The Annals of Statistics
1983, Vol. 11, No. 4, 1109–1118

ON INCONSISTENT BAYES ESTIMATES IN THE DISCRETE CASE

BY DAVID FREEDMAN[1] AND PERSI DIACONIS[2]

University of California, Berkeley and Stanford University

Consider sampling from an unknown probability distribution on the integers. With a tail-free prior, the posterior distribution is consistent. With a mixture of a tail-free prior and a point mass, however, the posterior may be inconsistent. This is likewise true for a countable mixture of tail-free priors. Similar results are given for Dirichlet priors.

1. Introduction. Dirichlet and tail free priors on the set of all probabilities were introduced to insure consistency of the corresponding Bayes estimates. The examples in this paper show that taking mixtures of such priors can lead to inconsistent estimators. We proceed to definitions and a historical review.

Let I be the positive integers. The parameter space Λ is the set of all probabilities on I. Write $\lambda(i)$ for the mass which λ assigns to $i \in I$. Let λ^∞ be product measure on I^∞, making the coordinates X_i independent with common distribution λ. Consider estimating λ from the data $X_1, \cdots, X_n$, in a Bayesian framework.

A *prior* μ is a probability on Λ; it induces a probability P_μ on $\Lambda \times I^\infty$ by the rule

$$P_\mu(A \times B) = \int_A \lambda^\infty(B)\mu(d\lambda).$$

This P_μ is the joint distribution of the parameter λ and the data $X_1, X_2, \cdots$, for someone holding the prior opinion μ. The *posterior* $\mu^{(n)}$ on Λ is the conditional distribution of λ given $X_1, \cdots, X_n$, computed from P_μ:

$$\mu^{(n)} = P_\mu\{A \mid X_1, \cdots, X_n\} = \frac{P_\mu\{A \text{ and } X_1, \cdots, X_n\}}{P_\mu\{X_1, \cdots, X_n\}}.$$

To avoid trivial complications, we suppose throughout that μ assigns positive mass to all weakly open subsets of Λ, so $P_\mu\{X_1, \cdots, X_n\} > 0$. The notation may be perplexing: $P_\mu\{X_1, \cdots, X_n\}$ is a random variable whose value on the set $\{X_1 = i_1, \cdots, X_n = i_n\}$ is the P_μ probability of that set.

The pair (λ, μ) is *consistent* if $\mu^{(n)} \to \delta_\lambda$ weakly as $n \to \infty$, with λ^∞-probability one; δ_λ is a point mass at λ. The prior μ is consistent if (λ, μ) is consistent for all λ. Informally, a Bayesian with the prior μ who samples repeatedly from λ will discover that fact.

Consistency has been investigated by Doob (1949), LeCam (1953), Freedman (1963), Schwartz (1965) and others. Doob (1949) showed that (λ, μ) is consistent for μ-almost all λ. LeCam (1953) and Schwartz (1965) proved consistency under strong regularity conditions, typically including the requirement that Λ be a finite-dimensional Euclidean space, ruling out nonparametric problems like the one considered here. When I is replaced by a finite set, Freedman (1963) proves that (λ, μ) is consistent if and only if λ is in the support of the prior μ and gives counterexamples to show that infinite-dimensional problems are basically different. For example, there is a prior μ on Λ which puts positive mass in every

Received January 1983; revised May 1983.

[1] Research partially supported by NSF Grant MCS80-02535.

[2] Research partially supported by NSF Grant MCS80-24649.

AMS 1980 *subject classifications.* Primary, 62A15; secondary, 62E20.

Key words and phrases. Consistent estimators, Dirichlet priors, Tail free priors, Bayes' rule, Inconsistent Bayes estimators.

neighborhood of $G_{1/4}$—the geometric distribution on I with parameter ¼. However, for almost all iid sequences drawn from $G_{1/4}$, the posterior converges to a point mass at $G_{3/4}$. Freedman (1965) showed that this behavior is generic: almost all pairs (λ, μ) are inconsistent in the sense of category. To insure consistency, Freedman (1963) introduced the class of Dirichlet and tail-free priors on Λ. For such priors (λ, μ) is consistent for all λ.

