic: the constant c' should actually be (close to) zero. We shall exploit more accurate bounds with $c' = c'_n \to 0$ only when discussing rates of convergence in Chapter 8. Then we shall take the *amount* of prior mass near p_0 into account; exploitation of only its positiveness (the Kullback-Leibler property) makes Schwartz's theory unnecessarily weak.

Example 6.19 (Finite-dimensional models) If the model is smoothly parameterized by a finite-dimensional parameter that varies over a bounded set, then consistent tests as required in Theorem 6.16 exist under mere regularity conditions. For unbounded Euclidean sets, some conditions are needed. (See e.g. Le Cam 1986, Chapter 16, or van der Vaart 1998, Lemmas 10.4 and 10.6.)

Example 6.20 (Weak topology) For the weak topology on the set of probability measures (corresponding to the densities in $\mathcal{P}$) consistent tests as in Theorem 6.16, always exist. Therefore, the posterior distribution is consistent for the weak topology at any density p_0 that has the Kullback-Leibler property for the prior.

To construct the tests, observe that sets of the type $\mathcal{U} = \{p: P\psi < P_0\psi + \epsilon\}$, for continuous functions $\psi: \mathfrak{X} \to [0,1]$ and $\epsilon > 0$, form a subbase for the weak neighborhood system at a probability measure P_0. By Hoeffding's inequality, Theorem K.1, the test $\phi_n = \mathbb{1}\{n^{-1}\sum_{i=1}^n \psi(X_i) > P_0\psi + \epsilon/2\}$ satisfies $P_0^n\phi_n \leq e^{-n\epsilon^2/2}$ and $P^n(1-\phi_n) \leq e^{-n\epsilon^2/2}$, for any $P \in \mathcal{U}^c$. A general neighborhood contains a finite intersection of neighborhoods from the subbase, and can be tested by the maximum of the tests attached to every of the subbasic neighborhoods. The errors of the first and second kind of this aggregate test are bounded by $Ne^{-n\epsilon^2/2}$ and $e^{-n\epsilon^2/2}$, respectively, for N the number of subbasic neighborhoods in the union.

Example 6.21 (Countable sample spaces) All the usual topologies (including total variation) on $\mathfrak{M}(\mathfrak{X})$ for a countable, discrete sample space $\mathfrak{X}$ coincide with the weak topology. The preceding example shows that the posterior distribution is strongly consistent at every parameter θ_0 such that

$$\Pi\Big(\theta = (\theta_1, \theta_2, \ldots) \colon \sum_{j=1}^{\infty} \theta_{0,j} \log \frac{\theta_{0,j}}{\theta_j} < \delta\Big) > 0. \tag{6.6}$$

In particular, for a prior with full support consistency pertains at any θ_0 that is essentially finite dimensional in the sense that $\{j \in \mathbb{N} \colon \theta_{0,j} > 0\}$ is finite.

Example 6.11 shows that consistency may not hold for arbitrary θ_0, and Theorem 6.12 suggests that (6.6) fails for "most" combinations (Π, θ_0). However, the condition is reasonable and many combinations seem to work.

In its original form Schwartz's theorem requires that the complement of every neighborhood of p_0 can be "tested away." For strong metrics, such as the $\mathbb{L}_1$-distance, such tests may not exist, but the posterior distribution may still be consistent. Theorem 6.17(b) can be very helpful here, as it shows that the posterior distribution will give negligible mass to a set of very small prior mass; such a set need not be tested. The following proposition shows that this possibility of splitting of the model in sets that are testable or of small prior mass is also necessary for posterior consistency (at an exponential rate).

Proposition 6.22 (Necessity of testing) *If $p_0 \in \mathrm{KL}(\Pi)$, then the following are equivalent for any sequence of sets $\mathcal{P}_n \subset \mathcal{P}$:*

(i) *There exists $C > 0$ such that $\Pi_n(\mathcal{P}_n | X_1, \ldots, X_n) \le e^{-nC}$, eventually a.s. $[P_0^\infty]$.*
(ii) *There exist partitions $\mathcal{P}_n = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$, test functions ϕ_n, and $C_1, C_2 > 0$ with:*

   (a) *$\phi_n \to 0$ a.s. $[P_0^\infty]$ and $\sup_{p \in \mathcal{P}_{n,1}} P^n(1 - \phi_n) \le e^{-C_1 n}$;*
   (b) *$\Pi(\mathcal{P}_{n,2}) \le e^{-C_2 n}$, eventually.*

Proof That (ii) implies that $\Pi_n(\mathcal{P}_n | X_1, \ldots, X_n) \to 0$ follows from Theorem 6.17. That this convergence happens at exponential speed can be seen by inspection of the proof.

To show that (i) implies (ii) let S_n be the event $\{\Pi_n(\mathcal{P}_n | X_1, \ldots, X_n) > e^{-Cn}\}$, and define tests $\phi_n = \mathbb{1}\{S_n\}$ and sets $\mathcal{P}_{n,2} = \{p \in \mathcal{P}_n \colon P^n(S_n^c) \ge e^{-Cn/2}\}$ and $\mathcal{P}_{n,1} = \mathcal{P}_n \setminus \mathcal{P}_{n,2}$. Then $\phi_n = 0$ for all sufficiently large n a.s. $[P_0^\infty]$ by (i), and $P^n(1 - \phi_n) = P(S_n^c) < e^{-Cn/2}$, for every $p \in \mathcal{P}_{n,1}$, by the definition of the set $\mathcal{P}_{n,1}$, which establishes (a) with $C_1 = C/2$. Because $e^{Cn/2} P^n(S_n^c) \ge 1$ for $p \in \mathcal{P}_{n,2}$,

$$e^{-Cn/2}\Pi(\mathcal{P}_{n,2}) \le \int_{\mathcal{P}_{n,2}} P^n(S_n^c)\, d\Pi(p) = \mathrm{E}\Pi_n(\mathcal{P}_{n,2} | X_1, \ldots, X_n)\mathbb{1}\{S_n^c\},$$

by Bayes's theorem. This is bounded above by e^{-Cn} by the definition of S_n and the fact that $\mathcal{P}_{n,2} \subset \mathcal{P}_n$, which yields (b) with $C_2 = C/2$. □

Appendix D discusses the construction of appropriate tests. It turns out that tests for *convex* alternatives exist as soon as these have a minimal, positive $\mathbb{L}_1$-distance from the null hypothesis P_0 (where a larger distance gives a bigger constant in the exponent). Unfortunately, the alternatives $\mathcal{U}^c$ that must be tested in the application of Schwartz's theorem are complements of balls around p_0 and hence are not convex. In general a positive distance to p_0 without convexity is not enough, and consistent tests may not exist.

Now tests for nonconvex alternatives can be constructed by covering them with convex sets, and aggregating the separate tests for each of the sets in the cover, simply by taking the maximum of the test functions: given nonrandomized tests the null hypothesis is rejected as soon as one of the tests rejects it. The probability of an error of the first kind of such an aggregated test depends on the number of sets in the cover. Typically we cover by balls for some metric, and then the number of tests is bounded by the *covering number* of the alternative. The ϵ-covering number of a set $\mathcal{P}$ is defined as the minimal number of d-balls of radius ϵ needed to cover $\mathcal{P}$, and denoted by $N(\epsilon, \mathcal{P}, d)$. Appendix C discusses these numbers in detail.

If combined with a partition $\mathcal{U}^c = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ in sets $\mathcal{P}_{n,1}$ that are tested and sets $\mathcal{P}_{n,2}$ of negligible prior mass, the appropriate cover refers to sets $\mathcal{P}_{n,1}$ that change with n. The covering numbers then depend on n. As shown in the following theorem they can be allowed to increase exponentially.

The theorem yields consistency relative to any distance d that is bounded above by (a multiple of) the Hellinger distance, for instance the $\mathbb{L}_1$-distance.

Theorem 6.23 (Consistency by entropy) *Given a distance d that generates convex balls and satisfies $d(p_0, p) \leq d_H(p_0, p)$ for every p, suppose that for every $\epsilon > 0$ there exist partitions $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ and a constant $C > 0$, such that, for sufficiently large n,*

(i) $\log N(\epsilon, \mathcal{P}_{n,1}, d) \leq n\epsilon^2$.
(ii) $\Pi(\mathcal{P}_{n,2}) \leq e^{-Cn}$.

Then the posterior distribution in the model $X_1, \ldots, X_n \,|\, p \overset{iid}{\sim} p$ and $p \sim \Pi$ is strongly consistent relative to d at every $p_0 \in \mathrm{KL}(\Pi)$.

Proof For given $\epsilon > 0$, cover the set $\mathcal{P}_{n,1}$ with a (minimal) collection of $N(\epsilon, \mathcal{P}_{n,1}, d)$ balls of radius ϵ. From every ball that intersects the set $\mathcal{U}^c := \{p: d(p, p_0) \geq 4\epsilon\}$, take an arbitrary point p_1 in $\mathcal{U}^c$. Then the balls of radius 2ϵ around the points p_1 cover $\mathcal{P}_{n,1} \cap \mathcal{U}^c$ and have minimal distance at least 2ϵ to p_0. By Proposition D.8 for every p_1 there exists a test ψ_n such that $P_0^n \psi_n \leq e^{-2n\epsilon^2}$ and $P^n(1-\psi_n) \leq e^{-2n\epsilon^2}$, for every p with $d(p, p_1) < 2\epsilon$. Define a test ϕ_n as the maximum of all tests ψ_n attached to some p_1 in this way. Then ϕ_n has power bigger than any of the individual tests, and hence satisfies $P^n(1-\phi_n) \leq e^{-2n\epsilon^2}$, for every $p \in \mathcal{P}_{n,1} \cap \mathcal{U}^c$. Furthermore, because ϕ_n is smaller than the sum of the tests ψ_n, its size is bounded by $(\#p_1)e^{-2n\epsilon^2} \leq e^{n\epsilon^2} e^{-2n\epsilon^2}$.

The assertion of the theorem is now a consequence of Theorem 6.17. □

Example 6.24 (Prior partition) If $p_0 \in \mathrm{KL}(\Pi)$ and for every $\epsilon > 0$ there exists a partition $\mathcal{P} = \cup_j \mathcal{P}_j$ in sets of Hellinger diameter smaller than ϵ and $\alpha < 1$ such that $\sum_j \Pi(\mathcal{P}_j)^\alpha < \infty$, then the posterior distribution is strongly Hellinger consistent at p_0.

This sufficient condition is attractive in that it refers to the prior only, and seemingly avoids a testing or entropy condition, but in fact it implies the conditions of Theorem 6.23, for $d = d_H$. To see this, given $\epsilon, c > 0$ let $J = \{j: \Pi(\mathcal{P}_j) \geq e^{-cn}\}$, and consider the partition of $\mathcal{P}$ into the two sets $\mathcal{P}_{n,1} = \cup_{j \in J} \mathcal{P}_j$ and $\mathcal{P}_{n,2} = \cup_{j \notin J} \mathcal{P}_j$. Because $\sum_j \Pi(\mathcal{P}_j) \leq 1$ the cardinality of J is at most e^{cn}, whence $\log N(\epsilon, \mathcal{P}_{n,1}, d_H) \leq \log \#J \leq cn$, as every partitioning set fits into a single ball of radius ϵ. Furthermore $\Pi(\mathcal{P}_{n,2}) \leq \sum_{j\notin J} \Pi(\mathcal{P}_j) \leq \sum_j \Pi(\mathcal{P}_j)^\alpha (e^{-cn})^{1-\alpha}$, which is exponentially small.

Sequence of priors

In the preceding, the prior is assumed not to depend on the number of observations. Schwartz's theorem extends to a sequence Π_n of prior distributions, but with weak rather than strong consistency as conclusion. We say that a density p_0 satisfies the *Kullback-Leibler property* relative to a sequence of priors Π_n if $\liminf_{n\to\infty} \Pi_n(p\colon K(p_0; p) < \epsilon) > 0$, for every $\epsilon > 0$. Weak consistency may be derived under this assumption in the setting of Theorem 6.17, and hence also in the setting of Theorem 6.23.

