The Annals of Statistics
1999, Vol. 27, No. 2, 536–561

THE CONSISTENCY OF POSTERIOR DISTRIBUTIONS IN NONPARAMETRIC PROBLEMS

BY ANDREW BARRON,[1] MARK J. SCHERVISH[2] AND LARRY WASSERMAN[3]

Yale University and Carnegie Mellon University

We give conditions that guarantee that the posterior probability of every Hellinger neighborhood of the true distribution tends to 1 almost surely. The conditions are (1) a requirement that the prior not put high mass near distributions with very rough densities and (2) a requirement that the prior put positive mass in Kullback–Leibler neighborhoods of the true distribution. The results are based on the idea of approximating the set of distributions with a finite-dimensional set of distributions with sufficiently small Hellinger bracketing metric entropy. We apply the results to some examples.

1. Introduction. Recent advances in statistical computing have generated renewed interest in nonparametric Bayesian inference. As argued by Diaconis and Freedman (1986), these nonparametric methods are of little value unless they possess reasonable consistency properties. Indeed, Diaconis and Freedman (1986) showed that even if the prior puts positive mass in weak neighborhoods of the true density, it does not follow that the posterior mass of every weak neighborhood of the true density tends to 1.

Doob (1949) showed consistency of the posterior under very weak conditions. However, his proof only gives consistency almost surely with respect to the prior. Consistency can fail on a null set and the theorem gives no guidance on what this null set looks like. For example, if the prior is a point mass at a single density g, then Doob's theorem applies, yet consistency fails at all densities except g. Schwartz (1965) showed that if the prior puts positive mass in each Kullback–Leibler neighborhood of the true density f_0, then asymptotically the posterior does accumulate in weak neighborhoods of f_0. However, weak neighborhoods contain many distributions that, in any practical sense, do not resemble f_0. Thus it seems useful to seek convergence in some stronger sense. The purpose of this paper is to provide a relatively simple, self-contained proof of consistency in Hellinger distance (which is

Received November 1996; revised December 1998.
[1]Supported in part by Office of Naval Research Contract N00014-86-K-0670 and by a National Science Foundation Postdoctor Fellowship.
[2]Supported in part by NSF Grant DMS-96-26181.
[3]Supported in part by NIH Grant RO1-CA54852 and NSF Grants DMS-93-03557 and DMS-93-57646.
AMS 1991 *subject classification.* 62G20.
Key words and phrases. Exponential families, Hellinger distance, nonparametric Bayesian inference, Pólya trees.

equivalent to consistency in total variation) using only a few conditions like those in Barron (1988) and Barron (1998). Related results may be found in Diaconis and Freedman (1997, 1998) and Ghoshal, Ghosh and Ramamoorthi (1999a).

A brief sketch of the main idea behind our results is as follows. Let $X^n = (X_1, \ldots, X_n)$ be n i.i.d. observations from a distribution P_0. The n-fold product measure of P_0 on the product space $(\mathscr{X}^n, \mathscr{B}^n)$ will be denoted P_0^n and the infinite product measure will be denoted P_0^∞. Let π be a prior on a set of distributions (described more formally in the next section) and let $\pi(A|X^n)$ be the posterior probability of A given X^n. Let $A_\varepsilon \equiv A_\varepsilon(P_0) = \{Q\colon d(P_0, Q) \leq \varepsilon\}$ where $d(\cdot, \cdot)$ is some metric. We say that consistency holds at P_0 if for every $\varepsilon > 0$, $\pi(A_\varepsilon|X^n) \to 1$ almost surely $[P_0^\infty]$.

Our strategy is to find a sequence $\{\mathscr{F}_n\}_{n=1}^\infty$ of sets of distributions such that the prior probability of $\mathscr{F}_n^c$ is exponentially small. The sequence $\{\mathscr{F}_n\}_{n=1}^\infty$ is essentially a sieve as in Grenander (1981) and Geman and Hwang (1982). Next, we find a finite set of *upper brackets* $\{f_i^U\colon i = 1, \ldots, N\}$ such that each $f \in \mathscr{F}_n$ satisfies $f \leq f_i^U$ for some i. The likelihood function is then bounded above by the f_i^U's and we show that the posterior is exponentially small outside A_ε as long as the number of brackets does not grow too quickly as a function of n. Bracketing methods have been used for many types of consistency results such as Wong and Shen (1995), van de Geer (1993) and Pollard (1991).

An outline of our paper is as follows. In Section 2 we present the notation and main results about consistency. In Section 3 we present some specific examples. The current paper builds on previous unpublished work by Barron.

2. General results. Let λ be a probability measure on a measurable space $(\mathscr{X}, \mathscr{B})$, where the σ-field $\mathscr{B}$ is separable. Let $\mathscr{Q}$ be the set of all finite measures on $(\mathscr{X}, \mathscr{B})$ that are absolutely continuous with respect to λ. Absolute continuity of all probability measures under consideration with respect to a common σ-finite measure allows us to use the familiar version of Bayes' theorem, (6) below, which we need for our results. It is well known that absolute continuity with respect to a σ-finite measure is equivalent to absolute continuity with respect to a probability measure. Let $d'(\cdot, \cdot)$ denote the Hellinger metric on $\mathscr{Q}$,

$$d'(Q_1, Q_2) = \left\{ \int \left[f_1(x)^{1/2} - f_2(x)^{1/2} \right]^2 d\lambda(x) \right\}^{1/2},$$

where $f_i = dQ_i/d\lambda$. Let $\mathscr{D}$ be the Borel σ-field of subsets of $\mathscr{Q}$ induced by open sets under the metric d'. Lemma 10, in the Appendix, shows that the Radon–Nikodym derivative $f_Q = dQ/d\lambda$ can be chosen so that $f_Q(x)$ is jointly measurable as a function of Q and x. Let $\mathscr{P}$ be the subset of $\mathscr{Q}$ consisting of all probability measures that are absolutely continuous with respect to λ and let $\mathscr{C}$ be the restriction of the σ-field $\mathscr{D}$ to $\mathscr{P}$. For the remainder of this paper, we will use the symbol f_P to stand for the jointly measurable Radon–Nikodym derivative of P with respect to λ when $P \in \mathscr{P}$.

Let $\mathcal{G}$ be the set of all nonnegative functions that are integrable with respect to λ. Let d denote the Hellinger pseudo-metric on $\mathcal{G}$,

$$d(f_1, f_2) = \left\{ \int \left[f_1(x)^{1/2} - f_2(x)^{1/2} \right]^2 d\lambda(x) \right\}^{1/2}. \tag{1}$$

If $f_i = dQ_i/d\lambda$, then $d(f_1, f_2) = d'(Q_1, Q_2)$. An alternative form of (1) is

$$d(f_1, f_2) = \left\{ c_1 + c_2 - 2\int \sqrt{f(x)g(x)}\, d\lambda(x) \right\}^{1/2} \leq \sqrt{c_1 + c_2}, \tag{2}$$

where $c_i = \int f_i(x)\, d\lambda(x)$.

Let $\mathcal{X}^n$ denote the product space of n copies of $\mathcal{X}$, and let λ^n denote product measure. Let $\mathcal{X}^\infty$ be the product space of countably many copies of $\mathcal{X}$. Suppose that $\{X_n\}_{n=1}^\infty$ is a sequence of i.i.d. random variables with distribution P_0 having density f_0 with respect to λ. Let E_0 stand for expectation under distribution P_0. Let $\mathcal{I}(\cdot;\cdot)$ be the Kullback–Leibler information

$$\mathcal{I}(P;Q) = \int \log \frac{f_P(x)}{f_Q(x)} f_P(x)\, d\lambda(x),$$

for $P, Q \in \mathcal{P}$. The integrand in the above expression is interpreted to be 0 whenever $f_P(x) = 0$. Note that $\mathcal{I}(P;Q) \geq 0$ with equality if and only if P and Q are the same probability. Also, $\mathcal{I}(P;Q) < \infty$ implies that $P \ll Q$. Lemma 11, in the Appendix, shows that $\mathcal{I}(P_0;P)$ is measurable as a function from $(\mathcal{P}, \mathcal{C})$ to $\mathbb{R}$. For each $\varepsilon > 0$, define

$$N_\varepsilon = \{P \in \mathcal{P} : \mathcal{I}(P_0;P) \leq \varepsilon\}, \tag{3}$$

$$A_\varepsilon = \{P \in \mathcal{P} : d'(P_0, P) \leq \varepsilon\}. \tag{4}$$

Let $X^\infty = (X_1, X_2, \ldots)$ be the sequence of observations of which the first n coordinates are denoted $X^n = (X_1, \ldots, X_n)$. Realizations of these random sequences are denoted $x^\infty = (x_1, x_2, \ldots)$ and $x^n = (x_1, \ldots, x_n)$, respectively. The density of the n-fold product measure of P_0 is denoted by

$$p_n(x^n) = \prod_{i=1}^n f_0(x_i). \tag{5}$$

For $P \in \mathcal{P}$, let

$$D_n(x^n; P) = \frac{1}{n} \log \frac{p_n(x^n)}{\prod_{i=1}^n f_P(x_i)} \tag{6}$$

be the sample Kullback–Leibler information so that $E_0 D_n(X^n; P) = \mathcal{I}(P_0; P)$. Lemma 2 shows that the denominator in (5) is finite and positive with probability 1. Let π be a prior distribution on $(\mathcal{P}, \mathcal{C})$.

The predictive density of X^n is given by

$$m_n(x^n) = \int \prod_{i=1}^n f_P(x_i)\, d\pi(P). \tag{7}$$

Bayes' theorem says that the posterior probability given $X^n = x^n$ of a measurable subset B of $\mathscr{P}$ is given by

$$\pi(B|x^n) = [m_n(x^n)]^{-1} \int_B \prod_{i=1}^n f_P(x_i)\, d\pi(P), \tag{8}$$

if $0 < m_n(x^n) < \infty$. Define $\pi(B|x^n) = I_B(\lambda)$ if $m_n(x^n) \in \{0, \infty\}$. (We will see in Lemma 1 that this case is essentially ignorable.) Schervish [(1995), Theorem 1.31] shows that (8) is a regular conditional distribution.

First, we state the assumptions under which we prove consistency. The following definition is based on one from Alexander (1984).

DEFINITION 1. For $\delta > 0$ and $C \subseteq \mathscr{P}$, define $\mathscr{H}(C, \delta)$ to be the logarithm of the infimum of the set of all k such that there exist nonnegative functions $f_1^U, \ldots, f_k^U$ satisfying:

(i) $\int f_i^U(x)\, d\lambda(x) \le 1 + \delta$ for all i;
(ii) for each $P \in C$ there exists i such that $f_P \le f_i^U$ a.e. $[\lambda]$.

We call $\mathscr{H}(C, \delta)$, the δ-upper metric entropy of C.

Also, the collection $f_1^U, \ldots, f_k^U$ is called a *δ-upper bracketing of C*. For the next assumption, recall that N_ε is an ε-Kullback–Leibler neighborhood of P_0 defined in (3).

ASSUMPTION 1. For every $\varepsilon > 0$, $\pi(N_\varepsilon) > 0$.