Dirichlet and tail-free priors have seen increasing use in recent years. Ferguson (1974) contains a review of the literature. Antoniak (1974) has discussed the need for mixtures of Dirichlet priors in routine problems. Good (1978), Dalal (1978), and Dalal and Hall (1980) suggest mixtures of Dirichlet priors as a rich family, dense in the set of priors, for routine use. The examples below suggest that caution is called for: countable mixtures of Dirichlet or tail free priors can lead to inconsistency. On the other hand, finite mixtures of consistent priors are consistent, and finite mixtures of Dirichlet priors are dense in the class of all priors. Diaconis and Freedman (1983) show that mixtures of Dirichlet priors can lead to inconsistent estimators in the problem of estimating a location parameter in a continuous setting. In general, it is not known when (λ, μ) is consistent, or when μ is consistent for all λ.

Define μ to be *tail-free* if the following random variables, called *cuts*, are independent under μ:

$$\lambda(1), \quad \lambda(2)/[1 - \lambda(1)], \quad \lambda(3)/[1 - \lambda(1) - \lambda(2)], \quad \cdots.$$

Informally, λ is chosen from μ by "stick-breaking:" start with a stick of unit length, break off a random length for $\lambda(1)$. Independently break off a piece $\lambda(2)$ from the remaining piece of length $1 - \lambda(1)$, etc. Dirichlet priors are tail free, the cuts having appropriate beta distributions. See Freedman (1963) or Ferguson (1974) for further details. The tail free prior with uniform cuts is consistent for any λ. Our first result shows that a mixture of this prior and a point mass can be inconsistent.

PROPOSITION 1. *Let ν be tail-free with all cuts uniformly distributed over* [0, 1]. *Let* $1 > \varepsilon > 0$. *There are distinct probabilities θ and ϕ in Λ such that the prior*

$$\mu = (1 - \varepsilon)\nu + \varepsilon\delta_\phi$$

makes (θ, μ) inconsistent: $\mu^{(n)} \to \delta_\phi$ with θ^∞ probability one.

In this example, if a Bayesian with prior μ observes a sample from θ, the posterior piles up near the "foil" $\phi \neq \theta$. By construction, θ and ϕ agree at all but finitely many places. However they are long-tailed—indeed both have infinite entropy. The prior ν, while supported on all of Λ, concentrates on short-tailed probabilities. When sampling from θ the "evidence" from the tail of the distribution overwhelms the evidence from the center of the distribution, causing convergence to ϕ. More specifically, under θ^∞, the maximum M_n of $X_1, X_2, \cdots, X_n$ is so large that it cannot reasonably come from λ's on which ν concentrates. This tilts the balance to ϕ. The construction in Proposition 1 is like that used by Bahadur (1958, 1971) to get inconsistent maximum likelihood estimators.

The prior used in Proposition 1 is an infinite dimensional version of a prior used by Bayesians to test hypotheses: a mixture of a point mass and a continuous prior. For example, when considering a sample from a multinomial distribution with unknown parameter vector λ, to test if $\lambda = \phi$, some Bayesians use a 50-50 mixture of a uniform prior and a point mass at ϕ. The test is based on the posterior mass at ϕ. For discussion, see Jeffreys (1967, Chapter 5).

Approximating δ_ϕ by a sequence of tail-free priors μ_i leads to the next result.

PROPOSITION 2. *There are tail-free priors μ_i and probabilities θ, ϕ in Λ such that setting $\mu = \sum_{i=1}^\infty \mu_j/j(j+1)$, the pair (θ, μ) is inconsistent: $\mu^{(n)} \to \delta_\phi$ with θ^∞ probability one.*

The inconsistent prior in Proposition 2 is a mixture of consistent priors. In Section 5,

it is shown that mixtures of Dirichlet priors can be inconsistent. Some positive results are also given: a mixture of Dirichlet priors is consistent if the mass of the parameter measures is uniformly bounded.

Tom Ferguson has told us about a nice example of inconsistency when I is the unit interval. Let λ be Lebesgue measure on I. Let the prior $\mu = \frac{1}{2} D(\lambda) + \frac{1}{2} \delta\lambda$. If the sampling is done from any continuous distribution, the posterior converges to point mass at λ. The basic reason is that $D(\lambda)$ concentrates on discrete (densely supported) distributions.

