The Annals of Statistics
1992, Vol. 20, No. 3, 1222–1235

SOME ASPECTS OF POLYA TREE DISTRIBUTIONS FOR STATISTICAL MODELLING[1]

BY MICHAEL LAVINE

Duke University

Polya tree distributions are defined. They are generalizations of Dirichlet processes that allow for the possibility of putting positive mass on the set of continuous distributions. Predictive and posterior distributions are explained. A canonical construction of a Polya tree is given so that the Polya tree has any desired predictive distribution. Choices of the Polya tree parameters are discussed. Mixtures of Polya trees are defined and examples are given.

1. Introduction. Polya trees form a class of distributions for a random probability measure $\mathscr{P}$ intermediate between Dirichlet processes [Ferguson (1973)] and tailfree processes [Freedman (1963) and Fabius (1964)]. Their advantage over Dirichlet processes is that they can be constructed to give probability 1 to the set of continuous or absolutely continuous probability measures, whereas their advantage over more general tailfree processes is their much greater tractability. Many of the ideas discussed later can also be found in Ferguson (1974) and Mauldin, Sudderth and Williams (1991) (MSW).

Let $E = \{0, 1\}$, $E^0 = \varnothing$, E^m be the m-fold product $E \times E \times \cdots \times E$, $E^* = \bigcup_0^\infty E^m$ and E^N be the set of infinite sequences of elements of E. Let Ω be a separable measureable space, $\pi_0 = \Omega$ and $\Pi = \{\pi_m;\ m = 0, 1, \ldots\}$ be a separating binary tree of partitions of Ω; that is, let $\pi_0, \pi_1, \ldots$ be a sequence of partitions such that $\bigcup_0^\infty \pi_m$ generates the measurable sets and such that every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two pieces. Let $B_\varnothing = \Omega$ and, for all $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in E^*$, let $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ be the two pieces into which B_ε is split. Degenerate splits are permitted, for example, $B_\varepsilon = B_{\varepsilon 0} \cup \varnothing$.

DEFINITION 1. A random probability measure $\mathscr{P}$ on Ω is said to have a Polya tree distribution, or a Polya tree prior, with parameter $(\Pi, \mathscr{A})$, written $\mathscr{P} \sim \mathrm{PT}(\Pi, \mathscr{A})$, if there exist nonnegative numbers $\mathscr{A} = \{\alpha_\varepsilon:\ \varepsilon \in E^*\}$ and random variables $\mathscr{Y} = \{Y_\varepsilon:\ \varepsilon \in E^*\}$ such that the following hold:

(i) all the random variables in $\mathscr{Y}$ are independent;
(ii) for every $\varepsilon \in E^*$, Y_ε has a Beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$;

Received November 1990; revised February 1992.
[1]Supported in part by NSF Grant DMS-89-03842.
AMS 1980 *subject classifications*. Primary 62A15; secondary 62G99.
Key words and phrases. Nonparametric Bayes, Dirichlet processes, tailfree processes.

(iii) for every $m = 1, 2, \ldots$ and every $\varepsilon \in E^m$,

$$\mathscr{P}(B_{\varepsilon_1 \cdots \varepsilon_m}) = \left(\prod_{j=1;\, e_j=0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \right) \left(\prod_{j=1;\, \varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}) \right),$$

where the first term in the products is interpreted as $Y_\varnothing$ or as $1 - Y_\varnothing$.

Degenerate Beta distributions are permitted, for example, $\alpha_{\varepsilon 0} = 0$ making the distribution degenerate at 0. The notation in Definition 1 follows that of Ferguson [(1974), page 620]. Polya trees were so named by MSW, who represent the Beta distributions by Polya urns. Our development differs from that of MSW in two other ways. First, MSW define Polya trees using the set $E = \{0, 1, \ldots, k\}$, rather than our $E = \{0, 1\}$, and therefore deal with Dirichlet rather than Beta distributions. Our definition loses no mathematical generality although it may be less convenient for some modelling problems. Second, MSW define Polya trees directly on E^N and induce Polya tree distributions on Ω through measurable functions $g\colon E^N \to \Omega$.

The random variables $\Theta_1, \Theta_2, \ldots$ are said to be a sample from $\mathscr{P}$ if, given $\mathscr{P}$, they are i.i.d. with distribution $\mathscr{P}$. The Y_ε's have the following interpretation: $Y_\varnothing$ and $1 - Y_\varnothing$ are, respectively, the probabilities that $\Theta_i \in B_0$ and $\Theta_i \in B_1$, and Y_ε and $1 - Y_\varepsilon$ are the conditional probabilities that $\Theta_i \in B_{\varepsilon 0}$ and $\Theta_i \in B_{\varepsilon 1}$ given that $\Theta_i \in B_\varepsilon$. A Polya tree prior can be elicited by questions about these probabilities.