ASSUMPTION 2. For every $\varepsilon > 0$, there exists a sequence $\{\mathscr{F}_n\}_{n=1}^\infty$ of subsets of $\mathscr{P}$, and positive, real numbers c, c_1, c_2, δ such that

$$c < \left(\left[\varepsilon - \sqrt{\delta}\right]^2 - \delta\right)\Big/2, \qquad \delta < \varepsilon^2/4$$

and such that:

(i) $\pi(\mathscr{F}_n^c) \le c_1 \exp(-nc_2)$ for all but finitely many n;
(ii) $\mathscr{H}(\mathscr{F}_n, \delta) \le nc$ for all but finitely many n.

The purpose of Assumption 1 is to avoid problems like those highlighted by Diaconis and Freedman (1986). The prior used by Diaconis and Freedman put positive probability on weak neighborhoods of the true distribution, but not on sets with finite Kullback–Leibler information. Since the likelihood function at P_0 divided by the likelihood at P is $\exp[nD_n(x^n; P)]$ and $D_n(x^n; P) \to \mathscr{I}(P_0; P)$ a.s. $[P_0^\infty]$, it seems plausible to expect the posterior distribution to concentrate on the set of probabilities P for which $\mathscr{I}(P_0; P)$ is small, but only if that set has positive prior probability. Assumption 2 is designed to prevent the prior from giving substantial mass to distributions that happen to have very rough densities. In Section 3.5, we give a detailed example in which Assumption 1 holds but the prior puts too much mass on distributions with densities that are allowed to jump up and down too often. What happens is

that, for too many data sequences there are densities with substantial prior probability that jump up just in the vicinity of each data value and then jump down just away from each data value. Assumption 2 is designed to force the prior probabilities of such distributions to be small enough so that only extremely large samples of highly clustered data will lead to their having large posterior probabilities. This same problem arises in nonparametric maximum likelihood estimation and it is often addressed in a similar fashion, namely by using sieves that satisfy a condition like part (ii) of Assumption 2.

To check part (ii) of Assumption 2, it is often convenient to set $\delta = \varepsilon^2/16$ and $c = \varepsilon^2/5$. Then one checks that $\mathscr{H}(\mathscr{F}_n, \varepsilon^2/16) \leq n\varepsilon^2/5$ for all large n. The main result of this paper is the following consistency result.

THEOREM 1. *Let A_ε be as defined in* (4). *Under Assumptions* 1 *and* 2, *for every* $\varepsilon > 0$,

$$\lim_{n\to\infty} \pi(A_\varepsilon | x^n) = 1 \quad a.s.\,[P_0^\infty].$$

The proof of Theorem 1 requires some lemmas, but the following simple consequence of Theorem 1 is easy to prove.

COROLLARY 1. *Define*

$$\hat{f}_n(\cdot) = \int f_P(\cdot)\, d\pi(P|x^n).$$

Under Assumptions 1 *and* 2, $\lim_{n\to\infty} d(f_0, \hat{f}_n) = 0$, *a.s.* $[P_0^\infty]$.

PROOF. For each $\varepsilon > 0$, we have

$$\begin{aligned} d(f_0, \hat{f}_n) &\leq \int d(f_0, f_P)\, d\pi(P|x^n) \\ &= \int_{A_\varepsilon} d(f_0, f_P)\, d\pi(P|x^n) + \int_{A_\varepsilon^c} d(f_0, f_P)\, d\pi(P|x^n) \quad \text{a.s. } [P_0^n], \end{aligned} \tag{9}$$

where the inequality follows from Jensen's inequality and the convexity of $d(f_0, \cdot)$. The first term on the right-hand side of (9) is at most ε by the definition of A_ε, and the second term goes to 0 a.s. $[P_0^\infty]$ by Theorem 1 and the fact that Hellinger distance is bounded. Since ε is arbitrary, the result follows. □

Note that $\hat{f}_n$ in Corollary 1 is the Bayes estimate of the density under a variety of loss functions.

The proof of Theorem 1 will appear after the next several lemmas. Before plunging into the lemmas and the main proof, here is an outline of the strategy. The posterior probability of A_ε^c may be written as the ratio $\int_{A_\varepsilon^c} R_n\, d\pi / \int R_n\, d\pi$ where $R_n = \prod_i f_P(x_i)/f_0(x_i) = \exp(-nD_n)$. Lemmas 3 and 4 establish that the denominator of the ratio is not exponentially small. Lemma 5 shows that a sequence of sets with exponentially small prior

probability has negligible posterior probability. This allows us to restrict attention to the sets $\mathscr{F}_n$. Lemmas 6 and 7 establish a large deviation inequality that will be used to show that the numerator of the ratio is exponentially small. Lemmas 1 and 2 establish that certain quantities that appear in several fractions during the course of the proof are finite and nonzero a.s. $[P_0^\infty]$. These facts are then combined in the proof of Theorem 1.

LEMMA 1. *Under Assumption* 1,

$$P_0^\infty(x^\infty\text{: there exists } n \text{ such that } m_n(x^n) \in \{0,\infty\}) = 0, \tag{10}$$

where m_n *is defined in* (7).

PROOF. We will prove that, for each n,

$$P_0^n(x^n\text{: } m_n(x^n) \in \{0,\infty\}) = 0. \tag{11}$$

The set in (10) occurs if and only if at least one of the sets in (11) occurs. For $A \in \mathscr{B}^n$, $P^n(A) = \int_A \prod_{i=1}^n f_P(x_i)\, d\lambda^n(x^n)$ and

$$\int P^n(A)\, d\pi(P) = \int\int_A \prod_{i=1}^n f_P(x_i)\, d\lambda^n(x^n)\, d\pi(P). \tag{12}$$

Since $f_P \geq 0$, we can change the order of integration in (12). The result is

$$\int P^n(A)\, d\pi(P) = \int_A m_n(x^n)\, d\lambda^n(x^n). \tag{13}$$

First, let $A = \{x^n\text{: } m_n(x^n) = 0\}$. Then the right-hand side of (13) equals 0 and this implies that $P^n(A) = 0$ a.s. $[\pi]$. Hence, $\pi(B) = 1$ where $B = \{P; P^n(A) = 0\}$. Choose any $\varepsilon > 0$. Assumption 1 says that $\pi(N_\varepsilon) > 0$. So $N_\varepsilon \cap B$ is nonempty. Choose some $P \in N_\varepsilon \cap B$. Since $P \in N_\varepsilon$, P_0 is absolutely continuous with respect to P, hence $P_0^n(A) = 0$.

Next, let $A = \{x^n\text{: } m_n(x^n) = \infty\}$. If $\lambda^n(A) > 0$, the integral on the right-hand side of (13) would be ∞, which would imply that $P^n(A)$ (on the left-hand side) was unbounded. This contradicts P^n being a probability. So $\lambda^n(A) = 0$ and hence $P_0^n(A) = 0$. □

LEMMA 2. *Let* p_n *be defined in* (5). *Then*

$$P_0^\infty(x^\infty\text{: there exists } n \text{ such that } p_n(x^n) \in \{0,\infty\}) = 0. \tag{14}$$

PROOF. As in lemma 1, we will prove that, for each n,

$$P_0^n(x^n\text{: } p_n(x^n) \in \{0,\infty\}) = 0. \tag{15}$$

For all $A \in \mathscr{B}^n$,

$$P_0^n(A) = \int_A p_n(x^n)\, d\lambda^n(x^n). \tag{16}$$

First, let $A = \{x^n\text{: } p_n(x^n) = 0\}$, then (16) clearly implies $P^n(A) = 0$. Next, let $A = \{x^n\text{: } p_n(x^n) = \infty\}$. If $\lambda^n(A) > 0$, then the integral on the right-hand side of (16) would be ∞ implying that $P_0^n(A) = \infty$, a contradiction. So $\lambda^n(A) = 0$ and $P_0^n(A) = 0$. □

LEMMA 3. *There exists a set* $B \subseteq \mathscr{X}^\infty$ *such that* $P_0^\infty(B) = 1$ *and such that for every* $x^\infty \in \mathscr{B}$, *there is a set* $G_{x^\infty} \in \mathscr{C}$ *such that* $\pi(G_{x^\infty}) = 1$ *and for every* $P \in G_{x^\infty}$, $\lim_{n\to\infty} D_n(x^n; P) = \mathscr{I}(P_0; P)$.

PROOF. Let $G = \{(x^\infty, P)\colon \lim_{n\to\infty} D_n(x^n; P) = \mathscr{I}(P_0; P)\}$. Since $\mathscr{I}(P_0; P)$ and $D_n(x^n; P)$ are jointly measurable from $\mathscr{X}^\infty \times \mathscr{P}$ to $\mathbb{R}$ (see Lemma 11 of the Appendix), we know that G is in the product σ-field $\mathscr{B}^\infty \otimes \mathscr{C}$. Let $G_P = \{x^\infty; (x^\infty, P) \in G\}$ and let $G_{x^\infty} = \{P; (x^\infty, P) \in G\}$ be the sections. Then $G_P \in \mathscr{B}^\infty$ for all P and $G_{x^\infty} \in \mathscr{C}$ for all x^∞. These facts are steps in any standard proof of Fubini's theorem, as are the facts that $P_0^\infty(G_P)$ and $\pi(G_{x^\infty})$ are measurable functions. By the strong law of large numbers, $P_0^\infty(G_P) = 1$, for every $P \in \mathscr{P}$. By Fubini's theorem we have

$$
\begin{aligned}
1 &= \int_{\mathscr{P}} P_0^\infty(G_P)\, d\pi(P) \\
&= \int_{\mathscr{P}} \int_{\mathscr{X}^\infty} I_G(x^\infty, P)\, dP_0^\infty(x^\infty)\, d\pi(P) \qquad (17) \\
&= \int_{\mathscr{X}^\infty} \int_{\mathscr{P}} I_G(x^\infty, P)\, d\pi(P)\, dP_0^\infty(x^\infty).
\end{aligned}
$$

Let B be the set of all x^∞ such that $\int I_G(x^\infty, P)\, d\pi(P) = \pi(G_{x^\infty}) = 1$. It follows from (17) that $P_0^\infty(B) = 1$. □

LEMMA 4. *Under Assumption* 1, *for every* $\varepsilon > 0$,

$$
P_0^\infty\left(x^\infty\colon \frac{m_n(x^n)}{p_n(x^n)} \le \exp(-n\varepsilon), \text{ i.o.} \right) = 0,
$$

where m_n *is defined in* (7) *and* p_n *is defined in* (5).