2. Some estimates. The same ϕ and θ can be used in both proposition, as follows. Let

$$\phi(i) = \theta(i) = 1/i(\log i)^2 \quad \text{for} \quad i = 10, 11, \cdots.$$

Choose $\phi(i) > 0$ and $\theta(i) > 0$ for $i = 1, \cdots, 9$ so $\sum_{i=1}^{\infty} \phi(i) = \sum_{i=1}^{\infty} \theta(i) = 1$. It is required that $\phi(i) \neq \theta(i)$ for some $i = 1, \cdots, 9$. Also required is that ϕ be close to θ in the sense of relative entropy.

$$-1 < \sum_{i=1}^{\infty} \theta(i) \log \frac{\phi(i)}{\theta(i)} < 0. \tag{1}$$

Some facts about θ and ϕ will now be developed. The estimates are crude, but sufficient. The first result shows that $M_n = \max\{X_1, \cdots, X_n\}$ is around e^n, with high θ^∞-probability.

LEMMA 1. *If $n \geq 25$, then*

$$M_n > \exp\left(\frac{n}{4 \log n}\right)$$

except on a set of θ^∞-probability at most $1/n^3$.

PROOF. Clearly,

$$\frac{1}{\log(t+1)} \leq \theta^\infty\{X_j \geq t\} \leq \frac{1}{\log(t-1)} \quad \text{for} \quad t \geq 10. \tag{2}$$

If $n \geq 25$, then $\exp(n/3 \log n) \geq 11$, so

$$\theta^\infty\left\{M_n < -1 + \exp\left(\frac{n}{3 \log n}\right)\right\} \leq \left(1 - \frac{3 \log n}{n}\right)^n \leq \frac{1}{n^3}.$$

Also, for $n \geq 25$,

$$\exp\left(\frac{n}{4 \log n}\right) < -1 + \exp\left(\frac{n}{3 \log n}\right). \quad \square$$

The next lemma shows that ϕ^∞ cannot be too much smaller than θ^∞.

LEMMA 2. *If $n \geq 1$,*

$$\phi^\infty\{X_1, \cdots, X_n\} > \exp(-n^3)\theta^\infty\{X_1, \cdots, X_n\}$$

except on a set of θ^∞-probability at most $1/n^2$.

PROOF. Clearly,

$$-\log \frac{\phi^\infty\{X_1, \cdots, X_n\}}{\theta^\infty\{X_1, \cdots, X_n\}} = \sum_{j=1}^{n} \left[-\log \frac{\phi(X_j)}{\theta(X_j)}\right]$$

is, relative to θ^∞, the sum of n independent, identically distributed random variables whose common mean is less than one, by condition (1). Now use Markov's inequality. $\square$

The next lemma puts a floor under $\theta^\infty\{X_1, \cdots, X_n\}$.

LEMMA 3. *If* $n \geq 3$,

$$\theta^\infty\{X_1, \cdots, X_n\} > \exp(-n^4)$$

except on a set of θ^∞*-probability at most* $2/n^2$.

PROOF. Let $\xi_j = -\log \theta(X_j)$. If $t \geq 10$, say, then

$$\begin{aligned}\theta^\infty\{\xi_j \geq t\} &= \sum_i \left\{\frac{1}{i(\log i)^2}: \log i + 2 \log\log i \geq t\right\} \\ &\leq \sum_i \left\{\frac{1}{i(\log i)^2}: \log i \geq \frac{1}{2}t + 1\right\} \\ &\leq 1/\log(e^{(1/2)t+1} - 1) \quad \text{by (2)} \\ &< 2/t.\end{aligned}$$

Now $n \geq 3$ makes $n^3 \geq 10$, so

$$\theta^\infty\{\xi_j < n^3 \text{ for } j = 1, \cdots, n\} > (1 - 2n^{-3})^n > 1 - 2n^{-2}.$$

But $\xi_j < n^3$ for $j = 1, \cdots, n$ entails $\sum_{j=1}^n \xi_j < n^4$, i.e., $\theta^\infty\{X_1, \cdots, X_n\} > \exp(-n^4)$. □

LEMMA 4. *If* $n \geq 3$,

$$\phi^\infty\{X_1, \cdots, X_n\} > \exp(-2n^4)$$

except on a set of θ^∞*-probability at most* $3/n^2$.