Polya trees are conjugate and easily updated. See Ferguson [(1974), page 620] or Theorem 4.3 of MSW. $Y_\varnothing$ is the unknown probability of the event $\Theta_i \in B_0$. In a Polya tree $Y_\varnothing$ has a Beta distribution. When Θ_1 is observed, so is the truth of $\Theta_1 \in B_0$. Therefore the conditional distribution of $Y_\varnothing$ given Θ_1 is a Beta distribution in which one of the parameters has been incremented by 1. A similar argument applies either to Y_0 if $\Theta_1 \in B_0$, or to Y_1 if $\Theta_1 \in B_1$, and so forth, down through the tree, adding 1 to every α_ε for which $\Theta_1 \in B_\varepsilon$.

The new Polya tree formed by updating gives the distribution of $\mathscr{P}|\Theta_1$; we write $\mathscr{P}|\Theta_1 \sim \mathrm{PT}(\Pi, \mathscr{A}|\Theta_1)$. Sometimes we will not have observed Θ_1 exactly but will only know that Θ_1 belongs to some set. If that set happens to be B_δ for some $\delta \in E^*$, then again the updating follows the same rule. The difference is that when Θ_1 is observed exactly there are infinitely many α_ε's to update; when we see $\Theta_1 \in B_\delta$, there are only finitely many.

Ferguson (1974) constructs a Polya tree prior on $(0, 1]$ using partitions comprised of the diadic intervals $(j/2^n, (j+1)/2^n]$. Except for minor details, the result is identical to one of MSW's Polya trees. Ferguson (personal communication) notes one such detail. The sets $(0, 1/2^n]$ decrease to $\varnothing$. The probability of the nth such set is $Y_\varnothing \times Y_0 \times \cdots \times Y_{00\ldots0}$. If this product does not converge to 0 then $\mathscr{P}$ will not be continuous and hence not countably additive. If $\mathscr{P}$ is to be countably additive with probability 1, then the α_ε's must be chosen so that this and other similar products converge to 0 with probability 1.

The following are important facts about Polya trees:

1. Dirichlet processes are special cases of Polya trees. A Polya tree is a Dirichlet process if, for every $\varepsilon \in E^*$, $\alpha_\varepsilon = \alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}$ [Ferguson (1974)].
2. Some Polya trees assign probability 1 to the set of continuous distributions. Kraft (1964), Metivier (1971), Ferguson (1974) and MSW give sufficient conditions for $\mathscr{P}$ to be continuous or absolutely continuous with probability 1.

Polya trees have advantages and disadvantages for modellers relative to Dirichlet process priors. An obvious advantage is that they can give probability 1 to the set of continuous random variables. A second advantage, as we shall see later, is that some sampling situations that lead to posterior mixtures of Dirichlet processes lead to just a single posterior Polya tree. A disadvantage is that, except for trivial special cases, the Dirichlet processes are the only tailfree processes in which Π does not play an essential role [Fabius (1973), Doksum (1974) and Ferguson (1974)].

2. Theory. This section discusses some computations with Polya trees, how to construct a Polya tree with any given marginal distribution for the data, how to choose the parameters Π and $\mathscr{A}$ and how to use mixtures of Polya trees. Throughout, $\mathscr{P}$ is a random probability measure, $\mathscr{P} \sim \mathrm{PT}(\Pi, \mathscr{A})$, and $\boldsymbol{\Theta} = \Theta_1, \Theta_2, \ldots$ is a sample from $\mathscr{P}$.

2.1. $E[\mathscr{P}]$. We wish to define the probability measure $Q = E[\mathscr{P}]$ by $Q(B) = \mathrm{E}[\mathscr{P}(B)]$ for any measurable set B. Definition 1(iii) gives, for any $\varepsilon \in E^*$,

$$
\begin{aligned}
Q(B_\varepsilon) &= \mathrm{E}\left[\prod_{j=1,\,\varepsilon_j=0}^{m} Y_{\varepsilon_1 \cdots \varepsilon_{j-1}} \prod_{j=1,\,\varepsilon_j=1}^{m} \left(1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}\right)\right] \\
&= \prod_{j=1;\,\varepsilon_j=0}^{m} \mathrm{E}\left[Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}\right] \prod_{j=1;\,\varepsilon_j=1}^{m} \mathrm{E}\left[1 - Y_{\varepsilon_1 \cdots \varepsilon_{j-1}}\right] \\
&= \prod_{j=1;\,\varepsilon_j=0}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}0}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}1}} \prod_{j=1;\,\varepsilon_j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}1}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}1}},
\end{aligned}
$$

which defines Q on the elements of $\bigcup_m \pi_m$. But $\bigcup_m \pi_m$ generates the measurable sets, so Q has a natural unique extension to the measurable sets. Note that Q is also the distribution of each Θ_i because $\Pr[\Theta_i \in B] = \mathrm{E}[\Pr[\Theta_i \in B|\mathscr{P}]] = \mathrm{E}[\mathscr{P}(B)] = Q(B)$. Let Z be a measurable real-valued function of Θ.