PROOF. Let $\varepsilon > 0$ be given and let $x^\infty \in B$, where B is the set with the same name guaranteed to exist by Lemma 3. Also, let G_{x^∞} be the set with the same name guaranteed to exist by Lemma 3. Then,

$$
\begin{aligned}
\exp(n\varepsilon)\frac{m_n(x^n)}{p_n(x^n)} &= \int \frac{\prod_{i=1}^n f(x_i)}{p_n(x^n)} \exp(n\varepsilon)\, d\pi(P) \\
&\ge \int_{N_{\varepsilon/2} \cap G_{x^\infty}} \frac{\prod_{i=1}^n f_P(x_i)}{p_n(x^n)} \exp(n\varepsilon)\, d\pi(P) \qquad (18) \\
&= \int_{N_{\varepsilon/2} \cap G_{x^\infty}} \exp\{n[\varepsilon - D_n(x^n; P)]\}\, d\pi(P).
\end{aligned}
$$

According to Lemma 3,

$$
\liminf_{n\to\infty} \exp\bigl(n[\varepsilon - D_n(x^n; P)]\bigr) = \infty \quad \text{for all } P \in N_{\varepsilon/2} \cap G_{x^\infty}. \qquad (19)
$$

Assumption 1 says that $\pi(N_{\varepsilon/2}) > 0$ and Lemma 3 says that $\pi(G_{x^\infty}) = 1$, so $\pi(N_{\varepsilon/2} \cap G_{x^\infty}) > 0$. Fatou's lemma and (19) imply that the integrals on the

far right-hand side of (18) go to ∞. The rest of (18) implies that

$$\lim_{n\to\infty} \exp(n\varepsilon)\frac{m_n(x^n)}{p_n(x^n)} = \infty,$$

hence $m_n(x^n)/p_n(x^n) > \exp(-n\varepsilon)$ all but finitely often. Since this holds for every $x^\infty \in B$ and $P_0^\infty(B) = 1$, the result is proved. □

LEMMA 5. *Suppose that Assumption* 1 *holds. Let* $c_1, c_2 > 0$. *Suppose that* $\{B_n\}_{n=1}^\infty$ *is a sequence of subsets of* $\mathscr{P}$ *such that* $\pi(B_n) < c_1 \exp(-c_2 n)$ *for all but finitely many* n. *Then* $\lim_{n\to\infty} \pi(B_n|x^n) = 0$ *a.s.* $[P_0^\infty]$.

PROOF. It suffices to prove that, for each $\delta > 0$,

$$P_0^\infty(x^\infty\colon \pi(B_n|x^n) > \delta \text{ i.o.}) = 0. \tag{20}$$

First, write

$$\begin{aligned}\pi(B_n|x^n) &= \frac{1}{m_n(x^n)}\int_{B_n}\prod_{i=1}^n f_P(x_i)\,d\pi(P)\\ &= \frac{p_n(x^n)}{m_n(x^n)}\int_{B_n}\frac{\prod_{i=1}^n f_P(x_i)}{p_n(x^n)}\,d\pi(P).\end{aligned}$$

For all but finitely many n, using the fact that $\prod_i f_P(x_i) = dP^n/d\lambda^n$,

$$\begin{aligned}&P_0^n\left(x^n\colon \int_{B_n}\frac{\prod_{i=1}^n f_P(x_i)}{p_n(x^n)}\,d\pi(P) > \exp\left[-n\frac{c_2}{2}\right]\right)\\ &\quad\le \exp\left(n\frac{c_2}{2}\right)\int_{\mathscr{X}^n}\int_{B_n}\prod_{i=1}^n f_P(x_i)\,d\pi(P)\,d\lambda^n(x^n)\\ &\quad= \exp\left(n\frac{c_2}{2}\right)\int_{B_n}\int_{\mathscr{X}^n}\prod_{i=1}^n f_P(x_i)\,d\lambda^n(x^n)\,d\pi(P)\\ &\quad= \exp\left(n\frac{c_2}{2}\right)\pi(B_n)\\ &\quad\le c_1\exp\left(-n\frac{c_2}{2}\right),\end{aligned}$$

where the first line follows from the Markov inequality, the second line follows from Fubini's theorem and the last line follows from the hypotheses of the lemma. Since $\sum_{n=1}^\infty \exp(-nc_2/2) < \infty$, the first Borel–Cantelli lemma implies that

$$P_0^\infty\left(x^\infty\colon \int_{B_n}\frac{\prod_{i=1}^n f_P(x_i)}{p_n(x^n)}\,d\pi(P) > \exp\left[-n\frac{c_2}{2}\right]\text{ i.o.}\right) = 0. \tag{21}$$

It follows from Lemma 4 that

$$P_0^\infty\left(x^\infty\colon \frac{p_n(x^n)}{m_n(x^n)} > \exp\left[n\frac{c_2}{4}\right] \text{ i.o.}\right) = 0.$$

Combining this with (21) yields

$$P_0^\infty\left(x^\infty\colon \pi(B_n|x^n) > \exp\left[-n\frac{c_2}{4}\right] \text{ i.o.}\right) = 0,$$

which implies (20). □

The following lemma is a modification of Lemma 1 of Wong and Shen (1995).

LEMMA 6. *Let g be a nonnegative, integrable function, $\beta > 0$, $d(f_0, g) = \sqrt{\gamma}$, $\int g(x)\,d\lambda(x) \le 1 + \delta$ and $\delta \le \gamma$. Then*

$$P_0^n\left(x^n\colon \prod_{i=1}^n \frac{g(x_i)}{f_0(x_i)} \ge \exp[-n\beta]\right) \le \exp\left(-n\frac{\gamma - \beta - \delta}{2}\right).$$

PROOF.

$$\begin{aligned} P_0^n\left(x^n\colon \prod_{i=1}^n \left(\frac{g(x_i)}{f_0(x_i)}\right)^{1/2} \ge \exp\left[-\frac{n\beta}{2}\right]\right) &\le \exp\left(\frac{n\beta}{2}\right)\left(E_0\left(\frac{g(X_1)}{f_0(X_1)}\right)^{1/2}\right)^n \\ &\le \exp\left(\frac{n\beta}{2}\right)\left(1 - \frac{\gamma - \delta}{2}\right)^n \\ &= \exp\left(\frac{n\beta}{2}\right)\exp\left(n\log\left[1 - \frac{\gamma - \delta}{2}\right]\right) \\ &\le \exp\left(n\left[\frac{\beta}{2} - \frac{\gamma - \delta}{2}\right]\right), \end{aligned}$$

where the first line follows from the Markov inequality, the second follows from (2) and the fourth follows from the facts that $\delta \le \gamma \le 2 + \delta$, according to (2), and $\log(1 - x) \le -x$ for $0 \le x \le 1$. □

LEMMA 7. *Let $P \in \mathscr{P}$ and $g \in \mathscr{G}$ be such that $f_P \le g$ a.e. $[\lambda]$ and $\int g(x)\,d\lambda(x) \le 1 + \delta$. Then $d(f_P, g) \le \sqrt{\delta}$.*

PROOF. It follows from (2) that

$$d(f, g)^2 = \int\left(\sqrt{f_P(x)} - \sqrt{g(x)}\right)^2 d\lambda(x) \le 2 + \delta - 2\int\sqrt{f_P(x)g(x)}\,d\lambda(x).$$

Write

$$\sqrt{f_P(x)g(x)} = f_P(x) + \sqrt{f_P(x)}\left(\sqrt{g(x)} - \sqrt{f_P(x)}\right) \ge f_P(x).$$

It follows that $\int\sqrt{f_P(x)g(x)}\,d\lambda(x) \geq \int f_P(x)\,d\lambda(x) = 1$. So $d(f_P, g)^2 \leq \delta$ and $d(f_P, g) \leq \sqrt{\delta}$. □

PROOF OF THEOREM 1. Let $\varepsilon > 0$ be given, and let $\{\mathscr{F}_n\}_{n=1}^{\infty}$, c, c_1, c_2 and δ be as guaranteed by Assumption 2. Recall that A_ε is a Hellinger neighborhood of the true density as defined in (4). Write

$$\pi(A_\varepsilon^c|x^n) = \pi(A_\varepsilon^c \cap \mathscr{F}_n|x^n) + \pi(A_\varepsilon^c \cap \mathscr{F}_n^c|x^n). \tag{22}$$

Since $\pi(\mathscr{F}_n^c) \leq c_1 \exp(-nc_2)$ for all but finitely many n, it follows from Lemma 5 that the second expression on the right-hand side of (22) goes to 0, a.s. $[P_0^\infty]$. So, it suffices to prove that the first expression on the right-hand side of (22) goes to 0, a.s. $[P_0^\infty]$. Let $C_n = A_\varepsilon^c \cap \mathscr{F}_n$. Now, write

$$\begin{aligned}\pi(C_n|x^n) &= \frac{\int_{C_n}\prod_{i=1}^n f_P(x_i)\,d\pi(P)}{\int\prod_{i=1}^n f_P(x_i)\,d\pi(P)} \\ &= \frac{p_n(x^n)}{m_n(x^n)}\int_{C_n}\prod_{i=1}^n\frac{f_P(x_i)}{f_0(x_i)}\,d\pi(P).\end{aligned} \tag{23}$$

Let $r \equiv r(n, \delta) = \exp(\mathscr{H}(\mathscr{F}_n, \delta))$ and let $\{f_1^U, \ldots, f_r^U\}$ be the brackets guaranteed by Assumption 2. For convenience, suppress the dependence of $r(n, \delta)$ on n and δ. For $j = 1, \ldots, r$, define

$$\tilde{E}_j = \left\{P \in \mathscr{F}_n\colon f_P \leq f_j^U \text{ a.e. } [\lambda]\right\}.$$

Let $E_1 = \tilde{E}_1$ and for $j > 1$ let $E_j = \{P \in \tilde{E}_j;\ P \notin \tilde{E}_s,\ s < j\}$. Hence, the sets $\{E_1, \ldots, E_r\}$ are disjoint and cover C_n, and if $P \in E_j$ then $f_P \leq f_j^U$ a.e. $[\lambda]$. We now write

$$\begin{aligned}\int_{C_n}\prod_{i=1}^n\frac{f_P(x_i)}{f_0(x_i)}\,d\pi(P) &= \sum_{j=1}^r\int_{E_j\cap C_n}\prod_{i=1}^n\frac{f_P(x_i)}{f_0(x_i)}\,d\pi(P) \\ &\leq \sum_{j=1}^r\int_{E_j\cap C_n}\prod_{i=1}^n\frac{f_j^U(x_i)}{f_0(x_i)}\,d\pi(P) \\ &= \sum_{j=1}^r\prod_{i=1}^n\frac{f_j^U(x_i)}{f_0(x_i)}\,\pi(E_j\cap C_n) \quad \text{a.e. } [\lambda^n].\end{aligned} \tag{24}$$

Since $\delta < \varepsilon^2/4$, by Assumption 2 there exists β and c such that

$$0 < \beta < \left(\varepsilon - \sqrt{\delta}\right)^2 - \delta - 2c. \tag{25}$$

Define

$$F_{n,j} = \left\{x^n\colon \prod_{i=1}^n\frac{f_j^U(x_i)}{f_0(x_i)} \geq \exp\left(-\frac{n\beta}{2}\right)\right\}.$$