PROOF. Combine Lemmas 2 and 3. □

Turn now to properties of the tail-free priors. The next result shows that tail-free priors with uniform cuts concentrate on short-tailed probabilities. It is the only estimate needed on tail-free priors.

LEMMA 5. *Suppose* ν *is tail-free, and past some index* k_0*, the cuts are uniformly distributed over* [0, 1]. *Then*

$$\int \lambda(k)\nu(d\lambda) \leq \left(\frac{1}{2}\right)^{k-k_0}.$$

PROOF. If $k \leq k_0$, the assertion is trivial. If $k > k_0$, then

$$\lambda(k) = [1 - \lambda(1) - \cdots - \lambda(k_0)]\{\textstyle\prod_{j=k_0+1}^{k-1}[1 - C_j(\lambda)]\}C_k(\lambda)$$

where C_j is the cut at index j. □

LEMMA 6. *Let* ν *be as in Lemma* 5. *If* $n \geq 25$ *and* $M_n > \exp\left(\dfrac{n}{4 \log n}\right)$ then

$$P_\nu\{X_1, \cdots, X_n\} < 2^{k_0} 2^{-\exp(n/4 \log n)}$$

NOTE. As Lemma 1 shows, $M_n > \exp(n/4 \log n)$ except on a set of θ^∞-probability at most $1/n^3$.

PROOF. Focus on the big M_n and use Lemma 5:

$$P_\nu\{X_1, \cdots, X_n\} = \int \lambda^\infty\{X_1, \cdots, X_n\}\, \nu(d\lambda) \leq \int \lambda(M_n)\nu(d\lambda) \leq 2^{k_0}2^{-\exp(n/4\log n)}. \quad \square$$

3. The first construction. Proposition 1 will now be proved. Fix a weak neighborhood N of the foil ϕ. It must be shown that $\mu^{(n)}(N) \to 1$ with θ^∞-probability one, where $\mu = \varepsilon\delta_\phi + (1 - \varepsilon)\nu$ and ν is tail-free with uniform cuts. Now

$$\mu^{(n)}(N) = P_\mu\{N \mid X_1, \cdots, X_n\} = \frac{P_\mu\{N \text{ and } X_1, \cdots, X_n\}}{P_\mu\{X_1, \cdots, X_n\}}.$$

The numerator is

$$\varepsilon\phi^\infty\{X_1, \cdots, X_n\} + (1 - \varepsilon)P_\nu\{N \text{ and } X_1, \cdots, X_n\}.$$

The denominator is

$$\varepsilon\phi^\infty\{X_1, \cdots, X_n\} + (1 - \varepsilon)P_\nu\{X_1, \cdots, X_n\}.$$

It suffices to show that as $n \to \infty$, with θ^∞-probability one,

$$\phi^\infty\{X_1, \cdots, X_n\}/P_\nu\{X_1, \cdots, X_n\} \to \infty.$$

Let $n \geq 25$. Except for a set of θ^∞-probability at most $3/n^2$,

$$\phi^\infty\{X_1, \cdots, X_n\} > \exp(-2n^4)$$

by Lemma 4; except for another set of θ^∞-probability at most $1/n^3$,

$$P_\nu\{X_1, \cdots, X_n\} < 2^{-\exp(n/4\log n)}$$

by Lemma 5, with $k_0 = 0$. Clearly,

$$2^{-\exp(n/4\log n)} \ll e^{-2n^4}. \quad \square$$

4. The second construction. The objects to be chosen are

N_k, a weak neighborhood of ϕ shrinking to ϕ,
S_k, a finite set of strings of positive integers of length k, with $\theta^\infty\{S_k\} > 1 - 1/k^2$,
μ_k, a tail-free prior whose cuts are, from index $k + 1$ on, uniformly distributed and $\mu_k \to \delta_\phi$ rapidly as $k \to \infty$.