THEOREM 1. *If* $\int |Z|\, dQ < \infty$, *then* $\int |Z|\, d\mathscr{P} < \infty$ *with probability* 1, *and* $E[\int Z\, d\mathscr{P}] = \int Z\, dQ$.

PROOF. $\int Z\, dQ = \mathrm{E}[Z] = \mathrm{E}[\mathrm{E}[Z|\mathscr{P}]] = \mathrm{E}[\int Z\, d\mathscr{P}]$. □

2.2. *Updating: predictive densities.* We show how to compute the density of $\Theta_{n+1}|\Theta_1, \ldots, \Theta_n$, with respect to a suitable dominating probability measure λ, assuming for the moment that this density exists. Generically, we refer to the distribution of $\Theta_{i+1}|\Theta_1, \ldots, \Theta_i$ as a predictive distribution. We begin with the predictive density of Θ_1 and proceed stepwise, always finding the predictive density of the next observation given all the previous ones. Because Polya trees are conjugate, it suffices to find the predictive density of Θ_2 given Θ_1.

To evaluate $g_{\Theta_1}(\theta)$, the predictive density of Θ_1 at a point θ, let $\varepsilon_1, \varepsilon_2, \ldots$ be the infinite sequence of 0's and 1's such that $\theta \in B_{\varepsilon_1 \cdots \varepsilon_m}$ for all $m = 1, 2, \ldots$.

THEOREM 2.

$$
\begin{aligned}
g_{\Theta_1}(\theta) &= \lim_{m \to \infty} \frac{\Pr\left[\Theta_1 \in B_{\varepsilon_1 \cdots \varepsilon_m}\right]}{\lambda\left(B_{\varepsilon_1 \cdots \varepsilon_m}\right)} \\
&= \lim_{m \to \infty} \frac{\prod_{j=1}^{m} \alpha_{\varepsilon_1 \cdots \varepsilon_j} / \left(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}1}\right)}{\lambda\left(B_{\varepsilon_1 \cdots \varepsilon_m}\right)},
\end{aligned} \tag{1}
$$

where the first equality holds for λ-almost all θ.

PROOF. The second equality is by construction. The first follows from a martingale argument which is given, for example, by Billingsley [(1986), Theorem 35.8, page 494]. □

As will be clarified later in this section, Polya trees can be constructed so that the limit in Theorem 2 exists and can be evaluated, and furthermore, so that the corresponding limit will exist after the tree has been updated. We assume for now that the limit exists.

Suppose that $\Theta_1 = \theta_1$ has been observed. Let $\delta_1, \delta_2, \ldots$ be the infinite sequence of 0's and 1's such that $\theta_1 \in B_{\delta_1, \ldots, \delta_m}$ for all $m = 1, 2, \ldots$, and let k be such that $B_{\delta_1 \cdots \delta_k} = B_{\varepsilon_1 \cdots \varepsilon_k}$ but that $B_{\delta_1 \cdots \delta_{k+1}} \neq B_{\varepsilon_1 \cdots \varepsilon_{k+1}}$. The predictive density $g_{\Theta_2|\Theta_1}(\theta|\theta_1)$ is the same as in (1), except that in the numerator of the right-hand side of (1), $\alpha_{\varepsilon_1}, \alpha_{\varepsilon_1\varepsilon_2}, \ldots, \alpha_{\varepsilon_1 \cdots \varepsilon_k}$ are each incremented by 1. This affects only the first $k + 1$ terms of the numerator. Otherwise, the formula is unchanged. Once $g_{\Theta_1}(\theta)$ has been evaluated, an updated predictive density after any arbitrary set of observations is easily calculated.

The predictive density after one observation is a piecewise rescaled version of the original predictive density. There is a separate rescaling factor for each B_ε containing θ_1; this is how Π plays a role in the Polya tree distribution. Thus, although the predictive density exists, it can be discontinuous, with infinitely many discontinuities in every neighborhood of θ_1.