If $P \in E_j$, Lemma 7 implies that $d(f_P, f_j^U) < \sqrt{\delta}$ and if $P \in C_n$, then $d(f_0, f_P) > \varepsilon$. Suppose there exists some P in $C_n \cap E_j$. By the triangle inequality, $d(f_0, f_j^U) \geq d(f_0, f_P) - d(f_j^U, f_P) \geq \varepsilon - \sqrt{\delta}$. Thus, for those j such

that $d(f_0, f_j^U) < \varepsilon - \sqrt{\delta}$, we can conclude that $E_j \cap C_n = \varnothing$. For every n and every j such that $d(f_0, f_j^U) \geq \varepsilon - \sqrt{\delta}$, apply Lemma 6 with $g = f_j^U$, $\gamma = d(f_0, f_j^U)^2 \geq [\varepsilon - \sqrt{\delta}]^2$, and β as defined in (25) to see that $P_0^n(F_{n,j}) \leq \exp(-nv)$, where $v > c$. So, because of (24) and the fact that $\sum_j \pi(E_j \cap \mathscr{C}_n) \leq 1$, we have

$$
\begin{aligned}
&P_0^n\left(x^n: \int_{C_n} \prod_{i=1}^n \frac{f_P(x_i)}{f_0(x_i)}\, d\pi(P) \geq \exp\left\{-\frac{n\beta}{2}\right\}\right) \\
&\quad \leq P_0^n\left(x^n: \sum_{j=1}^r \left[\prod_{i=1}^n \frac{f_j^U(x_i)}{f_0(x_i)}\right] \pi(E_j \cap C_n) \geq \exp\left\{-\frac{n\beta}{2}\right\}\right) \\
&\quad \leq \sum_{j=1}^r P_0^n(F_{n,j}) \\
&\quad \leq r \exp(-nv) \\
&\quad = \exp(\mathscr{H}(\mathscr{T}_n, \delta) - nv) \\
&\quad \leq \exp(-n[v - c]),
\end{aligned}
$$

for all but finitely many n. The last inequality follows from part (ii) of Assumption 2. The first Borel–Cantelli lemma implies that

$$
P_0^\infty\left(x^\infty: \int_{C_n} \prod_{i=1}^n \frac{f_P(x_i)}{f_0(x_i)}\, d\pi(P) \geq \exp\left\{-\frac{n\beta}{2}\right\}, \text{i.o.}\right) = 0.
$$

Lemma 4 says that

$$
P_0^\infty\left(x^\infty: \frac{p_n(x^n)}{m_n(x^n)} \geq \exp\left\{\frac{n\beta}{4}\right\}, \text{i.o.}\right) = 0.
$$

Combining these last two equations with (23) gives that

$$
P_0^\infty\left(x^\infty: \pi(C_n | x^n) \geq \exp\left\{-\frac{n\beta}{4}\right\}, \text{i.o.}\right) = 0.
$$

Hence $\lim_{n\to\infty} \pi(C_n | x^n) = 0$, a.s. $[P_0^\infty]$. □

Verifying part (ii) of Assumption 2 can be awkward. The next lemma gives a specific condition that can be checked to verify this assumption.

LEMMA 8. *Let $\{\mathscr{T}_n\}_{n=1}^\infty$ be a sequence of finite measurable partitions of $\mathscr{X}$ and let N_n be the cardinality of $\mathscr{T}_n$. For each n, let $a_n > 0$ and suppose that $\lambda(A) = 1/N_n$ for every $A \in \mathscr{T}_n$. Define*

$$
\mathscr{F}_n = \{P \in \mathscr{P}: \text{for every } A \in \mathscr{T}_n \text{ and for every } x, y \in A, \; |f_P(x) - f_P(y)| \leq a_n\}.
$$

Then $\mathscr{H}(\mathscr{F}_n, 2a_n) \leq N_n[1 + \log(1 + 1/[2N_n a_n])]$.

PROOF. For each n and each vector $\mathbf{l}$ of nonnegative integers $\mathbf{l} = (l_1, \ldots, l_{N_n})$ such that

$$a_n \sum_{i=1}^{N_n} l_i \le N_n \le a_n \sum_{i=1}^{N_n} (l_i + 2), \tag{26}$$

define

$$f_l^U(x) = a_n \sum_{i=1}^{N_n} I_{A_i}(x)/(l_i + 2).$$

It is easy to see that for every $P \in \mathscr{F}_n$, there exists f_l^U such that $f_P \le f_l^U$ a.e. $[\lambda]$. Note that

$$\int f_l^U(x)\, d\lambda(x) = a_n \sum_{i=1}^{N_n} (l_i + 2)\frac{1}{N_n} \le 1 + 2a_n.$$

So the collection of $f_{\mathbf{l}}^U$ functions for all $\mathbf{l}$ that satisfy (26) forms a $2a_n$-upper bracketing of $\mathscr{F}_n$. The cardinality of this upper bracketing is the number of hypercubes with sides of length $2a_n$ needed to cover the $N_n - 1$-dimensional simplex in $\mathbb{R}^{N_n}$. An upper bound on this number is $(2a_n)^{-N_n}$ times the volume of $C_n = \{x \in \mathbb{R}^{N_n}: \forall i\, x_i \ge 0, \Sigma_{i=1}^{N_n} x_i \le 1 + 2N_n a_n\}$. It is easy to see that C_n is just $1 + 2N_n a_n$ times the set where the Dirichlet density with all parameters equal to 1 is positive. Hence the volume of C_n is equal to $(1 + 2N_n a_n)^{N_n}/N_n!$. It follows that

$$\begin{aligned}
\mathscr{H}(\mathscr{F}_n, 2a_n) &\le N_n \log(1 + 2N_n a_n) - \log(N_n!) - N_n \log(2a_n) \\
&\le N_n \log(N_n) + N_n \log\left(1 + \frac{1}{2N_n a_n}\right) - N_n \log(N_n) + N_n \\
&= N_n\left[1 + \log\left(1 + \frac{1}{2N_n a_n}\right)\right],
\end{aligned}$$

since $\log(x!) \ge x \log(x) - x$ for all x. □

A simple corollary helps to verify part (ii) of Assumption 2.

COROLLARY 2. *For each* $\varepsilon > 0$, *let* $N_n \le n\varepsilon^2/10$, $a_n = \varepsilon^2/32$ *and* $\delta = \varepsilon^2/16$. *If* $\lim_{n\to\infty} N_n = \infty$, *then the sequence* $\{\mathscr{F}_n\}_{n=1}^{\infty}$ *from Lemma* 8 *satisfies*

$$\mathscr{H}(\mathscr{F}_n, \delta) \le n\varepsilon^2/5.$$

for all but finitely many n.

To verify part (i) of Assumption 2, one must show that $\pi(\mathscr{F}_n^c)$ is exponentially small.

3. Some prior distributions. In this section, we present some prior distributions that satisfy Assumptions 1 and 2. In Section 3.5, we also give an example to show how failure of Assumption 2 can lead to an inconsistent posterior.

All of the examples in this section deal with real-valued random variables. If π is a prior over $\mathscr{P}$, then the prior marginal distribution of each X_i is

$$P_*(A) = \int_{\mathscr{P}} P(A)\, d\pi(P) \quad \text{for } A \in \mathscr{B}.$$

We find it convenient to construct examples in which P_* is the uniform distribution on $[0,1]$. It is easy to create priors for distributions on the real line. Let P_* be a distribution with cumulative distribution function (cdf) F_*. Of course, we can transform each observation X_i to the unit interval by $Y_i = F_*(X_i)$. Use one of our priors for the unknown distribution P of the Y_i random variables and then map the prior back to a prior on the set of distributions on $\mathbb{R}$. More precisely, let $\mathscr{P}$ be the set of all probabilities on $\mathbb{R}$ that are absolutely continuous with respect to Lebesgue measure λ and let $\mathscr{P}_{[0,1]}$ be the subset consisting of distributions on $[0, 1]$. Define $h_*: \mathscr{P}_{[0,1]} \to \mathscr{P}$ by saying that $h_*(P)$ is the probability with cdf $F_P(F_*)$ where F_P is the cdf of the distribution P. It is easy to see that the function h_* is continuous in the Hellinger topology (hence measurable) since $d'(P, Q) = d'(h_*(P), h_*(Q))$. Therefore, any prior π on $\mathscr{P}_{[0,1]}$ induces a prior π_* on $\mathscr{P}$ by means of the function h_*. The induced prior π_* has the property that the marginal prior distribution of each observation is P_*.

3.1. *Histograms.* One prior distribution for continuous distributions that satisfies the conditions of Theorem 1 is a prior concentrated on a collection of distributions with step-function densities. For each n, we construct a collection $\mathscr{U}_n$ of distributions whose densities are constant on each of the finitely many intervals in a partition $\mathscr{T}_n$. We use Corollary 2 to ensure that part (ii) of Assumption 2 holds. We assign the set $\mathscr{U}_n$ prior probability p_n, which is chosen so that part (i) of Assumption 2 holds.

Suppose that $\mathscr{P}$ consists of all probability measures on $[0,1]$ that are absolutely continuous with respect to Lebesgue measure λ. Assume that $\mathscr{I}(P_0; \lambda) < \infty$. For each integer n, let $p_n > 0$ be such that $\sum_{n=1}^{\infty} p_n = 1$. For each n, let N_n be an integer and let $\mathscr{T}_n$ be the partition

$$\mathscr{T}_n = \left\{ \left[0, \frac{1}{N_n}\right), \left[\frac{1}{N_n}, \frac{2}{N_n}\right), \ldots, \left[\frac{N_n - 1}{N_n}, 1\right] \right\}.$$

Let $\mathscr{U}_n$ be the collection of all distributions that have constant density on every interval in $\mathscr{T}_n$. Our prior distribution will place probability p_n on the set $\mathscr{U}_n$ and distribute the probability as follows. Let $a_n > 0$ and denote a random element of $\mathscr{P}$ as P. If $P \in \mathscr{U}_n$, we can write $P = (f_1, \ldots, f_{N_n})$ where $\sum_{i=1}^{N_n} f_i \lambda(A_i) = 1$ and A_i is the interval $[(i-1)/N_n, i/N_n)$. This makes $f_i = f_P(x)$ for all $x \in A_i$. Conditional on $P \in \mathscr{U}_n$, we assign P/N_n the Dirichlet distribution $\mathrm{Dir}(a_n, \ldots, a_n)$.

We now prove that, by careful choice of N_n and p_n as functions of n this prior distribution will satisfy the conditions of Theorem 1. We will let $N_n = 2^{m_n}$ with $\{m_n\}_{n=1}^{\infty}$ a nondecreasing sequence of positive integers that goes to ∞. The following result is proved in the Appendix.