The N_k may be chosen arbitrarily, subject to the given conditions; S_k may also be chosen arbitrarily, subject to the conditions given. The choice of μ_k is inductive. Fix $k \geq 1$, and suppose μ_j chosen for $j < k$; this is vacuous if $k = 1$. Now choose μ_k so close to δ_ϕ that the following conditions are satisfied:

$$P_{\mu_k}\{N_j \mid X_1, \cdots, X_j\} > 1 - (1/j) \quad \text{on} \quad S_j \quad \text{for} \quad j \leq k \tag{3}$$

$$P_{\mu_k}\{N_k \text{ and } X_1, \cdots, X_k\} > \tfrac{1}{2}\phi^\infty\{X_1, \cdots, X_k\} \quad \text{on} \quad S_k. \tag{4}$$

Condition (3) is feasible because S_j is finite; for any string $x_1, \cdots, x_j$, the numerator and denominator of

$$\frac{P_{\mu_k}\{N_j \text{ and } X_1 = x_1, \cdots, X_j = x_j\}}{P_{\mu_k}\{X_1 = x_1, \cdots, X_j = x_j\}}$$

are both nearly $\phi^\infty\{X_1 = x_1, \cdots, X_j = x_j\}$ because $\phi \in N_j$ and μ_k is nearly δ_ϕ. Likewise for (4). To get μ_k near δ_ϕ from the point of view of (3) and (4), make the first k cuts in μ_k practically equal to the corresponding cuts in ϕ; the remaining cuts are to be uniformly

distributed over [0, 1]. The first k cuts can have continuous, strictly positive densities: but these densities will be highly concentrated.

This completes the construction, and Proposition 2 must now be proven. It is helpful to rewrite (3), interchanging j and k:

(5) $$P_{\mu_j}\{N_k \mid X_1, \cdots, X_k\} > 1 - (1/k) \quad \text{on} \quad S_k \quad \text{for all} \quad j \geq k.$$

The following two observations will be helpful:

(6) For $k \geq 1000$, simultaneously for all $j < k$,

$$P_{\mu_j}\{X_1, \cdots, X_k\} < \frac{1}{k}\frac{1}{2}\frac{1}{k(k+1)}\exp(-2k^4)$$

except on a set Q_k of θ^∞-probability at most $1/k^3$.

(7) For $k \geq 3$, $\phi^\infty\{X_1, \cdots, X_k\} > \exp(-2k^4)$ except on a set R_k of θ^∞-probability at most $3/k^2$.

Relation (6) follows from Lemmas 1 and 6, with k for n and μ_j for ν. Let Q_k be the event that $M_k \leq \exp(k/4 \log k)$. As Lemma 1 shows, $\theta^\infty(Q_k) \leq 1/k^3$. Off Q_k, by Lemma 6,

$$P_{\mu_i}\{X_1, \cdots, X_k\} < 2^j 2^{-\exp(k/4\log k)} < 2^k 2^{-\exp(k/4\log k)} < \frac{1}{k}\frac{1}{2}\frac{1}{k(k+1)}\exp(-2k^4)$$

provided k is large; 1000 will do. This completes the proof of (6), and relation (7) is just Lemma 4.

The next step is to estimate the ratio

(8) $$P_\mu\{N \mid X_1, \cdots, X_k\} = \frac{P_\mu\{N \text{ and } X_1, \cdots, X_k\}}{P_\mu\{X_1, \cdots, X_k\}}$$

for $N = N_k$, and $k \geq 1000$. As will be seen, this ratio exceeds $(k-1)/(k+1)$ except on a set of θ^∞-probability at most $3/k^2$. Since N_k shrinks to δ_ϕ, the proposition follows.

The denominator in (8) is $a_k + b_k$, where

$$a_k = \sum_{j=1}^{k-1} \frac{1}{j(j+1)} P_{\mu_j}\{X_1, \cdots, X_k\}, \quad b_k = \sum_{j=k}^{\infty} \frac{1}{j(j+1)} P_{\mu_j}\{X_1, \cdots, X_k\}.$$

Likewise, the numerator is $a'_k + b'_k$, where

$$a'_k = \sum_{j=1}^{k-1} \frac{1}{j(j+1)} P_{\mu_j}\{N_k \text{ and } X_1, \cdots, X_k\},$$

$$b'_k = \sum_{j=k}^{\infty} \frac{1}{j(j+1)} P_{\mu_j}\{N_k \text{ and } X_1, \cdots, X_k\}.$$

The terms a_k and a'_k are negligible:

(9) On $S_k - Q_k - R_k$, a set of θ^∞-probability at least $1\text{–}5/k^2$,

$$P_{\mu_j}\{X_1, \cdots, X_k\} < \frac{1}{k}\frac{1}{k(k+1)} P_{\mu_k}\{N_k \text{ and } X_1, \cdots, X_k\}.$$