Theorem 6.25 (Schwartz, sequence of priors) *If p_0 satisfies the Kullback-Leibler property relative to the sequence of priors Π_n and for every neighborhood $\mathcal{U}$ of p_0 there exists a constant $C > 0$, a measurable partition $\mathcal{P}_n = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$, and tests ϕ_n such that $P_0^n \phi_n < e^{-Cn}$ and $\sup_{p\in\mathcal{P}_{n,1}\cap\mathcal{U}^c} P^n(1-\phi_n) < e^{-Cn}$ and $\Pi_n(\mathcal{P}_{n,2}) < e^{-Cn}$, then the posterior distribution $\Pi_n(\cdot|\, X_1, \ldots, X_n)$ in the model $X_1 \ldots, X_n |\, p \overset{iid}{\sim} p$ and $p \sim \Pi_n$ is weakly consistent at p_0.*

The theorem can be proved by proceeding as in the proof of Theorem 6.17, except that the a.s. lower bound (6.5) on the denominator of the posterior distribution is replaced by the lower bound given in the following lemma, for a sufficiently large constant D and ϵ such that $2D\epsilon < C$.

Lemma 6.26 (Evidence lower bound) *For any probability measure Π on $\mathcal{P}$, and any constants $D > 1$ and $\epsilon \geq n^{-1}$, with P_0^n-probabiity at least $1 - D^{-1}$,*

$$\int \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p) \geq \Pi(p\colon K(p_0; p) < \epsilon) e^{-2Dn\epsilon}. \tag{6.7}$$

Proof The integral becomes smaller by restricting it to the set $B := \{p\colon K(p_0; p) < \epsilon\}$. By next dividing the two sides of the equation by $\Pi(B)$, we can write the inequality in terms of the prior restricted and renormalized to a probability measure on B. Thus without loss of generality we may assume that Π is concentrated on B. By Jensen's inequality applied to the logarithm,

$$\log \int \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p) \geq \int \log \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}\, d\Pi(p) =: Z.$$

Hence, the probability of the complement of the event in (6.7) is bounded above by $\mathrm{P}(Z < -2Dn\epsilon) \leq \mathrm{E}(-Z)^+/(2Dn\epsilon)$, by Markov's inequality. By another application of Jensen's inequality and since $(-\log x)^+ = \log_+(1/x)$,

$$\mathrm{E}(-Z)^+ \leq \mathrm{E}\Big[\int \log_+ \prod_{i=1}^n \frac{p_0(X_i)}{p(X_i)}\, d\Pi(p)\Big] = \int K^+(p_0^n; p^n)\, d\Pi(p).$$

Because $K^+ \leq K + \sqrt{K/2}$, by Lemma B.13, and the integral is over B by construction, the right is bounded above by $n\epsilon + \sqrt{n\epsilon/2} \leq 2n\epsilon$, whence $\mathrm{P}(Z < -2Dn\epsilon) \leq 1/D$. □

6.5 Tail-Free Priors

The supports of Dirichlet and Pólya tree priors are (often) not dominated, which precludes a direct application of Schwartz's theorem. However, they were among the earliest examples of consistent priors on measures. The key to their (weak) consistency is that the posterior distribution of the masses of the sets in a partition depends only on the vector of counts of the sets (see Lemma 3.14), which reduces the setting to one involving a finite-dimensional multinomial vector.

Theorem 6.27 *Let the prior distribution* Π *be tail-free with respect to a sequence* $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m\}$ *of successive binary partitions whose union generates the Borel* σ*-field and is convergence-forcing for weak convergence, and with splitting variables* $(V_{\varepsilon 0} : \varepsilon \in \mathcal{E}^*)$ *of which every finite-dimensional subvector has full support (a unit cube). Then the posterior distribution* $\Pi_n(\cdot| X_1, \ldots, X_n)$ *in the model* $X_1, X_2, \ldots | P \overset{iid}{\sim} P$ *and* $P \sim \Pi$ *is strongly consistent with respect to the weak topology, at any probability measure* P_0.

Proof Because $\cup_m \mathcal{T}_m$ is convergence-forcing for weak convergence, the topology generated by the metric $d(P, Q) = \sum_{\varepsilon \in \mathcal{E}^*} |P(A_\varepsilon) - Q(A_\varepsilon)| 2^{-2|\varepsilon|}$ is at least as strong as the weak topology. Therefore, it is certainly sufficient to prove consistency relative to d. Since every d-ball of P_0 contains a set of the form $\cap_{A \in \mathcal{T}_m} \{P : |P(A) - P_0(A)| < \delta\}$ (for sufficiently large m and small δ), it suffices to show that the posterior distribution of $(P(A) : A \in \mathcal{T}_m)$ is consistent (relative to the Euclidean metric on $\mathbb{R}^{2^m}$).

By Lemma 3.14, this posterior distribution is the same as the posterior distribution of $(P(A) : A \in \mathcal{T}_m)$, given the vector of counts of the cells in $\mathcal{T}_m$. Thus the problem reduces to a finite-dimensional multinomial problem, with parameter $\theta = (P(A) : A \in \mathcal{T}_m)$, with induced prior distribution θ and with true parameter $\theta_0 = (P_0(A) : A \in \mathcal{T}_m)$. Because $\Pi(P : |P(A) - P_0(A)| < \epsilon, A \in \mathcal{T}_m) > 0$, for any $\epsilon > 0$ by the condition of full support of the splitting variables (cf. the proof of Theorem 3.10), the true vector θ_0 is in the support of the prior for θ. The proof can now be completed by an application of Example 6.21. □

The above theorem implies in particular that the posterior based on a Pólya tree prior is consistent with respect to the weak topology if all of its parameters α_ε, for $\varepsilon \in \mathcal{E}^*$, are positive. For the special case of a Dirichlet process this conclusion was reached in Corollary 4.17 even without the restriction on the support.

6.6 Permanence of the Kullback-Leibler Property

The Kullback-Leibler property is useful in many consistency studies, in particular in semiparametric problems, also beyond density estimation, because it is stable under various operations, unlike the tail-free property. In this section we show that the property is preserved under forming mixtures, symmetrization, averaging and conditioning.

Proposition 6.28 (Mixture priors) *In the model $p|\xi \sim \Pi_\xi$ and $\xi \sim \rho$ we have $p_0 \in \mathrm{KL}(\int \Pi_\xi \, d\rho(\xi))$ for any p_0 such that $\rho\{\xi\colon p_0 \in \mathrm{KL}(\Pi_\xi)\} > 0$.*

Proof The mixture measure $\Pi = \int \Pi_\xi \, d\rho(\xi)$ satisfies, for any set B,

$$\Pi\{p\colon K(p_0; p) < \epsilon\} \geq \int_B \Pi_\xi(p\colon K(p_0; p) < \epsilon) \, d\rho(\xi).$$

For B the set of ξ such that $p_0 \in \mathrm{KL}(\Pi_\xi)$ the integrand is positive. This set has positive ρ-probability by assumption. □

Proposition 6.29 (Products) *If $p \sim \Pi_1$, $q \sim \Pi_2$, then $p_0 \times q_0 \in \mathrm{KL}(\Pi_1 \times \Pi_2)$ whenever $p_0 \in \mathrm{KL}(\Pi_1)$ and $q_0 \in \mathrm{KL}(\Pi_2)$.*

Proof This is a consequence of the additivity of Kullback-Leibler divergence: $K(p_0 \times q_0; p \times q) = K(p_0; p) + K(q_0; q)$. □

Many preservation properties follow from the fact that the Kullback-Leibler divergence decreases as the information in an experiment decreases. Specifically, if Y is an observation with density f, then the induced density p_f of a given measurable transformation $X = T(Y, U)$, where U is a random variable with a fixed distribution independent of Y, satisfies $K(p_f; p_g) \leq K(f; g)$, for any pair of densities f, g (see Lemma B.11). Equivalently, this is true if X is a randomization of Y through a Markov kernel, in the sense that p_f is a density of a variable X drawn from a conditional distribution $\Psi(\cdot|\, Y)$, and $Y \sim f$. (These two forms of randomization are in fact equivalent if $\mathfrak{X}$ is Polish.)

Now if $K(p_f; p_g) \leq K(f; g)$, then $\{g\colon K(f; g) < \epsilon\} \subset \{g\colon K(p_f; p_g) < \epsilon\}$, and hence $f \in \mathrm{KL}(\Pi)$ for a prior on f implies that $p_f \in \mathrm{KL}(\tilde\Pi)$, for $\tilde\Pi$ the induced prior on p_f.

Proposition 6.30 (Transformation) *If $p \sim \Pi$ and $\tilde\Pi$ is the induced prior on $p \circ T^{-1}$ for a given measurable transformation, then $p_0 \in \mathrm{KL}(\Pi)$ implies $p_0 \circ T^{-1} \in \mathrm{KL}(\tilde\Pi)$.*

This follows from the above discussion as a special case where $X = T(Y)$, not depending on the auxiliary random variable U.

Proposition 6.31 (Mixtures) *If $f \sim \Pi$ and $\tilde\Pi$ is the induced prior on the mixture density $p_f(x) = \int \psi(x; \theta) f(\theta) \, d\nu(\theta)$, for a given probability density kernel ψ, then $f_0 \in \mathrm{KL}(\Pi)$ implies $p_{f_0} \in \mathrm{KL}(\tilde\Pi)$.*

This follows from the above discussion with the aid of Example B.12.

Proposition 6.32 (Symmetrization) *If $p \sim \Pi$ is a density on $\mathbb{R}$ and $\bar\Pi$ is the induced prior on its symmetrization $\bar p(x) = (p(x) + p(-x))/2$, then $p_0 \in \mathrm{KL}(\Pi)$ implies that $\bar p_0 \in \mathrm{KL}(\bar\Pi)$. The same result is true for the symmetrization $\bar p(x) = \bar p(-x) = \frac{1}{2} p(x)$, for $x \geq 0$, of a density p on $[0, \infty)$.*

The conclusions follow from the above discussion by considering the transformation $T(X, U) = |X|\text{sign}(U - 1/2)$.

Proposition 6.33 (Invariance) *If $p \sim \Pi$ is a density on $\mathfrak{X}$ and $\bar{p}$ is a density of the invariant measure $\bar{P}(A) = \int P(gA)\,d\mu(g)$ induced by p under the action of a compact metrizable group $\mathfrak{G}$ on $\mathfrak{X}$ having normalized Haar measure μ, then $p_0 \in \mathrm{KL}(\Pi)$ implies that $\bar{p}_0 \in \mathrm{KL}(\bar{\Pi})$, for $\bar{\Pi}$ the induced prior.*[2] *If the action of the group $\mathfrak{G}$ is free and $\mathfrak{X}$ is Borel isomorphic to $\mathfrak{R} \times \mathfrak{G}$, where $\mathfrak{R}$ is a subset of $\mathfrak{X}$ isomorphic to the space of orbits as in Subsection 4.6.1, then the same is true for the invariant densities $\bar{p}$ induced by densities p on $\mathfrak{R}$ with respect to an arbitrary σ-finite measure ν through $\bar{p}(x) = p(y)$, where y is the orbit of x.*

This result clearly follows by considering the transformation $T(X, g) = gX$ with g having the known distribution ν, the normalized Haar measure on $\mathfrak{G}$.

Proposition 6.34 (Domination) *If for every $C > 1$ there exists a density $p_1 \in \mathrm{KL}(\Pi)$ with $p_0 \le Cp_1$ a.e., then $p_0 \in \mathrm{KL}(\Pi)$.*

Proof By Lemma B.14 the Kullback-Leibler divergence $K(p_0; p)$ is bounded above by $\log C + C(\delta + \sqrt{\delta/2})$ if $K(p_1; p) < \delta$. We can make this arbitrarily small by first choosing C close to 1 and next δ close to zero. □

Corollary 6.35 *If $\nu(\mathfrak{X}) < \infty$ and every strictly positive continuous probability density belongs to the Kullback-Leibler support of Π, then $p_0 \in \mathrm{KL}(\Pi)$ for every continuous p_0.*

Proof For given $\eta > 0$ the probability density $p_1 = C^{-1}(p_0 \vee \eta)$, for $C = \int (p_0 \vee \eta)\,d\nu$, is well defined, strictly positive and continuous, and $p_0 \le Cp_1$. Furthermore, $C \downarrow 1$ as $\eta \downarrow 0$ by the dominated convergence theorem. Since $p_1 \in \mathrm{KL}(\Pi)$, the result follows by Proposition 6.34. □

In the following proposition we write p_θ for the shifted density $p(\cdot - \theta)$ (with $p_{0,\theta}$ the shifted version of a density p_0), and $\bar{p}$ for the symmetrized density $x \mapsto \frac{1}{2}(p(x) + p(-x))$. Combined these notations yield the density $\overline{p_\theta}$ given by $x \mapsto \frac{1}{2}(p(x - \theta) + p(-x - \theta))$, which is symmetric about 0 (and not θ). If p is itself symmetric, then the latter density can also be written $\frac{1}{2}(p_\theta + p_{-\theta})$.