For example, let $\mathscr{P}$ be a random distribution on the unit interval with a Polya tree distribution for which π_m consists of the diadic intervals $\{(j/2^m, (j+1)/2^m)\}$ and for which $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = m^2$. By symmetry, $\Theta_1 \sim U(0, 1)$ so $g_{\Theta_1}(\theta) = 1$ for almost all $\theta \in (0, 1)$. After $\Theta_1 = 0.1$ has been observed, $g_{\Theta_2|\Theta_1}(\theta|0.1)$ is evaluated using (1) to modify $g_{\Theta_1}(\theta)$. Initially, $\alpha_0 = \alpha_1 = 1$

and $\alpha_{00} = \alpha_{01} = \alpha_{10} = \alpha_{11} = 4$. After observing $\Theta_1 = 0.1$, α_0 becomes 2 and α_{00} becomes 5. Thus, for $\theta > 0.5$, $\varepsilon_1 = 1$, $k = 0$, the first term in the numerator of the RHS of (1), which was initially $1/2$, gets changed to $1/3$, and $g_{\Theta_2|\Theta_1}(\theta|0.1) = (2/3)g_{\Theta_1}(\theta) = 2/3$. For $\theta \in (0.25, 0.5)$, $\varepsilon_1 = 0$, $\varepsilon_2 = 1$, $k = 1$, the first two terms in the numerator of the RHS of (1), which were initially $(1/2)(4/8)$, get changed to $(2/3)(4/9)$, and $g_{\Theta_2|\Theta_1}(\theta|0.1) = (4/3)(8/9)g_{\Theta_1}(\theta) = 32/27$. Similarly, $g_{\Theta_2|\Theta_1}(\theta|0.1)$ can be calculated for any $\theta \neq 0.1$. It is piecewise constant, has infinitely many discontinuities near 0.1, and is nonincreasing away from 0.1. In contrast, for Dirichlet process priors, the predictive distribution after one observation is a rescaled version of the original predictive distribution, plus a point mass at θ_1.

2.3. *Constructing Polya trees.* We give a canonical construction of a Polya tree such that $\Theta_1 \sim Q$, where Q is specified in advance and is continuous. Begin by choosing B_0 and B_1 to satisfy $Q(B_0) = Q(B_1) = 1/2$. Then, for every $\varepsilon \in E^*$, choose $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ to satisfy $Q(B_{\varepsilon 0}|B_\varepsilon) = Q(B_{\varepsilon 1}|B_\varepsilon) = 1/2$. Any choice of $\mathscr{A}$ satisfying $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ will satisfy $\Theta_1 \sim Q$. Of course other Polya trees also satisfy $\Theta_1 \sim Q$; the preceding construction is suggested for convenience and when there is no cogent reason for choosing Π otherwise. An important example is when $\Omega = \mathbb{R}$ and Q has cumulative distribution function G, in which case the elements of π_m can be taken to be the intervals $(G^{-1}(k/2^m), G^{-1}((k+1)/2^m)]$ for $k = 0, \ldots, 2^m - 1$, with the obvious interpretations for $G^{-1}(0)$ and $G^{-1}(1)$.

One reason for choosing Π otherwise is convenience in updating with a sample of censored observations. Suppose we know only $\Theta_1 > \theta_1$, $\Theta_2 > \theta_2, \ldots, \Theta_k > \theta_k$, but do not know the exact values of $\Theta_1, \Theta_2, \ldots, \Theta_k$, and where we assume without loss of generality that $\theta_1 < \theta_2 < \cdots < \theta_k$. Let Π be chosen so that $B_1 = (\theta_1, \infty)$, $B_{11} = (\theta_2, \infty), \ldots, B_{11\cdots 1} = (\theta_k, \infty)$, where there are k 1's in $11\cdots 1$. Then the distribution of $\mathscr{P}$ given the data is PT(Π, $\mathscr{A}^*$), where $\alpha_1^* = \alpha_1 + k$, $\alpha_{11}^* = \alpha_{11} + k - 1, \ldots, \alpha_{11\cdots 1}^* = \alpha_{11\cdots 1} + 1$ and $\mathscr{A}^*$ is otherwise identical to $\mathscr{A}$. If we also have some uncensored observations, then we still can choose Π so that updating is convenient for the censored observations because Π is irrelevant to the ease of updating for the exact observations. When $\mathscr{P}$ initially has a Dirichlet process distribution, although the posterior distribution is a single Polya tree, it is a mixture of Dirichlet processes.

There are three considerations in choosing the values of α_ε's.