LEMMA 9. *Let $(\mathscr{X}, \mathscr{B}, \lambda)$ be a probability space, and let $\mathscr{R}$ be a collection of measurable real-valued functions defined on $\mathscr{X}$. For each $b > 0$, define*

$$\mathscr{R}_b = \{f \in \mathscr{R}: \operatorname{ess\,sup}|f| \leq b\},$$
$$L_b = \{f: \operatorname{ess\,sup}|f| \leq b\},$$

where the essential supremum is relative to λ. Suppose that there exists $r > 1$ such that $\mathscr{R}_{rb}$ is dense [in the sense of $L^1(\lambda)$] in L_b for all large b. Let $P_0 \ll \lambda$ be another probability on $(\mathscr{X}, \mathscr{B})$ such that $\mathscr{I}(P_0; \lambda) < \infty$. Then for every $\varepsilon > 0$, there is a bounded function $g \in \mathscr{R}$ such that $\mathscr{I}(P_0; P_g) < \varepsilon$, where P_g is the distribution with density

$$p_g(x) = \frac{\exp(g(x))}{\int \exp(g(y))\, d\lambda(y)}.$$

Let $\mathscr{R}$ be the set of all step functions that are constant on all of the intervals in at least one of the $\mathscr{T}_n$ partitions, and let λ be Lebesgue measure. Since step functions are dense in the collection of bounded measurable functions and $\mathscr{R}$ is dense in the collection of step functions, it follows that for each ε, there exists n and $P_\varepsilon \in \mathscr{U}_n$ such that $\mathscr{I}(P_0; P_\varepsilon) < \varepsilon/2$. Since the Dirichlet distribution over $\mathscr{U}_n$ assigns positive probability to every open neighborhood of P_ε and $\mathscr{I}(P_0; P)$ is continuous as a function of P for distributions with densities in $\mathscr{R}$, it follows that Assumption 1 holds.

Next, construct the sets $\{\mathscr{F}_n\}_{n=1}^{\infty}$ as in Corollary 2. Since each $a_n > 0$ and the probabilities in $\mathscr{U}_n$ have constant density on every $A \in \mathscr{T}_n$, it follows that $\mathscr{U}_n \subseteq \mathscr{F}_n$. Also, $\mathscr{U}_n \subseteq \mathscr{U}_{n+1}$ for all n, so $\pi(\mathscr{F}_n^c) \leq \sum_{l=n+1}^{\infty} p_l$. Setting $p_n = (1 - a)a^n$ for some $0 < a < 1$ will satisfy part (i) of Assumption 2. Finally, let $N_n = n/\log(n)$ (that is, $m_n = \lfloor \log_2(n) - \log_2(\log(n)) \rfloor$) in Corollary 2 so that part (ii) of Assumption 2 holds.

3.2. *Pólya Tree Priors.* The class of Pólya tree distributions was described by Mauldin, Sudderth and Williams (1992) and Lavine (1992). Pólya trees are special cases of tailfree processes [see Freedman (1963) and Fabius (1964)]. Consider the sequence of partitions $\{\mathscr{S}_k\}_{k=1}^{\infty}$ of $[0, 1]$ such that $\mathscr{S}_1 = \{[0, 1/2], (1/2, 1]\}$ and each $\mathscr{S}_k$ contains the left and right halves of all intervals in $\mathscr{S}_{k-1}$ for $k > 1$. (For convenience, let $\mathscr{S}_0 = \{[0, 1]\}$.) For an interval $I \in \mathscr{S}_k$ for $k = 0, 1, \ldots$ create a random variable V_I taking values in $[0, 1]$ and having mean $1/2$. Make all of the V_I independent of each other. For each $I \in \mathscr{S}_k$ for $k \geq 1$, define $p^1(I) \in \mathscr{S}_{k-1}$ to be the interval J in $\mathscr{S}_{k-1}$ such that $I \subseteq J$ and set

$$W_I = \begin{cases} V_{p^1(I)}, & \text{if } I \text{ is the left subinterval of } p^1(I), \\ 1 - V_{p^1(I)}, & \text{if } I \text{ is the right subinterval of } p^1(I). \end{cases}$$

For $I \in \mathcal{S}_k$ with $k \geq 2$, define $p^2(I) = p^1(p^1(I))$, and similarly define $p^i(I)$ for $i = 1, \ldots, k$. [For convenience, let $p^0(I) = I$ for all I.] Let $\mathcal{S} = \bigcup_{k=0}^{\infty} \mathcal{S}_k$, and for each $I \in \mathcal{S}$, define $K(I)$ to be that k such that $I \in \mathcal{S}_k$, and set $P(I) = \prod_{i=1}^{K(I)} W_{p^{i-1}(I)}$. The set function P extends uniquely to the smallest σ-field containing $\mathcal{S}$, which is the Borel σ-field, and becomes a random probability measure on the unit interval. Kraft (1964) shows that, if the distribution of V_I becomes concentrated around $1/2$ sufficiently rapidly as I shrinks (moves through later partitions $\mathcal{S}_k$) then P will have a density with respect to Lebesgue measure with probability 1. (For example, if each $V_I \in \mathcal{S}_k$ has the Beta(a_k, a_k) distribution, then the conditions of Kraft (1964) will be met if $\sum_{k=1}^{\infty} a_k^{-1} < \infty$.) Lavine (1994), Theorem 2), and Ghosal, Ghosh and Ramamoorthi (1999b), Theorem 3.1, prove results like the following.

PROPOSITION 1. *Suppose that for every k and every $I \in \mathcal{S}_k$, V_I has the* Beta(a_k, a_k) *distribution and that $\mathcal{A}(P_0; P_*) < \infty$. If $\sum_{k=1}^{\infty} a_k^{-1/2} < \infty$, then $\pi(N_\varepsilon) > 0$ for every $\varepsilon > 0$.*

In words, Assumption 1 can be satisfied by a Pólya tree distribution so long as the prior predictive distribution is not infinitely far away from the true distribution.

We show next that Assumption 2 can also be satisfied. We will construct a sequence of sets $\{\mathcal{F}_n\}_{n=1}^{\infty}$ as in Corollary 2. Let P have the Pólya tree prior distribution on $[0, 1]$ that has a density, f_P, with probability 1. For each $y \in [0, 1]$ and $k = 0, 1, \ldots,$ let $I_k(y)$ be that interval in $\mathcal{S}_k$ that contains y. Then $p(y) = \lim_{k \to \infty} 2^k \prod_{i=1}^{k} W_{I_i(y)}$. [See Kraft (1964)]. Let $\hat{f}_k(y) = 2^k \prod_{i=1}^{k} W_{I_i(y)}$, the approximation to f_P based on the first k partitions. Suppose that $W_I \sim$ Beta(a_k, a_k) for all $I \in \mathcal{S}_k$. Let $\{g_k\}_{k=1}^{\infty}$ be a sequence of numbers such that $\sum_{k=1}^{\infty} 2^k g_k < \infty$, and let $e_k = \Pr(|2W_I - 1| > g_k)$ for $I \in \mathcal{S}_k$. Let $E_k = \sum_{i=k+1}^{\infty} 2^i e_i$ and $G_k = \sum_{i=k+1}^{\infty} 2^{i-1} g_i$. Then

$$\pi\left(\exists y: \left|f_P(y) - \hat{f}_k(y)\right| > G_k\right) \leq E_k, \tag{27}$$

because $W_{I_k(y)} < 2$ for all k and y. If $G_k \leq \varepsilon^2/16$, then the event whose probability is bounded in (27) contains $\mathcal{F}_n^c$. Hence, (27) provides a bound on $\pi(\mathcal{F}_n^c)$. The partition $\mathcal{S}_k$ plays the role of $\mathcal{T}_n$ in Lemma 8, and the cardinality of $\mathcal{S}_k$ is $N_n = 2^k$. To satisfy the conditions of Corollary 2, we need $k \leq \log_2(\varepsilon^2 n/10)$. So, let

$$k_n(\varepsilon) = \left\lfloor \log_2\left(\frac{\varepsilon^2 n}{10}\right) \right\rfloor,$$

and set $\mathcal{T}_n = \mathcal{S}_{k_n(\varepsilon)}$ to guarantee that part (ii) of Assumption 2 holds. Choose the a_k large enough so that $\sum_{k=1}^{\infty} a_k^{-1/2} < \infty$ (so Proposition 1 says that Assumption 1 holds) and large enough so that $\log E_k \leq b_1 - b_2 2^k$ for some constants b_1, b_2 with $b_2 > 0$. This makes part (i) of Assumption 2 hold.

As an example of how to set the numbers a_k, consider the following. Let $W \sim \text{Beta}(a, a)$ and let $Z = |2W - 1|$. Then, for $g \in (0, 1)$,

$$\begin{aligned}
\Pr(Z > g) &= 2\Pr\left(W > \frac{1+g}{2}\right) \\
&= 2\frac{\Gamma(2a)}{\Gamma(a)^2}\int_{(1+g)/2}^{1} w^{a-1}(1-w)^{a-1}\,dw \\
&\le 2\frac{\Gamma(2a)}{\Gamma(a)^2}\left(\frac{1+g}{2}\right)^{a-1}\left(\frac{1-g}{2}\right)^{a} \\
&\le \frac{\sqrt{a}}{\sqrt{\pi}}(1-g^2)^a \\
&< \frac{\sqrt{a}}{\sqrt{\pi}}\exp(-g^2 a),
\end{aligned}$$

where the third line follows from the monotonicity of the Beta density on $[1/2, 1]$, the fourth line follows from the facts that

$$\Gamma(2a) = \frac{1}{\sqrt{\pi}}2^{2a-1}\Gamma(a)\Gamma\left(a + \frac{1}{2}\right)$$

and $\Gamma(a + \frac{1}{2}) \le \Gamma(a)\sqrt{a}$ and the fifth line follows from the fact that $(1 - x)^y < \exp(-xy)$ for $0 < x < 1$ and $y > 0$. So, we can let $g_k = 2^{-k}$ for $\alpha > 0$ and let $a_k = 1/g_k^3$. It now follows that

$$e_k \le \frac{(\sqrt{8})^k}{\sqrt{\pi}}\exp(-2^k),$$

so

$$\begin{aligned}
E_k &\le \sum_{i=k+1}^{\infty} 2^i \frac{(\sqrt{8})^i}{\sqrt{\pi}}\exp(-2^i) \\
&= \frac{(2\sqrt{8})^{k+1}\exp(-2^{k+1})}{\sqrt{\pi}} \sum_{i=k+1}^{\infty} (2\sqrt{8})^{i-(k+1)}\exp(-2^i + 2^{k+1}).
\end{aligned}$$

Since the last sum is finite it follows that $\log(E_k) \le b_1 - b_2 2^k$. In summary, a Pólya tree prior with every W_I for $I \in \mathscr{S}_k$ having a $\text{Beta}(a_k, a_k)$ distribution with $a_k = 8^k$ will satisfy the assumptions of Theorem 1 with the sequence $\{\mathscr{T}_n\}_{n=1}^{\infty}$ from Corollary 2 so long as $\mathscr{I}(P_0; P_*) < \infty$.

3.3. *Infinite-dimensional exponential families*

Let $\Psi = \{\psi_j\}_{j=1}^{\infty}$ be a sequence of independent random variables with $\psi_j \sim N(0, \tau_j^2)$. Let $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ be a sequence of orthogonal polynomials on $[0, 1]$.

Define

$$f_\Psi(x) = \exp\left(\sum_{j=1}^{\infty} \psi_j \phi_j(x) - c(\Psi) \right),$$

where $c(\cdot)$ makes f_Ψ a density. Let P_Ψ stand for the distribution with density f_Ψ. This model for infinite dimensional parameter spaces has been studied by Leonard (1978) and Lenk (1988, 1991).