Proposition 6.36 *Given priors Π on symmetric densities on $\mathbb{R}$ and μ on $\mathbb{R}$, let $\bar{\Pi}$ denote the induced prior on $p_\theta(\cdot) = p(\cdot - \theta)$ when $(p, \theta) \sim \Pi \times \mu$. If p_0 is symmetric and $\overline{p_{0,\theta}} \in \mathrm{KL}(\Pi)$ for every sufficiently small $\theta > 0$, θ_0 is in the support of μ and $K(p_0; p_{0,\theta}) \to 0$ as $\theta \to 0$, then $p_{0,\theta_0} \in \mathrm{KL}(\bar{\Pi})$.*

Proof Since $K(p_{0,\theta_0}; p_\theta) = K(p_0; p_{\theta-\theta_0})$, we can assume that $\theta_0 = 0$. It can be verified that $K(p_0; p_\theta) = K(p_{0,\theta}; p)$ for symmetric densities p_0 and p. The right side can be

[2] If the dominating measure is invariant then $\bar{p}(x) = \int p\big(g(x)\big)\,d\mu(g)$.

decomposed as $K(p_{0,\theta}; \overline{p_{0,\theta}}) + P_{\theta_0}\log(\overline{p_{0,\theta}}/p) = K(p_{0,\theta}; \overline{p_{0,\theta}}) + K(\overline{p_{0,\theta}}; p)$, because the function $\log(\overline{p_{0,\theta}}/p)$ is symmetric. Because $\overline{p_{0,\theta}} = \frac{1}{2}(p_{0,\theta} + p_{0,-\theta})$, the first term on the right is bounded above by $\frac{1}{2}0 + \frac{1}{2}K(p_{0,\theta}; p_{0,-\theta})$. Combining all inequalities we find $K(p_0; p_\theta) \le \frac{1}{2}K(p_{0,\theta}; p_{0,-\theta}) + K(\overline{p_{0,\theta}}; p)$. The first term on the right is $\frac{1}{2}K(p_0; p_{0,-2\theta})$ and is arbitrarily small as θ is close to $\theta_0 = 0$; the second term is small with positive Π probability by the assumption that $\overline{p_{0,\theta}} \in \mathrm{KL}(\Pi)$. □

The Kullback-Leibler property is also preserved when one passes from the prior to the posterior given an observation, and uses the posterior as a new prior in combination with future data. An interesting aspect is that the prior can be improper, as long as the posterior is proper. (The Kullback-Leibler property for an improper prior is defined in the obvious way.)

Let $X_1, \ldots, X_m$ be i.i.d. with density p following a possibly improper prior distribution Π.

Proposition 6.37 *If* $p_0 \in \mathrm{KL}(\Pi)$, *then* $p_0 \in \mathrm{KL}(\Pi_n(\cdot|\, X_1, \ldots, X_m))$ *a.s.* $[P_0^m]$ *on the event* $\{\int \prod_{i=1}^m p(X_i)\, d\Pi(p) < \infty\}$.

Proof For any $\epsilon > 0$, on

$$E_\epsilon = \{\Pi_n(p\colon K(p_0; p) < \epsilon|\, X_1, \ldots, X_m) = 0\} \cap \{\int \prod_{i=1}^m p(X_i)\, d\Pi(p) < \infty\}$$

we have that

$$0 = \Pi_n(p\colon K(p_0; p) < \epsilon|\, X_1, \ldots, X_m) = \frac{\int_{K(p_0;p)<\epsilon} \prod_{i=1}^m p(X_i)\, d\Pi(p)}{\int \prod_{i=1}^m p(X_i) d\Pi(p)}.$$

Thus the numerator vanishes a.s. on E_ϵ. Integrating it over this event and applying Fubini's theorem, we find that $\int_{K(p_0;p)<\epsilon} P^m(E_\epsilon)\, d\Pi(p) = 0$. This implies that there exists $\tilde{p}$ with $K(p_0; \tilde{p}) < \epsilon$ such that $\tilde{P}^m(E_\epsilon) = 0$. Since, by the definition of Kullback-Leibler divergence, $K(p_0; \tilde{p}) < \epsilon < \infty$ is possible only if $P_0 \ll \tilde{P}^*$, it follows that $P_0^m(E_\epsilon) = 0$.

Finally let ϵ run through the positive rational numbers, and accumulate the corresponding null sets E_ϵ in a single null set. □

Remark 6.38 The preceding result implies that Schwartz's consistency theorem, Theorem 6.16, generalizes to improper priors Π for which the posterior distribution is proper for some m a.s. $[P_0^\infty]$. It suffices to apply the theorem conditionally on the first m observations, for m an index which gives a proper posterior, which can be viewed as the prior for the model with the remaining observations.

6.7 General Observations

In this section we return to the general set-up of the introduction, and consider a sequence of statistical experiments $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}\colon \theta \in \Theta_n)$, with parameter sets Θ_n, and observations $X^{(n)}$. Let Π_n stand for a sequence of prior distributions for θ. We assume that every $P_\theta^{(n)}$ admits a density $p_\theta^{(n)}$ with respect to a σ-finite measure $\nu^{(n)}$ on $\mathfrak{X}^{(n)}$, which is jointly measurable in the parameter and the observation, so that (a version of) the posterior distribution $\Pi_n(\cdot|\, X^{(n)})$ is given by Bayes's formula (1.1).

All entities may depend on n, including the prior Π_n and the "true" values $\theta_{n,0}$ of the parameter.

For probability densities p and q, let $K(p;q) = \int p\log(/q)$ and $V^+_{k,0}(p;q) = \int p((\log(p/q) - K(p;q))^+)^k$, respectively, the Kullback-Leibler divergence and kth order positive Kullback-Leibler variation between two probability densities p and q, as defined in (B.2).

Theorem 6.39 *If for some $k \geq 2$ and $c > 0$ there exist measurable sets $B_n \subset \Theta_n$ with $\liminf \Pi_n(B_n) > 0$ and*

$$\sup_{\theta\in B_n} \frac{1}{n} K(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta) \leq c, \qquad \sup_{\theta\in B_n} \frac{1}{n^k} V^+_{k,0}(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta) \to 0, \tag{6.8}$$

then $\Pi_n(\tilde\Theta_n|\, X^{(n)}) \to 0$ in $P^{(n)}_{\theta_{n,0}}$-probability for any sets $\tilde\Theta_n \subset \Theta_n$ such that either (a) *or* (b) *holds for a constant $C > c$,*

(a) *there exist tests ϕ_n with $P^{(n)}_{\theta_{n,0}}\phi_n \to 0$ and $\int_{\tilde\Theta_n} P^{(n)}_\theta(1-\phi_n)\, d\Pi_n(\theta) \leq e^{-Cn}$;*
(b) $\Pi_n(\tilde\Theta_n) \leq e^{-Cn}$.

The convergence is also in the almost sure sense if ϕ_n in (a) and $\mathbb{1}\{A^c_n\}$ for A_n defined in (6.10) *below tend to zero almost surely under $P^{(n)}_{\theta_{n,0}}$.*

Proof The proof is similar to that of Theorem 6.17. Let $\tilde\Pi_n$ be the renormalized restriction of Π_n to the set B_n. Then $\Pi_n \geq \Pi_n(B_n)\tilde\Pi_n$, and hence, by Jensen's inequality,

$$\begin{aligned} &\frac{1}{n}\log\Big[e^{c'n}\int \frac{p^{(n)}_\theta}{p^{(n)}_{\theta_{n,0}}}(X^{(n)})\, d\Pi_n(\theta)\Big] \\ &\qquad\qquad \geq c' + \int \frac{1}{n}\log\frac{p^{(n)}_\theta}{p^{(n)}_{\theta_{n,0}}}(X^{(n)})\, d\tilde\Pi_n(\theta) + \frac{1}{n}\log \Pi_n(B_n). \end{aligned}\tag{6.9}$$

Because the integrand has expectation $-n^{-1}K(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta) \geq -c$, for $\theta \in B_n$, the right-hand side is bigger than $c' - c - \epsilon + o(1)$, on the event

$$A_n := \Big\{\frac{1}{n}\int \Big(\log\frac{p^{(n)}_\theta}{p^{(n)}_{\theta_{n,0}}}(X^{(n)}) - P^{(n)}_{\theta_{n,0}}\log\frac{p^{(n)}_\theta}{p^{(n)}_{\theta_{n,0}}}\Big)\, d\tilde\Pi_n(\theta) > -\epsilon\Big\}. \tag{6.10}$$

By Markov's inequality followed by Jensen's inequality, we obtain

$$\begin{aligned} P^{(n)}_{\theta_{n,0}}(A^c_n) &\leq \frac{1}{(n\epsilon)^k}\mathrm{E}_{\theta_{n,0}}\Big|\int\Big(\log\frac{p^{(n)}_{\theta_{n,0}}}{p^{(n)}_\theta}(X^{(n)}) - P^{(n)}_{\theta_{n,0}}\log\frac{p^{(n)}_{\theta_{n,0}}}{p^{(n)}_\theta}\Big)^+ d\tilde\Pi_n(\theta)\Big|^k \\ &\leq \frac{1}{(n\epsilon)^k}\int V^+_{k,0}(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta)\, d\tilde\Pi_n(\theta). \end{aligned}$$

The right side tends to zero by assumption, for any $\epsilon > 0$.

If $c' > c$, then we can choose $\epsilon > 0$ so that $c' - c - \epsilon > 0$, and then the left side of (6.9) is bounded below by a positive number on the event A_n. Then the multiple

$$e^{c'n} \int (p_\theta^{(n)}/p_{\theta_{n,0}})^{(n)}(X^{(n)})\, d\Pi_n(\theta)$$

of the denominator in Bayes's formula tends to infinity on A_n, and hence, eventually,

$$P_{\theta_{n,0}}^{(n)} \Pi_n(\tilde{\Theta}_n | X^{(n)})(1-\phi_n)\mathbb{1}\{A_n\} \le e^{c'n} \int_{\tilde{\Theta}_n} P_{\theta_{n,0}}^{(n)}\Big[(1-\phi_n)\frac{p_\theta^{(n)}}{p_{\theta_{n,0}}^{(n)}}\Big]\, d\Pi_n(\theta).$$

This can be further bounded above by both $e^{c'n}\int_{\tilde{\Theta}_n} P_\theta^{(n)}(1-\phi_n)\, d\Pi_n(\theta)$ and $e^{c'n}\Pi_n(\tilde{\Theta}_n)$. These upper bounds tend to zero under (a) and (b), respectively, if c' is chosen such that $c < c' < C$.

Thus $\Pi_n(\tilde{\Theta}_n | X^{(n)})(1-\phi_n)\mathbb{1}\{A_n\}$ tends to zero in probability. We combine this with the convergence to zero of $\Pi_n(\tilde{\Theta}_n | X^{(n)})(\phi_n + \mathbb{1}\{A_n^c\}) \le \phi_n + \mathbb{1}\{A_n^c\}$, which is true by assumption and by construction.

The strengthening to almost sure convergence is valid as the convergence in probability is exponentially fast. □

Typically the preceding theorem will be applied with B_n equal to a neighborhood of the true parameter, chosen small so that c can be small, and $\tilde{\Theta}_n$ the complement of another neighborhood, which should be testable versus the true parameter or possess small prior mass in order to satisfy (a) or (b).