Indeed, by (4) and (7), on $S_k - R_k$

$$\frac{1}{k(k+1)} P_{\mu_k}\{N_k \text{ and } X_1, \cdots, X_k\} > \frac{1}{2}\frac{1}{k(k+1)}\exp(-2k^4).$$

By (6), except on Q_k,

$$P_{\mu_j}\{X_1, \cdots, X_k\} < \frac{1}{k}\frac{1}{2}\frac{1}{k(k+1)}\exp(-2k^4).$$

Since $\theta^\infty\{Q_k\} < 1/k^3$ and $\theta^\infty\{R_k\} < 3/k^2$, the proof of (9) is complete.

In particular, $a_k < b_k/k$ on $S_k - R_k$, so

$$P_\mu\{N_k \mid X_1, \cdots, X_{n_k}\} = \frac{a'_k + b'_k}{a_k + b_k} > \frac{b'_k}{a_k + b_k} > \frac{k}{k+1}\frac{b'_k}{b_k}$$

$$> \frac{k}{k+1} \inf_{j \geq k} \frac{P_{\mu_j}\{N_k \text{ and } X_1, \cdots, X_k\}}{P_{\mu_j}\{X_1, \cdots, X_k\}} > \frac{k-1}{k+1} \quad \text{by (5).} \quad \square$$

5. Dirichlet priors. This section indicates the modifications needed to obtain Propositions 1 and 2 for Dirichlet priors. A Dirichlet prior μ is characterized by a measure α on the positive integers, of arbitrary finite mass $\| \alpha \|$. The expected value of μ is $\alpha/\| \alpha \|$ in the sense

$$\int \lambda(i)\mu(d\lambda) = \alpha(i)/\| \alpha \|. \tag{10}$$

The Dirichlet is a tail-free prior with independent beta cuts:

$$\lambda(1) \text{ is } B[\alpha(1), \| \alpha \| - \alpha(1)], \quad \lambda(2)/1 - \lambda(1) \text{ is } B[\alpha(2), \| \alpha \| - \alpha(1) - \alpha(2)], \cdots.$$

The variability of μ around its expected value decreases as $\| \alpha \|$ increases. For example, the variance of $\lambda(i)$ under μ is

$$\frac{\alpha(i)(\| \alpha \| - \alpha(i))}{\| \alpha \|^2(1 + \| \alpha \|)}. \tag{11}$$

Dirichlet examples for Propositions 1 and 2 may be constructed as follows: For Proposition 1, choose $\alpha(i) = 2^{-i}$, with θ and ϕ as given in Section 2. The argument is essentially as given in Section 3; Lemma 5 follows from (10).

For the second construction, a sequence of α's is needed; index them by j so $\alpha_j(i)$ is the mass that the jth measure α assigns to the integer i. Again, θ and ϕ are as given in Section 2. Let $\| \alpha_j \|$ be large and $\alpha_j(i)/\| \alpha_j \| = \phi(i)$ for $i = 1, 2, \cdots, j$. For $i > j$, let $\alpha_j(i)/\| \alpha_j \| = 2^{-i}$. The rest of the argument is as given in Section 4.

The examples just presented have parameter measures with exponential tails. This is not crucial. Let ν be Dirichlet with parameter α, where $\alpha(i) = 1/i^2$ is the mass assigned to the integer $i \geq 1$. To get the analogue of Proposition 1, take

$$\theta(i) = \phi(i) = 1/i(\log i)(\log \log i)^2 \quad \text{for all large } i.$$

Assume (1) is satisfied. Let $X_1, X_2, \cdots$ be independent with common distribution θ and $M_n = \max\{X_1, \cdots, X_n\}$. As before,

$$P_\nu(X_1, \cdots, X_n) \leq \int \lambda(M_n)\nu(d\lambda) = \frac{\| \alpha \|}{M_n^2}.$$