Let $a_j = \sup_{0 \le x \le 1} |\psi_j|$ and $b_j = \sup_{0 \le x \le 1} |\phi_j'(x)|$, which are finite since the ϕ_j are polynomials. Now, choose the τ_j's so that $\sum_j a_j \tau_j < \infty$ to assure that f_Ψ is a density with probability 1 and $\sum_j b_j \tau_j < \infty$. Let $A_{n,i} = [(i-1)/N, i/N)$, where $N = \lfloor n/\log(n) \rfloor$, for $i = 1, \ldots, N$. Let $\mathscr{T}_n = \{A_{n,1}, \ldots, A_{n,N}\}$. Define

$$Y_{n,i} = \sup_{x, y \in A_{n,i}} |f_\Psi(x) - f_\Psi(y)|,$$

$$X_{n,i} = \sup_{x, y \in A_{n,i}} \log \frac{f_\Psi(x)}{f_\Psi(y)}.$$

Let $\varepsilon > 0$. Since $X_{n,i} \le \Delta$ implies $Y_{n,i} \le \exp(\Delta) - 1$, we need to show that with $\delta = a_n = \varepsilon^2/32$ as in Corollary 2, there exists $c_1, c_2 > 0$ such that $\Pr(X_{n,j} > \log(1+\delta)) \le c_1 \exp(-c_2 n)$ for all but finitely many n. Let $\Delta = \log(1+\delta)$.

Now, for $x, y \in A_{n,i}$, write

$$\begin{aligned} \log \frac{f_\Psi(x)}{f_\Psi(y)} &= \sum_{j=1}^{\infty} \psi_j \left[\phi_j(x) - \phi_j(y) \right] \\ &= \sum_{j=1}^{\infty} \psi_j \int_y^x \phi_j'(t)\, dt \\ &\le \sum_{j=1}^{\infty} b_j |\psi_j| \, |x - y| \le \frac{1}{N} \sum_{j=1}^{\infty} b_j |\psi_j|. \end{aligned}$$

Let $Z \sim N(0,1)$. Then

$$\begin{aligned} \pi(X_{n,j} > \Delta) &\le \pi\left(\sum_{j=1}^{\infty} b_j |\psi_j| > N\Delta \right) \\ &= \pi\left(\exp\left[\sum_{j=1}^{\infty} b_j |\psi_j| \right] > \exp[N\Delta] \right) \\ &\le \inf_{t \ge 0} \exp(-N\Delta t) \mathrm{E} \exp\left(t \sum_{j=1}^{\infty} b_j |\psi_j| \right) \end{aligned}$$

$$= \inf_{t \geq 0} \exp(-N\Delta t) \prod_{j=1}^{\infty} \mathrm{E} \exp\bigl(t b_j |Z|\bigr)$$

$$= \inf_{t \geq 0} \exp(-N\Delta t) \prod_{j=1}^{\infty} \bigl[2\Phi\bigl(b_j \tau_j t\bigr) \exp\bigl(t^2 b_j^2 \tau_j^2 / 2\bigr)\bigr]$$

$$\leq \exp\left(-N\Delta t_0 + t_0^2 \sum_{j=1}^{\infty} b_j^2 \tau_j^2 / 2\right) \prod_{j=1}^{\infty} \bigl[2\Phi\bigl(b_j \tau_j t_0\bigr)\bigr],$$

where Φ is the standard normal distribution function and $t_0 \geq 0$. Suppose that we choose the τ_j so that $\sum_{j=1}^{\infty} b_j \tau_j < \infty$. Let $t_0 = N\Delta / \sum_{j=1}^{\infty} b_j^2 \tau_j^2$. Then $\prod_{j=1}^{\infty} [2\Phi(b_j \tau_j t_0)] = c_1 < \infty$, and

$$\Pr\bigl(X_{n,j} > \Delta\bigr) \leq c_1 \exp\left(-\frac{N^2 \Delta^2}{\sum_{j=1}^{\infty} b_j^2 \tau_j^2}\right).$$

Since $N^2 > n$, this completes the proof that the prior probability of $\mathscr{F}_n^c$ is exponentially small, where $\mathscr{F}_n$ is as in Corollary 2.

The uniform density on $[0, 1]$ is Lebesgue measure λ. We want to show that if $\mathscr{I}(P_0; \lambda) < \infty$ then, for every $\varepsilon > 0$, $\pi(N_\varepsilon) > 0$. First, suppose that there exist m and $\alpha_1, \ldots, \alpha_m$ such that $\log f_0 = \sum_{j=1}^{m} \alpha_j \phi_j - c(\alpha)$. Let

$$\rho(P, Q) = \sup_{0 \leq x \leq 1} \bigl|\log f_P(x) - \log f_Q(x)\bigr|,$$

and let $B_\varepsilon = \{P\colon \rho(P_0, P) \leq \varepsilon\}$. A simple calculation shows that $\mathscr{I}(P; Q) \leq \rho(P, Q)$. So it suffices to show that $\pi(B_\varepsilon) > 0$.

Let $Z_\xi(r) = \{\psi = (\psi_1, \psi_2, \ldots)\colon \sum_{j=r}^{\infty} a_j |\psi_j| < \xi\}$. Then $\pi(B_\varepsilon) \geq \pi(B_\varepsilon | Z_\xi(r)) \pi(Z_\xi(r))$. Recall that the τ_j's have been chosen so that $\sum_j a_j |\psi_j|$ is finite with probability 1. It follows that for any $\xi > 0$, there exists r_0 such that $\pi(Z_\xi(r_0)) > 0$. Choose $\xi = \varepsilon/2$ and choose $r = \max\{r_0, m + 1\}$. For $\psi \in Z_\xi(r)$ and defining $\alpha_j = 0$ for $j > m$, we see that

$$\rho(P_0, P_\Psi) = \sup_{0 \leq x \leq 1} \left|\sum_j (\psi_j - \alpha_j) \phi_j(x)\right|$$

$$\leq \sum_j a_j |\psi_j - \alpha_j|$$

$$\leq \sum_{j=1}^{r} a_j |\psi_j - \alpha_j| + \xi.$$

So

$$\pi\bigl(B_\varepsilon | Z_\xi(r)\bigr) \geq \pi\left(\sum_{j=1}^{r} a_j |\psi_j - \alpha_j| + \xi < \varepsilon \middle| Z_\xi(r)\right)$$

$$\geq \pi\left(\sum_{j=1}^{r} a_j |\psi_j - \alpha_j| < \frac{\varepsilon}{2} \middle| Z_\xi(r)\right)$$

$$\geq \pi\left(\sum_{j=1}^{r} a_j |\psi_j - \alpha_j| < \frac{\varepsilon}{2}\right).$$

Since the marginal distribution of $(\psi_1, \ldots, \psi_r)$ has support over all $\mathbb{R}^r$, we see that this latter event has positive probability. Thus, $\pi(B_\varepsilon | Z_\xi(r)) > 0$.

Now consider any P_0 such that $\mathcal{I}(P_0; \lambda) < \infty$. Lemma 9 says that for any $a > 0$ there exists a distribution P with density f such that $\log f$ is a polynomial of finite degree and such that $\mathcal{I}(P_0; P) < a$. Further,

$$\mathcal{I}(P_0; P_\Psi) \le \mathcal{I}(P_0; P) + \sup_{0 \le x \le 1} \log \frac{f(x)}{f_\Psi(x)} \le a + \rho(P, P_\Psi).$$

Choose $a = \varepsilon/2$ and note that $B_{\varepsilon/2} \subset N_\varepsilon$. Since we have already shown that $B_{\varepsilon/2}$ has positive prior probability, the proof is complete. □

3.4. *Mixtures of priors.* In the examples earlier in this section, the posterior distributions are somewhat sensitive to the choice of prior predictive distribution P_*. In particular, the posterior predictive densities of future observations computed from histogram and Pólya tree priors tend to have noticeable jumps at the boundaries of the sets in the partitions $\mathcal{T}_n$ and $\mathcal{S}_k$. Also, a choice of P_* that is particularly unlike the sample distribution of the data will make the convergence of the posterior very slow. One way to alleviate these problems is to use a mixture of prior distributions.

Suppose that we replace P_* by a family of distributions $\{P_\theta : \theta \in \Omega\}$ where (Ω, τ) is a measurable space and each $P_\theta \ll \lambda$. (One typical choice is a location/scale family that one thinks of as a first-order approximation to the distribution of the data.) Let ν be a prior probability measure on (Ω, τ). Let Θ be a random variable such that, conditional on $\Theta = \theta$, the prior distribution on $(\mathcal{P}, \mathcal{C})$ is π_θ, where π_θ is constructed just like a π in one of the other examples in this section with P_θ replacing P_*. Let $\pi_\theta(\cdot | x^n)$ denote the conditional posterior distribution on $(\mathcal{P}, \mathcal{C})$ given $\Theta = \theta$ after observing $X^n = x^n$. Let $\nu(\cdot | x^n)$ denote the posterior distribution of Θ given $X^n = x^n$. Then the posterior on $(\mathcal{P}, \mathcal{C})$ is

$$\pi(B | x^n) = \int_\Omega \pi_\theta(B | x^n)\, d\nu(\theta | x^n).$$

To prove consistency of this posterior, we will make some additional assumptions. Assume that $\nu(\{\theta : \mathcal{I}(P_0; P_\theta) < \infty\}) = 1$. Suppose that $\pi_\theta(\cdot | x^n)$ is uniformly consistent a.s. $[\nu]$, that is, there is a subset B of $\mathcal{X}^\infty$ with $P_0(B) = 1$ such that for every $x^\infty \in B$ and $\delta > 0$, there exists $B_{x^\infty} \in \tau$ and $N(x^\infty)$ such that $\nu(B_{x^\infty}) = 1$ and $n \ge N(x^\infty)$ implies $\pi_\theta(A_\varepsilon | x^n) > 1 - \delta$ for all $\theta \in B_{x^\infty}$.

First, note that the conditional distribution of X^n given $\Theta = \theta$ is absolutely continuous with respect to λ with density $g_n(x^n | \theta) = \int_{\mathcal{P}} \prod_{i=1}^n f_P(x_i)\, d\pi_\theta(P)$. It follows from the measure-theoretic version of Bayes' theorem that, for each n, $\nu(\cdot | x^n) \ll \nu$ with probability 1 under M_n, the marginal distribution of X^n. [See Schervish (1995), Theorem 1.3.1.]

Next, note that in the earlier examples in this section, we transformed the data to the interval $(0, 1)$ using F_*. The resulting distribution for P, the

distribution of the transformed data, gave probability 1 to the set of probabilities with densities that are strictly positive on all of $(0, 1)$. This, together with $\mathcal{I}(P_0; P_\theta) < \infty$ implies that for all n, the n-fold product of P_0 on $\mathcal{X}^n$ is absolutely continuous with respect to M_n.

