Convergence rates of posterior distributions.

Ghosal, S.; Ghosh, J.K.; van der Vaart, A.W.

published in
Annals of Statistics
2000

DOI (link to publisher)
10.1214/aos/1016218228

document version
Publisher's PDF, also known as Version of record

The Annals of Statistics
2000, Vol. 28, No. 2, 500–531

CONVERGENCE RATES OF POSTERIOR DISTRIBUTIONS

BY SUBHASHIS GHOSAL, JAYANTA K. GHOSH AND AAD W. VAN DER VAART

Free University Amsterdam, Indian Statistical Institute and Free University Amsterdam

We consider the asymptotic behavior of posterior distributions and Bayes estimators for infinite-dimensional statistical models. We give general results on the rate of convergence of the posterior measure. These are applied to several examples, including priors on finite sieves, log-spline models, Dirichlet processes and interval censoring.

1. Introduction. Suppose that we observe a random sample $X_1, \ldots, X_n$ from a distribution P with density p relative to some reference measure on the sample space $(\mathscr{X}, \mathscr{A})$. The unknown distribution is known to belong to some model $\mathscr{P}$, a set of probability measures on the sample space. Given some *prior* distribution Π_n on the set $\mathscr{P}$, the *posterior* distribution is the random measure given by

$$\Pi_n(B|X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(P)}{\int \prod_{i=1}^n p(X_i)\, d\Pi_n(P)}.$$

If the distribution P is considered random and distributed according to Π, as it is in Bayesian inference, then the posterior distribution is the conditional distribution of P given the observations. The prior is, of course, a measure on some σ-field on $\mathscr{P}$ and we must assume that the expressions in the display are well defined. In particular, we assume that the map $(x, p) \mapsto p(x)$ is measurable for the product σ-field on $\mathscr{X} \times \mathscr{P}$. It will be silently understood that the sets of which we compute prior or posterior measures are measurable.

In this paper we study the frequentist properties of the posterior distribution as $n \to \infty$, assuming that the observations are a random sample from some fixed measure P_0. In particular, we study the rate at which this random distribution converges to P_0. The posterior is said to be *consistent* if, as a random measure, it concentrates on arbitrarily small neighborhoods of P_0, with probability tending to 1 or almost surely, as $n \to \infty$. We study the rate at which such neighborhoods may decrease to zero meanwhile still capturing most of the posterior mass.

If $\mathscr{P} = \{P_\theta\colon \theta \in \Theta\}$ is parametrized by a parameter θ, then usually the prior is constructed by putting a measure on the parameter set Θ. If Θ is a

Received July 1998; revised December 1999.
AMS 1991 *subject classifications*. 62G15, 62G20, 62F25
Key words and phrases. Infinite dimensional model, posterior distribution, rate of convergence, sieves, splines

subset of a finite-dimensional Euclidean space and the dependence of $\theta \mapsto P_\theta$ is sufficiently regular, then it is well known that the posterior distribution of θ achieves the optimal rate of convergence, as $n \to \infty$ [see, for example, Le Cam (1973) and Ibragimov and Has'minskii (1981)]. In particular, if the model $\theta \mapsto P_\theta$ is suitably differentiable, then the rate of the posterior mean for θ is $\sqrt{n}$ and the posterior distribution, when rescaled, tends to a normal distribution with covariance the inverse Fisher information, according to the Bernstein–von Mises theorem. In that case the posterior expectation is an asymptotically efficient estimator for the parameter under some integrability conditions.

Much less is known on the behavior of posterior distributions for infinite-dimensional models. Most of the known results in this area address consistency issues. A famous theorem by Doob (1949) shows that consistency obtains on a set of prior measure 1, but his result concludes nothing on consistency at a particular true distribution of interest. Schwartz (1965) gives results that do apply to a particular true distribution. She shows that the posterior distribution is consistent if the true distribution P_0 can be suitably tested versus the complements of neighborhoods of P_0 and Kullback–Leibler neighborhoods of P_0 receive positive probabilities under the prior. Examples by, among others, Freedman (1963, 1965) and Diaconis and Freedman (1986) show that the situation is more complicated, even though, perhaps, these examples put too much emphasis on the situations where Bayes estimation does not work. A number of recent papers consider the consistency with a particular interest in the infinite-dimensional situation. Barron, in an unpublished paper, refines Schwartz's theorem [see Proposition 2 of Barron, Schervish and Wasserman (1999)] in a way that is particularly suitable for prior measures on infinite-dimensional spaces of densities. Ghosal, Ghosh and Ramamoorthi (1999a) use this extension to study consistency in the variation distance for Dirichlet mixture priors. For reviews on posterior consistency in infinite dimensions, see Ghosal, Ghosh and Ramamoorthi (1999b) and Wasserman (1998).

Le Cam [(1986), pages 509–529], addresses rates of convergence of Bayes estimators in an abstract setting. Our methods are clearly related to the methods used by Le Cam. A crucial distinction appears to be that Le Cam appears to base his argument on the prior mass present in fairly small balls (the sets V in his Lemma 1 on page 510, later chosen such that P_V is close to P_0^n), whereas our result is based on having sufficient prior mass in balls of radius equal to the rate of convergence that we wish to obtain. The behavior of product densities in these bigger balls appears not to be determined by the metric distance of the marginal components alone. Instead we use a combination of Kullback–Leibler numbers and distances on the log likelihood ratio ratios. Another distinction is that we consider rates of posterior measures, whereas Le Cam considers the rate of "formal Bayes estimators." For these reasons our results appear not to be covered by Le Cam's Theorem 1 on page 513 (in which distances and other quantities are in terms of the product measures P^n). However, we would like to acknowledge the great importance of Le Cam's work to

the present paper. In particular, we are indebted to the part of Le Cam's work as it was extended by Birgé (1983) to general metric spaces.

Shortly after completing this paper, we learned of independent work by Shen and Wasserman (1998), who also address rates of convergence.

The construction of prior measures on infinite-dimensional models is not a trivial matter and has also received recent attention. This development started with the introduction of Dirichlet processes by Ferguson (1973, 1974). Given computing algorithms such as Markov chain Monte Carlo methods and powerful computing machines, implementation of Bayesian methods has now become feasible even for many complicated priors and infinite dimensional models.

In Section 2 we present a main result and several variations concerning the rate of convergence of the posterior relative to the total variation, Hellinger and L_2-metrics. Every time the two main elements characterizing the rate of convergence are the size of the model (measured by covering numbers or existence of certain tests) and the amount of prior mass given to a shrinking ball around the true measure. Actually, the size of the model comes in only to guarantee the existence of certain tests of the true measure versus the complement of a shrinking ball around it, and conditions can be put in terms of such tests instead. Conditions of this form go back to Schwartz (1965) and Le Cam (1973). We discuss testing in Section 7, and reformulate our main result in terms of tests in this section. The proofs of the main results are contained in Section 8 following the discussion of the existence of tests. In Section 2 we also note that a rate of convergence for the posterior automatically entails the existence of point estimators with the same rate.

We apply the general result to several examples. In Section 3 we consider discrete priors constructed on ε-nets over the model. In Section 4 we discuss Bayes estimators based on the log spline models for density estimation discussed by Stone (1986). In Section 6 we consider finite-dimensional models. In Section 6 we discuss applications to Dirichlet priors.

The notation $\lesssim$ is used to denote inequality up to a universal multiplicative constant, or up to a constant that is fixed throughout. We define the *Hellinger distance* $h(p, q)$ or $h(P, Q)$ between two probability densities or measures by the L_2-distance between the root densities $\sqrt{p}$ and $\sqrt{q}$. The total variation distance is the L_1-distance. (Some authors define these distances with an additional factor 1/2.)

2. Main results. Let $X_1, \ldots, X_n$ be distributed according to some distribution P_0 and let Π_n be a sequence of prior probability measures supported on some set of probability measures $\mathscr{P}$. Let d be either the variation or the Hellinger metric on $\mathscr{P}$. If the set of densities is uniformly bounded, then we may also choose d equal to the L_2-distance. This metric is used in condition (2.2) of the following theorem and also in its assertion.

Let $D(\varepsilon, \mathscr{P}, d)$ denote the *ε-packing number* of $\mathscr{P}$. This is the maximal number of points in $\mathscr{P}$ such that the distance between every pair is at least ε. It is easy to see that this is related to the *ε-covering number* $N(\varepsilon, \mathscr{P}, d)$,

which is the minimal number of balls of radius ε needed to cover $\mathscr{P}$, by the inequalities

$$N(\varepsilon, \mathscr{P}, d) \leq D(\varepsilon, \mathscr{P}, d) \leq N(\varepsilon/2, \mathscr{P}, d). \tag{2.1}$$

Because we are only interested in rates of convergence, the additional constant 2 is of no real importance in the following, and covering numbers may replace packing numbers throughout. The set of centers of a minimal set of balls of radius ε covering $\mathscr{P}$ is called an *ε-net*.

A good early reference on entropy numbers is the paper Kolmogorov and Tikhomirov (1961). Alternative references are Dudley (1984) and van der Vaart and Wellner (1996).

We use the notation Pf to abbreviate $\int f\,dP$, and, later on, $\mathbb{P}_n f$ for $n^{-1}\sum_{i=1}^n f(X_i)$.

THEOREM 2.1. *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and sets $\mathscr{P}_n \subset \mathscr{P}$, we have*

$$\log D(\varepsilon_n, \mathscr{P}_n, d) \leq n\varepsilon_n^2, \tag{2.2}$$

$$\Pi_n(\mathscr{P} \setminus \mathscr{P}_n) \leq \exp\bigl(-n\varepsilon_n^2(C+4)\bigr), \tag{2.3}$$

$$\Pi_n\Bigl(P\colon -P_0\Bigl(\log\frac{p}{p_0}\Bigr) \leq \varepsilon_n^2, P_0\Bigl(\log\frac{p}{p_0}\Bigr)^2 \leq \varepsilon_n^2\Bigr) \geq \exp\bigl(-n\varepsilon_n^2 C\bigr). \tag{2.4}$$

Then for sufficiently large M, we have that $\Pi_n(P\colon d(P, P_0) \geq M\varepsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability.

The first and third conditions of the theorem are the essential ones. Condition (2.3) allows some additional flexibility, but should first be understood as expressing that $\mathscr{P}_n$ is almost the support of the prior (in which case its left side is zero and the condition is trivially satisfied).

Condition (2.2) requires that the "model" $\mathscr{P}_n$ be not too big. It is true for every $\varepsilon_n' \geq \varepsilon_n$ as soon as it is true for ε_n and can thus be seen as defining a minimal possible value of ε_n. Condition (2.2) ensures the existence of certain tests, as discussed in Section 7, and could be replaced by a testing condition. Note that the metric d used here reappears in the assertion of the theorem. Since the total variation metric is bounded above by twice the Hellinger metric, the assertion of the theorem using the Hellinger metric is stronger, but also condition (2.2) will be more restrictive, so that we really have two theorems. In the case that the densities are uniformly bounded, we even have a third theorem, when using the L_2-distance, which in that case will be bounded above by a multiple of the Hellinger distance. If the densities are also uniformly bounded and uniformly bounded away from zero, then these three distances

are equivalent and are also equivalent to the Kullback–Leibler number and L_2-norm appearing in condition (2.4). See, for example, Lemmas 8.2 and 8.3 and (8.6).

A rate ε_n satisfying (2.2) for $\mathscr{P} = \mathscr{P}_n$ and d the Hellinger metric is often viewed as giving the "optimal" rate of convergence for estimators of P relative to the Hellinger metric, given the model $\mathscr{P}$. Under certain conditions, such as likelihood ratios bounded away from zero and infinity, this is proved as a theorem by Birgé (1983) and Le Cam (1973, 1986). From Birgé's work it is clear that condition (2.2) is the correct expression of the complexity of the model, as relating to estimating the true density relative to the Hellinger distance, if this is to be given in terms of metric entropy. A weaker, but more involved, condition is in terms of the existence of certain tests. We give a generalization of the theorem using tests in Section 7.

Condition (2.4) is the other main determinant of the posterior rate given by the theorem. It requires that the prior measures put a sufficient amount of mass near the true measure P_0. Here "near" is measured through a combination of the Kullback–Leibler divergence of p and p_0 and the $L_2(P_0)$-norm of $\log(p/p_0)$. Again this condition is satisfied for $\varepsilon_n' \geq \varepsilon_n$ if it is satisfied for ε_n and thus is another restriction on a minimal value of ε_n. The form of this condition can be motivated from entropy considerations. Suppose that we wish to satisfy (2.4) for the minimal ε_n satisfying (2.2) with $\mathscr{P}_n = \mathscr{P}$, that is, for the optimal rate of convergence for the model. Furthermore, for the sake of the argument assume that all distances used are equivalent. Then a minimal ε_n-cover of $\mathscr{P}$ consists of $\exp(n\varepsilon_n^2)$ balls. If the prior Π_n would spread its mass uniformly over $\mathscr{P}$, then every ball would obtain mass approximately $\exp(-Cn\varepsilon_n^2)$. (The constant C expresses the constants in comparing the distances and the fact that the balls of radius ε_n may overlap.) On the other hand, if Π_n is not "uniform," then we should expect (2.4) to fail for some $P_0 \in \mathscr{P}$. Here we must admit that "uniform" priors do not exist in infinite-dimensional models and actually condition (2.4) is stronger than needed and will be improved ahead in Theorem 2.4. However, a rough implication of the condition is that Π_n should be "uniformly spread" in order for the posterior distribution to attain the optimal rate of convergence.