It will be seen that $(1/n) \log \log M_n$ is asymptotically distributed as $1/V$, where V has an exponential distribution. Thus, M_n is of size $\exp[\exp(n/V)]$. On the other hand, with high probability,

$$\phi^\infty\{X_1, \cdots, X_n\} \geq \phi(M_n)^{1.1} \geq 1/M_n^{1.2} \gg \| \alpha \|/M_n^2$$

by an argument to be sketched. The main idea is that $\theta^\infty\{X_j > t\} \approx 1/\log \log t$ as $t \to \infty$, in the sense that the ratio converges to one. So X_j can be replaced for present purposes by $\exp[\exp(1/U_j)]$, the U_j being independent and uniformly distributed over $[0, 1]$. Let $U_{(1)}, \cdots, U_{(n)}$ be the order statistics. As usual, these can be realized as

$$V_1/S, (V_1 + V_2)/S, \cdots, (V_1 + V_2 + \cdots + V_n)/S$$

where $V_1, V_2, \cdots, V_n, V_{n+1}$ are independent exponential random variables with sum S.

Now $S \approx n$ and in effect M_n is $\exp[\exp(1/U_{(1)})] = \exp[\exp(S/V_1)]$, explaining the assertion about $(1/n)\log\log M_n$. Also

$$\sum_{j=1}^n \exp(1/U_j) = \sum_{j=1}^n \exp(1/U_{(j)}) \leq 1.1 \exp(1/U_{(1)})$$

with high probability. Now $\sum_{j=1}^n \log \phi(X_j)$ behaves like $\sum_{j=1}^n \exp(1/U_j)$ and with high probability is bounded above by $-1.1 \log \phi(M_n)$, that is,

$$\phi^\infty\{X_1, \cdots, X_n\} \geq \phi(M_n)^{1.1},$$

as required. The argument just sketched gives $\mu^{(n)} \to \delta_\phi$ in θ^∞ probability. We do not know if the convergence is a.e. A similar modification gives the analogue of Proposition 2, again in probability.

To get inconsistent behavior with a mixture of Dirichlet priors, the masses of the parameter measures must tend to infinity. On the other hand, if the masses stay bounded, the Bayes' estimates are consistent. A sharp version of this result will now be presented. For simplicity, discrete mixtures and discrete data are considered first.

Let λ and λ' be probabilities on the positive integers. Let

$$|\lambda - \lambda'|^2 = \sum_{i=1}^\infty [\lambda(i) - \lambda'(i)]^2. \tag{12}$$

Let μ be Dirichlet with parameter α, abbreviated $\mu = D(\alpha)$. Let $\|\alpha\|$ be the mass of α, and $\bar{\alpha} = \alpha/\|\alpha\|$, so

$$\int \lambda(i)\mu(d\lambda) = \bar{\alpha}(i).$$

The next two bounds are straightforward.

LEMMA 7. *If $\mu = D(\alpha)$, then $\int |\lambda - \bar{\alpha}|^2 \mu(d\lambda) \leq 1/1 + \|\alpha\|$.*

PROOF. Use (10) and (11). □

LEMMA 8. *If $\mu = D(\alpha + \beta)$, then*

$$\int |\lambda - \bar{\beta}|^2 \mu(d\lambda) \leq \frac{2}{1 + \|\beta\|} + \frac{8\|\alpha\|^2}{\|\alpha\|^2 + \|\beta\|^2}.$$

PROOF. Plainly,

$$|\lambda - \bar{\beta}|^2 \leq 2|\lambda - \overline{\alpha + \beta}|^2 + 2|\overline{\alpha + \beta} - \bar{\beta}|^2.$$

The μ-integral of the first term may be estimated by Lemma 7, and is at most $2/1 + \|\alpha + \beta\| \leq 2/1 + \|\beta\|$. For the second term,

$$\begin{aligned} |\overline{\alpha + \beta} - \bar{\beta}|^2 &= \sum_i \left[\frac{\alpha(i) + \beta(i)}{\|\alpha + \beta\|} - \frac{\beta(i)}{\|\beta\|}\right]^2 \\ &\leq 2\sum_i \frac{\alpha^2(i)}{\|\alpha + \beta\|^2} + 2\sum_i \beta^2(i)\left(\frac{1}{\|\alpha + \beta\|} - \frac{1}{\|\beta\|}\right)^2 \\ &\leq \frac{4\|\alpha\|^2}{\|\alpha + \beta\|^2} \end{aligned}$$

because $\sum_i \alpha^2(i) \leq \|\alpha\|^2$, and likewise for β. Of course $\|\alpha + \beta\|^2 \geq \|\alpha\|^2 + \|\beta\|^2$. □