The Kullback-Leibler variations $V_{k,0}$ help to control the variability of the log-likelihood around its mean, so that the probability of the events A_n defined in (6.10) tends to one. Theorems 6.16 and 6.17 illustrate that consideration of only the Kullback-Leibler divergence itself may suffice. These results are limited to a fixed prior $\Pi_n = \Pi$ and based on the law of large numbers for i.i.d. variables, which is valid under existence of first moments only. A similar simplification is possible for other special cases, the crux being to ensure that the events A_n are asymptotically of probability one.

By a slightly more elaborate proof we also have the following variation of the theorem. It is restricted to a fixed prior and parameter set, and assumes that the experiments are linked across n, with the observation $X^{(n)}$ defined on a single underlying probability space and its law $P_\theta^{(n)}$ the image of a fixed measure $P_\theta^{(\infty)}$ on this space.

Theorem 6.40 (Fixed prior) *If there exist $c > 0$ and a measurable set $B \subset \Theta$ with $\Pi(B) > 0$ and, for every $\theta \in B$,*

$$\frac{1}{n}K(p_{\theta_0}^{(n)}; p_\theta^{(n)}) \le c, \qquad \frac{1}{n}\Big(\log\frac{p_\theta^{(n)}}{p_{\theta_0}^{(n)}}(X^{(n)}) - P_{\theta_0}^{(n)}\log\frac{p_\theta^{(n)}}{p_{\theta_0}^{(n)}}\Big) \to 0,\ a.s.\ [P_{\theta_0}^{(\infty)}], \quad (6.11)$$

then $\Pi_n(\tilde{\Theta}_n | X^{(n)}) \to 0$ almost surely $[P_{\theta_0}^{(\infty)}]$ for any sets $\tilde{\Theta}_n \subset \Theta_n$ such that either (a) with $\phi_n \to 0$ a.s. $[P_{\theta_0}^{(\infty)}]$ or (b) of Theorem 6.39 holds for some constant $C > c$.

Proof For given $\epsilon > 0$, let $\hat{\Pi}_n$ be the renormalized restriction of Π to the set

$$\hat{B}_n = \left\{\theta \in B\colon \inf_{k\geq n} \frac{1}{k}\Bigl(\log \frac{p_\theta^{(k)}}{p_{\theta_0}^{(k)}}(X^{(k)}) - P_{\theta_0}^{(k)} \log \frac{p_\theta^{(k)}}{p_{\theta_0}^{(k)}}\Bigr) \geq -\epsilon\right\}.$$

Although the set $\hat{B}_n \subset \Theta$ and measure $\hat{\Pi}_n$ are presently random, the argument leading to (6.9) in the proof of Theorem 6.39 still applies, and yields that the left side of (6.9) is bounded below by

$$c' - c - \epsilon + \frac{1}{n} \log \Pi(\hat{B}_n).$$

On the event $E_n = \{\Pi(\hat{B}_n) \geq \frac{1}{2}\Pi(B)\}$ this is further bounded below by $c' - c - \epsilon + o(1)$, and the remainder of the proof of Theorem 6.39 can be copied to show that the posterior mass tends to zero on the event $\liminf E_n$. It suffices to prove that $P_{\theta_0}^{(\infty)}(\limsup E_n^c) = 0$. Because $E_1^c \supset E_2^c \supset \cdots$ this is equivalent to the probability of the events E_n^c tending to zero.

If we write $\hat{B}_n$ as $\{\theta \in B\colon \inf_{k\geq n} Z_k(\theta, X^{(k)}) \geq -\epsilon\}$, then

$$\Pi(\hat{B}_n) = \Pi(B) - \Pi\Bigl(\theta \in B\colon \inf_{k\geq n} Z_k(\theta, X^{(k)}) < -\epsilon\Bigr),$$

and the event E_n^c can be written in the form $\{\Pi(\theta \in B\colon \inf_{k\geq n} Z_k(\theta, X^{(k)}) < -\epsilon) > \frac{1}{2}\Pi(B)\}$. By Markov's inequality followed by Fubini's theorem,

$$P_{\theta_0}^{(\infty)}(E_n^c) \leq \frac{2}{\Pi(B)}(\Pi \times P_{\theta_0}^{(\infty)})\Bigl(\theta \in B, \inf_{k\geq n} Z_k(\theta, X^{(k)}) < -\epsilon\Bigr).$$

By assumption the sequence $Z_n(\theta, X^{(n)})$ tends to zero almost surely $[P_{\theta_0}^{(\infty)}]$, for every fixed $\theta \in B$. Therefore $P_{\theta_0}^{(\infty)}(\inf_{k\geq n} Z_k(\theta, X^{(k)}) < -\epsilon) \to 0$ for each such θ, whence the right side tends to zero, by Fubini's theorem. □

In Section 8.3 we evaluate the Kullback-Leibler divergence and variation for several statistical settings, in the framework of obtaining rates of convergence. Here we consider only independent observations and ergodic Markov processes.

6.7.1 Independent Observations

For $X^{(n)}$ consisting of independent observations the Kullback-Leibler divergence and variation can be expressed or bounded in the same quantities of the individual observations. Furthermore, tests exist automatically relative to the *root average square Hellinger metric*

$$d_{n,H}(\theta_1, \theta_2) = \sqrt{\frac{1}{n}\sum_{i=1}^n d_H^2(P_{n,\theta_1,i}, P_{n,\theta_2,i})}.$$

This yields the following corollary of Theorem 6.39, and extension of Theorem 6.23.

Theorem 6.41 *Let $P_\theta^{(n)}$ be the product of measures $P_{n,\theta,1}, \ldots, P_{n,\theta,n}$. If for every $\epsilon > 0$ there exist measurable sets $B_n \subset \Theta_n$ with $\liminf \Pi_n(B_n) > 0$ and*

$$\sup_{\theta\in B_n} \frac{1}{n}\sum_{i=1}^n K(p_{n,\theta_{n,0},i};\, p_{n,\theta,i}) \leq \epsilon, \qquad \sup_{\theta\in B_n} \frac{1}{n^2}\sum_{i=1}^n V_{2,0}(p_{n,\theta_{n,0},i};\, p_{n,\theta,i}) \to 0,$$

then $\Pi_n(\tilde\Theta_n \mid X^{(n)}) \to 0$ in $P_{\theta_{n,0}}^{(n)}$-probability for any sets $\tilde\Theta_n \subset \Theta_n$ such that either (a) or (b) of Theorem 6.39 holds for a constant $C > 0$. In particular, if for every $\epsilon > 0$ there exists a partition $\Theta_n = \Theta_{n,1} \cup \Theta_{n,2}$ and $C > 0$ such that

(i) $\log N(\epsilon, \Theta_{n,1}, d_{n,H}) \leq 3n\epsilon^2$,
(ii) $\Pi(\Theta_{n,2}) \leq e^{-Cn}$,

then $\Pi_n(d_{n,H}(\theta, \theta_{n,0}) > \epsilon \mid X^{(n)}) \to 0$ in $P_{\theta_{n,0}}^{(n)}$-probability, for every $\epsilon > 0$.

Proof The first assertion is an immediate corollary of Theorem 6.39. For the proof of the second we construct the tests for $\Theta_{n,1} = \{\theta\colon d_{n,H}(\theta, \theta_{n,0}) > \epsilon\}$ by the same arguments as in the proof of Theorem 6.23, where we obtain the basic tests from Proposition D.9. □

The preceding theorem applies to general triangular arrays of independent observations. For sequences $X^{(n)} = (X_1, \ldots, X_n)$ of independent variables with $X_i \sim p_{\theta,i}$ and a fixed prior Π, we may apply Theorem 6.40, with the condition of almost sure convergence of the log-likelihood ratios verified by Kolmogorov's theorem: for $\theta \in B$,

$$\sum_{i=1}^\infty \frac{1}{i^2} V_{2,0}(p_{\theta_0,i};\, p_{\theta,i}) < \infty.$$

Other versions of the law of large numbers may be brought in as well.

6.7.2 Markov Processes

Consider $X^{(n)} = (X_1, \ldots, X_n)$ for a Markov process $\{X_n; n \geq 0\}$ with stationary transition density $(x, y) \mapsto p_\theta(y \mid x)$ relative to a dominating measure ν on the state space $(\mathfrak{X}, \mathscr{X})$. For Q_θ the invariant distribution corresponding to p_θ (assumed to exist and to be unique) define the *Kullback-Leibler divergence between two transition densities* as

$$K(p_{\theta_0}; p_\theta) = \int \log \frac{p_{\theta_0}}{p_\theta}(y \mid x)\, p_{\theta_0}(y \mid x)\, d\nu(y)\, dQ_{\theta_0}(x).$$

Say that a prior Π on the parameter set Θ possesses the *KL property* at θ_0, and written $p_{\theta_0} \in \mathrm{KL}(\Pi)$, if $\Pi(\theta\colon K(p_{\theta_0}; p_\theta) < \epsilon) > 0$ for every $\epsilon > 0$.

Theorem 6.42 (Markov chain, fixed prior) *Let $P_\theta^{(n)}$ be the distribution of $(X_1, \ldots, X_n)$ for a Markov chain $(X_t\colon t \geq 0)$. If the chain is positive Harris recurrent under θ_0 and $p_{\theta_0} \in \mathrm{KL}(\Pi)$, then $\Pi_n(\tilde\Theta_n \mid X^{(n)}) \to 0$ in $P_{\theta_0}^{(n)}$-probability for any sets $\tilde\Theta_n \subset \Theta_n$ such that either (a) or (b) of Theorem 6.39 holds for a constant $C > 0$.*

Proof The condition that the Markov chain is Harris recurrent is equivalent to validity of the strong law for all functions that are Q_{θ_0}-integrable, and also equivalent to triviality of the invariant σ-field relative to any initial condition x for the chain (see Theorem 17.1.7 of Meyn and Tweedie 1993). The last shows that the property is inherited by the chain of pairs (X_{t-1}, X_t), and hence

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}}{p_\theta}(X_i|\, X_{i-1}) \to K(p_{\theta_0}; p_\theta). \qquad \text{a.s. } [P_{\theta_0}^{(\infty)}].$$

Thus, Theorem 6.40 applies. □

6.8 Alternative Approaches

In this section we discuss alternative approaches to formulate and prove posterior consistency. These complement Schwartz's theory as expounded in Sections 6.4 and 6.7, but actually never give weaker conditions.

6.8.1 Separation

The existence of exponentially consistent tests, used in Section 6.4, is equivalent to *separation* of measures (see Le Cam 1986). Schwartz's original proof of her theorem implicitly used this concept.

For a probability density p write p^k for its k-fold product.

Definition 6.43 A probability density p_0 and a set $\mathcal{V}$ of probability densities are *strongly δ-separated* at stage $k \in \mathbb{N}$ if $\rho_{1/2}(p_0^k, \int p^k\, d\mu(p)) < \delta$, for every probability measure μ on $\mathcal{V}$.

For example, the ball $\mathcal{V} = \{p \in \mathcal{P}\colon \|p - p_1\|_1 < \delta\}$ of radius δ around a point p_1 with $\|p_1 - p_0\|_1 > 2\delta$ is strongly δ-separated at stage 1. Strong separation for some k automatically entails exponential separation of product measures. The following lemma is essentially a restatement of Lemma D.6, and allows a variation of Schwartz's theorem in terms of strong δ-separation.

Lemma 6.44 *If p_0 and $\mathcal{V}$ are strongly δ-separated at stage k, then $\rho_{1/2}(p_0^n, \int p^n\, d\mu(p)) < e^{-\lfloor n/k\rfloor \log_- \delta}$, for any probability measure μ on $\mathcal{V}$.*

Theorem 6.45 *If $p_0 \in \mathrm{KL}(\Pi)$, then $\Pi(\mathcal{V}|\, X_1, \ldots, X_n) \to 0$ a.s. $[P_0^\infty]$, for any set $\mathcal{V}$ that is strongly δ separated from p_0 at stage k, for some δ and k.*

Proof Inequality (6.5) in the proof of Theorem 6.17, applied with $\mathcal{P}_0$ an ϵ-Kullback-Leibler neighborhood of p_0, shows that $\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)$ is bounded below by $e^{-\epsilon n}$, for any $\epsilon > 0$, eventually almost surely. It suffices that $e^{\epsilon n} L_n$, for $L_n = \int_{\mathcal{V}} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)$, tends to zero almost surely for some $\epsilon > 0$. Now $P_0^n \sqrt{L_n} = \rho_{1/2}(p_0^n, \int_{\mathcal{V}} p^n\, d\Pi(p)) \le e^{-cn}$, for some $c > 0$, by Lemma 6.44. □

6.8.2 Le Cam's Inequality

Le Cam's inequality allows to verify posterior consistency by a testing argument, in the spirit of Schwartz's theorem, but it substitutes the total variation distance for the Kullback-Leibler divergence. Thus, avoiding the Kullback-Leibler property, it is applicable to models that are not dominated.