Condition (2.3), combined with (2.2), can be interpreted as saying that a part of $\mathscr{P}$ that barely receives prior mass need not be small. The sets $\mathscr{P}_n$ may be thought of as "sieves" approximating the parameter space, which capture most of the prior probability. This type of condition has received much attention in the discussion of consistency issues [see Barron, Schervish and Wasserman (1998)], but plays a smaller role in the present paper. Of course, condition (2.3) is trivially satisfied for $\mathscr{P}_n = \mathscr{P}$; we can make this choice if condition (2.2) holds with $\mathscr{P}_n = \mathscr{P}$ itself.

The assertion of the theorem is an in-probability statement that the posterior mass outside a large ball of radius proportional to ε_n is approximately zero. The in-probability statement can be improved to an almost sure assertion, but under stronger conditions. We present two results.

Let h be the Hellinger distance and write $\log_+ x$ for $(\log x) \vee 0$.

THEOREM 2.2. *Suppose that conditions* (2.2) *and* (2.3) *hold as in the preceding theorem and in addition* $\sum_n \exp(-Bn\varepsilon_n^2) < \infty$ *for every* $B > 0$ *and*

$$\Pi_n\Big(P\colon h^2(P, P_0)\Big\|\frac{p_0}{p}\Big\|_\infty \le \varepsilon_n^2\Big) \ge \exp\big(-n\varepsilon_n^2 C\big). \tag{2.5}$$

Then for sufficiently large M, *we have that* $\Pi_n(P\colon d(P, P_0) \ge M\varepsilon_n | X_1, \ldots, X_n) \to 0$ *in* P_0^n*-almost surely.*

THEOREM 2.3. *Suppose that conditions* (2.2) *and* (2.3) *hold as in the preceding theorem and in addition* $\sum_n \exp\big(-Bn\varepsilon_n^2\big) < \infty$ *for every* $B > 0$ *and for a given function* m *with* $P_0 m < \infty$,

$$\Pi_n\Big(P\colon 18h^2(P, P_0)\Big(\log_+\big(\sqrt{P_0 m}/h(P, P_0)\big) + \Phi^{-1}(h^2(P, P_0))\Big) \le \varepsilon_n^2, \frac{p_0}{p} \le m\Big) \ge \exp\big(-n\varepsilon_n^2 C\big), \tag{2.6}$$

where $\Phi^{-1}(\varepsilon) = \sup\{M\colon \Phi(M) \ge \varepsilon\}$ *is the inverse of the function* $\Phi(M) = P_0 m 1\{m \ge M\}/M$. *Then for sufficiently large* M, *we have that* $\Pi_n(P\colon d(P, P_0) \ge M\varepsilon_n | X_1, \ldots, X_n) \to 0$ *in* P_0^n*-almost surely.*

If the quotients p_0/p are uniformly bounded, then condition (2.5) simply requires that shrinking Hellinger balls possess a sufficient amount of prior mass. Then a fairly symmetric statement is obtained when combined with condition (2.2) for the Hellinger metric d: if we can cover the model with not too many Hellinger balls and the Hellinger ball around P_0 contains a sufficient amount of mass, then the rate of convergence relative to the Hellinger distance is ε_n.

Lemmas 8.2 and 8.3 in Section 8 relate the Kullback–Leibler divergence and L_2-norm of $\log(p/p_0)$ to $h^2(P, P_0)\|p_0/p\|_\infty$ and imply that the conditions of Theorem 2.2 are essentially stronger than those of Theorem 2.1.

Condition (2.6) is milder in its control of p/p_0 than (2.5) by allowing a general bound m that need only satisfy a moment condition. However, in comparison with (2.5) it will be satisfied only for somewhat bigger ε_n, due to the presence of the term involving log and Φ^{-1}.

In general, good control on the quotients p/p_0 is needed next to the closeness of p to p_0 relative to, for example, the Hellinger metric, because the product measures P^n and P_0^n can be arbitrarily far apart as $n \to \infty$ within balls of radii ε_n, for the values of ε_n bigger than $n^{-1/2}$ that we are considering here. The bound on p/p_0 together with the distance ensures that P^n and P_0^n are still "close" enough on an exponential scale. Only prior mass on such close alternatives helps to increase the rate of convergence of the posterior.

One deficit of the theorems as presented so far is that they do not satisfactorily cover finite-dimensional models. When applied to such models, they would yield the rate $1/\sqrt{n}$ times a logarithmic factor rather than $1/\sqrt{n}$ itself. Similarly, the theorems may also yield unnecessary logarithmic factors when applied to priors constructed on a sequence of finite-dimensional sieves. To

improve this situation we must refine both the entropy condition (2.2) and the prior mass condition (2.4). The following generalization of Theorem 2.1 is more complicated but does yield the right result in the finite-dimensional situation. It is essential for our examples using spline approximations in Section 4. Theorems 2.2 and 2.3 can be generalized similarly. Let

$$B_n(\varepsilon) = \Big\{P\colon\ -P_0\Big(\log\frac{p}{p_0}\Big) \le \varepsilon^2,\ \ P_0\Big(\log\frac{p}{p_0}\Big)^2 \le \varepsilon^2\Big\}.$$

THEOREM 2.4. *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and such that $n\varepsilon_n^2$ is bounded away from zero, every sufficiently large j and sets $\mathscr{P}_n \subset \mathscr{P}$, we have*

$$\log D\Big(\frac{\varepsilon}{2}, \{P \in \mathscr{P}_n\colon \varepsilon \le d(P, P_0) \le 2\varepsilon\}, d\Big) \le n\varepsilon_n^2 \quad \text{for every } \varepsilon \ge \varepsilon_n, \tag{2.7}$$

$$\frac{\Pi_n(\mathscr{P} - \mathscr{P}_n)}{\Pi_n\big(B_n(\varepsilon_n)\big)} = o\big(\exp\big(-2n\varepsilon_n^2\big)\big), \tag{2.8}$$

$$\frac{\Pi_n\big(P\colon\ j\varepsilon_n < d(P, P_0) \le 2j\varepsilon_n\big)}{\Pi_n\big(B_n(\varepsilon_n)\big)} \le \exp\big(Kn\varepsilon_n^2 j^2/2\big). \tag{2.9}$$

Here K is the universal testing constant appearing in (7.1) *and* (7.2). *Then for every $M_n \to \infty$, we have that $\Pi_n(P\colon d(P, P_0) \ge M_n\varepsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability.*

Convergence of the posterior distribution at the rate ε_n implies the existence of point estimators, which are Bayes in that they are based on the posterior distribution, that converge at least as fast as ε_n in the frequentist sense. One possible construction is to define $\hat{P}_n$ as the (near) maximizer of

$$Q \mapsto \Pi_n\big(P\colon d(P, Q) < \varepsilon_n | X_1, \ldots, X_n\big).$$

THEOREM 2.5. *Suppose that $\Pi_n(P\colon d(P, P_0) \ge \varepsilon_n | X_1, \ldots, X_n)$ converges to* 0, *almost surely (respectively, in probability) under P_0^n and let $\hat{P}_n$ maximize, up to $o(1)$, the function $Q \mapsto \Pi_n\big(P\colon d(P, Q) < \varepsilon_n | X_1, \ldots, X_n\big)$. Then $d(\hat{P}_n, P_0) \le 2\varepsilon_n$ eventually almost surely (respectively, in probability) under P_0^n.*

PROOF. By definition, the ε_n-ball around $\hat{P}_n$ contains at least as much posterior probability as the ε_n-ball around P_0. The latter, by posterior convergence at rate ε_n, has posterior probability close to unity. Therefore, these two balls cannot be disjoint, for otherwise, the total posterior mass would exceed unity. Now apply the triangle inequality. □

The preceding construction actually applies to general statistical models and posterior distributions, and the theorem is well-known. [See, e.g., Le Cam (1986) or Le Cam and Yang (1990).] If we use the Hellinger or total variation metric (or some other bounded metric whose square is convex), then

an alternative is to use the posterior expectation, which typically has a similar property. By Jensen's inequality and convexity of $P \mapsto d^2(P, P_0)$,

$$d^2\Big(\int P\, d\Pi_n(P|X_1, \ldots, X_n), P_0\Big) \le \int d^2(P, P_0)\, d\Pi_n(P|X_1, \ldots, X_n)$$
$$\le \varepsilon_n^2 + d_\infty^2 \Pi_n\big(P\colon d(P, P_0) > \varepsilon_n | X_1, \ldots, X_n\big),$$

where d_∞ is a bound on the maximal distance ($\sqrt{2}$ and 2, respectively, for Hellinger and variation distance). To obtain the desired result, we now need that the posterior probability of the complement of the ε_n-ball around p_0 converges to zero at least at the order ε_n^2. This is usually the case, in particular under the conditions of Theorems 2.2 and 2.3, whose proofs yield the exponential order $\exp(-Bn\varepsilon_n^2)$. (We use the square of the distance, because the Hellinger distance is not convex. With the total variation distance the argument would work also with the distance itself.)

More generally, we could use the minimizer $\hat{P}_n$ of

$$Q \mapsto \int \ell_n\big(d(Q, P)\big)\, d\Pi_n(P|X_1, \ldots, X_n),$$

for appropriate loss functions ℓ_n. Such estimators are called *formal Bayes estimators* in Le Cam (1986).

On the one hand, Theorem 2.5 shows that we can construct good estimators from the posterior if the posterior converges at a good rate. On the other hand, it shows that the posterior cannot converge at a rate faster than the optimal rate of convergence for point estimators. We use this argument in a number of examples to show that the posterior converges at the best possible rate. Of course, our arguments have nothing to say about the best possible constants. Furthermore, for many priors the rate may be suboptimal.

3. Priors based on finite approximating sets. In this section, we construct, under bracketing entropy conditions, priors based on uniform distributions on carefully chosen finite sets for which the posterior converges at the best possible rate. Priors based on uniform distributions on finite subsets are introduced by Ghosal, Ghosh and Ramamoorthi (1997) as the Bayesian default priors for nonparametric problems. They establish posterior consistency for such priors under mild entropy conditions. In the present case, the prior is constructed more carefully to achieve the optimal rate of convergence as well.

Given two functions $l, u\colon \mathscr{X} \mapsto \mathbb{R}$ the *bracket* $[l, u]$ is defined as the set of all functions $f\colon \mathscr{X} \mapsto \mathbb{R}$ such that $l \le f \le u$ everywhere. The bracket is said to be of size ε relative to the distance d if $d(l, u) < \varepsilon$. In this section we use the Hellinger distance h as the distance d and restrict the brackets to consisting of nonnegative functions, which are assumed to be integrable relative to a reference measure μ. Let $N_{[]}(\varepsilon, \mathscr{P}, h)$ be the minimal number of brackets of size ε needed to cover $\mathscr{P}$. Because a bracket of size ε is contained in the ball of radius $\varepsilon/2$ around its midpoint, it follows that $N(\varepsilon/2, \mathscr{P}, h) \le N_{[]}(\varepsilon, \mathscr{P}, h)$

and hence the present bracketing numbers are bigger than the packing numbers $D(\varepsilon, \mathscr{P}, h)$ defined previously [see (2.1)]. However, in many examples there is also an equality in the other direction, up to a constant, and bracketing and packing numbers give equivalent results. The corresponding bracketing entropy is defined as the logarithm of the bracketing number $N_{[]}(\varepsilon, \mathscr{P}, h)$.

We shall construct a discrete prior supported on densities constructed from minimal sets of brackets for the Hellinger distance. For a given number $\varepsilon_n > 0$, let Π_n be the uniform discrete measure on the $N_{[]}(\varepsilon_n, \mathscr{P}, h)$ densities obtained by covering $\mathscr{P}$ with a minimal set of ε_n-brackets and next renormalizing the upper bounds of the brackets to integrate to 1. Thus if $[l_1, u_1], \ldots, [l_N, u_N]$ are the $N = N_{[]}(\varepsilon_n, \mathscr{P}, h)$ brackets, then Π_n is the uniform measure on the N functions $u_j / \int u_j \, d\mu$. Next set

$$\Pi = \sum_{n \in \mathbb{N}} \lambda_n \Pi_n$$

for a given sequence λ_n with $\lambda_n \geq 0$ and $\sum_n \lambda_n = 1$.