PROPOSITION 3. *For each j, let α_j be a measure on the positive integers with finite mass $\|\alpha\| \leq F < \infty$. Suppose that α_j assigns positive mass $\alpha_j(i)$ to each positive integer i. Let $\pi(j) > 0$ with $\sum_j \pi(j) = 1$. Let $\mu_j = D(\alpha_j)$ and $\mu = \sum_j \pi(j)\mu_j$. Let θ be any probability on the*

positive integers. Then (θ, μ) *is consistent. Indeed, let* $X_1, X_2, \cdots$ *be independent with common distribution* θ. *Let* d_n *be the empirical distribution*

$$d_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k} \tag{13}$$

where δ_x *is point mass at* x. *Then*

$$\int |\lambda - d_n|^2 \mu^{(n)}(d\lambda) \leq \frac{2}{1+n} + \frac{8F^2}{F^2 + n^2}. \tag{14}$$

Thus, the posterior concentrates near the empirical.

PROOF. By Bayes' rule,

$$\mu^{(n)} = \sum_j w_j \mu_j^{(n)}$$

where the weights w_j sum to 1 but are data-dependent:

$$w_j = \pi(j) P_{\mu_j}\{\Lambda \text{ and } X_1, \cdots, X_n\} / \sum_j \pi(j) P_{\mu_j}\{\Lambda \text{ and } X_1, \cdots, X_n\}.$$

Thus,

$$\int |\lambda - d_n|^2 \mu^{(n)}(d\lambda) = \sum_j w_j \int |\lambda - d_n|^2 \mu_j^{(n)}(d\lambda).$$

As usual,

$$\mu_j^{(n)} = D(\alpha_j + n d_n).$$

Now use Lemma 8 on each summand, with α_j for α and nd_n for β. □

The argument extends to continuous mixtures of Dirichlet priors defined for general spaces: if $\| \alpha_\cdot \|$ is bounded the mixture is consistent. Continuous mixtures are easily handled by standard arguments: j is replaced by a general index and π by a general probability. Continuous data are harder to handle, because the L_2 norm on probabilities defined in (12) does not generalize. One way to proceed is by discretization. Suppose the observation space I is a general separable measure space. Let $\mathscr{A} = \cup_{i=1}^{\infty} A_i$ be a partition of I. Let

$$|\lambda - \lambda'|_{\mathscr{A}} = \{\sum_{i=1}^{\infty} [\lambda(A_i) - \lambda'(A_i)]^2\}^{1/2}.$$

This defines a semi-norm on the probabilities, and (14) holds with $|\ |$ replaced by $|\ |_{\mathscr{A}}$. The argument is completed by considering a generating sequence of partitions $\mathscr{A}$.

REMARK. While the class of finite mixtures of Dirichlet priors $D(\alpha)$ is dense in all priors, the class of mixtures with $\| \alpha \| \leq F$ is not dense.

REFERENCES

[1] ANTONIAK, C. (1974). Mixtures of Dirichlet processes with application to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

[2] BAHADUR, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā* **20** 207–210.

[3] BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

[4] DALAL, S. R. (1978). On the adequacy of mixtures of Dirichlet processes. *Sankhyā* **40** 185–191.

[5] DALAL, S. R. and HALL, G. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.* **8** 664–672.

[6] DIACONIS, P. and FREEDMAN, D. (1983). On inconsistent Bayes estimates of location in the continuous case. Department of Statistics, Stanford University, Technical Report #200.

[7] DOOB, J. L. (1949). Application of the theory of martingales. *Coll. Int. du CNRS* 22–28.

[8] FERGUSON, T. (1974). Prior distributions on the space of all probability measures. *Ann. Statist.* **2** 615–629.
[9] FREEDMAN, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403.
[10] FREEDMAN, D. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **35** 454–456.
[11] GOOD, I. J. (1978). Review of Ferguson, Thomas S., "Prior distributions on spaces of probability measures". *Math. Rev.* **55** 1546–1547.
[12] JEFFREYS, H. (1967). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford.
[13] LECAM, L. (1953). On some asymptotic properties of the maximum likelihood estimate and related Bayes estimates. *Univ. of Calif. Pub. Statist.* 277–330.
[14] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. verw. Gebiete* **4** 10–26.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA BERKELEY
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305