Let C be the set of sequences x^∞ such that $\nu(\cdot|x^n) \ll \nu$ for all n. Since $M_n(C) = 1$, then $P_0^\infty(C) = 1$ and $P_0^\infty(C \cap B) = 1$. For each $x^\infty \in C \cap B$, $n \geq N(x^\infty)$ implies

$$\pi(A_\varepsilon|x^n) = \int_{B_{x^\infty}} \pi_\theta(A_\varepsilon|x^n)\, d\nu(\theta|x^n) > (1 - \delta)\nu(B_{x^\infty}|x^n) = 1 - \delta.$$

Since δ is arbitrary, this proves that $P_0^\infty(\lim_{n\to\infty} \pi(A_\varepsilon|x^n) = 1$.

Uniform consistency is difficult to verify in general, and we do not believe that it is a necessary condition. [For example, Ghosal, Ghosh and Ramamoorthi (1999b) use a continuity condition instead of uniform consistency to prove a weaker form of consistency for location mixtures of symmetric Pólya trees.]

On the other hand, uniform consistency does hold in the simple case in which Ω is a finite set. For example, Ω might consist of a Pólya tree, an exponential family and a histogram. The posterior, as the prior, would then be a mixture of these three types of distributions.

3.5. *A counterexample.* In this section, we present an example in which Assumption 1 holds but Assumption 2 fails and the posteriori is inconsistent. The idea of the example is that the prior π is split evenly between disjoint sets of probabilities $\mathcal{P}_0$ and $\mathcal{P}_*$. There are distributions $P \in \mathcal{P}_0$ such that $\mathcal{I}(P_0; P)$ is arbitrarily small and $\pi(N_\varepsilon) > 0$ for all $\varepsilon > 0$. The densities in $\mathcal{P}_*$ however, are very far from P_0 in Hellinger distance, yet they track the data sufficiently well to acquire significant posterior probability infinitely often.

For each positive integer N, let

$$\mathcal{T}_N = \left\{ \left[0, \frac{1}{2N^2}\right), \left[\frac{1}{2N^2}, \frac{2}{2N^2}\right), \ldots, \left[\frac{2N^2 - 1}{2N^2}, 1\right] \right\}$$

be a partition of $[0, 1]$. Let $\mathcal{P}_N$ be the set of probabilities with density functions that are constant on every interval in $\mathcal{T}_N$ and that assume only the two values 0 and 2. The cardinality of $\mathcal{P}_N$ is $q_N = \binom{2N^2}{N^2}$. Let $a_N = N^{-2}/c_0$, where

$$c_0 = \sum_{N=1}^{\infty} 1/N^2. \tag{28}$$

Our prior distribution will place probability $a_N/[2q_N]$ on each distribution in $\mathcal{P}_N$ for all N. The other half of the probability is distributed as follows. Let $\mathcal{U}_0$ be the set of all normal distributions with variance 1 and mean μ where $\mu \in [0, \sqrt{2}]$. Let $\Theta = \mu^2/2$. Let Θ have prior density

$$\frac{1}{c_1}\exp\left(-\frac{1}{\theta}\right) I_{(0,1)}(\theta) \quad \text{where } c_1 = \int_0^1 \exp\left(-\frac{1}{\theta}\right) d\theta. \tag{29}$$

Let $\mathscr{P}_0$ be the collection of all distributions on $[0, 1]$ obtained by transforming the distributions in $\mathscr{U}_0$ by the standard normal cumulative distribution function Φ. That is, each $P_\theta \in \mathscr{P}_0$ has density of the form $f_{P_\theta}(x) = \exp(-\theta + \sqrt{2\theta}\,\Phi^{-1}(x))$ for $0 < x < 1$ and some $\theta \in (0, 1)$.

Now, suppose that $f_0(x) \equiv 1$ is the uniform density on $[0, 1]$ and that $\{X_n\}_{n=1}^{\infty}$ are i.i.d. with this density. We will show that Assumption 1 holds but that the posterior is not consistent in the Hellinger distance metric. Using the well-known formula for Kullback–Leibler distance between Normals, it follows that $\mathscr{I}(P_0; P_\theta) = \theta$ and the prior for Θ puts positive probability on every neighborhood of 0, which means that π puts positive probability on every set N_ε. So Assumption 1 holds. Second, let $\mathscr{P}_* = \bigcup_{N=1}^{\infty} \mathscr{P}_N$. Each $P \in \mathscr{P}_*$ satisfies $d(P_0, P) = \sqrt{2 - \sqrt{2}}$, so $A_\varepsilon \cap \mathscr{P}_* = \varnothing$ for all $\varepsilon < \sqrt{2 - \sqrt{2}}$. We will prove that $\limsup_{n\to\infty} \pi(\mathscr{P}_* | x^n) = 1$ a.s., hence the conclusion to Theorem 1 fails.

If $N^2 \geq n$, then for all $x_1, \ldots, x_n$, there are at least $\binom{2N^2 - n}{N^2 - n}$ distributions $P \in \mathscr{P}_N$ such that $\prod_{i=1}^n f_P(x_i) = 2^n$. So, for every x^∞,

$$
\begin{aligned}
\int_{\mathscr{P}_*} \prod_{i=1}^n f_P(x_i)\, d\pi(P) &\geq 2^n \sum_{N \geq \sqrt{n}} \frac{a_N}{2q_N} \binom{2N^2 - n}{N^2 - n} \\
&= 2^{n-1} \sum_{N \geq \sqrt{n}} a_N \frac{\binom{2N^2 - n}{N^2 - n}}{\binom{2N^2}{N^2}} \\
&\geq \sum_{N \geq \sqrt{n}} \frac{a_N}{2} \left(1 - \frac{n-1}{N^2}\right)^n \\
&\geq \sum_{N \geq n} \frac{a_N}{2} \left(1 - \frac{n-1}{N^2}\right)^n \\
&\geq \left(1 - \frac{n-1}{n^2}\right)^n \sum_{N \geq n} \frac{a_N}{2} \\
&\geq \left(1 - \frac{n-1}{n^2}\right)^n \frac{1}{2c_0 n} \\
&\geq \frac{\exp(-2)}{2c_0 n},
\end{aligned}
\tag{30}
$$

for all but finitely many n, where c_0 is defined in (28). The inequality on the combinatorial terms follows from noting that the ratio of the two terms consists of the product of n fractions, each of which is greater than or equal to $(N^2 - n + 1)/(2N^2)$.

Next, consider the integral over $\mathscr{P}_0$. Recalling the definition of c_1 in (29), we have that for all x^∞ and all n

$$\int_{\mathscr{P}_0} \prod_{i=1}^n f_P(x_i)\, d\pi(P) = \frac{1}{2c_1}\int_0^1 \exp\left(-\frac{1}{\theta} - n\theta + \sqrt{2\theta}\sum_{i=1}^n \Phi^{-1}(x_i)\right) d\theta$$

$$\leq \frac{1}{2c_1}\exp\left(-2\sqrt{n} + \sqrt{2}\max\left\{0, \sum_{i=1}^n \Phi^{-1}(x_i)\right\}\right),$$

because $1/\theta + n\theta \geq 2\sqrt{n}$ for all $\theta \in (0,1)$ and all n. We know that with probability 1, $\max\{0, \sum_{i=1}^n \Phi^{-1}(x_i)\} = 0$ infinitely often according to the law of the iterated logarithm. It follows that, with probability 1,

$$\int_{\mathscr{P}_0} \prod_{i=1}^n f_P(x_i)\, d\pi(P) \leq \frac{1}{2c_1}\exp(-2\sqrt{n}), \quad \text{i.o.} \tag{31}$$

It follows from (30) and (31) that

$$P_0^\infty\left(\frac{\pi(\mathscr{P}_*|x^n)}{\pi(\mathscr{P}_0|x^n)} \geq \frac{\exp\left(2\left[\sqrt{n} - 1\right]\right)c_1}{c_0 n} \text{ i.o.}\right) = 1. \tag{32}$$

Since the fraction on the right-hand side of the inequality in (32) goes to infinity and $\pi(\mathscr{P}_0 \cup \mathscr{P}_*|x^n) = 1$, we have that

$$\limsup_{n\to\infty} \pi(\mathscr{P}_*|x^n) = 1 \quad \text{a.s. } [P_0^\infty].$$

4. Discussion. We have given two conditions that imply consistency of the posterior, and we have shown how to verify the conditions in a few examples. Geman and Hwang (1982) and Wong and Shen (1995) give results on consistency of sieve maximum likelihood estimators (MLE's). Some of their conditions are similar to ours. Our major difference between proving consistency for MLEs and posterior distributions using sieves is that for MLEs the sieve plays a crucial role in the definition of the MLE. That is, the MLE is the element of $\mathscr{F}_n$ that leads to the largest value of the likelihood function. If one changes to a different sieve, the sequence of MLEs will change. On the other hand, when using sieves to prove consistency of posterior distributions, only the prior distribution and likelihood affect the posterior. The particular sieve used to prove consistency is only a tool for the proof. Of course some sieves are easier to work with than others, but they do not figure in the computation of posterior probabilities.

We have not discussed rates of convergence in this paper. It is possible to compute rates of convergence by replacing the fixed ε in Theorem 1 with a decreasing sequence $\{\varepsilon_n\}_{n=1}^\infty$. See Shen and Wasserman (1998) and Ghoshal, Ghosh and van der Vaart (1998).

APPENDIX

Proof that Radon-Nikodym derivatives are jointly measurable. We use the following lemma in much of this paper. The notation comes from Section 2.

LEMMA 10. *For every $Q \in \mathscr{Q}$, there is a version f_Q of $dQ/d\lambda$ such that the function $g\colon \mathscr{Q} \times \mathscr{X} \to \mathbb{R}$ defined by $g(Q, x) = f_Q(x)$ is jointly measurable.*

PROOF. Since $(\mathscr{X}, \mathscr{B})$ is separable, there exists a countable collection $\mathscr{B}_0 = \{B_n\}_{n=1}^{\infty}$ of elements of $\mathscr{B}$ such that $\mathscr{B}$ is the smallest σ-field containing $\mathscr{B}_0$. Create a sequence $\{\rho_n\}_{n=1}^{\infty}$ of partitions of $\mathscr{X}$ as follows. Let $\rho_1 = \{B_1, B_1^C\}$. For $n > 1$, let ρ_n consist of the intersections of B_n and B_n^C with all of the elements of ρ_{n-1}. In this way we have that ρ_n is a refinement of ρ_{n-1} for all $n > 1$ and $\bigcup_{n=1}^{\infty} \rho_n$ generates the σ-field $\mathscr{B}$. Let m_n be the number of distinct nonempty sets in ρ_n and let $\rho_n = \{B_{n,1}, \ldots, B_{n,m_n}\}$.

For each $Q \in \mathscr{Q}$ each n and each $x \in \mathscr{X}$, define

$$f_{Q,n}(x) = \sum_{i=1}^{m_n} \frac{Q(B_{n,i})}{\lambda(B_{n,i})} I_{B_{n,i}}(x), \tag{33}$$

where we can let the fraction be 0 whenever $\lambda(B_{n,i}) = 0$. We will prove that $g_n\colon \mathscr{Q} \times \mathscr{X} \to \mathbb{R}$ defined by $g_n(Q, x) = f_{Q,n}(x)$ is jointly measurable, that $g' = \limsup_{n\to\infty} g_n$ is a jointly measurable and that $g'(Q, \cdot)$ is a version of $dQ/d\lambda$ for each Q.