THEOREM 3.1. *Suppose that ε_n are numbers decreasing in n such that $\log N_{[]}(\varepsilon_n, \mathscr{P}, h) \leq n\varepsilon_n^2$ for every n and $n\varepsilon_n^2 / \log n \to \infty$. Construct the prior Π as given previously for a sequence λ_n such that $\lambda_n > 0$ for all n and $\log \lambda_n^{-1} = O(\log n)$. Then the conditions of Theorem 2.2 are satisfied for ε_n a sufficiently large multiple of the present ε_n and hence the corresponding posterior converges at the rate ε_n almost surely, for every $P_0 \in \mathscr{P}$, relative to the Hellinger distance.*

PROOF. The prior Π gives probability 1 to the set $\mathscr{Q} = \bigcup_{j=1}^{\infty} \mathscr{P}_j$ for $\mathscr{P}_j$ the $N_{[]}(\varepsilon_j, \mathscr{P}, h)$ functions in the support of Π_j. We claim that

$$D\big(8\varepsilon_n, \mathscr{Q}, h\big) \leq \exp\big(2n\varepsilon_n^2\big). \tag{3.1}$$

To see this, we first note that, given an ε-bracket $[l, u]$ that contains a probability density p, with $\|\cdot\|_2$ the norm in $L_2(\mu)$,

$$1 \leq \Big(\int u \, d\mu\Big)^{1/2} = \|\sqrt{u}\|_2 \leq \|\sqrt{u} - \sqrt{p}\|_2 + \|\sqrt{p}\|_2 = h(u, p) + 1 \leq 1 + \varepsilon,$$

$$h\Big(p, \frac{u}{\int u \, d\mu}\Big) \leq h(p, u) + h\Big(u, \frac{u}{\int u \, d\mu}\Big) \leq 2\varepsilon.$$

Therefore, $\mathscr{P}_n$ is a $2\varepsilon_n$-net over $\mathscr{P}$: every point of $\mathscr{P}$ is within distance $2\varepsilon_n$ of some point in $\mathscr{P}_n$. Since for $j > n$ every point of $\mathscr{P}_j$ is within distance $2\varepsilon_j \leq 2\varepsilon_n$ of $\mathscr{P}$, it follows that $\mathscr{P}_n$ is also a $4\varepsilon_n$-net over $\mathscr{P}_j$. This being true for every $j > n$ it follows that $\mathscr{P}_n$ is a $4\varepsilon_n$-net over $\bigcup_{j \geq n} \mathscr{P}_j$ and hence, trivially, $\bigcup_{j \leq n} \mathscr{P}_j$ is a $4\varepsilon_n$-net over $\mathscr{Q}$. The cardinality of the latter net is at most $nN_{[]}(\varepsilon_n, \mathscr{P}, h) \leq \exp(n\varepsilon_n^2 + \log n) \leq \exp(2n\varepsilon_n^2)$ for sufficiently large n. By

virtue of the relationship (2.1) between covering numbers and packing numbers, we obtain (3.1). This verifies condition (2.2) with $\mathscr{P}_n$ taken equal to $\mathscr{Q}$ and ε_n taken equal to eight times the present ε_n.

If u is the upper limit of the ε_n-bracket containing p_0, then

$$\frac{p_0}{(u/\int u\,d\mu)} \le \int u\,d\mu \le (1+\varepsilon_n)^2.$$

It follows that for large n the set of points p such that $h^2(p, p_0)\|p_0/p\|_\infty \le 8\varepsilon_n^2$ contains at least the function $u/\int u\,d\mu$ and hence has prior mass at least

$$\lambda_n \frac{1}{N_{[\,]}(\varepsilon_n, \mathscr{P}, h)} \ge \exp[-n\varepsilon_n^2 - O(\log n)] \ge \exp(-2n\varepsilon_n^2),$$

for large n. This verifies condition (2.5) for ε_n a multiple of the present ε_n.

Since condition (2.3) is trivially satisfied for $\mathscr{P}_n = \mathscr{Q}$, the proof is complete. □

There are many specific examples in which the preceding theorem applies. The situation here is similar to that in recent papers on rates of convergence of (sieved) maximum likelihood estimators, as in Birgé and Massart (1993, 1997, 1998), Wong and Shen (1995) or Chapter 3.4 of van der Vaart and Wellner (1996). It is interesting to note that these authors also use brackets, whereas Birgé (1983), in his study of the metric entropy of statistical models, uses ε-nets. This is because the cited papers are concerned with a particular type of estimator (namely, minimum contrast estimators), whereas Birgé (1983) uses special constructs, called *d-estimators*. It appears that for good behavior of Bayes estimators on nets we also need some special property of the nets, such as available from nets obtained from brackets.

We include two concrete examples.

EXAMPLE 3.1 (Smooth densities). Suppose that $\mathscr{P}$ consists of all measures with densities whose roots $\sqrt{p}$ belong to a fixed multiple of the unit ball of the Hölder class $C^\alpha[0,1]$, for some fixed $\alpha > 0$. [See, e.g., van der Vaart and Wellner (1996) for a precise definition of this space of functions.] By results of Kolmogorov and Tihomirov (1961), the ε-entropy numbers of this unit ball relative to the uniform norm are bounded by a multiple of $(1/\varepsilon)^{1/\alpha}$. [Their result is reproduced in Theorem 2.7.1 of van der Vaart and Wellner (1996).] Because we can construct upper and lower brackets from uniform approximations, this shows that the bracketing Hellinger entropies grow like $\varepsilon^{-1/\alpha}$, so that we can take ε_n of the order $n^{-\alpha/(2\alpha+1)}$ to satisfy the relation $\log N_{[\,]}(\varepsilon_n, \mathscr{P}, h) \le n\varepsilon_n^2$. This rate is known to be the frequentist optimal rate for estimators. From Theorem 2.5, we therefore conclude that the prior constructed above achieves the optimal rate of convergence for the posterior.

Upper brackets are, in principle, available from the classical proof of Kolmogorov and Tihomirov (1961). Alternatively, we may use more modern classes of approximating functions, such as wavelets or splines.

EXAMPLE 3.2 (Monotone densities). Suppose that $\mathscr{P}$ consists of all monotone decreasing densities on a compact interval in $\mathbb{R}$, bounded above by a fixed constant. The root of a monotone density is monotone and hence the bracketing entropy of $\mathscr{P}$ for the Hellinger distance is bounded by the L_2-entropy for the set of monotone functions. This is of the order $1/\varepsilon$ [e.g., van der Vaart and Wellner (1996), Theorem 2.7.5], whence we obtain a $n^{-1/3}$-rate of convergence of the posterior. Again this rate cannot be improved.

Inspection of the proof of the theorem shows that the lower bounds of the brackets are not really needed. The theorem can be generalized by defining *upper bracketing numbers* $N_{]}(\varepsilon, \mathscr{P}, h)$ as the minimal number of functions $u_1, \ldots, u_m$ such that for every $p \in \mathscr{P}$ there exist a function u_i such that both $p \leq u_i$ and $h(u_i, p) < \varepsilon$. Next we construct a prior Π as before. These upper bracketing numbers are clearly smaller than the bracketing numbers $N_{[]}(\varepsilon, \mathscr{P}, h)$. We have formulated the theorem using the better known bracketing numbers, because we do not know any example where this generalization could be useful.

The preceding theorem implicitly requires that the model $\mathscr{P}$ be totally bounded for the Hellinger metric. A simple modification works for countable unions of totally bounded models, provided that we use a sequence of priors. Suppose that the bracketing numbers of $\mathscr{P}$ are infinite, but there exist subsets $\mathscr{P}_n \uparrow \mathscr{P}$ with finite bracketing numbers. Let ε_n be numbers such that $\log N_{[]}(\varepsilon_n, \mathscr{P}_n, h) \leq n\varepsilon_n^2$. Then we construct Π_n as the discrete uniform distribution on renormalized upper brackets of a minimal set of ε_n-brackets over $\mathscr{P}_n$, as before. Then the posterior relative to prior Π_n achieves the convergence rate ε_n. (Note that this time we do not construct a fixed prior $\Pi = \sum_n \lambda_n \Pi_n$, but use the prior Π_n when n observations are available.)

In the preceding we start with a condition on the entropies with bracketing even though we apply Theorem 2.2, which demands control over metric entropies only. This is because Theorem 2.2 also requires control over the likelihood ratios. If, for instance, the densities would be uniformly bounded away from zero and infinity, so that the quotients p_0/p are uniformly bounded, then we can replace the bracketing entropy in Theorem 3.1 by ordinary entropy. Alternatively, if the set of densities $\mathscr{P}$ possesses an integrable envelope function, then we can construct priors achieving the rate ε_n determined by the covering numbers up to logarithmic factors. Here we define ε_n as the minimal solution of the equation $\log N(\varepsilon, \mathscr{P}, h) \leq n\varepsilon^2$ and $N(\varepsilon, \mathscr{P}, h)$ denotes the Hellinger covering number (without bracketing). The construction, described briefly below, parallels Theorem 6 of Wong and Shen (1995) for sieved maximum likelihood estimators.

We assume that the set of densities $\mathscr{P}$ has a μ-integrable envelope function: a measurable function m with $\int m \, d\mu < \infty$ such that $p \leq m$ for every $p \in \mathscr{P}$. Given $\varepsilon_n > 0$ let $\{s_{1,n}, \ldots, s_{N_n,n}\}$ be a minimal ε_n-net over $\mathscr{P}$ [hence $N_n = N(\varepsilon_n, \mathscr{P}, h)$] and put

$$g_{j,n} = (s_{j,n}^{1/2} + \varepsilon_n m^{1/2})^2 / c_{j,n},$$

where $c_{j,n}$ is a constant ensuring that $g_{j,n}$ is a probability density. Finally, let Π_n be the uniform discrete measure on $g_{1,n}, \ldots, g_{N_n,n}$ and let $\Pi = \sum_{n=1}^{\infty} \lambda_n \Pi_n$ be a convex combination of the Π_n as before.

THEOREM 3.2. *Suppose that ε_n are numbers decreasing in n such that $\log N(\varepsilon_n, \mathscr{P}, h) \leq n\varepsilon_n^2$ for every n and $n\varepsilon_n^2/\log n \to \infty$. Construct the prior Π as given previously for a sequence λ_n such that $\lambda_n > 0$ for all n and $\log \lambda_n^{-1} = O(\log n)$. Assume m is a μ-integrable envelope. Then the corresponding posterior converges at the rate $\varepsilon_n \log(1/\varepsilon_n)$ in probability, relative to the Hellinger distance.*

PROOF. The proof follows as before, but this time we apply Theorem 2.1, using the observation of Wong and Shen (1995) that for any $p \in \mathscr{P}$ such that $h(p, s_{j,n}) \leq \varepsilon_n$ we have that $h(p, g_{j,n}) = O(\varepsilon_n)$ and that $p/g_{j,n}$ is bounded above by a multiple of ε_n^{-2}. This verifies (2.4) with ε_n replaced by a multiple of $\varepsilon_n \log(1/\varepsilon_n)$ through a use of Theorem 5 of Wong and Shen (1995), the relevant part of which is reproduced below as Lemma 8.6. □

4. Log spline models. In this section we apply the general results to prior distributions on log spline models for densities. Log spline models for density estimation have been used, among others, by Stone (1990), who shows that the sieved maximum likelihood estimator attains the optimal rate of convergence for estimating a smooth density. As shown by Stone (1994) they can be extended to higher dimensions by using tensor splines, but following Stone (1990), we restrict ourselves to the one-dimensional case.

We assume that the observations are sampled from a density p_0 on the unit interval $[0, 1]$ in the real line that is bounded away from zero and infinity. Our choice of priors will yield the optimal rate of convergence of the posterior if the density p_0 belongs to the Hölder space $C^{\alpha}[0, 1]$. (This is the set of all functions that have α_0 derivatives, for α_0 the largest integer strictly smaller than α, with the α_0th derivative being Lipschitz of order $\alpha - \alpha_0$.)

Our prior measures will not be supported on the set of smooth functions, but on exponential families constructed from a spline basis. Fix some "order" q, a natural number, throughout this section. Let K be another natural number, which will increase with n, and partition the half-open unit interval $[0, 1)$ into K subintervals $\big[(k-1)/K, k/K\big)$ for $k = 1, \ldots, K$. Consider the linear space of splines of order q relative to this partition, that is, all functions $f\colon [0, 1) \mapsto \mathbb{R}$ whose restriction to every of the partioning intervals $\big[(k-1)/K, k/K\big]$ is a polynomial of degree strictly less than q and, in the case that $q \geq 2$, that are $q-2$ times continuously differentiable on $[0, 1)$. It can be shown that this is a $J = q + K - 1$-dimensional vector space. A convenient basis is the set of B-splines $B_1, \ldots, B_J$, defined, for example, in de Boor (1978). More precisely, let $B_1, \ldots, B_J$ be the B-splines of order q for the known sequence

$$\overbrace{0, 0, \ldots, 0}^{q \text{ times}}, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, \overbrace{1, 1, \ldots, 1}^{q \text{ times}},$$

as defined on page 108 of de Boor (1978). The exact nature of these functions does not matter to us here, except for the following properties [cf. de Boor (1978), pages 109 and 110]:

1. $$B_j \geq 0, \qquad j = 1, \ldots, J$$

2. $$\sum_{j=1}^{J} B_j \equiv 1$$

3. B_j is supported inside an interval of length q/K
4. at most q functions B_j are nonzero at every given x.

The first two properties express that the basis elements form a partition of unity, and the third and fourth properties mean that their supports are close to being disjoint if K is very large relative to q.