The first is that α_ε controls how quickly the updated predictive distribution moves from the prior predictive distribution to the sample distribution. If the α_ε's are large, then the distribution of $\Theta_{n+1}|\Theta_1, \Theta_2, \ldots, \Theta_n$ is close to Q. However, if the α_ε's are small, then the distribution of $\Theta_{n+1}|\Theta_1, \Theta_2, \ldots, \Theta_n$ is close to the sample distribution function. For example, suppose that for each i, a separate coin toss determines whether Θ_i is in B_0 or B_1. If we have a strong initial belief that the coin toss is fair, then we can choose α_0 and α_1 very large, making $\Pr[\Theta_2 \in B_j]$ approximately independent of the event $\Theta_1 \in B_i$, for $i, j \in \{1, 2\}$. However, if we believe the coin is biased, but do not know whether

heads or tails is favored, then we can let α_0 and α_1 be small. The choice $\alpha_0 = \alpha_1 = 1$ yields $\Pr[\Theta_2 \in B_j | \Theta_1 \in B_j] = 2/3$, for $j \in \{1, 2\}$.

Choosing α_ε for small m typically involves judgments about sets which have large initial probability. However, when m is large, α_ε is controlling the conditional probabilities of small and initially unlikely sets, and it is unrealistic to expect to elicit meaningful values. The next two considerations provide guidance in choosing α_ε for large m.

The second consideration is that α_ε can express a belief about the smoothness of $\mathscr{P}$. Heuristically, if $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ is large, then $\mathscr{P}[B_{\varepsilon 0} | B_\varepsilon]$ has a distribution tightly concentrated around $1/2$, which makes $\mathscr{P}$ smooth because with high probability $\mathscr{P}(B_{\varepsilon 0})$ and $\mathscr{P}(B_{\varepsilon 1})$ will be roughly equal. The same is true for $\mathscr{P}[B_{\varepsilon 0} | \Theta_1, \ldots, \Theta_n]$ and $\mathscr{P}[B_{\varepsilon 1} | \Theta_1, \ldots, \Theta_n]$, as long as $n \ll \alpha_{\varepsilon 0}$. Ferguson [(1974), page 621] remarks that for Polya trees on $\mathbb{R}$, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 1$ yields $\mathscr{P}$ that is continuous singular with probability 1, that is, without point masses yet singular with respect to Lebesgue measure, and that $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = m^2$ yields $\mathscr{P}$ that is absolutely continuous with probability 1. For that reason, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = m^2$ would often be a sensible canonical choice. The following two theorems support the belief that sufficiently large $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ for large m implies $\mathscr{P}$ is smooth.

Let $\bar{\sigma}_m = \sup\{\operatorname{var}(Y_\varepsilon) : \varepsilon \in E^m\}$.

THEOREM 3 [Kraft (1964)]. *If $E[Y_\varepsilon] = 1/2$ for all ε, and if $\sum_{m=1}^{\infty} \bar{\sigma}_m < \infty$ then, with probability 1, $\mathscr{P}$ is absolutely continuous with respect to Lebesgue measure.*

If $E[Y_\varepsilon] = 1/2$, then $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ and the variance of Y_ε is $1/(4(2\alpha_{\varepsilon 0} + 1))$. Therefore, as long as the α_ε's increase sufficiently rapidly with m, $\mathscr{P}$ will be absolutely continuous.

THEOREM 4 (MSW). *A sufficient condition for $\mathscr{P}$ to be continuous with probability 1 is $\Pr[\Theta_2 = \theta | \Theta_1 = \theta] = 0$ for every θ.*

For $\theta \in \Omega$ let $\varepsilon_1, \varepsilon_2, \ldots$ be the infinite sequence of 0's and 1's satisfying $\theta \in B_{\varepsilon_1 \cdots \varepsilon_m}$ for every $m = 1, 2, \ldots$ and note that

$$\Pr[\Theta_2 = \theta | \Theta_1 = \theta] = \prod_{m=1}^{\infty} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_m} + 1}{(\alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{m-1} 1} + 1)}.$$

Therefore, as long as the α_ε's do not decrease too rapidly with m, $\mathscr{P}$ will be continuous.

The third consideration is that α_ε controls how closely the distribution of $\mathscr{P}$ is concentrated about its mean. The next theorem shows that a Polya tree for $\Omega = \mathbb{R}$ can be constructed that concentrates arbitrarily closely around any desired mean. Dalal and Hall (1980) show that Dirichlet process priors can be constructed that concentrate arbitrarily closely around any desired mean, in

the sense of weak convergence. Our criterion for closeness is different. Let F be the cumulative distribution function of $\mathscr{P}$, let G be the cumulative distribution function of $Q = E[\mathscr{P}]$, let $\mathscr{E}(\theta, \delta)$ be the event $|F(\theta) - G(\theta)| < \delta$ and let $\bar{\mathscr{E}}$ be the complement of $\mathscr{E}$.

THEOREM 5. *Let* $\Omega = \mathbb{R}$. *For every* $\delta > 0$ *and* $\eta \in (0, 1)$, *there exists a Polya tree such that* $\Pr[\bigcap_\theta \mathscr{E}(\theta, \delta)] > \eta$.