We state the inequality in abstract form, for the posterior distribution based on a single observation X, in the Bayesian setting where a probability distribution P is drawn from a prior Π on the collection $\mathfrak{M}$ of probability measures on the sample space, and next X is drawn according to P. The posterior distribution $\Pi(\cdot|\, X)$ is the conditional distribution of P given X. It is not assumed that Π concentrates on a dominated set of measures, and hence Bayes's formula may not be available.

We write $P_{\mathcal{U}}$ for the average $P_{\mathcal{U}} = \int_{\mathcal{U}} P\, d\Pi(P)/\Pi(\mathcal{U})$ of the measures P in a measurable subset $\mathcal{U} \subset \mathfrak{M}$.

Lemma 6.46 (Le Cam) *For any pair of measurable sets $\mathcal{U}, \mathcal{V} \subset \mathfrak{M}$, test ϕ, and measure $P_0 \in \mathfrak{M}$,*

$$P_0\Pi(\mathcal{V}|\, X) \le d_{TV}(P_0, P_{\mathcal{U}}) + P_0\phi + \frac{1}{\Pi(\mathcal{U})}\int_{\mathcal{V}} P(1-\phi)\, d\Pi(P).$$

Proof Since both $0 \le \phi \le 1$ and $0 \le \Pi(\mathcal{V}|\, X) \le 1$, for any probability measure Q,

$$P_0\Pi(\mathcal{V}|\, X) \le P_0\phi + P_0[\Pi(\mathcal{V}|\, X)(1-\phi)] \le P_0\phi + d_{TV}(P_0, Q) + Q[\Pi(\mathcal{V}|\, X)(1-\phi)].$$

The measure $Q = P_{\mathcal{U}}$ is bounded above by $1/\Pi(\mathcal{U})$ times $\int P\, d\Pi(P)$. The latter is the marginal distribution of X in the model $P \sim \Pi$ and $X|\, P \sim P$. Therefore, $\Pi(\mathcal{U})$ times the third term on the right is bounded above by $\mathrm{E}[\Pi(\mathcal{V}|\, X)(1-\phi)(X)] = \mathrm{E}[\mathbb{1}\{P \in \mathcal{V}\}(1-\phi)(X)]$, by the orthogonality (projection property) of conditional expectations. Since $\mathrm{E}[(1-\phi)(X)|\, P] = P(1-\phi)$, this can be further rewritten as $\int_{\mathcal{V}} P(1-\phi)\, d\Pi(P)$, leading to the third term in the bound of the lemma. □

In a consistency proof the set $\mathcal{U}$ in Le Cam's inequality is taken equal to a small neighborhood of P_0, to ensure that the first term on the right $d_{TV}(P_0, P_{\mathcal{U}})$ is small. A useful bound is then obtained if there exists a test of P_0 versus $\mathcal{V}$ with integrated error probability that is significantly smaller than $\Pi(\mathcal{U})$.

In applications where $X = (X_1, \ldots, X_n)$ is a sample of n i.i.d. observations, the measure P_0 in the inequality, which is the true distribution of X, must be replaced with a product measure P_0^n, and $\mathcal{U}$ will consist of product measures P^n as well. By Lemmas B.1 and B.8 the total variation distance $d_{TV}(P_0^n, P^n)$ between product measures is bounded above by $2\sqrt{n}d_H(P_0, P)$, and hence a good choice for $\mathcal{U}$ is the set of products of measures in a Hellinger ball of radius a small multiple of $1/\sqrt{n}$. This gives the following theorem, which applies to every topology on P for any sequence of priors Π_n.

Theorem 6.47 *If for every neighborhood $\mathcal{U}$ of P_0 and every $\epsilon > 0$ there exist tests ϕ_n and $C > 0$ such that $P_0^n\phi_n \le e^{-Cn}$ and*

$$\int_{\mathcal{U}^c} P^n(1-\phi_n)\,d\Pi_n(P) \le e^{-Cn}\Pi_n\Bigl(P\colon d_H(P,P_0) < \frac{\epsilon}{\sqrt{n}}\Bigr),$$

then the posterior distribution $\Pi_n(\cdot|\,X_1,\ldots,X_n)$ in the model $X_1,\ldots,X_n|\,P \overset{iid}{\sim} P$ and $P \sim \Pi_n$ is strongly consistent at P_0.

Similar as in Schwartz's theorem, the requirements are existence of good tests and sufficient prior mass in a neighborhood of the true measure. Presently the neighborhood is relative to the Hellinger distance rather than the Kullback-Leibler divergence, and it shrinks with n. If

$$\frac{1}{n}\log\frac{1}{\Pi_n(P\colon d_H(P,P_0) < \epsilon/\sqrt{n})} \to 0, \tag{6.12}$$

then $e^{-Cn}\Pi_n(P\colon d_H(P,P_0) < \epsilon/\sqrt{n})$ is of the order $e^{-Cn+o(n)}$, and the testing condition takes a similar form as in Section 6.4.

Some insight in the size of $\Pi_n(P\colon d_H(P,P_0) < \epsilon/\sqrt{n})$ is obtained by considering a fixed prior Π that distributes its mass "uniformly." If we could place $N(\epsilon/\sqrt{n})$ disjoint balls of radius $\epsilon/\sqrt{n}$ in the support $\mathcal{P}$ of the prior, then every ball would obtain prior mass $1/N(\epsilon/\sqrt{n})$. The number of balls ought to be comparable to the covering number $N(\epsilon/\sqrt{n},\mathcal{P},d_H)$ of the model, and then (6.12) would be equivalent to $\log N(\epsilon/\sqrt{n},\mathcal{P},d_H) = o(n)$. This is true if the entropy $\log N(\epsilon,\mathcal{P},d_H)$ is of smaller order than $(1/\epsilon)^2$, meaning that the model could be big, but not too big.

For further illustration, we derive posterior consistency for a *default prior* constructed by mixing finite uniform distributions on ϵ_m-nets, for a sequence of approximation levels $\epsilon_m \downarrow 0$. Consider a target model $\mathcal{P}$ that is totally bounded relative to the Hellinger distance, let Π_m be the uniform discrete distribution on a minimal ϵ_m-net over $\mathcal{P}$ for the Hellinger distance, and let $\Pi = \sum_{m=1}^\infty \lambda_m\Pi_m$ be a fixed overall prior on $\mathcal{P}$, for a given infinite probability vector $(\lambda_1,\lambda_2,\ldots)$.

Theorem 6.48 *For a set $\mathcal{P}$ of probability measures with $\epsilon^2\log N(\epsilon,\mathcal{P},d_H) \to 0$ as $\epsilon \to 0$, a sequence $\epsilon_m \downarrow 0$ such that $\epsilon_{m-1}/\epsilon_m = O(1)$ and weights such that $\epsilon_m^2\log(1/\lambda_m) \to 0$, construct a prior Π as indicated. Then the posterior distribution $\Pi_n(\cdot|\,X_1,\ldots,X_n)$ in the model $X_1,\ldots,X_n|\,P \overset{iid}{\sim} P$ and $P \sim \Pi$ is strongly consistent relative to the Hellinger distance at any $P_0 \in \mathcal{P}$.*

Proof Given $\epsilon > 0$ and m such that $\epsilon_m < \epsilon/\sqrt{n} \le \epsilon_{m-1}$ a Hellinger ball of radius $\epsilon/\sqrt{n}$ contains at least one support of Π_m, whence $\Pi(P\colon d_H(P,P_0) < \epsilon/\sqrt{n}) \ge \lambda_m/N(\epsilon_m,\mathcal{P},d_H)$. It follows that (6.12) is satisfied if both $n^{-1}\log N(\epsilon_m,\mathcal{P},d_H)$ and $n^{-1}\log(1/\lambda_m)$ tend to zero. As $n^{-1} \le (\epsilon_{m-1}/\epsilon)^2 \lesssim \epsilon_m^2$, this follows from the assumptions.

By the compactness of $\mathcal{P}$ there exist tests of P_0 versus the complement of a Hellinger ball of positive radius with error probabilities bounded by e^{-Cn}, for some $C > 0$, by combination of Lemma D.3 and Proposition D.8, or by Example 6.20 and the fact that the weak and strong topologies coincide on strongly compact sets. □

Example 6.49 Consider the set $\mathcal{P}$ of all probability measures on $[0,1]^d$ that possess a Lebesgue density p such that $\sqrt{p}$ possesses Hölder norm of order α (see Definition C.4) bounded by a fixed constant. By Proposition C.5 this has Hellinger entropy of the order $(1/\epsilon)^{d/\alpha}$, whence the entropy condition is satisfied for $\alpha > d/2$. Thus, consistency pertains unless λ_m decays too rapidly relative to ϵ_m. For instance, if $\epsilon_m = m^{-1}$, then λ_m with tails thicker than any distribution proportional to e^{-am^2} suffices. The choice $\epsilon_m = 2^{-m}$ allows even double-exponential tails for λ_m.

6.8.3 Predictive Consistency

Given a sequence of priors Π_n for a density p and i.i.d. observations $X_1, X_2, \ldots$, the *predictive density* $\hat{p}_i$ for X_i is the posterior mean based on $X_1, \ldots, X_{i-1}$:

$$\hat{p}_i(x) = \mathrm{E}(p(x)|\, X_1, \ldots, X_{i-1}) = \frac{\int p(x) \prod_{j=1}^{i-1} p(X_j)\, d\Pi_n(p)}{\int \prod_{j=1}^{i-1} p(X_j)\, d\Pi_n(p)}. \tag{6.13}$$

The conditional mean in the central expression is in the Bayesian setup with $p \sim \Pi_n$ and $X_1, \ldots, X_n|\, p \overset{\text{iid}}{\sim} p$.

The posterior mean of the usual posterior distribution is $\hat{p}_{n+1}$, and ordinarily we would be interested in its consistency (and the contraction of the full posterior distribution). We have seen that this pertains under a combination of the Kullback-Leibler property and a testing (or entropy) condition, where the latter cannot be omitted in general (see Example 7.12), and is necessary for consistency at an exponential rate (see Proposition 6.22). Interestingly, the Kullback-Leibler property alone is sufficient for consistency of the Cesàro averages of the predictive densities, also for strong metrics.

Theorem 6.50 *If* $n^{-1} \log \Pi_n(p\colon K(p_0; p) < \epsilon) \to 0$ *for every* $\epsilon > 0$, *then it holds that* $K(p_0; n^{-1}\sum_{i=1}^n \hat{p}_i) \le n^{-1}\sum_{i=1}^n K(p_0; \hat{p}_i) \to 0$ *in mean under* P_0^n. *In particular, this is true for a fixed prior* $\Pi_n = \Pi$ *with* $p_0 \in \mathrm{KL}(\Pi)$.

We shall prove a more general version of the theorem, within a setup with arbitrary, not necessarily i.i.d. observations. Let Π_n be a prior on densities on the sample space $(\mathfrak{X}^\infty, \mathscr{X}^\infty)$, and consider an infinite vector of observations in the Bayesian model

$$p_\infty \sim \Pi_n, \qquad (X_1, X_2, \ldots)|\, p_\infty \sim p_\infty.$$

The *predictive density* at stage i is defined in this setting as the conditional density of X_i given $X_1, \ldots, X_{i-1}$. As before, we denote it by $\hat{p}_i$. In terms of the density p_i of $(X_1, \ldots, X_i)$, it can be written

$$\hat{p}_i(x) = p_i(x|\, X_1, \ldots, X_{i-1}) = \frac{\int p_i(X_1, \ldots, X_{i-1}, x)\, d\Pi_n(p)}{\int p_{i-1}(X_1, \ldots, X_{i-1})\, d\Pi_n(p)}. \tag{6.14}$$

A true density $p_{0,\infty}$ on $(\mathfrak{X}^\infty, \mathscr{X}^\infty)$ similarly defines densities $p_{0,i}$ for $(X_1, \ldots, X_i)$ and conditional densities $\hat{p}_{0,i}$ of X_i, given $X_1, \ldots, X_{i-1}$. For i.i.d. observations $\hat{p}_{0,i} = p_0$, for every i, and the preceding display reduces to (6.13).