For $\theta \in \mathbb{R}^J$ let $\theta^T B = \sum_j \theta_j B_j$ and define

$$p_\theta(x) = \exp\bigl(\theta^T B(x) - c(\theta)\bigr), \qquad e^{c(\theta)} = \int_0^1 \exp\bigl(\theta^T B(x)\bigr)\, dx.$$

Thus p_θ belongs to a J-dimensional exponential family, with the B-spline functions as sufficient statistics. Since the B-splines add up to unity, the family is actually of dimension $J-1$ and we could restrict θ to the subset $\Theta_0 = \{\theta \in \mathbb{R}^J\colon \theta^T 1 = 0\}$. The true density p_0 of the observations need not be of the form p_θ for some θ. (Hence we make a difference between p_0 and $p_{\mathbf{0}}$ for $\mathbf{0} \in \mathbb{R}^J$; this should not lead to confusion as $p_{\mathbf{0}}$ does not play a role.) In the following we construct a prior measure Π_n on the set of probability densities on $[0, 1]$ by choosing a prior on Θ_0, which next induces a prior on the probability densities p_θ through the map $\theta \mapsto p_\theta$.

For $q = 1$ the linear space of splines consists of histograms with cell boundaries k/K for $k = 0, 1, \ldots, K$. Since exponentials of histograms are histograms, our construction therefore contains priors constructed on histograms as a special case.

Since the true density p_0 need not belong to this "log spline model," we must ensure that it is approximated sufficiently closely by some p_θ. To approximate sufficiently many p_0 it is necessary to let the dimension $J-1$ of the log spline models tend to infinity with n. Here we fix the order q and let the number K of partioning sets tend to infinity. If we focus on α-smooth densities p_0, then the minimal rate at which $J = J_n$ must grow is determined by the following lemma, taken from de Boor [(1978), page 170]. Let $\|f\|_\infty = \sup_{0\leq x\leq 1}|f(x)|$ be the supremum norm, and let $\|f\|_\alpha$ be the seminorm

$$\|f\|_\alpha = \sup_{x\neq y} \frac{\bigl|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y)\bigr|}{|x-y|^{\alpha-\alpha_0}}.$$

Because we assume that p_0 is bounded away from zero (and infinity) the function p_0 is in $C^\alpha[0, 1]$ if and only if $\log p_0 \in C^\alpha[0, 1]$.

LEMMA 4.1. *Let $q \geq \alpha > 0$. There exists a constant C depending only on q and α such that, for every $p_0 \in C^\alpha[0,1]$ that is bounded away from zero,*

$$\inf_{\theta \in \mathbb{R}^J} \| \theta^T B - \log p_0 \|_\infty \leq C J^{-\alpha} \| \log p_0 \|_\alpha.$$

It is easy to see from this, as we show in part below, that the root of the Kullback–Leibler divergence and the Hellinger distance between p_0 and the closest p_θ are of the order $J^{-\alpha}$ as well. Since a ball of radius ε_n around p_0 must contain prior mass in order to satisfy (2.9), the rate of convergence ε_n of the posterior can certainly not be faster than $J^{-\alpha}$. The minimum distance of alternatives to allow appropriate tests, determined by (2.7), will be shown to satisfy $n\varepsilon_n^2 \geq J_n$. Together with the previous restriction on ε_n this will yield a rate of convergence of $n^{-\alpha/(2\alpha+1)}$, for $J_n \sim n^{1/(2\alpha+1)}$. This is also the rate of convergence of the sieved maximum likelihood estimator, found by Stone (1990). It is well known that this rate is optimal for α-smooth densities.

To make this precise we start with stating some lemmas that connect distances and norms on the densities p_θ with the J-dimensional Euclidean norm $\|\theta\|$ and infinity norm $\|\theta\|_\infty = \max_j |\theta_j|$. Let $\|f\|$ be the $L_2[0,1]$ norm of f and write $a \lesssim b$ if $a \leq Cb$ for a constant C that is universal or depends only on q (which is fixed throughout) and not on K. Most of these are known from or implicit in Stone (1986, 1990) or the literature on approximation theory.

LEMMA 4.2. *For any $\theta \in \mathbb{R}^J$,*

$$\|\theta\|_\infty \lesssim \|\theta^T B\|_\infty \leq \|\theta\|_\infty,$$
$$\|\theta\| \lesssim \sqrt{J}\, \|\theta^T B\| \lesssim \|\theta\|.$$

PROOF. The first inequality is proved by de Boor [(1978), page 156, Corollary 3]. The second is immediate from the fact that the B-spline basis forms a partition of unity. The third and fourth inequalities are stated in Stone [(1986), equation (12)]. As their full proofs are not in one place, we sketch the argument for completeness.

Let I_i be the interval $\big[(i-q)/K \vee 0,\ \ i/K \wedge 1\big]$. By (2) on page 155 of de Boor (1978), we have

$$\sum_i \theta_i^2 \lesssim \sum_i \|\theta^T B_{|I_i}\|_\infty^2 \lesssim \sum_i K \|\theta^T B_{|I_i}\|^2.$$

The last inequality follows, because $\theta^T B_{|I_i}$ consists of at most q polynomial pieces, each on an interval of length $1/K$, and the supremum norm of a polynomial of order q on an interval of length L is bounded by $1/\sqrt{L}$ times the L_2-norm, up to a constant depending on q. [To see the last: the squared $L_2[0,1]$-norm of the polynomial $x \mapsto \sum_{j=0}^{q-1} \alpha_j x^j$ on $[0,1]$ is the quadratic form $\alpha^T \mathrm{E} U_q U_q^T \alpha$ for $U_q = (1, U, \ldots, U^{q-1})$ and U a uniform $[0,1]$ variable. The second moment matrix $\mathrm{E} U_q U_q^T$ is nonsingular and hence the quadratic form is bounded below by a constant times $\|\alpha\|^2 \geq \|\alpha\|_\infty^2$.) This yields the third inequality.

By property (3) of the B-spline basis at most q elements $B_j(x)$ are nonzero for every given x, say for $j \in J(x)$. Therefore,

$$\big(\theta^T B(x)\big)^2 = \Big(\sum_{j\in J(x)} \theta_j B_j(x)\Big)^2 \le \sum_{j\in J(x)} \theta_j^2 B_j^2(x)\, q,$$

by the Cauchy–Schwarz inequality. Since each B_j is supported on an interval of length proportional to $1/J$ and takes its values in $[0,1]$, its $L_2[0,1]$-norm is of the order $1/\sqrt{J}$. Combined with the preceding display this yields

$$\int_0^1 \big(\theta^T B(x)\big)^2\, dx \lesssim \frac{q}{J}\|\theta\|^2.$$

This yields the fourth inequality. □

LEMMA 4.3. *For any $\theta \in \mathbb{R}^J$ such that $\theta^T 1 = 0$,*

$$\|\theta\|_\infty \lesssim \|\log p_\theta\|_\infty \le 2\|\theta\|_\infty.$$

PROOF. By the second inequality in Lemma 4.2 we have that $\|\theta^T B\|_\infty \le \|\theta\|_\infty$, whence $e^{c(\theta)}$ is contained in the interval $[e^{-M}, e^M]$ for $M = \|\theta\|_\infty$, by its definition, so that $|c(\theta)| \le \|\theta\|_\infty$. Consequently, by the triangle inequality

$$\|\log p_\theta\|_\infty = \|\theta^T B - c(\theta)\|_\infty \le 2\|\theta\|_\infty.$$

This yields the inequality on the right.

For the inequality on the left, we note that, since $\theta^T 1 = 0$,

$$\begin{aligned} |c(\theta)| &= |(\theta - c(\theta)1)^T 1|\frac{1}{J} \le \|\theta - c(\theta)1\|_\infty \|1\|_1 \frac{1}{J} \\ &\lesssim \|(\theta - c(\theta)1)^T B\|_\infty = \|\log p_\theta\|_\infty, \end{aligned}$$

by Lemma 4.2. Consequently, by Lemma 4.2 and the triangle inequality,

$$\|\theta\|_\infty \lesssim \|\theta^T B\|_\infty \le \|\theta^T B - c(\theta)\|_\infty + |c(\theta)| \lesssim 2\|\log p_\theta\|_\infty.$$

This concludes the proof. □

As a consequence of the preceding lemma, a set of densities p_θ is uniformly bounded away from 0 and ∞ if and only if the norms $\|\theta\|_\infty$ of the corresponding set of θ are bounded. This is true uniformly in $J \in \mathbb{N}$.

LEMMA 4.4. *For every θ_1, θ_2 such that $1^T(\theta_1 - \theta_2) = 0$,*

$$\inf_{x,\,\theta} p_\theta(x)\Big(\frac{\|\theta_1 - \theta_2\|^2}{J} \wedge 1\Big) \lesssim h^2(P_{\theta_1}, P_{\theta_2}) \lesssim \sup_{x,\,\theta} p_\theta(x)\Big(\frac{\|\theta_1 - \theta_2\|^2}{J}\Big),$$

where the infimum and supremum are taken over all θ on the line segment between θ_1 and θ_2 and all $x \in [0,1]$.

PROOF. By direct calculation and Taylor's theorem, we have

$$h^2(P_{\theta_1}, P_{\theta_2}) = 2\Big(1 - \exp\Big[c\Big(\tfrac{1}{2}\theta_1 + \tfrac{1}{2}\theta_2\Big) - \tfrac{1}{2}c(\theta_1) - \tfrac{1}{2}c(\theta_2)\Big]\Big)$$
$$= 2\Big(1 - \exp\big[-(\tfrac{1}{16})(\theta_1 - \theta_2)^T\big(\ddot{c}(\tilde{\theta}) + \ddot{c}(\tilde{\tilde{\theta}})\big)(\theta_1 - \theta_2)\big]\Big),$$

for $\tilde{\theta}$ and $\tilde{\tilde{\theta}}$ vectors on the line segment between θ_1 and θ_2 and $\ddot{c}(\theta)$ the Hessian of c. By the well-known properties of exponential families, we have

$$\tau^T \ddot{c}(\theta)\tau = \operatorname{var}_\theta \tau^T B = \inf_{\mu\in\mathbb{R}} \int_0^1 \big((\tau(x) - \mu 1)^T B(x)\big)^2 p_\theta(x)\,dx,$$

since $1^T B \equiv 1$. Up to bounds below and above on p_θ the right side is equivalent to the infimum over μ of the squared $L_2[0,1]$-norm of $(\tau - \mu 1)^T B$. By Lemma 4.2 the latter is comparable to the infimum over μ of $\|\tau - \mu 1\|^2/J$, which is equal to $\|\tau\|^2/J$ if $1^T\tau = 0$.

We can finish the proof by applying this in the first display, together with the inequalities $1 - e^{-x} \le x$ for $x \ge 0$ and $1 - e^{-x} \ge \frac{1}{2}(x \wedge 1)$ for $x \ge 0$ and $(cx) \wedge 1 \ge c(x \wedge 1)$ for $x \ge 0$ and $c \le 1$. □

By combining these lemmas we see that the Hellinger distance $h(P_{\theta_1}, P_{\theta_2})$ and $1/\sqrt{J}$ times the J-dimensional Euclidean distance $\|\theta_1 - \theta_2\|$ are proportional, uniformly in J and in θ_1, θ_2 having uniformly bounded coordinates. This combined with the estimate on the distance of p_0 to the set of p_θ given by Lemma 4.1 reduces the verification of (2.7) and (2.9) to calculations in the Euclidean setting.

We are now ready to prove the following theorem. By Lemma 4.3 there exists a constant d such that $d\|\theta\|_\infty \le \|\log p_\theta\|_\infty$ for every $\theta \in \mathbb{R}^J$ with $\theta^T 1 = 0$ and every $J \in \mathbb{N}$. We shall assume that the prior is chosen as roughly uniform on a large box $[-M, M]^J$. This corresponds to densities p_θ that are bounded and bounded away from zero by at least a small constant.