PROOF. The proof is by construction and begins with the case where G is continuous and G^{-1} is well defined on $(0, 1)$. For $m = 1, 2, \ldots$ and $k = 1, 2, \ldots, 2^m - 1$, let $\theta_{k,m} = G^{-1}(k/2^m)$, $\theta_{0,m} = -\infty$, $\theta_{2^m,m} = \infty$ and let $\eta_1, \eta_2, \ldots$ be a sequence of numbers in $(0, 1)$ satisfying $\Pi_i \eta_i > \eta$. Let the partitions of the Polya tree be defined by $\pi_m = \{(\theta_{0,m}, \theta_{1,m}], \ldots, (\theta_{2^m-1,m}, \theta_{2^m,m})\}$ and let the Beta parameters satisfy $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$, so that $E[Y_\varepsilon] = 0.5$ for $\varepsilon \in E^*$, where $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$ are chosen to satisfy conditions given later. Because $\{\theta_{k,m}\}$ is dense in $\mathbb{R}$, it suffices to show

$$\Pr\left[\bigcap_{m=1}^{\infty} \bigcap_{k=1,2,\ldots,2^m-1} \mathscr{E}(\theta_{k,m}, \delta)\right] > \eta.$$

However,

$$\begin{aligned}
&\Pr\left[\bigcap_{m=1}^{\infty} \bigcap_{k=1,2,\ldots,2^m-1} \mathscr{E}(\theta_{k,m}, \delta)\right] \\
&\quad = \Pr\left[\bigcap_{m=1}^{\infty} \bigcap_{k=1,3,\ldots,2^m-1} \mathscr{E}(\theta_{k,m}, \delta)\right] \\
&\quad \geq \Pr\left[\bigcap_{m=1}^{\infty} \bigcap_{k=1,3,\ldots,2^m-1} \mathscr{E}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right)\right] \\
&\quad = \Pr\left[\mathscr{E}\left(\theta_{1,1}, \frac{\delta}{2}\right)\right] \\
&\qquad \times \prod_{m=2}^{\infty} \Pr\left[\bigcap_k \mathscr{E}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right) \middle| \bigcap_{i=1}^{m-1} \bigcap_j \mathscr{E}\left(\theta_{j,i}, \frac{2^i - 1}{2^i}\delta\right)\right] \qquad (2) \\
&\quad = \Pr\left[\mathscr{E}\left(\theta_{1,1}, \frac{\delta}{2}\right)\right] \\
&\qquad \times \prod_{m=2}^{\infty} \left(1 - \Pr\left[\bigcup_k \bar{\mathscr{E}}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right) \middle| \bigcap_{i=1}^{m-1} \bigcap_j \mathscr{E}\left(\theta_{j,i}, \frac{2^i - 1}{2^i}\delta\right)\right]\right) \\
&\quad \geq \Pr\left[\mathscr{E}\left(\theta_{1,1}, \frac{\delta}{2}\right)\right] \\
&\qquad \times \prod_{m=2}^{\infty} \left(1 - \sum_k \Pr\left[\bar{\mathscr{E}}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right) \middle| \bigcap_{i=1}^{m-1} \bigcap_j \mathscr{E}\left(\theta_{j,i}, \frac{2^i - 1}{2^i}\delta\right)\right]\right).
\end{aligned}$$

Finally, $\Pr[\mathscr{E}(\theta_{1,1}, \delta/2)] = \Pr[|Y_\phi - 0.5|] < \delta/2$ can be made larger than η_1 by choosing α_0 and α_1 sufficiently large and, for $k \in \{1, 3, \ldots, 2^m - 1\}$,

$$\Pr\left[\bar{\mathscr{E}}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right) \middle| \bigcap_i \bigcap_j \mathscr{E}\left(\theta_{j,i}, \frac{2^i - 1}{2^i}\delta\right)\right]$$

$$\leq \Pr\left[\bar{\mathscr{E}}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right) \middle| \mathscr{E}\left(\theta_{(k-1)/2,\,m-1}, \frac{2^{m-1} - 1}{2^{m-1}}\delta\right) \cap \mathscr{E}\left(\theta_{(k+1)/2,\,m-1}, \frac{2^{m-1} - 1}{2^{m-1}}\delta\right)\right]$$

$$\leq \Pr\left[\bar{\mathscr{E}}\left(\theta_{k,m}, \frac{2^m - 1}{2^m}\delta\right) \middle| F(\theta_{(k-1)/2,\,m-1}) = G(\theta_{(k-1)/2,\,m-1}) + \frac{2^{m-1} - 1}{2^{m-1}}\delta;\right.$$

$$\left. F(\theta_{(k+1)/2,\,m-1}) = G(\theta_{(k+1)/2,\,m-1}) + \frac{2^{m-1} - 1}{2^{m-1}}\delta\right]$$

$$\leq \Pr\left[|Y_\varepsilon - 0.5| > \frac{\delta}{2^m}\right]$$

for some $\varepsilon \in E^{m-1}$. So, by choosing $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$ sufficiently large, each summand in (2) can be made arbitrarily small, and the mth factor in the infinite product in (2) can be made larger than η_m.