Theorem 6.51 *If* $n^{-1} \log \Pi_n(p_\infty\colon n^{-1} K(p_{0,n}; p_n) < \epsilon) \to 0$ *for every* $\epsilon > 0$, *then it holds that* $K(n^{-1}\sum_{i=1}^n \hat{p}_{0,i}; n^{-1}\sum_{i=1}^n \hat{p}_i) \le n^{-1}\sum_{i=1}^n K(\hat{p}_{0,i}; \hat{p}_i) \to 0$ *in mean under* $P_{0,\infty}$.

Proof The first inequality is a consequence of Jensen's inequality and the convexity of the function $(u, v) \mapsto u \log(u/v)$. To prove the convergence to zero, observe that

$$\frac{\hat{p}_i(X_i)}{\hat{p}_{0,i}(X_i)} = \frac{\int (p_i/p_{0,i})(X_1, \ldots, X_i)\, d\Pi_n(p)}{\int (p_{i-1}/p_{0,i-1})(X_1, \ldots, X_{i-1})\, d\Pi_n(p)} = \frac{I_{i,n}}{I_{i-1,n}},$$

for $I_{i,n} = \int (p_i/p_{0,i})(X_1, \ldots, X_i)\, d\Pi_n(p)$ for $i = 1, \ldots, n$, and $I_{0,n} = 1$. It follows that

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{\hat{p}_{0,i}(X_i)}{\hat{p}_i(X_i)} = -\frac{1}{n} \log I_{n,n}. \tag{6.15}$$

The expectation of the left side under $P_{0,\infty}$ is also the expectation of $n^{-1} \sum_{i=1}^{n} K(\hat{p}_{0,i}; \hat{p}_i)$. It is clearly nonnegative, whence it suffices to show that the expectation of the right side is asymptotically bounded above by ϵ, for every $\epsilon > 0$. If $\Pi_{n,\epsilon}$ is the renormalized restriction of Π_n to $B_{n,\epsilon} := \{p\colon n^{-1} K(p_{0,n}; p_n) < \epsilon\}$, then

$$I_{n,n} \geq \Pi_n(B_{n,\epsilon}) \int \frac{p_n}{p_{0,n}}(X_1, \ldots, X_n)\, d\Pi_{n,\epsilon}(p).$$

By Jensen's inequality,

$$\begin{aligned} P_{0,\infty}\Big[-\frac{1}{n} \log I_{n,n}\Big] &\leq -\frac{1}{n} \log \Pi_n(B_{n,\epsilon}) - P_{0,n} \frac{1}{n} \int \log \frac{p_n}{p_{0,n}}(X_1, \ldots, X_n)\, d\Pi_{n,\epsilon}(p) \\ &= -\frac{1}{n} \log \Pi_n(B_{n,\epsilon}) + \int \frac{1}{n} K(p_{0,n}; p_n)\, d\Pi_{\epsilon,n}(p) \leq -\frac{1}{n} \log \Pi_n(B_{n,\epsilon}) + \epsilon. \end{aligned}$$

The right side tends to ϵ as $n \to \infty$, by assumption. □

As the Kullback-Leibler divergence is larger than the squared Hellinger distance (see Lemma B.1), it also follows that, in P_0^n-probability,

$$d_H^2\Big(\frac{1}{n} \sum_{i=1}^{n} \hat{p}_{0,i}; \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i\Big) \to 0.$$

In other words, the distance between the Cesàro averages of the predictive and true conditional densities tends to zero. In particular, for i.i.d. observations the Cesàro mean of the predictive densities is a weakly consistent density estimator at the true density p_0 whenever $p_0 \in \mathrm{KL}(\Pi)$, or more generally if a sequence of priors satisfies the condition of the theorem. (In Lemma 6.52 below the consistency is shown to be in the almost sure sense for a fixed prior.)

Besides that its consistency can be proved under weak conditions, there appears to be no good motivation for the Cesàro estimator $n^{-1} \sum_{i=1}^{n} \hat{p}_i$. Its asymmetric, sequential use of i.i.d. observations can be remedied by applying the procedure of Rao-Blackwell, which leads to the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\binom{n}{i}}\sum_{1\le j_1<\cdots<j_i\le n}\mathrm{E}(p|\,X_{j_1},\ldots,X_{j_i}). \tag{6.16}$$

This estimator contains 2^n-1 terms, and hence is computationally prohibitive, but it inherits the Hellinger consistency, by the convexity of d_H^2.

6.8.4 Martingale Approach

In Section 6.8.3 the predictive density $\hat{p}_i$, defined in (6.13) for i.i.d. observations and in (6.14) in general, was seen to approximate the true conditional density $\hat{p}_{0,i}$ (or p_0 in the i.i.d. case), as $n\to\infty$. When using a fixed prior Π, for a given function $\ell\colon[0,\infty)\to\mathbb{R}$ we may study this further through the variables

$$M_n=\sum_{i=1}^{n}\Bigl[\ell\Bigl(\frac{\hat{p}_i}{\hat{p}_{0,i}}(X_i)\Bigr)-\mathrm{E}_0\Bigl[\ell\Bigl(\frac{\hat{p}_i}{\hat{p}_{0,i}}(X_i)\Bigr)|\,X_1,\ldots,X_{i-1}\Bigr]\Bigr].$$

Note here that the hats in $\hat{p}_i$ and $\hat{p}_{0,i}$ refer to the conditioning on the "past" observations $X_1,\ldots,X_{i-1}$.

The variables M_n are sums of variables with zero conditional means given the past and hence form a martingale. We take the conditional expectations E_0 relative to the "frequentist" model $(X_1,X_2,\ldots)\sim p_{0,\infty}$, whence the martingale property is also relative to this model. The leading term of M_n can be viewed as a cumulative sum of discrepancies $\ell((\hat{p}_i/\hat{p}_{0,i})(X_i))$ between the predictive and true density, and the "compensator"

$$\mathrm{E}_0\bigl[\ell((\hat{p}_i/\hat{p}_{0,i})(X_i))|\,X_1,\ldots,X_{i-1}\bigr]$$

as the integrated version of this discrepancy. For the special cases $\ell(x)=\log x$ and $\ell(x)=2(\sqrt{x}-1)$ this compensator is

$$\mathrm{E}_0\Bigl(\log\frac{\hat{p}_i}{\hat{p}_{0,i}}(X_i)|\,X_1,\ldots,X_{i-1}\Bigr)=-K(\hat{p}_{0,i};\,\hat{p}_i), \tag{6.17}$$

$$\mathrm{E}_0\Bigl(2\Bigl(\sqrt{\frac{\hat{p}_i}{\hat{p}_{0,i}}(X_i)}-1\Bigr)|\,X_1,\ldots,X_{i-1}\Bigr)=-d_H^2(\hat{p}_{0,i};\,\hat{p}_i). \tag{6.18}$$

Martingale convergence theory gives a handle on these objects.

Lemma 6.52 *If $\sum_{n=1}^{\infty}n^{-2}\,\mathrm{var}_0\,\ell((\hat{p}_n/\hat{p}_{0,n})(X_n))<\infty$, then $n^{-1}M_n\to 0$ almost surely. This condition is automatic for $\ell(x)=2(\sqrt{x}-1)$. For i.i.d. observations, assuming the Kullback-Leibler property of the prior Π at p_0, this implies that $n^{-1}\sum_{i=1}^{n}d_H^2(\hat{p}_i,p_0)\to 0$, almost surely $[P_0^\infty]$.*

Proof The martingale differences $\Delta M_n=M_n-M_{n-1}$ have conditional mean zero and conditional variances satisfying $\mathrm{E}_0\,\mathrm{var}_0\,(\Delta M_n|\,X_1,\ldots,X_{n-1})\le\mathrm{var}_0\,\ell((\hat{p}_n/\hat{p}_{0,n})(X_n))$. Hence the convergence of $n^{-1}M_n$ follows from the strong law for martingale differences (which follows from the martingale convergence theorem and Kronecker's lemma).

The variance is bounded above by the second moment $\mathrm{E}_0\ell^2((\hat{p}_n/\hat{p}_{0,n})(X_n))$, which in turn is bounded above by $4\mathrm{E}_{p_0}d_H^2(\hat{p}_n;\,\hat{p}_{0,n})\le 8$ for $\ell(x)=2(\sqrt{x}-1)$.

In view of (6.18), $n^{-1}M_n \to 0$ implies that $\limsup n^{-1}\sum_{i=1}^n d_H^2(\hat p_i, \hat p_{0,i})$ is bounded above by

$$\limsup_{n\to\infty} -\frac{2}{n}\sum_{i=1}^n\Bigl(\sqrt{\frac{\hat p_i}{\hat p_{0,i}}(X_i)}-1\Bigr) \le \limsup_{n\to\infty} -\frac{1}{n}\sum_{i=1}^n \log\frac{\hat p_i}{\hat p_{0,i}}(X_i),$$

since $2(\sqrt{x}-1)\ge \log x$, for every $x\ge 0$. By (6.15) the variable on the right is equal to $-n^{-1}\log I_n$, where in the case of i.i.d. observations for $I_n = \int\prod_{i=1}^n (p/p_0)(X_i)\,d\Pi(p)$. This tends to zero almost surely by Remark 6.18. □

The denominator $\int p_n(X_1,\ldots,X_n)\,d\Pi(p)$ of the posterior distribution is the marginal density of the observations in the Bayesian model. It can be factorized as the product of the predictive densities, as $p_1(X_1)p_2(X_2|\,X_1)\cdots p_n(X_n|\,X_1,\ldots,X_{n-1}) = \prod_{i=1}^n \hat p_i(X_i)$. The posterior probability of a set $\mathcal{U}$ can be factorized similarly. If $\Pi_{\mathcal{U}}$ is the renormalized restriction of Π to a set $\mathcal{U}$, then the predictive density of X_i in the model $p_\infty\sim\Pi_{\mathcal{U}}$ and $(X_1,X_2,\ldots)|\,p_\infty\sim p_\infty$ is

$$\hat p_i(X_i|\mathcal{U}) = p_i(X_i|\,X_1,\ldots,X_{i-1},\mathcal{U}) = \frac{\int p_i(X_1,\ldots,X_i)\,d\Pi_{\mathcal{U}}(p)}{\int p_{i-1}(X_1,\ldots,X_{i-1})\,d\Pi_{\mathcal{U}}(p)}.$$

With this notation we can write

$$\Pi_n(\mathcal{U}|\,X_1,\ldots,X_n) = \frac{\prod_{i=1}^n \hat p_i(X_i|\mathcal{U})\,\Pi(\mathcal{U})}{\prod_{i=1}^n \hat p_i(X_i)} = \frac{1}{I_n}\prod_{i=1}^n\frac{\hat p_i(X_i|\mathcal{U})}{\hat p_{0,i}(X_i)}\Pi(\mathcal{U}),$$

for $I_n = \int (p_n/p_{0,n})(X_1,\ldots,X_n)\,d\Pi(p)$. An intuitive interpretation of the center formula is that the posterior mass of a set $\mathcal{U}$ will tend to one if and only if in the limit the priors $\Pi_{\mathcal{U}}$ and Π lead to the same predictive densities.