First, define $h_{n,i}(Q) = Q(B_{n,i})$. This function is clearly continuous in the total variation topology and hence in the Hellinger topology. Since continuous functions are measurable, $h_{n,i}$ is measurable as a function from $\mathscr{Q}$ to $\mathbb{R}$. Since $h_{n,i}$ depends only on Q, it can also be considered as a measurable function from $\mathscr{Q} \times \mathscr{X}$ to $\mathbb{R}$. Since $B_{n,i} \in \mathscr{B}$, we know that $I_{B_{n,i}}$ is measurable as a function from $\mathscr{X}$ to $\mathbb{R}$ and can also be considered as a measurable function from $\mathscr{Q} \times \mathscr{X}$ to $\mathbb{R}$. Hence $h_{n,i} I_{B_{n,i}}$ is a measurable function from $\mathscr{Q} \times \mathscr{X}$ to $\mathbb{R}$. Since all $\lambda(B_{n,i})$ are constants, we have that each term in the sum in (33) is measurable, so the sum is measurable. Hence, $g_n(\cdot, \cdot)$ is jointly measurable. Which terms in the sum are 0 for all x and Q is predetermined by the values of $\lambda(B_{n,i})$, so this does not affect measurability. Since the lim sup of a sequence of measurable functions is measurable, it follows that $g' = \limsup_{n\to\infty} g_n$ is measurable.

All that remains is to show that $g'(Q, \cdot)$ is a version of $dQ/d\lambda$ for each Q. For each $Q \in \mathscr{Q}$, let f'_Q be an arbitrary version of $dQ/d\lambda$. Since we have assumed that λ is a probability measure we can think of f'_Q as a finite-mean random variable on the probability space $(\mathscr{X}, \mathscr{B}, \lambda)$. Let $\mathscr{B}_n$ stand for the finite σ-field generated by the partition ρ_n. It follows that $\mathscr{B}$ is the smallest σ-field containing $\bigcup_{n=1}^{\infty} \mathscr{B}_n$. If $A \in \mathscr{B}_n$, then A is a union of some of the $B_{n,i}$, say $B_{n,i_1}, \ldots, B_{n,i_k}$. It follows that

$$\int_A f'_Q(x)\, d\lambda(x) = Q\left(\bigcup_{j=1}^{k} B_{n,i_j}\right) = \sum_{j=1}^{k} Q(B_{n,i_j}) = \int_A g_n(Q, x)\, d\lambda(x).$$

It follows that $g_n(Q, \cdot) = \mathrm{E}(f'_Q | \mathscr{B}_n)$ for all n. Since $\{\mathscr{B}_n\}_{n=1}^{\infty}$ is an increasing sequence of σ-fields, we have that $g_n(Q, \cdot)$ is a martingale adapted to that sequence of σ-fields. Since $\mathrm{E}(|f'_Q|) = 1$, Lévy's theorem [Schervish (1995), Theorem B.118] applies and $g_n(Q, \cdot)$ converges a.s. $[\lambda]$ to $\mathrm{E}(f'_Q | \mathscr{B}_\infty)$, where $\mathscr{B}_\infty$ is the σ-field generated by $\bigcup_{n=1}^{\infty} \mathscr{B}_n$. We already saw that $\mathscr{B}_\infty = \mathscr{B}$. Since f'_Q

is $\mathscr{B}$-measurable, we have that $g_n(Q, \cdot)$ converges to f'_Q a.s. $[\lambda]$. This implies that $\limsup_{n\to\infty} g_n(Q, x) = f'_Q(x)$ a.s. $[\lambda]$, hence $g'(Q, \cdot)$ is a version of $dQ/d\lambda$. □

Measurability of $\mathscr{I}(P_0; P)$.

LEMMA 11. *$\mathscr{I}(P_0; P)$ is measurable as a function of P.*

PROOF. Let $g(P, x) = f_P(x)$ be the function constructed in the proof of Lemma 10. Define h: $\mathscr{P} \times \mathscr{X} \to \mathbb{R} \cup \{\pm\infty\}$ by

$$h(P, x) = f_0(x) \log \frac{f_0(x)}{g(P, x)}.$$

This h is also jointly measurable and $\mathscr{I}(P_0; P) = \int_{\mathscr{X}} h(P, x)\, d\lambda(x)$. Let $h = h^+ - h^-$ where $h^+ \geq 0$ and $h^- \geq 0$ are known as the positive and negative parts of h. Since $\mathscr{I}(P_0; P) \geq 0$ for all P and since

$$\begin{aligned} \mathscr{I}(P_0; P) &= \int_{\mathscr{X}} h(P, x)\, d\lambda(x) \\ &= \int_{\mathscr{X}} h^+(P, x)\, d\lambda(x) - \int_{\mathscr{X}} h^-(P, x)\, d\lambda(x), \end{aligned} \tag{34}$$

we must have that $\int_{\mathscr{X}} h^-(P, x)\, d\lambda(x) < \infty$ for all P. So, if we can prove that both integrals on the far right-hand side of (34) are measurable functions of P, we are done. The proofs are identical. Let h^* be a nonnegative measurable function of (P, x) and approximate it from below by a sequence $\{h_n\}_{n=1}^{\infty}$ of nonnegative simple functions, where each $h_n(P, x) = \sum_{i=1}^{m_n} a_{n,i} I_{A_{n,i}}(P, x)$. The monotone convergence theorem implies that for every P,

$$\lim_{n\to\infty} \int_{\mathscr{X}} h_n(P, x)\, d\lambda(x) = \int_{\mathscr{X}} h^*(P, x)\, d\lambda(x).$$

Since the limit of measurable functions is measurable, all we need to prove is that the integral of each h_n is measurable. But this will follow if the integral of each indicator function is measurable. This last fact is proven in Schervish (1995), Lemma A.61. □

PROOF OF LEMMA 9. Fix $\varepsilon \in (0, 1)$. Let $h = \log dP_0/d\lambda$ and let $A_b = \{x: |h| < b\}$, $A_b^+ = \{x: h \geq b\}$ and $A_b^- = \{x: h < -b\}$. Define $\rho(x) = h(x) I_{A_b}(x) + bI_{A_b^+}(x) - bI_{A_b^-}$. Choose b such that $\mathrm{E}(|h| I_{|h|\geq b}) < \varepsilon$ where the expectation is with respect to λ. Thus, $\mathrm{E}(|h - \rho|) < \varepsilon$ and by Markov's inequality, $P_0(|h| >$

$b) \leq \varepsilon/b$. Choose $g \in \mathscr{C}_{rb}$ such that $\int |\rho - g|\, d\lambda < \varepsilon^2 e^{-rb}$. Let $a = e^b$. Then,

$$\begin{aligned}\int |\rho - g|\, dP_0 &= \int |\rho - g| \frac{dP_0}{d\lambda}\, d\lambda \\ &\leq a \int |\rho - g|\, d\lambda + (b + rb) P_0\left(\frac{dP_0}{d\lambda} > a\right) \\ &< \varepsilon + (r+1) b P_0\left(\log \frac{dP_0}{dQ} > b\right) \\ &< \varepsilon + (r+1)\varepsilon.\end{aligned}$$

Let $c = \int e^g\, d\lambda$. Then

$$\begin{aligned}\mathscr{I}(P_0; P_g) &= \mathrm{E}\left(\log \frac{dP_0}{d\lambda} - g\right) + \log c \\ &\leq \mathrm{E}\left|\log \frac{dP_0}{d\lambda} - \rho\right| + \mathrm{E}|\rho - g| + \log c \\ &\leq (r+3)\varepsilon + \log c.\end{aligned}$$

Finally,

$$\begin{aligned}c &\leq \int e^{\rho + \varepsilon}\, d\lambda + e^{rb} \lambda(g - \rho > \varepsilon) \\ &\leq (1 + e^{-b}) e^{\varepsilon} + e^{rb} \varepsilon^{-1} \int |g - \rho|\, d\lambda \\ &\leq (1 + e^{-b}) e^{\varepsilon} + \varepsilon \\ &\leq 1 + e^{-(b-1)} + e\varepsilon.\end{aligned}$$

For large b, $\log c \leq e^{-(b-1)} + e\varepsilon \leq 3\varepsilon$ so that $\mathscr{I}(P_0; P_g) < (r+6)\varepsilon$. □

Acknowledgments. The authors thank R. V. Ramamoorthi and J. K. Ghosh for helpful comments on an earlier draft. Also, the authors thank the referee for carefully reading the manuscript several times, for suggesting many improvements and for suggesting the proof of Lemma 10.

REFERENCES

ALEXANDER, K. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.

BARRON, A. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6* (J. Bernardo, J. Berger, A. Dawid and A. Smith, eds.) 27–52. Oxford, New York.

BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. Statistics, Univ. Illinois, Champaign.

DIACONIS, P. and FREEDMAN, D. A. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–26.

DIACONIS, P. and FREEDMAN, D. A. (1997). Consistency of Bayes estimates for nonparametric regression: a review. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.) 157–165. Springer, New York.

DIACONIS, P. and FREEDMAN, D. A. (1998). Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli* **4** 411–444.

DOOB, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications* 23–27. Colloques Internationaux du Centre National de la Recherche Scientifique, Paris.

FABIUS, J. (1964). Asymptotic behavior of Bayes estimates. *Ann. Math. Statist.* **35** 846–856.

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

GEMAN, S. and HWANG, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GHOSHAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999A). Consistency issues in Bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time series* (S. Ghosh, ed.) 639–667. Dekker, New York.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999B). Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference*. To appear.

GHOSHAL, S., GHOSH, J. and VAN DER VAART, A. (1998). Convergence rates of posterior distribution. Technical Report 504, Dept. Mathematics, Free University, Amsterdam.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

KRAFT, C. H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probab.* **1** 385–388.

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235.

LAVINE, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **22** 1161–1176.

LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83** 509–516.

LENK, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78** 531–543.

LEONARD, T. (1978). Density estimation, stochastic processes, and prior information. *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

MAULDIN, R. D., SUDDERTH, W. D. and WILLIAMS, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.* **20** 1203–1221.

POLLARD, D. (1991). Bracketing methods in statistics and econometrics. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics*: *Proceedings of the Fifth International Symposium in Economic Theory and Econometrics* (W. A. Barnett, J. Powell and G. E. Tauchen, eds.) 337–355. Cambridge Univ. Press.

SCHERVISH, M. J. (1995). *Theory of Statistics*. Springer, New York.

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

SHEN, X. and WASSERMAN, L. (1998). Rates of convergence of posterior distribution. Technical Report 678, Dept. Statistics, Carnegie Mellon Univ.

VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

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

A. BARRON
DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520
E-MAIL: barron@stat.yale.edu

M. J. SCHERVISH
L. WASSERMAN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
E-MAIL: mark@stat.cmu.edu,
larry@stat.cmu.edu