For G not invertible on $(0, 1)$, that is, not strictly increasing, let $I \subset \mathbb{R}$ be a G null set with G strictly increasing on $\mathbb{R} - I$. Then construct a Polya tree with $B_0 = I$, $B_1 = \mathbb{R} - I$, $\alpha_0 = 0$ and $\{B_{1\varepsilon}\colon \varepsilon \in E^*\}$ and $\{\alpha_{1\varepsilon}\colon \varepsilon \in E^*\}$ chosen as for invertible G. If G is not continuous, let $\theta_1, \ldots, \theta_n$ be a collection of atoms of G with masses $w_1, \ldots, w_n$, and J be the collection of all the remaining atoms of G, divided so that G assigns mass less than $\delta/(n + 2)$ to J. Then construct a Polya tree such that the following hold:

(i) $B_0 = J$ and $\Pr[Y_\phi < \delta/(n + 2)]$ is large.
(ii) $B_{10} = \{\theta_1, \ldots, \theta_n\}$.
(iii) Y_1 is close to $\sum w_i$ with high probability.
(iv) $\{B_{11\varepsilon}\}$ and $\{\alpha_{11\varepsilon}\}$ are constructed as for continuous G.
(v) $\{B_{10\varepsilon}\}$ and $\{\alpha_{10\varepsilon}\}$ are constructed so that the jumps of F match the jumps of G, for example, $B_{100} = \theta_1$ and Y_{10} is close to $w_1/\sum w_i$ with high probability. □

2.4. *Mixtures of Polya trees.* The distribution of a random probability measure $\mathscr{P}$ is said to be a mixture of Polya trees if there is a random variable U, called the index variable, with distribution H, called the mixing distribution, and Polya tree parameters $\{\Pi_u, \mathscr{A}_u\}$ such that $[\mathscr{P}|U = u] \sim \mathrm{PT}(\Pi_u, \mathscr{A}_u)$.

For any measurable set S of probability measures on Ω, $\Pr[\mathscr{P} \in S] = \int \Pr[\mathscr{P} \in S|u]H(du)$. A single Polya tree is a mixture of Polya trees where H is degenerate. When Θ_1 is an observation from a mixture of Polya trees, then each $\mathscr{A}_u$ must be updated, exactly as for a single Polya tree, and the mixing distribution H must be updated as well. Using the usual notation for densities, $h(u|\Theta_1 = \theta_1)$ is proportional to $h(u)g_{\Theta_1|U}(\theta_1|u)$, where $g_{\Theta_1|U}(\theta_1|u)$ is the predictive density from the Polya tree with parameter $(\Pi_u, \mathscr{A}_u)$. Updating the mixture and calculating the distribution of $\Theta_2|\Theta_1 = \theta_1$ are easily handled by a computer.

With mixtures of Polya trees the problem of dependence on the partitions is not as critical. Roughly speaking, if the partition elements are different in each Π_u and if $H(u) = 0$ for all u, then the partition effects get smoothed out and updated predictive densities can be continuous. More specifically, suppose Θ_1, a sample of size 1 from a mixture of Polya trees has been observed. The question is, for which values of θ is the predictive density $g_{\Theta_2|\Theta_1}(\theta|\theta_1)$ continuous?

THEOREM 6. *Suppose that the following hold*:

(i) *The conditional c.d.f.* $G_{\Theta_1|U}(\theta|u)$ *is a measurable function of* u *for each* θ *in* (a, b).

(ii) *For* $u \in A$, *where* $H(A) = 1$, $G_{\Theta_1|U}(\theta|u)$ *has in* (a, b) *a derivative* $g_{\Theta_1|U}(\theta|u)$ *with respect to* θ.

(iii) $g_{\Theta_1|U}(\theta|u) \le r(u)$ *for* $u \in A$ *and* $\theta \in (a, b)$, *where* r *is integrable*.

(iv) $g_{\Theta_2|\Theta_1, U}(\theta|\theta_1, u)$ *is* H*-almost everywhere continuous in* θ *at* θ_0.