The formula on the far right invites an alternative consistency proof: the posterior mass of a set $\mathcal{U}$ such that $\hat p_i(\cdot|\mathcal{U})$ is unlike $\hat p_{0,i}$ will tend to zero. For i.i.d. observations this idea can be made precise by noting that I_n is bounded below by $e^{-\epsilon n}$, for any $\epsilon>0$, if $p_0\in \mathrm{KL}(p_0)$, by (6.5) or Remark 6.18. Thus the far right expression tends to zero if the product is exponentially small. Now, for any $\alpha\in(0,1)$ and ρ_α the Hellinger transform,

$$\mathrm{E}_0\biggl(\Bigl(\frac{\hat p_i(X_i|\mathcal{U})}{\hat p_{0,i}(X_i)}\Bigr)^\alpha |\,X_1,\ldots,X_i\biggr) = \rho_\alpha(\hat p_i(\cdot|\mathcal{U}),\hat p_{0,i}).$$

If $\mathcal{U}$ is convex, then $\hat p_i(\cdot|\mathcal{U})$ is contained in $\mathcal{U}$ by Jensen's inequality, and hence the Hellinger transform is bounded by $\rho_\alpha(\mathcal{U},\hat p_{0,i}) = \sup_{p\in\mathcal{U}}\rho_\alpha(p,\hat p_{0,i})$. By applying this argument for $i=n,n-1,\ldots,1$ and with $\hat p_{0,i}=p_0$, we can "peel off" the observations one by one and find

$$\mathrm{E}_0\Bigl(\prod_{i=1}^n\frac{\hat p_i(X_i|\mathcal{U})}{\hat p_{0,i}(X_i)}\Pi(\mathcal{U})\Bigr)^\alpha \le \rho_\alpha(\mathcal{U},p_0)^n\,\Pi(\mathcal{U})^\alpha.$$

Thus the numerator in the posterior is exponentially small for any convex $\mathcal{U}$ such that $\rho_\alpha(\mathcal{U},p_0)<1$. A slight extension of this argument allows to recover the result of Example 6.24 (see Problem 6.17), which was obtained by Schwartz's testing approach. Of course, these methods are closely related, as the Hellinger affinities provide bounds on the testing rates (see Appendix D).

6.8.5 α-Posterior

Another way to obtain Hellinger consistency under only the Kullback-Leibler property is to change the definition of the posterior distribution. For $\alpha \geq 0$ the *α-posterior distribution* based observations $X_1, \ldots, X_n$ from a density p following a prior Π is defined by

$$\Pi_n^{(\alpha)}(p \in B \mid X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p^\alpha(X_i)\, d\Pi_n(p)}{\int \prod_{i=1}^n p^\alpha(X_i)\, d\Pi_n(p)}. \tag{6.19}$$

The α-posterior can be interpreted as the usual posterior distribution with respect to the data-dependent prior Π^* defined by

$$d\Pi^*(p) \propto \prod_{i=1}^n p^{\alpha-1}(X_i)\, d\Pi(p). \tag{6.20}$$

Example 6.53 For p the density of the normal Nor(θ, σ^2)-distribution with unknown mean θ and known variance σ^2, and a Nor(μ, τ^2) prior, the α-posterior distribution is the normal distribution with mean and variance

$$\frac{(\alpha n \bar{X}/\sigma^2) + (\mu/\tau^2)}{(\alpha n/\sigma^2) + (1/\tau^2)}, \qquad \frac{1}{(\alpha n/\sigma^2) + (1/\tau^2)}.$$

This leads to the interpretation that either the variance is misspecified as σ^2/α, or the sample size misspecified as $n\alpha$. The α-posterior mean is equal to the usual posterior mean with respect to the Nor$(\mu, \alpha\tau^2)$-prior, but the corresponding posterior variances are different. The latter implies that a credible region of the α-posterior does not have asymptotically correct frequentist coverage.

Theorem 6.54 *If $p_0 \in \mathrm{KL}(\Pi)$ for a fixed prior Π, then for any $0 < \alpha < 1$ the α-posterior distribution $\Pi_n^{(\alpha)}(\cdot \mid X_1, \ldots, X_n)$ in the model $X_1, \ldots, X_n \mid p \overset{iid}{\sim} p$ and $p \sim \Pi$ is strongly consistent at p_0. In particular, the α-posterior mean is strongly Hellinger consistent at p_0.*

Proof By the same argument as in the proof of Schwartz's theorem (see (6.5)) we can derive that $e^{\epsilon n} \int \prod_{i=1}^n (p/p_0)^\alpha(X_i)\, d\Pi(p) \to \infty$ almost surely $[P_0^\infty]$, for any $\epsilon > 0$. Furthermore, since $P_0(p/p_0)^\alpha = \rho_\alpha(p, p_0)$ is the Hellinger transform, for any neighborhood $\mathcal{U}$ of p_0,

$$P_0^n\Big[\int_{\mathcal{U}^c} \prod_{i=1}^n \frac{p^\alpha}{p_0^\alpha}(X_i)\, d\Pi(p)\Big] = \int_{\mathcal{U}^c} \rho_\alpha(p, p_0)^n\, d\Pi(p).$$

For $\mathcal{U}$ a Hellinger ball of radius ϵ, the integrand is bounded above by e^{-Cn}, for $C = \epsilon^2 \min(\alpha, 1-\alpha)$, by Lemma B.5. The theorem follows upon choosing $\epsilon < C$. □

In the above result, we can even allow $\alpha = \alpha_n$ to increase to 1 slowly; see Problem 6.15. A generalization to non-identically distributed observations is described in Problem 6.16.

Example 6.55 (Pólya tree) Let $X_1, X_2, \ldots \mid P \overset{\text{iid}}{\sim} P$, where $P \sim \mathrm{PT}(\mathcal{T}_m, \alpha_\varepsilon)$. Assume that the condition (3.16) holds, so that the prior is absolutely continuous almost surely, and the

α-posterior is well defined. We claim that this is again a Pólya tree process with parameters updated as

$$\alpha_\varepsilon \mapsto \alpha_\varepsilon + \alpha \sum_{i=1}^n \mathbb{1}\{X_i \in B_\varepsilon\}, \qquad \varepsilon \in \mathcal{E}^*. \tag{6.21}$$

The result will follow from the posterior updating rule for the ordinary Pólya tree process prior and posterior (the display with $\alpha = 1$) if the data-dependent prior Π^* in (6.20) is a Pólya tree process with updated parameters $\alpha_\varepsilon + (\alpha - 1)n \sum_{i=1}^n \mathbb{1}\{X_i \in B_\varepsilon\}$. To see this, it is convenient to describe the Pólya tree process in terms of the splitting variables $V = (V_{\varepsilon 0})_{\varepsilon \in \mathcal{E}^*}$, which are independent Be$(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$-variables. The density of the Pólya tree process Π can be expressed as $p = \phi(V)$ for some measurable map $\phi\colon [0,1]^{\mathcal{E}^*} \to \mathcal{P}$, and hence for every measurable function h and λ the uniform measure on $[0,1]^{\mathcal{E}^*}$,

$$\int h(p)\, d\Pi(p) = \int_{[0,1]^{\mathcal{E}^*}} h \circ \phi(v) \prod_{\varepsilon \in \mathcal{E}^*} \frac{v_{\varepsilon 0}^{\alpha_{\varepsilon 0}-1}(1 - v_{\varepsilon 0})^{\alpha_{\varepsilon 1}-1}}{B(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})}\, d\lambda(v).$$

The density $\prod_{i=1}^n p(X_i)^{\alpha-1}$ of the prior Π^* relative to Π is expressed in the splitting variables through (3.18), and it follows that

$$\begin{aligned}\int h(p)\, d\Pi^*(p) &= \frac{\int h(p) \prod_{i=1}^n p(X_i)^{\alpha-1}\, d\Pi(p)}{\int \prod_{i=1}^n p(X_i)^{\alpha-1}\, d\Pi(p)}\\ &= \frac{1}{c} \int_{[0,1]^{\mathcal{E}^*}} h \circ \phi(v) \prod_{i=1}^n \prod_{j=1}^\infty (2v_{X_{i,1}\cdots X_{i,j}})^{\alpha-1} \prod_{\varepsilon \in \mathcal{E}^*} \frac{v_{\varepsilon 0}^{\alpha_{\varepsilon 0}}(1 - v_{\varepsilon 0})^{\alpha_{\varepsilon 1}}}{B(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})}\, d\lambda(v).\end{aligned}$$

Here $X_{i,1}X_{i,2}, \cdots$ is the binary expansion of X_i over the partitioning tree, and c is the normalizing constant (which depends on $X_1, \ldots, X_n$). Upon comparing the preceding displays we see that Π^* is indeed a Pólya tree process as indicated.

The canonical Pólya tree PT$^*(\lambda, a_m)$ with $\sum_{m=1}^\infty a_m^{-1} < \infty$ possesses the Kullback-Leibler property at every p_0 with $K(p_0; \lambda) < \infty$, by Theorem 7.1. Consequently the α-posterior described by (6.21) and the resulting α-posterior mean are Hellinger consistent by Theorem 6.54. As in (3.23), the α-posterior mean of $p(x)$ is given by

$$\prod_{j=1}^\infty \frac{2a_j + 2\alpha \sum_{i=1}^n \mathbb{1}\{X_i \in B_{\epsilon_1 \cdots \epsilon_j}\}}{2a_j + \alpha \sum_{i=1}^n \mathbb{1}\{X_i \in B_{\epsilon_1 \cdots \epsilon_{j-1}}\}}. \tag{6.22}$$

Interestingly, this coincides with the expression for the usual posterior mean (3.23) with parameters αa_m. We can conclude that the ordinary posterior mean for PT$^*(\lambda, a_m)$ is also consistent with respect to d_H.

6.9 Historical Notes

Consistency, as the most basic asymptotic property, has been long studied in the frequentist literature. Doob (1949) and Le Cam (1953) were apparently the first to formalize the concept of posterior consistency. Le Cam (1953) used uniform strong laws of large numbers to address posterior consistency in abstract spaces, similar to Wald's method for consistency of the maximum likelihood estimator. Freedman (1963, 1965) first addressed posterior consistency issues in concrete infinite-dimensional spaces. The importance of selecting an appropriate version of the posterior was emphasized by Diaconis and Freedman (1986b). Theorem 6.6 about merging in the weak*-sense, due to Diaconis and Freedman (1986b), is based on a similar idea of merging in the variation distance by Blackwell and Dubins (1962) (see Problem 6.1). Theorem 6.9 and Proposition 6.10 are due to Doob (1949). The example of inconsistency in Problem 6.11 for infinite-dimensional spaces with priors having the true distribution in their weak support is due to Freedman (1963), who was the first to construct such examples. Theorem 6.12 appeared in Freedman (1965), and Example 6.21 in Freedman (1963). Example 6.14 on inconsistency in the location model was constructed by Diaconis and Freedman (1986b,a). The inconsistency is related to the inconsistency of an M-estimator shown in Freedman and Diaconis (1982). Parallel results when the error distribution is constrained to median zero instead of being symmetric were obtained by Doss (1985a,b). The concept of Kullback-Leibler support and the role of exponentially consistent tests were first pointed out by Schwartz (1965). Her theorem was partly inspired by Freedman (1963)'s results for discrete sample spaces. The weak consistency result of Example 6.20 is also due to Schwartz (1965). The observation in Theorem 6.17 that the test can be restricted to a part of the parameter space of overwhelming prior mass appeared in an unpublished technical report by Barron in 1988. Theorem 6.23 is a strengthening of a result due to Ghosal et al. (1999b), who extended the scope of Schwartz's theory to address density estimation problems. A similar result was obtained by Barron et al. (1999) under stronger conditions involving bracketing entropy numbers and a direct proof bounding likelihood ratios. The first part of Example 6.24 is due to Walker (2004). Stability of the Kullback-Leibler property seems to have first been noticed by Ghosal et al. (1999a), especially for symmetrization and small location shift. Proposition 6.37 is due to Choi and Ramamoorthi (2008). Consistency for independent, non-identically distributed observations was addressed by Amewou-Atisso et al. (2003), who treated linear regression models without the provision of a sieve. Subsequently, many other authors addressed the issue, including Choudhuri et al. (2004a), Choi and Schervish (2007) and Wu and Ghosal (2008b). Dependent cases were addressed by Tang and Ghosal (2007b,a) for Markov processes (Theorem 6.42) and for general models by Roy et al. (2009). The results as presented here borrow in form from the rate results in Ghosal and van der Vaart (2007a). The idea of separation goes back to Schwartz (1965), and was later considered in a paper by Choi and Ramamoorthi (2008). Barron (1999) seems to have first observed that just the Kullback-Leibler property is sufficient to imply the convergence of the Cesàro Kullback-Leibler risk. Theorem 6.54 (with $\alpha = \frac{1}{2}$) as well as Example 6.55 are due to Walker and Hjort (2001). The martingale approach to consistency was introduced by Walker (2003, 2004); an extension to ergodic Markov processes can be found in Ghosal and Tang (2006). The prior constructed in Theorem 6.48 was suggested as an automatic prior for density estimation by Ghosal et al. (1997), who showed consistency of the posterior.