THEOREM 4.5. *Suppose that Π_n has a density with respect to Lebesgue measure on $\{\theta \in \mathbb{R}^{J_n}\colon \theta^T 1 = 0\}$ for $J_n \sim n^{1/(2\alpha+1)}$ whose minimum and maximal values on $[-M, M]^{J_n}$ are bounded below and above by terms of the orders c^{J_n} and C^{J_n}, respectively, for positive constants c, C, and which vanishes outside $[-M, M]^{J_n}$. Let $M \ge 1$. Then for every $p_0 \in C^\alpha[0,1]$ for $q \ge \alpha \ge 1/2$ such that $\|\log p_0\|_\infty \le \frac{1}{2}dM$ the conditions of Theorem 2.4 are satisfied for ε_n a large multiple of $n^{-\alpha/(2\alpha+1)}$ and $\mathscr{P}_n$ the support of Π_n, and hence the posterior rate of convergence is $n^{-\alpha/(2\alpha+1)}$.*

PROOF. Let θ_0 minimize $\theta \mapsto \|\log p_\theta - \log p_0\|_\infty$ over $\theta \in \mathbb{R}^J$ such that $\theta^T 1 = 0$. We first show that, for constants C_1, C depending on p_0, α and q only,

$$h(p_{\theta_0}, p_0) \le C_1\|\log p_{\theta_0} - \log p_0\|_\infty \le CJ^{-\alpha}. \tag{4.1}$$

By Lemma 4.1 there exists θ^* such that $\|(\theta^*)^T B - \log p_0\|_\infty \lesssim J^{-\alpha}$. Taking exponentials we see that this implies that $\|\exp(\theta^*)^T B - p_0\|_\infty \lesssim J^{-\alpha}$, and next, by integrating this inequality, that $|\exp c(\theta^*) - 1| \lesssim J^{-\alpha}$, whence $|c(\theta^*)| \lesssim J^{-\alpha}$. Consequently, $\|\log p_{\theta^*} - \log p_0\|_\infty \lesssim J^{-\alpha}$, whence θ^* minimizes $\theta \mapsto \|\log p_\theta - \log p_0\|_\infty$ up to a multiple of $J^{-\alpha}$. Since the set of p_θ is the same whether θ is restricted to satisfy $\theta^T 1 = 0$ or not, the second inequality in the display follows by the definition of θ_0. The first now follows easily, since p_0 and hence p_{θ_0} is bounded away from zero and infinity.

Thus the Hellinger ball of radius ε around P_0 is contained in a multiple of the Hellinger ball of radius $\varepsilon + J^{-\alpha}$ around P_{θ_0}, whence by Lemma 4.4, for any $\varepsilon > 0$ and suitable constants A, B, C, since $\|\theta_0\|_\infty \le \frac{1}{2}M + o(1)$ by the assumption that $\|\log p_0\|_\infty \le \frac{1}{2}dM$,

$$\begin{aligned}
&\{P_\theta\colon h(P_\theta, P_0) \le \varepsilon,\ \|\theta\|_\infty \le M\} \\
&\quad \subset \{P_\theta\colon Ah(P_\theta, P_{\theta_0}) \le \varepsilon + J^{-\alpha},\ \|\theta\|_\infty \le M\} \\
&\quad \subset \Big\{P_\theta\colon B\frac{\|\theta - \theta_0\|}{\sqrt{J}} \wedge 1 \le \varepsilon + J^{-\alpha},\ \|\theta - \theta_0\| \le 2\sqrt{J}M,\ \|\theta\|_\infty \le M\Big\} \\
&\quad \subset \Big\{P_\theta\colon \|\theta - \theta_0\| \le C\sqrt{J}(\varepsilon + J^{-\alpha})M\Big\},
\end{aligned}$$

since $\{x\colon x \wedge 1 \le \varepsilon,\ x \le M\} \subset \{x\colon x \le \varepsilon M\}$ for $M \ge 1$. Hence, in view of Example 7.1 [or Pollard (1990), Lemma 4.1] and Lemma 4.4, for constants E, F,

$$\begin{aligned}
&D\Big(\frac{\varepsilon}{2}, \{P_\theta\colon h(P_\theta, P_0) \le 2\varepsilon,\ \|\theta\|_\infty \le M\}, h\Big) \\
&\quad \le D\Big(E\varepsilon\sqrt{J}, \{\theta\colon \|\theta - \theta_0\| \le 2C\sqrt{J}(\varepsilon + J^{-\alpha})M\}, \|\cdot\|\Big) \\
&\quad \le \Big(\frac{F\sqrt{J}(\varepsilon + J^{-\alpha})M}{\varepsilon\sqrt{J}}\Big)^J.
\end{aligned}$$

Therefore, we can verify (2.7) for $\mathscr{P}_n = \{P_\theta\colon \|\theta\|_\infty \le M\}$ and every ε_n such that

$$J_n \log\Big(1 + \frac{J_n^{-\alpha}}{\varepsilon_n}\Big) \lesssim n\varepsilon_n^2.$$

Next, we have, with vol_J the volume of the $(J-1)$-dimensional unit ball,

$$\begin{aligned}
&\Pi_n(P_\theta\colon h(P_\theta, P_0) \le 2j\varepsilon, \|\theta\|_\infty \le M) \\
&\quad \le \sup_{\|\theta\|_\infty \le M} \pi_n(\theta)\big(2C\sqrt{J}(j\varepsilon + J^{-\alpha})M\big)^J \mathrm{vol}_J.
\end{aligned}$$

By Lemma 4.3 and the assumption that p_0 is bounded, the norms $\|p_0/p_\theta\|_\infty$ are uniformly bounded over θ ranging over a set of bounded $\|\theta\|_\infty$. Therefore,

in view of Lemma 4.4 and (4.1), uniformly in $\|\theta\|_\infty \le M$,

$$h^2(p_\theta, p_0)\left\|\frac{p_0}{p_\theta}\right\|_\infty \lesssim h^2(p_\theta, p_{\theta_0}) + h^2(p_{\theta_0}, p_0) \lesssim \frac{\|\theta-\theta_0\|^2}{J} + J^{-2\alpha}.$$

We conclude that for ε_n bigger than a sufficiently large multiple of $J^{-\alpha}$,

$$\begin{aligned}
&\Pi_n\Big(P_\theta\colon h^2(p_\theta, p_0)\left\|\frac{p_0}{p_\theta}\right\|_\infty \lesssim \varepsilon_n^2\Big)\\
&\quad\ge \Pi_n\Big(\theta\colon \|\theta\|_\infty \le M,\ \|\theta-\theta_0\|/\sqrt{J} \le \varepsilon_n - J^{-\alpha}\Big)\\
&\quad\ge \inf_{\|\theta\|_\infty \le M} \pi_n(\theta)\mathrm{vol}\big\{\theta\colon \|\theta\|_\infty \le M, \|\theta-\theta_0\| \le \frac{1}{2}\varepsilon_n\sqrt{J}\big\}\\
&\quad= \inf_{\|\theta\|_\infty \le M} \pi_n(\theta)\left(\frac{1}{2}\varepsilon_n\sqrt{J}\right)^J \mathrm{vol}_J,
\end{aligned}$$

since $\|\theta\|_\infty \le \|\theta_0\|_\infty + \|\theta-\theta_0\|\sqrt{J} \le \|\theta_0\|_\infty + \frac{1}{2}\varepsilon_n\sqrt{J} \le M$ eventually, if $\|\theta-\theta_0\| \le \frac{1}{2}\varepsilon_n\sqrt{J}$. By assumption, the first term is of the order c^J. Thus condition (2.9) is satisfied if, for all sufficiently large j,

$$J\log j \lesssim n\varepsilon_n^2 j^2 \quad\text{and}\quad \varepsilon_n \gtrsim J^{-\alpha}.$$

This gives ε_n of the order $(1/n)^{\alpha/(2\alpha+1)}$ for J_n of the order $\varepsilon_n^{-1/\alpha}$. □

5. Finite-dimensional models. Although in this paper we are primarily interested in infinite-dimensional models, it is desirable to have a unified theory applicable to both finite- and infinite-dimensional models. In this section we show that Theorem 2.4 yields the right rate of convergence for finite-dimensional models.

Let $\{p_\theta\colon \theta \in \Theta\}$ be a family of densities parametrized by a Euclidean parameter θ running through a set $\Theta \subset \mathbb{R}^d$. Assume that for every $\theta, \theta_1, \theta_2 \in \Theta$ and some $\alpha > 0$,

$$\begin{aligned}
-P_{\theta_0}\log\frac{p_\theta}{p_{\theta_0}} &\lesssim \|\theta-\theta_0\|^{2\alpha},\\
P_{\theta_0}\Big(\log\frac{p_\theta}{p_{\theta_0}}\Big)^2 &\lesssim \|\theta-\theta_0\|^{2\alpha},\\
\|\theta_1-\theta_2\|^\alpha &\lesssim h(P_{\theta_1}, P_{\theta_2}) \lesssim \|\theta_1-\theta_2\|^\alpha.
\end{aligned}$$

Assume that the prior measure Π possesses a density that is uniformly bounded away from zero and infinity on Θ. In this situation the posterior rate of convergence is $1/\sqrt{n}$ relative to the Hellinger distance h. Under the assumptions, this translates into a $n^{1/(2\alpha)}$-rate of convergence of the posterior for θ in the Euclidean distance.

THEOREM 5.1. *Under the conditions listed previously and θ_0 interior to Θ, the conditions of Theorem 2.4 are satisfied for $\mathscr{P} = \mathscr{P}_n = \{P_\theta\colon \theta \in \Theta\}$, the Hellinger distance d and ε_n a sufficiently large multiple of $1/\sqrt{n}$.*

PROOF. The left side of condition (2.7) is seen to be bounded by a constant in Example 7.1 in the case that $\alpha = 1$. The case of general α is not different. It follows that (2.7) is satisfied for $\varepsilon_n = M/\sqrt{n}$ and sufficiently large M.

In order to verify (2.9) we calculate

$$\frac{\Pi_n\big(P_\theta\colon h(P_\theta, P_{\theta_0}) \le j\varepsilon_n\big)}{\Pi_n\big(P_\theta\colon -P_{\theta_0}\log(p_\theta/p_{\theta_0}) \le \varepsilon_n^2,\ P_{\theta_0}\big(\log(p_\theta/p_{\theta_0})\big)^2 \le \varepsilon_n^2\big)}$$
$$\le \frac{\Pi_n\big(\theta\colon \|\theta-\theta_0\| \le A(j\varepsilon_n)^{1/\alpha}\big)}{\Pi_n\big(\theta\colon \|\theta-\theta_0\| \le B\varepsilon_n^{1/\alpha}\big)} \le C\Big(\frac{A}{B}\Big)^d j^{d/\alpha},$$

for constants A, B defined by the conditions preceding the theorem, and a constant C depending on the prior density only. It follows that (2.9) is satisfied easily for $\varepsilon_n = M/\sqrt{n}$ and sufficiently large M. □

It may be noted that our conditions preclude unbounded parameter spaces Θ: we cannot have that the Hellinger distance is bounded below by a multiple of the Euclidean distance unless the latter is bounded, since the Hellinger distance is uniformly bounded above. This could be improved by replacing condition (2.7) by a testing condition. The lower bound on the Hellinger distance is used only to verify (2.7), which in turn is used only to ensure the existence of tests of θ_0 versus the complements of balls of ε around θ_0. For most classical parametric models such tests exist. In fact, existence of uniformly consistent tests of the outside of a compact neigborhood of θ_0 already implies existence of tests with exponential error probabilities (see Lemma 7.2), and this would be sufficient to reduce the problem to a bounded parameter set, to which the preceding theorem applies. Note that the conditions are very reasonable for Θ equal to a small neighborhood of θ_0. See also Le Cam (1973) and Le Cam and Yang (1990).

6. Priors based on Dirichlet processes. In this section we apply the general theorems to priors based on Dirichlet processes. A major difficulty is the computation of the prior mass, as in conditions (2.4) or (2.5). We present one such computation and expect that future papers will address more problems of this sort. We shall need an estimate of the probability of an L_1-ball under a Dirichlet distribution given by the following lemma.

LEMMA 6.1. *Let $(X_1, \ldots, X_N)$ be distributed according to the Dirichlet distribution on the N-simplex with parameters $(m; \alpha_1, \ldots, \alpha_N)$, where $A\varepsilon \le \alpha_i \le 1$ and $\sum_{i=1}^N \alpha_i = m$ for some constant A. Let $(x_{10}, \ldots, x_{N0})$ be any point on the N-simplex. There exist positive constants c and C depending only on A such that, for $\varepsilon \le 1/N$,*

$$\Pr\Big(\sum_{i=1}^N |X_i - x_{i0}| \le 2\varepsilon\Big) \ge C\exp\Big(-cN\log\frac{1}{\varepsilon}\Big). \tag{6.1}$$

PROOF. Find an index i such that $x_{i0} \geq 1/N$. By relabelling, we can assume that $i = N$. If $|x_i - x_{i0}| \leq \varepsilon^2$ for $i = 1, \ldots, N-1$, then

$$\sum_{i=1}^{N-1} x_i \leq 1 - x_{N0} + (N-1)\varepsilon^2 \leq (N-1)(\varepsilon^2 + 1/N) \leq 1 - \varepsilon^2 < 1.$$

Hence there exists $x = (x_1, \ldots, x_N)$ in the simplex with these first $N-1$ coordinates. Furthermore, $\sum_{i=1}^{N} |x_i - x_{i0}| \leq 2\sum_{i=1}^{N-1} |x_i - x_{i0}| \leq 2\varepsilon^2(N-1) \leq 2\varepsilon$. Therefore the probability on the left-hand side of (6.1) is bounded below by

$$\mathrm{P}\Big(|X_i - x_{i0}| \leq \varepsilon^2, i = 1, \ldots, N-1\Big)$$

$$\geq \frac{\Gamma(m)}{\prod_{i=1}^{N}\Gamma(\alpha_i)} \prod_{i=1}^{N-1} \int_{\max((x_{i0}-\varepsilon^2),\,0)}^{\min((x_{i0}+\varepsilon^2),\,1)} x_i^{\alpha_i - 1}\, dx_i.$$

We use here that $(1 - \sum_{i=1}^{N-1} x_i)^{\alpha_N - 1} \geq 1$, since $\alpha_N \leq 1$. Similarly, since $\alpha_i \leq 1$ for every i, we can lower bound the integrand by 1 and note that the interval of integration contains at least an interval of length ε^2. Since $\alpha\Gamma(\alpha) = \Gamma(\alpha+1) \leq 1$ for $0 < \alpha \leq 1$ we can bound the last display from below by

$$\Gamma(m)\varepsilon^{2(N-1)} \prod_{i=1}^{N} \alpha_i \geq \Gamma(A)\varepsilon^{2(N-1)}(A\varepsilon)^N \geq C \exp\left(-cN \log \frac{1}{\varepsilon}\right).$$

This concludes the proof. □

EXAMPLE 6.1 (Current status censoring). Let $Y_1, \ldots, Y_n$ be an i.i.d. sample from a distribution F and $C_1, \ldots, C_n$ be an independent i.i.d. sample from a distribution G, both on $(0, \infty)$. Suppose that we observe $X_i = (\Delta_i, C_i)$ for $i = 1, \ldots, n$, where $\Delta_i = 1\{Y_i \leq C_i\}$ and would like to estimate F. The density function p_F of X with respect to the product of counting measure on $\{0, 1\}$ and a dominating measure for G at (δ, c) is given by

$$p_F(\delta, c) = F(c)^{\delta}\big(1 - F(c)\big)^{1-\delta} g(c).$$

Since this factorizes in parts depending on F and G only, if we put a product prior on the pair (F, G) and next compute the posterior for F only, then the part involving G will cancel out. Therefore, it is equivalent to treat g as a known density and put no prior on g.

We assume that G is supported on some compact interval $[a, b]$ and that the true distribution F_0 is continuous and has support which extends to the left and the right of $[a, b]$. [Hence $F_0(a-) > 0$ and $F_0(b) < 1$.] As a prior measure on F we consider a Dirichlet prior with base measure α that has a positive, continuous density on a compact interval containing $[a, b]$. We shall show that the conditions of Theorem 2.2 are satisfied for ε_n a large multiple of $n^{-1/3}(\log n)^{1/3}$. This is very close to the optimal rate of convergence in this model, which is $n^{-1/3}$. We do not exclude the possibility that this small discrepancy is due to suboptimal estimates of the prior mass in the following, and not a deficit of Dirichlet priors. We note that the priors based on ε-nets

given in Section 3 do lead to a posterior rate of convergence of $n^{-1/3}$, as the bracketing entropy for this model is of the order $1/\varepsilon$.

Since the roots of the densities p_F are essentially pairs of two bounded monotone functions and the Hellinger distance is the L_2-distance between the root densities, the Hellinger entropy of the model $\{P_F\colon F \in \mathscr{F}\}$, where $\mathscr{F}$ is the set of all distribution functions on $(0, \infty)$, can be estimated by the estimate of the entropy of the space of uniformly bounded monotone functions. Thus it is of the order $1/\varepsilon$ [see Theorem 2.7.5 of van der Vaart and Wellner (1996)]. Therefore condition (2.2) is verified for ε_n equal or bigger than $n^{-1/3}$.

Under our conditions, F_0 is bounded away from 0 and 1 on the interval $[a, b]$ that contains all observation times C_i. Consequently, the quotients $p_{F_0}/p_F(x)$ are uniformly bounded away from zero and infinity, uniformly in F that are uniformly close to F_0 on the interval $[a, b]$. The squared Hellinger distance is equal to

$$\begin{aligned} h^2(P_F, P_{F_0}) &= \int \bigl|F^{1/2}(c) - F_0^{1/2}(c)\bigr|^2 \, dG(c) \\ &\quad + \int \bigl|(1 - F(c))^{1/2} - (1 - F_0(c))^{1/2}\bigr|^2 \, dG(c) \\ &\le C \sup_{c \in [a, b]} \bigl|F(c) - F_0(c)\bigr|^2, \end{aligned}$$

for a constant depending on F_0. Thus, to verify (2.5) it suffices to estimate the prior mass of a Kolmogorov–Smirnov ball of radius ε_n around F_0. Given $\varepsilon > 0$, partition the positive half line in intervals $E_1, \ldots, E_N$ such that $F_0(E_i) \le \varepsilon$ and $A\varepsilon \le \alpha(E_i) \le 1$ for every i and some fixed constant A. We can achieve this with $N = O(1/\varepsilon)$ intervals. By Lemma 6.1, the set $\bigl\{F\colon \sum_{i=1}^N |F(E_i) - F_0(E_i)| \le \varepsilon\bigr\}$ has probability of the order $\exp(-c(1/\varepsilon)\log(1/\varepsilon))$. For every F in this set, the Kolmogorov–Smirnov distance to F_0 is of the order ε. We conclude that the prior mass in a Hellinger ball of radius a large multiple of ε is of the order $\exp(-c(1/\varepsilon)\log(1/\varepsilon))$. Thus condition (2.5) is verified for ε_n a large multiple of $n^{-1/3}(\log n)^{1/3}$.

7. Existence of tests. In this section we consider some results on the existence of tests of P_0 versus the complement $\bigl\{P\colon d(P, P_0) > \varepsilon\bigr\}$ of the ball of radius ε around P_0. The existence of certain tests is a main element in the proofs of Theorems 2.1–2.3 and is guaranteed by entropy bounds. At the end of this section we state a theorem on the rate of convergence of posteriors directly in terms of tests.

Appropriate tests can be built up from tests of P_0 versus balls $\bigl\{P\colon d(P, P_1) \le \eta\bigr\}$ for given P_1. Throughout this section we use a distance d such that for every pair P_0 and P_1 in the model $\mathscr{P}$ there exist tests ϕ_n such that, for some universal constant K,

$$P_0^n \phi_n \le \exp\bigl(-Knd^2(P_0, P_1)\bigr), \tag{7.1}$$

$$\sup_{d(P, P_1) < d(P_0, P_1)/2} P^n(1 - \phi_n) \le \exp\bigl(-Knd^2(P_0, P_1)\bigr). \tag{7.2}$$

This is true both for d equal to the total variation distance and for d equal to the Hellinger distance. (The constant 2 has no particular interest and is not optimal; any constant bigger than 1 is possible and would do for our purposes.)

More generally, it is known from Birgé (1984) and Le Cam (1986) (see Lemma 4 on page 478) that given any two *convex* sets $\mathscr{P}_0$ and $\mathscr{P}_1$ of probability measures, there exist tests ϕ_n such that

$$\sup_{P\in\mathscr{P}_0} P^n\phi_n \le \exp\bigl(n\log\rho(\mathscr{P}_0,\mathscr{P}_1)\bigr), \tag{7.3}$$

$$\sup_{P\in\mathscr{P}_1} P^n(1-\phi_n) \le \exp\bigl(n\log\rho(\mathscr{P}_0,\mathscr{P}_1)\bigr), \tag{7.4}$$

where $\rho(\mathscr{P}_0,\mathscr{P}_1) = 1 - \frac{1}{2}h^2(\mathscr{P}_0,\mathscr{P}_1)$ is the *Hellinger affinity*, and $h(\mathscr{P}_0,\mathscr{P}_1)$ is the minimum of $h(P_0,P_1)$ over $P_0\in\mathscr{P}_0$ and $P_1\in\mathscr{P}_1$. Because $\log\rho \le -\frac{1}{2}h^2$ this gives exponential decrease of the error probabilities, with the exponent proportional to $-nh^2(\mathscr{P}_0,\mathscr{P}_1)$. This general result brings out the special role of the Hellinger distance (even though in some situations it may be preferable to work with the log Hellinger affinity directly).

If a distance d is bounded above by the Hellinger distance, then the ball $\{P\colon d(P,P_1) < d(P_0,P_1)/2\}$ is at Hellinger distance at least $d(P_0,P_1)/2$ from P_0. Thus if this ball is a convex set of probability measures, then (7.1) and (7.2) is satisfied for d (with $K=\frac{1}{2}$), by the general results of Birgé and Le Cam. This argument immediately gives (7.1) and (7.2) for the Hellinger distance itself and the total variation distance, which satisfies $\int|dP-dQ| \le 2h(P,Q)$, by the Cauchy–Schwarz inequality. If the set of probability densities under consideration is uniformly bounded, then it also gives (7.1) and (7.2) for the L_2-distance, because this is then also bounded by a multiple of the Hellinger distance.

The next step is to combine the tests for balls (which are convex) into a test for the complements of balls, which are nonconvex. The following result is related to Lemma 2.1 in Birgé (1983). The number $D(\varepsilon)$ in its first condition is related to the measure of metric dimension used by Birgé and Le Cam. The number $\sup_{\varepsilon\ge\varepsilon_n} D(\varepsilon)$ is almost identical to what Le Cam (1986) calls the *dimension of $\mathscr{P}$ for the pair (d,ε_n)*.

THEOREM 7.1. *Suppose that for some nonincreasing function $D(\varepsilon)$, some $\varepsilon_n \ge 0$ and every $\varepsilon > \varepsilon_n$,*

$$D\Bigl(\frac{\varepsilon}{2}, \bigl\{P\colon \varepsilon \le d(P,P_0) \le 2\varepsilon\bigr\}, d\Bigr) \le D(\varepsilon).$$

Then for every $\varepsilon > \varepsilon_n$ there exist tests ϕ_n (depending on $\varepsilon > 0$) such that, for a universal constant K and every $j\in\mathbb{N}$,

$$P_0^n\phi_n \le D(\varepsilon)\exp\bigl(-Kn\varepsilon^2\bigr)\frac{1}{1-\exp(-Kn\varepsilon^2)}, \tag{7.5}$$

$$\sup_{d(P,P_0)>j\varepsilon} P^n(1-\phi_n) \le \exp\bigl(-Kn\varepsilon^2 j^2\bigr). \tag{7.6}$$

PROOF. For a given $j \in \mathbb{N}$ choose a maximal $j\varepsilon/2$ separated set of points in $S_j = \{P\colon j\varepsilon < d(P, P_0) \le (j+1)\varepsilon\}$. This yields a set S_j' of at most $D(j\varepsilon)$ points and every $P \in S_j$ is within distance $j\varepsilon/2$ of at least one of these points. (Take S_j' empty and adapt the following in the obvious way if S_j is empty.) For every such point $P_1 \in S_j'$ there exists a test ω_n with the properties as in (7.1) and (7.2). Let ϕ_n be the maximum of all tests attached in this way to some point $P_1 \in S_j'$ for some $j \in \mathbb{N}$. Then

$$P_0^n \phi_n \le \sum_j \sum_{P_1 \in S_j'} \exp\bigl(-Knj^2\varepsilon^2\bigr) \le \sum_{j \in \mathbb{N}} D(j\varepsilon) \exp\bigl(-Knj^2\varepsilon^2\bigr),$$

$$\sup_{P \in \bigcup_{i \ge j} S_i} P^n(1-\phi_n) \le \sup_{i \ge j} \exp\bigl(-Kni^2\varepsilon^2\bigr).$$

The right sides can be further bounded as desired. [Note that $D(j\varepsilon) \le D(\varepsilon)$ for every $j \in \mathbb{N}$, by assumption.] □

One possible choice for $D(\varepsilon)$ is the ε-packing number $D(\varepsilon/2, \mathscr{P}, d)$. This is a bigger number, but in many infinite-dimensional situations this does not appear to yield a real loss. On the other hand, the theorem is needed as stated if $\mathscr{P}$ is finite-dimensional.