Problems

6.1 (Blackwell and Dubins 1962) In the setup of Theorem 6.6, assume in addition that two priors Π and Γ are such that P_Π^∞ and P_Γ^∞ are mutually absolutely continuous. Show that $\|Q_{\Pi,n} - Q_{\Gamma,n}\|_{TV} \to 0$ a.s. $[P_\Pi^\infty]$.

6.2 (Breiman et al. 1964) Consider real-valued i.i.d. observations from P_θ, $\theta \in \Theta = [0,1]$. Show that there exists $\tilde{f}(X_1, X_2, \dots)$ such that $P_\theta\{\tilde{f}(X_1, X_2, \dots) = \theta\} = 1$ for all $\theta \in \Theta$ if and only if $\{\theta \mapsto P_\theta(X \le x) \colon x \in \mathfrak{X}\}$ generates the Borel σ-field.

6.3 (Ghosh et al. 1994) Consider a sequence $\{(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, p^{(n)}(x^{(n)}, \theta)); \theta \in \Theta\}$ of dominated statistical experiments, where Θ is a finite-dimensional Euclidean space and the true value θ_0 is an interior point of Θ. Consider prior densities π_1 and π_2 which are positive and continuous at θ_0. Suppose that the posterior distributions π_{1n} and π_{2n} are strongly consistent (respectively, weakly consistent) at θ_0. Show that the two posterior distributions merge in total variation sense, i.e., $d_{TV}(\pi_{1n}, \pi_{2n}) \to 0$ a.s. (respectively, in probability) under the true distribution.

6.4 (Breiman et al. 1964) Let $0 < C < 1$ and $1 = a_0 > a_1 > a_2 > \cdots$ be defined by $\int_{a_k}^{a_{k-1}} (e^{1/x^2} - C)\, dx = 1 - C$. Consider the parameter space $\Theta = \mathbb{N}$ and for every $k \in \mathbb{N}$, let p_k be a probability density on $[0,1]$ defined by $p_k(x) = e^{x^{-2}}$ for $a_k < x < a_{k-1}$, and $p_k(x) = C$ otherwise. Consider i.i.d. observations $X_1, X_2, \dots$ from this model.

(a) Show that $a_k \to 0$ as $k \to \infty$.

(b) Show that the likelihood function is maximized on $\{k \colon a_k > \min(X_1, \dots, X_n)\}$ and hence the MLE exists.

(c) Let $I_n = I_n(X_1, \dots, X_n)$ be the unique k such that $\min(X_1, \dots, X_n) \in (a_k, a_{k-1})$. Show that for any $j \in \mathbb{N}$, $\sum_{i=1}^n \log(p_{I_n}/p_j) \to \infty$ in P_j probability, and hence the MLE tends to infinity. In particular, the MLE is inconsistent everywhere. (Note that the posterior is consistent everywhere by Doob's theorem.)

6.5 (Doksum and Lo 1990) Consider the setup of Example 6.14 without any symmetrization. Show that the "partial posterior" of θ given the median is consistent.

6.6 (Barron 1986) Let p_0 be the true density of i.i.d. observations and let $q(x_1, \dots, x_n | \mathcal{N}) = \int_{\mathcal{N}} \prod_{i=1}^n p(x_i; \theta)\, d\Pi_n(p)/\Pi_n(\mathcal{N})$, where Π_n is the prior for p and $\mathcal{N}$ is a neighborhood of p_0 with $\Pi_n(\mathcal{N}) > 0$. We say that the *conditional mixture likelihood* $q(X_1, \dots, X_n | \mathcal{N})$ matches with the true likelihood $\prod_{i=1}^n p_0(X_i)$ if for every $\epsilon > 0$,

$$e^{-n\epsilon} q(X_1, \dots, X_n | \mathcal{N}) \le \prod_{i=1}^n p_0(X_i) \le e^{n\epsilon} q(X_1, \dots, X_n | \mathcal{N}) \qquad (6.23)$$

for all sufficiently large n a.s. $[P_0^\infty]$. Show that the second condition in (6.23) is always satisfied, and first holds if $p_0 \in \mathrm{KL}_n(\Pi_n)$ and $\mathcal{N} \supset \{p \colon K(p_0; p) < \epsilon\}$ for sufficiently small $\epsilon > 0$.

6.7 Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} p$. Show that a prior for p of the form $(1-\epsilon)\Pi_0 + \epsilon\delta_{p_0}$ is consistent at p_0 with respect to d_H or the $\mathbb{L}_1$-distance.

6.8 Consider a model where a $[0,1]$-values observation has density $p := \pi u_0 + (1-\pi) f$, where u_0 is a fixed density on $[0,1]$ positive everywhere and f is an arbitrary density on $[0,1]$. Let $\pi \sim \mu$ and $f \sim \Pi$ be a priori independent. Suppose that the true value of π is π_0, $0 < \pi_0 < 1$ and $\pi_0 \in \mathrm{supp}(\mu)$. Let the true value of f be f_0, where f_0 belongs to the $\mathbb{L}_1$-support of Π. Show that $p_0 := \pi_0 u_0 + (1-\pi_0) f_0$ is in the Kullback-Leibler support of the prior induced on p.

6.9 (Salinetti 2003) Let a family of densities be parameterized by $\theta \in \Theta$. If the Kullback-Leibler property holds at the true density p_{θ_0}, then consistency holds if

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{U}^c} n^{-1}\sum_{i=1}^{n} \log \frac{p(X_i;\theta)}{p(X_i;\theta_0)} < 0 \tag{6.24}$$

for every neighborhood $\mathcal{U}$ of θ_0. Show that (6.24) may be derived from a weaker form of uniform strong law of large numbers such as *hypo-convergence* or Wald's method of deriving consistency of MLE.

6.10 (Ghosal and van der Vaart 2003) Show that (6.24) implies existence of uniformly consistent tests in Theorem 6.16.

6.11 (Choi and Ramamoorthi 2008) Often, especially in the i.i.d. case, one derives a stronger conclusion than consistency, namely: for any fixed neighborhood $\mathcal{U}$ of the true density p_0, the posterior probability $\Pi_n(\mathcal{U}^c| X_1,\ldots,X_n) \to 0$ exponentially fast a.s. $[P_0]$. Call this property strong *exponential consistency*. By the following example, show that the assertion is strictly stronger than consistency:

Let Π be a prior such that the posterior is inconsistent at f_0 with respect to the Hellinger (equivalently, $\mathbb{L}_1$) distance, and consider the prior $\Pi^* = \frac{1}{2}\Pi + \frac{1}{2}\delta_{p_0}$. Show that the posterior based on Π^* is strongly Hellinger consistent, but is not strongly exponentially consistent.

6.12 The Kullback-Leibler property is not necessary for consistency, even when the family is dominated. Consider the family $\mathrm{Unif}[\theta-1,\theta+1]$, $\theta\in\mathbb{R}$, i.e., density given by $p_\theta(x) = \frac{1}{2}\mathbb{1}\{\theta-1\le x\le\theta+1\}$, and the prior $\theta\sim\mathrm{Nor}(0,1)$. Show that $K(p_\theta; p_{\theta'}) = \infty$ whenever $\theta\neq\theta'$, so the Kullback-Leibler support of the prior (or any other prior) is empty; yet the posterior is consistent at any value of θ.

6.13 Prove the following generalization of Lemma 6.26 for general observations: if $p_\theta^{(n)}$ is the joint density of $X^{(n)}$, $\theta\in\Theta$, θ_0 is the true parameter and Π any prior on θ, then show that

$$P_{\theta_0}^{(n)}\left\{\int \frac{p_\theta^{(n)}}{p_{\theta_0}^{(n)}} d\Pi(\theta) < \Pi(\theta\colon K(p_{\theta_0}^{(n)}; p_\theta^{(n)}) < n\delta)e^{-2bn\delta}\right\} \le b^{-1}.$$

Hence show that in Theorem 6.39, the second condition on Kullback-Leibler variations can be dropped at the expense of strengthening the second assumption for arbitrary $C>0$. State a posterior consistency result based on the revised assertions.

6.14 (Choi and Ramamoorthi 2008) Derive Example 6.24 using Theorem 6.45.

6.15 Show that Theorem 6.54 holds for $\alpha=\alpha_n\to 1$ sufficiently slowly.

6.16 For any sequence of independent random observations, define the α-posterior in an analogous way. Show that Theorem 6.54 holds with the squared Hellinger distance d_H^2 replaced by the average squared Hellinger distance $n^{-1}\sum_{i=1}^n d_H^2(p_{0i},p_i)$ under the assumptions that there exists a sequence of sets B_n, such that $\liminf \Pi_n(B_n) > 0$ and (6.8) holds on B_n.

6.17 Use the argument at the end of Section 6.8.4 to show: If $p_0\in\mathrm{KL}(\Pi)$ and for every $\epsilon>0$ there exists a partition $\mathcal{P}=\cup_j\mathcal{P}_j$ and $\alpha<1$ such that $\sum_j \Pi(\mathcal{P}_j)^\alpha<\infty$, then the posterior distribution is strongly Hellinger consistent at p_0. [Hint: Given a Hellinger ball $\mathcal{U}$ of radius $2\epsilon>0$ around p_0 let $\mathcal{P}=\cup_j\mathcal{P}_j$ be a partition for ϵ as given, and let J be the set of indices of the partitioning sets that intersect $\mathcal{U}^c$. Because the diameters are smaller than ϵ, the convex hull of each of the latter sets has distance at least ϵ to p_0, whence

$$\rho_\alpha(\mathrm{conv}(\mathcal{P}_j), p_0) \leq 1 - \min(\alpha, 1-\alpha)\epsilon^2 \leq e^{-Cn}.$$

Finally, $\Pi_n(\cup_{j\in J} \mathcal{P}_j | X_1, \ldots, X_n) \leq \sum_{j\in J} \Pi_n(\mathcal{P}_j | X_1, \ldots, X_n)^\alpha$, and we can apply the preceding to each term of the sum to find a bound of the form $e^{c'n} \sum_j e^{-Cn} \Pi(\mathcal{P}_j)^\alpha$.]

6.18 (Walker 2004, Ghosal and Tang 2006) Consider an ergodic Markov chain with transition density f given a prior Π. Let the Kullback-Leibler property hold at the true value f_0. Assume that $\mathcal{A}_n \supset \mathcal{A}_{n+1}$ are random sets and $\mathbb{1}\{f \in \mathcal{A}_n\}$ is an $\mathscr{X}_n$-measurable random variable for all f and n. Show that $\Pi(\mathcal{A}_n | X_0, \ldots, X_n) \to 0$ a.s. if

$$\liminf_{N\to\infty} N^{-1} \sum_{n=0}^{N-1} d_H^2(f_0(\cdot| X_n), f_{n,\mathcal{A}_n}(\cdot| X_n)| X_n) > 0 \text{ a.s.}, \tag{6.25}$$

where

$$f_{n,\mathcal{A}_n}(y|X_n) = \frac{\int_{\mathcal{A}_n} f(y|X_n) \prod_{i=1}^n f(X_i|X_{i-1})\, d\Pi(f)}{\int_{\mathcal{A}_n} \prod_{i=1}^n f(X_i|X_{i-1})\, d\Pi(f)}$$

and $d_H^2(f(\cdot|x), g(\cdot|x)|x) = \int (\sqrt{f(y|x)} - \sqrt{g(y|x)})^2 d\nu(y)$.

Specialize the result to density estimation based on i.i.d. variables.

6.19 (Walker 2004) Show that for i.i.d. observations the Kullback-Leibler property implies consistency with respect to the weak topology using the i.i.d. case of Problem 6.18. [Hint: Consider a subbasic neighborhood for the weak topology $\mathcal{U} = \{p\colon P\psi < P_0\psi + \epsilon\}$, where ψ is a bounded continuous function. Observe that its complement is also convex!].