EXAMPLE 7.1. Suppose that $\mathscr{P} = \{P_\theta\colon \theta \in \Theta \subset \mathbb{R}^m\}$ and, for given constants A and B and $\|\cdot\|$ the m-dimensional Euclidean norm,

$$d(P_\theta, P_{\theta_0}) \ge A\|\theta - \theta_0\|,$$
$$d(P_{\theta_1}, P_{\theta_2}) \le B\|\theta_1 - \theta_2\|.$$

(Since both the Hellinger and total variation metric are bounded, the first can be true with d one of these distances only if Θ is bounded.) The ε-packing number of the m-dimensional unit ball is bounded above by $(6/\varepsilon)^m$ [e.g., Pollard (1990), Lemma 4.1]. Thus

$$D\Bigl(k\varepsilon, \bigl\{\theta \in \mathbb{R}^m\colon \|\theta - \theta_0\| \le l\varepsilon\bigr\}, \|\cdot\|\Bigr) \le \left(\frac{6l}{k}\right)^m.$$

It follows that

$$D\bigl(\varepsilon, \{P_\theta\colon d(P_\theta, P_{\theta_0}) \le 2\varepsilon\},\ d\bigr) \le D\left(\frac{\varepsilon}{B}, \left\{\theta\colon \|\theta - \theta_0\| \le \frac{2\varepsilon}{A}\right\}, \|\cdot\|\right) \le \left(\frac{12B}{A}\right)^m.$$

Thus we can take $D(\varepsilon)$ in Theorem 7.1 independent of ε, but increasing exponentially with the dimension (if A/B is fixed). In comparison, the numbers $D(\varepsilon, \mathscr{P}, h)$ are of the order ε^{-m}.

It is known from Le Cam (1973) that even for a fixed ε there need not exist a consistent sequence of tests of P_0 versus $\{P \in \mathscr{P}\colon d(P, P_0) > \varepsilon\}$. The preceding theorem shows that total boundedness of $\mathscr{P}$ [which is equivalent to $D(\varepsilon, \mathscr{P}, d)$ being finite for every $\varepsilon > 0$] is sufficient for the existence of such a test. However, this is not necessary. One example showing this is given by

(7.1) and (7.2) applied with $\mathscr{P} = \{P_0\} \bigcup \{P\colon d(P, P_1) < d(P_0, P_1)/2\}$, because a total variation or Hellinger ball is usually not totally bounded. A classical example is as follows.

EXAMPLE 7.2. The collection of all normal distributions $N(\theta, 1)$ on $\mathbb{R}$ is not totally bounded for the Hellinger or total variation distances, but certainly there are good tests of $H_0\colon \theta = 0$ versus $H_1\colon |\theta| > \varepsilon$. [Actually, in this case the affinity satisfies $\log \rho\big(N(\theta, 1), N(\theta', 1)\big) = -(\theta - \theta')^2/8$ and hence we could apply the general form of (7.3) and (7.4) in combination with the Euclidean distance to obtain good tests through the approach of Theorem 7.1, even for unbounded alternatives. For other parametric models the log affinity is typically not nicely related to the Euclidean distance and this approach would fail.]

On the other hand, if a fixed part of $\mathscr{P}$ can be uniformly consistently tested versus P_0, then it can also be tested with exponentially small error probabilities. This implies that such a fixed part can be ignored for our purposes, in that it is not a loss of generality in the main result to assume that the prior only charges the remaining part of $\mathscr{P}$ (and P_0). The error probabilities of the tests ϕ_n given in the following lemma are of smaller order than the error probabilities of the tests in Theorem 7.1 if $\varepsilon = \varepsilon_n \to 0$ in the latter theorem. This can be useful to reduce the model $\mathscr{P}$ to a totally bounded submodel, by trimming away parts that can easily be tested by ad hoc arguments. The following lemma is a consequence of results of Le Cam (1973).

LEMMA 7.2. *Suppose that there exist tests ω_n such that for fixed sets $\mathscr{P}_0$ and $\mathscr{P}_1$ of probability measures*

$$\sup_{P_0 \in \mathscr{P}_0} P_0^n \omega_n \to 0, \qquad \sup_{P \in \mathscr{P}_1} P^n(1 - \omega_n) \to 0.$$

Then there exist tests ϕ_n and constants $K > 0$ such that

$$\sup_{P_0 \in \mathscr{P}_0} P_0^n \phi_n \le e^{-Kn}, \qquad \sup_{P \in \mathscr{P}_1} P^n(1 - \omega_n) \le e^{-Kn}.$$

In view of the fact that, apparently, entropy conditions are not always appropriate to ensure the existence of tests, it is fruitful to formulate a theorem on rates of convergence directly in terms of existence of tests. The following is a result of this type.

THEOREM 7.3. *Suppose that* (2.8) *and* (2.9) *hold for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2$ bounded away from zero and sets $\mathscr{P}_n \subset \mathscr{P}$, and in addition suppose that their exists a sequence of tests ϕ_n such that for some constant $K > 0$ and for every sufficiently large j,*

$$P_0^n \phi_n \to 0, \tag{7.7}$$

$$\sup_{P \in \mathscr{P}_n\colon\, \varepsilon_n j < d(P, P_0) \le 2j\varepsilon_n} P^n(1 - \phi_n) \le \exp\big(-Kn\varepsilon_n^2 j^2\big). \tag{7.8}$$

Then for any $M_n \to \infty$, we have that $\Pi_n(P\colon d(P, P_0) \geq M_n\varepsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability.

8. Proof of Theorems 2.1–2.3. In the proof of Theorem 2.1 we use the following simple lemma, which will need to be replaced by more complicated results for the proofs of Theorems 2.2 and 2.3.

LEMMA 8.1. *For every $\varepsilon > 0$ and probability measure Π on the set*

$$\{P\colon -P_0 \log(p/p_0) \leq \varepsilon^2, P_0(\log(p/p_0))^2 \leq \varepsilon^2\}$$

we have, for every $C > 0$,

$$P_0^n\left(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \leq \exp\bigl(-(1+C)n\varepsilon^2\bigr)\right) \leq \frac{1}{C^2 n\varepsilon^2}.$$

PROOF. By Jensen's inequality applied to the logarithm,

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

Thus the probability is bounded by, with $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ the empirical process,

$$P_0^n\left(\mathbb{G}_n \int \log \frac{p}{p_0}\, d\Pi(P) \leq -\sqrt{n}(1+C)\varepsilon^2 - \sqrt{n}P_0 \int \log \frac{p}{p_0}\, d\Pi(P)\right).$$

By Fubini's theorem and the assumption on Π the expression on the right of the inequality sign is bounded by $-\sqrt{n}\varepsilon^2 C$. An application of Chebyshev's inequality yields the upper bound

$$\frac{\operatorname{var} \int \log(p/p_0)(X_1)\, d\Pi(P)}{C^2 n\varepsilon^4} \leq \frac{P_0 \int (\log(p/p_0))^2\, d\Pi(P)}{C^2 n\varepsilon^4}.$$

by another application of Jensen's inequality. The right side is bounded by $(C^2 n\varepsilon^2)^{-1}$ by the assumption on Π. This concludes the proof. □

PROOF OF THEOREM 2.1. For every $\varepsilon > 2\varepsilon_n$ we have by (2.2),

$$\log D\left(\frac{\varepsilon}{2}, \mathscr{P}_n, d\right) \leq \log D(\varepsilon_n, \mathscr{P}_n, d) \leq n\varepsilon_n^2.$$

Therefore, by Theorem 7.1, applied with $D(\varepsilon) = \exp(n\varepsilon_n^2)$ (constant in ε) and $\varepsilon = M\varepsilon_n$ and $j = 1$ in its assertion, where $M \geq 2$ is a large constant to be chosen later, there exist tests ϕ_n that satisfy

$$P_0^n \phi_n \leq \exp\bigl(n\varepsilon_n^2\bigr) \times \exp\bigl(-KnM^2\varepsilon_n^2\bigr) \frac{1}{1 - \exp\bigl(-KnM^2\varepsilon_n^2\bigr)}, \tag{8.1}$$

$$\sup_{P \in \mathscr{P}_n\colon\, d(P, P_0) > M\varepsilon_n} P^n(1 - \phi_n) \leq \exp\bigl(-KnM^2\varepsilon_n^2\bigr). \tag{8.2}$$

By the first condition (8.1) it follows that, if $KM^2 - 1 > K$, as $n \to \infty$,

$$\mathrm{E}_{P_0}\Pi_n\bigl(P\colon d(P, P_0) \geq \varepsilon_n | X_1, \ldots, X_n\bigr)\phi_n \leq P_0^n\phi_n \leq 2e^{-Kn\varepsilon_n^2}. \tag{8.3}$$

By Fubini's theorem and the fact that $P_0(p/p_0) \leq 1$,

$$\mathrm{E}_{P_0}\int_{\mathscr{P}\ \mathscr{P}_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(P) \leq \Pi_n(\mathscr{P} - \mathscr{P}_n).$$

Combining the above assertion with (8.2) we see that

$$\begin{aligned}
&\mathrm{E}_{P_0}\int_{P\colon\ d(P, P_0) > M\varepsilon_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(P)(1-\phi_n)\\
&\quad\leq \Pi_n(\mathscr{P}\ \mathscr{P}_n) + \int_{P\in\mathscr{P}_n\colon\ d(P, P_0) > M\varepsilon_n} P^n(1-\phi_n)\,d\Pi_n(P)\\
&\quad\leq \Pi_n(\mathscr{P}\ \mathscr{P}_n) + \exp(-KnM^2\varepsilon_n^2) \leq 2\exp(-n\varepsilon_n^2(C+4)),
\end{aligned}$$

for $M \geq \sqrt{(C+4)/K}$. By Lemma 8.1, we have with probability tending to 1, with $B_n = \bigl\{P\colon\ -P_0\log(p/p_0) \leq \varepsilon_n^2,\ P_0(\log(p/p_0))^2 \leq \varepsilon_n^2\bigr\}$,

$$\int\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(P) \geq \exp\bigl(-2n\varepsilon_n^2\bigr)\Pi_n(B_n) \geq \exp\bigl(-n\varepsilon_n^2(2+C)\bigr),$$

by assumption (2.4). If A_n is the event that this inequality is true, so that $P_0^n(A_n) \to 1$, then it follows that

$$\begin{aligned}
&\mathrm{E}_{P_0}\Pi_n\bigl(P\colon d(P, P_0) > M\varepsilon_n | X_1, \ldots, X_n\bigr)(1-\phi_n)1_{A_n}\\
&\qquad\leq \exp\bigl(n\varepsilon_n^2(2+C)\bigr)2\exp\bigl(-n\varepsilon_n^2(C+4)\bigr) \to 0.
\end{aligned}$$

This concludes the proof. □

For the proof of Theorem 2.2 we need a replacement of Lemma 8.1 that gives a faster rate of convergence in its statement. We can achieve this by controlling the quotients p/p_0. First, if one has uniform control from below, then the Hellinger distance and the Kullback–Leibler information are comparable. The following lemma can be found in Birgé and Massart (1998) [see their (7.6)].

LEMMA 8.2. *For any pair of probability measures P and P_0,*

$$h^2(P, P_0) \leq -P_0\log\frac{p}{p_0} \leq 2h^2(P, P_0)\left[1 + \log\left\|\frac{p_0}{p}\right\|_\infty\right] \leq 2h^2(P, P_0)\left\|\frac{p_0}{p}\right\|_\infty.$$

A second lemma is a comparison of a certain exponential moment and the Hellinger distance. This exponential moment [called the "Bernstein norm" in van der Vaart and Wellner (1996) even though it is not a norm] is essential in Bernstein's inequality. Birgé and Massart (1993) used this "norm" to derive results on rates of convergence of minimum contrast estimators.

LEMMA 8.3. *For any pair of probability measures P and P_0,*

$$P_0\big(\exp(|\log(p/p_0)|) - 1 - |\log(p/p_0)|\big) \le 2h^2(P, P_0)\left\|\frac{p_0}{p}\right\|_\infty. \tag{8.4}$$

PROOF. For every $c \le 0$ and $x \ge c$ we have the inequality

$$\Big(e^{|x|} - 1 - |x|\Big) \le 2e^{|c|}\Big(e^{x/2} - 1\Big)^2. \tag{8.5}$$

If $e^{-c} = \|p_0/p\|_\infty$, then $\log(p/p_0) \ge c$ and hence the integrand on the left side of (8.4) is bounded above by

$$2e^{-c}\left(\exp\left(\frac{1}{2}\log(p/p_0)\right) - 1\right)^2 = 2e^{-c}\left(\sqrt{\frac{p}{p_0}} - 1\right)^2.$$

The integral of the right side with respect to P_0 is equal to $2e^{-c}$ times the squared Hellinger distance. □

The "Bernstein norm" of $\log(p/p_0)$ dominates all moments of order greater than or equal to 2 of $\log(p/p_0)$ up to constants, including the second moment up to a factor 2. Therefore, when combined the preceding two lemmas show that

$$\begin{aligned}&\left\{P\colon h^2(P, P_0)\left\|\tfrac{p_0}{p}\right\|_\infty \le \varepsilon^2\right\}\\ &\qquad\subset \left\{P\colon P_0\log\tfrac{p_0}{p} \le 2\varepsilon^2,\ P_0\left(\log\tfrac{p}{p_0}\right)^2 \le 4\varepsilon^2\right\}.\end{aligned} \tag{8.6}$$

This shows that condition (2.4) is weaker than condition (2.5), up to constants. Actually controlling all moments is more than what is needed. Another possible extension of Lemma 8.1 would be to replace the second moment of $\log(p/p_0)$ by a higher moment (and use Markov's inequality at the end of the proof). This would give a result good enough for the proof of Theorem 2.2 provided the higher moment is chosen "high enough" (dependent on the order ε_n, faster convergence to zero needing a higher moment). We have chosen here to forego such refinements and obtain an exponential inequality under a somewhat stronger assumption.

We are ready for an adaptation of Lemma 8.1

LEMMA 8.4. *For every $\varepsilon > 0$ and probability measure Π on the set*

$$\big\{P\colon h^2(P, P_0)\big\|p_0/p\big\|_\infty \le \varepsilon^2\big\}, \tag{8.7}$$

we have, for a universal constant $B > 0$,

$$P_0^n\left(\int\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(P) \le \exp(-3n\varepsilon^2)\right) \le \exp(-Bn\varepsilon^2). \tag{8.8}$$

PROOF. Lemma 8.2 gives that $-P_0 \log p/p_0 \leq 2\varepsilon^2$ for every P in the set (8.7), which has Π-probability 1. Furthermore, by Lemma 8.3,

$$P_0(\exp|\log(p/p_0)| - 1 - |\log(p/p_0)|) \leq 2\varepsilon^2.$$

By monotonicity and convexity of the function $y \mapsto e^y - 1 - y$ on $[0, \infty)$ and Jensen's inequality,

$$P_0\left(\exp\left(\left|\int \log(p/p_0)\, d\Pi(P)\right|\right) - 1 - \left|\int \log(p/p_0)\, d\Pi(P)\right|\right)$$
$$\leq P_0 \int (\exp(|\log(p/p_0)|) - 1 - |\log(p/p_0)|)\, d\Pi(P) \leq 2\varepsilon^2,$$

by Fubini's theorem. By the lemma below the same bound is true for $\frac{1}{2}$ times the variable $\int \log(p/p_0)\, d\Pi(P)$ centered at its expectation. Therefore, rewriting the probability on the left side of (8.8) as in the proof of Lemma 8.1, we see that it is bounded above by

$$P_0^n\left(\mathbb{G}_n \int \left(\log \frac{p}{p_0}\right) d\Pi(P) \leq -3\sqrt{n}\varepsilon^2 + \sqrt{n}2\varepsilon^2\right) \leq \exp\left(-D\frac{n\varepsilon^4}{\varepsilon^2 + \sqrt{n}\varepsilon^2/\sqrt{n}}\right),$$

by (the refined version of) Bernstein's inequality. [see, e.g., Lemma 2.2.11 of van der Vaart and Wellner (1996).] □

LEMMA 8.5. *If* ψ: $[0, \infty) \to \mathbb{R}$ *is convex and nondecreasing, then* $\mathrm{E}\psi(|X - \mathrm{E}X|) \leq \mathrm{E}\psi(2|X|)$ *for every random variable* X.

PROOF. The map $y \mapsto \psi(|y|)$ is convex on $\mathbb{R}$. If X' is an independent copy of X, then the left side is equal to $\mathrm{E}\psi(|X - \mathrm{E}X'|) \leq \mathrm{E}\psi(|X - X'|)$, by Jensen's inequality. Next bound $|X - X'| \leq |X| + |X'|$ and use the monotonicity and convexity of ψ again to bound the expectation by $\mathrm{E}\frac{1}{2}(\psi(2|X|) + \psi(2|X'|))$, which is the right side. □

PROOF OF THEOREM 2.2. The proof of Theorem 2.2 follows the same lines as the proof of Theorem 2.1. The difference is that we use Lemma 8.4 instead of Lemma 8.1 to ensure that the probability of the events A_n converges to 1 at an exponential rate. By inspecting the proof, we conclude that for some $B_1, B_2 > 0$ and M chosen as before,

$$P_0\big(\Pi_n(P\colon d(P, P_0) > M\varepsilon_n | X_1, \ldots, X_n) \geq \exp(-B_1 n\varepsilon_n^2)\big)$$

converges to zero at the rate $\exp(-B_2 n\varepsilon_n^2)$. Since $\sum_n \exp(-B_2 n\varepsilon_n^2) < \infty$, almost sure convergence follows by the Borel–Cantelli lemma. □

For the proof of Theorem 2.3 we need other variations on the preceding lemmas. The following lemma follows from Theorem 5 of Wong and Shen (1995). Let $\log_+ x = (\log x) \vee 0$.

LEMMA 8.6. *For any pair of probability measures P and P_0 such that $h(P, P_0) \le 0.44$ and $P_0(p_0/p) < \infty$,*

$$-P_0 \log \frac{p}{p_0} \le 18h^2(P, P_0)\left(1 + \log_+ \frac{\sqrt{P_0(p_0/p)}}{h(P, P_0)}\right),$$

$$P_0\left(\log \frac{p}{p_0}\right)^2 \le 5h^2(P, P_0)\left(1 + \log_+ \frac{\sqrt{P_0(p_0/p)}}{h(P, P_0)}\right)^2.$$

LEMMA 8.7. *For any pair of probability measures P and P_0 such that $P_0(p_0/p) < \infty$,*

$$P_0(\exp(|\log(p/p_0)|) - 1 - |\log(p/p_0)|) \le 4h^2(P, P_0)(1 + \Phi^{-1}(h^2(P, P_0))).$$

for $\Phi^{-1}(\varepsilon) = \sup\{M\colon \Phi(M) \ge \varepsilon\}$ the inverse of the function $\Phi(M) = P_0(p_0/p)1\{p_0/p \ge M\}/M$.

PROOF. Set $m = p_0/p$. By inequality (8.5) in the proof of Lemma 8.3, the left side is bounded above by

$$2P_0\left(\sqrt{\frac{p}{p_0}} - 1\right)^2 1\{p \ge p_0\} + 2P_0\frac{p_0}{p}\left(\sqrt{\frac{p}{p_0}} - 1\right)^2 1\{p < p_0\}$$
$$\le 2h^2(P, P_0)(1 + M) + 2P_0 m 1\{m > M\},$$

for every $M > 0$. The function Φ is left continuous and strictly decreasing from infinity at 0 to 0 at a point $\tau \le \infty$. If we choose $M = \Phi^{-1}(h^2(P, P_0))$, then $M\Phi(M) \ge Mh^2(P, P_0) \ge M\Phi(M+) = P_0 m 1\{m > M\}$. The right side of the last display can now be bounded by an expression as in the lemma. □

LEMMA 8.8. *For a given function m let $\Phi^{-1}(\varepsilon) = \sup\{M\colon \Phi(M) \ge \varepsilon\}$ be the inverse function of $\Phi(M) = P_0 m 1\{m \ge M\}/M$. For every $\varepsilon \in (0, 0.44)$ and probability measure Π on the set*

$$\left\{P\colon p_0/p \le m, 18h^2(P, P_0)\left(1 + \log_+ \frac{\sqrt{P_0 m}}{h(P, P_0)} + \Phi^{-1}\big(h^2(P, P_0)\big)\right) \le \varepsilon^2\right\}$$

we have, for a universal constant $B > 0$,

$$P_0^n\left(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \le \exp(-2n\varepsilon^2)\right) \le \exp(-Bn\varepsilon^2). \tag{8.9}$$

PROOF. This follows the same lines as the proof of Lemma 8.4, now substituting Lemmas 8.6 and 8.7 for Lemmas 8.2–8.3.

PROOF OF THEOREM 2.3. This is identical to the proof of Theorem 2.2, except that we use Lemma 8.8 instead of Lemma 8.4.

PROOF OF THEOREM 2.4. The first part of the proof is identical to the first part of the proof of Theorem 2.1, except that we choose the tests ϕ_n to satisfy (8.1) and [instead of 8.2] for every $j \in \mathbb{N}$,

$$\sup_{P \in \mathscr{P}_n:\ d(P, P_0) > M\varepsilon_n j} P^n(1 - \phi_n) \le \exp(-KnM^2\varepsilon_n^2 j^2). \tag{8.10}$$

We also choose M large enough to ensure that the right side of (8.1) and hence the left side of (8.3) converges to zero. Defining $S_{n,j} = \{P \in \mathscr{P}_n:\ M\varepsilon_n j < d(P, P_0) \le M\varepsilon_n(j+1)\}$ and using (8.10), we obtain

$$\mathrm{E}_{P_0} \int_{S_{n,j}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(P)(1 - \phi_n) \le \exp(-KnM^2\varepsilon_n^2 j^2)\Pi_n(S_{n,j}).$$

Fix some $C_0 \ge 1$. By Lemma 8.1, we have on an event A_n with probability at least $1 - (n\varepsilon_n^2 C_0^2)^{-1}$,

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(P) \ge \exp\bigl(-2C_0 n\varepsilon_n^2\bigr)\Pi_n\bigl(B_n(\varepsilon_n)\bigr).$$

Hence, by assumption (2.9), for every sufficiently large J,

$$\begin{aligned}
&\mathrm{E}_{P_0}\Pi_n\bigl(P \in \mathscr{P}_n:\ d(P, P_0) > J\varepsilon_n M | X_1, \ldots, X_n\bigr)(1 - \phi_n)1_{A_n} \\
&\quad \le \sum_{j \ge J} \frac{\exp\bigl(-KnM^2\varepsilon_n^2 j^2\bigr)\Pi_n(S_{n,j})}{\exp\bigl(-2C_0 n\varepsilon_n^2\bigr)\Pi_n\bigl(B_n(\varepsilon_n)\bigr)} \\
&\quad \le \sum_{j \ge J} \exp\bigl(-n\varepsilon_n^2(KM^2 j^2 - 2C_0 - \tfrac{1}{2}KM^2 j^2)\bigr).
\end{aligned}$$

This converges to zero as $J \to \infty$ if $n\varepsilon_n^2$ is bounded away from zero. Next

$$\mathrm{E}_{P_0}\Pi_n(P \notin \mathscr{P}_n | X_1, \ldots, X_n)(1 - \phi_n)1_{A_n} \le \frac{\Pi_n(\mathscr{P} \setminus \mathscr{P}_n)}{\exp(-2C_0 n\varepsilon_n^2\Pi_n)\bigl(B_n(\varepsilon_n)\bigr)}.$$

We may assume that either $n\varepsilon_n^2$ is bounded or $n\varepsilon_n^2 \to \infty$; otherwise we argue along subsequences. If $n\varepsilon_n^2$ is bounded, then we first choose C_0 large but fixed so as to make $P_0^n(A_n)$ as large as desired. Then the right side of the preceding display converges to zero by assumption (2.8). If $n\varepsilon_n^2 \to \infty$, then we choose $C_0 = 1$, in which case $P_0^n(A_n) \to 1$ and again the right side of the preceding display converges to zero. □

PROOF OF THEOREM 7.3. This is essentially contained in the proof of Theorem 2.4 (take $M = 1$).

Acknowledgment. We thank Lucien Birgé for insightful discussions that have led to an improved presentation (and some corrections), in particular relating to Section 7.

REFERENCES

BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–238.

BIRGÉ, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* (G. Yang and D. Pollard, eds.) 55–87. Springer, New York.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

DOOB, J. L. (1949). Le Calcul des Probabilités et ses Applications. *Coll. Int. du CNRS* **13** 23–27.

DUDLEY, R. M. (1984). A course on empirical processes. *Lectures Notes in Math.* **1097** 2–141. Springer, Berlin.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distribution on the spaces of probability measures. *Ann. Statist.* **2** 615–629.

FREEDMAN, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1194–1216.

FREEDMAN, D. A. (1965). On the asymptotic behavior of Bayes' estimates in the discrete case II. *Ann. Math. Statist.* **36** 454–456.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1997). Non-informative priors via sieves and packing numbers. In *Advances in Statistical Decision Theory and Applications* (S. Panchapakeshan and N. Balakrishnan eds.) 129–140. Birkhäuser, Boston.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI R. V. (1999a). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999b). Consistency issues in Bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri* (Subir Ghosh, ed.) 639–667. Dekker, New York.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1961). Epsilon-entropy and epsilon-capacity of sets in function spaces. *Amer. Math. Soc. Trans. Ser. 2* **17** 277–364.

LE CAM, L. M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

LE CAM, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

LE CAM, L. M. and YANG, G. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York.

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. IMS, Hayward, CA and Amer. Statist. Assoc., Alexandria, VA.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26.

SHEN, X. and WASSERMAN, L. (1999). Rates of convergence of posterior distributions. Preprint.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

STONE, C. J. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 293–304. Springer, New York.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362.

S. GHOSAL
A. W. VAN DER VAART
DEPARTMENT OF MATHEMATICS
FREE UNIVERSITY
DE BOELELAAN 1081A
1081 HV AMSTERDAM
NETHERLANDS
E-MAIL: aad@cs.vu.nl

J. K. GHOSH
STATISTICS AND MATHEMATICS UNIT
INDIAN STATISTICAL INSTITUTE
203 B.T. ROAD
CALCUTTA 700 035
INDIA