(v) *There exists a positive number* M *such that for* H*-almost every* u, *for every nonnegative integer* m *and for every* $\varepsilon \in E^m$,

$$\prod_{j=1}^{m} \frac{\alpha_{u,\varepsilon_1 \cdots \varepsilon_{j-1}0} + \alpha_{u,\varepsilon_1 \cdots \varepsilon_{j-1}1}}{\alpha_{u,\varepsilon_1 \cdots \varepsilon_j}} \frac{\alpha_{u,\varepsilon_1 \cdots \varepsilon_j} + 1}{\alpha_{u,\varepsilon_1 \cdots \varepsilon_{j-1}0} + \alpha_{u,\varepsilon_1 \cdots \varepsilon_{j-1}1+1}} \le M.$$

Then $g_{\Theta_2|\Theta_1}(\theta|\theta_1)$ *exists and is continuous at* θ_0.

Condition 5 relates to the RHS of (1) and says that observing Θ_1 can change the conditional predictive density at θ given u by no more than a factor of M.

PROOF OF THEOREM 6. The theorem follows from Theorem 16.8 of Billingsley [(1986), page 215]. □

Mixtures of Polya trees can be useful when a standard parametric Bayesian analysis is suspect because the family of sampling densities is not known exactly. The standard Bayesian analysis requires specification of a family of sampling densities $g(\theta|u)$ and a prior density $h(u)$ for the parameter u. There is a substantial body of work studying sensitivity to and modelling uncertainty about the prior. In contrast, although the issue is more critical, the state of the

art is less advanced in modelling uncertainty about the family of sampling densities. Such uncertainty can be modelled with a mixture of Polya trees where $U \sim h(u)$, $\mathscr{P}|u \sim \mathrm{PT}(\Pi_u, \mathscr{A}_u)$ and $\Theta_1|u \sim g(\theta|u)$.

In the parametric analysis, the data are modelled as though u is chosen according to $h(u)$; then $\theta_1, \theta_2, \ldots$ are chosen according to $g(\theta|u)$. In the mixture of Polya trees the data are modelled as though u is chosen according to $h(u)$; then P is chosen according to $\mathrm{PT}(\Pi_u, \mathscr{A}_u)$; then $\theta_1, \theta_2, \ldots$ are chosen according to P. Uncertainty about $g(\theta|u)$ is quantified by $\mathscr{A}$; large values of α_ε give $\mathscr{P}$ a distribution that is tightly concentrated around $g(\theta|u)$, small values of α_ε give $\mathscr{P}$ a more diffuse distribution.

3. Examples.

3.1. *Estimation of a distribution function.* Let $\mathscr{P}$ have cumulative distribution function F and $G = E[F]$.

An artificial example was constructed that illustrates how predictive densities evolve as more data are accumulated and illustrates the effect of different choices for α_ε. Two Polya trees were created, each having predictive density $g(\theta) = \exp(-\theta)$. The partitions were chosen by the canonical method of Section 2.3. In the first tree, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 2m$; in the second tree, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 2$. The first tree gives more smoothness, the second tree gives closer adherence to the data. A sample of size 100 was generated from the density $2\exp(-2\theta)$. For each tree, predictive densities were calculated after 5, 25, and 100 observations.

Figure 1 shows the densities as solid lines. The top row shows $f_{\Theta_6|\Theta_1,\ldots,\Theta_5}(\theta|\theta_1,\ldots,\theta_5)$, the middle row shows $f_{\Theta_{26}|\Theta_1,\ldots,\Theta_{25}}(\theta|\theta_1,\ldots,\theta_{25})$ and the bottom row shows $f_{\Theta_{101}|\Theta_1,\ldots,\Theta_{100}}(\theta|\theta_1,\ldots,\theta_{100})$. Densities from the first tree are on the left; densities from the second tree are on the right. In each plot the initial predictive density $\exp(-\theta)$ is shown as a dotted line and the true sampling density $2\exp(-2\theta)$ is shown as a dashed line. Going from top to bottom shows how the predictive density moves away from $\exp(-\theta)$ toward $2\exp(-2\theta)$. Going from left to right shows that the predictive density is smoother when α_ε is larger.

3.2. *Estimation of the mean.* Let $Z(\theta)$ be a measurable real-valued function. We investigate conditions under which the mean, $E[\int Z\,d\mathscr{P}] = \int Z\,dQ$, can be evaluated. Let $r(\varepsilon) = \sup_{\theta \in B_\varepsilon} Z(\theta) - \inf_{\theta \in B_\varepsilon} Z(\theta)$, $t(m) = \max_{\varepsilon \in E^m} r(\varepsilon)$ and $v(B_\varepsilon) = Z(\theta)$ for some $\theta \in B_\varepsilon$. If $\lim_{m\to\infty} t(m) = 0$, then $\int Z\,dQ = \lim_{m\to\infty} \sum_{\varepsilon \in E^m} v(B_\varepsilon) Q(B_\varepsilon)$, which can be evaluated to arbitrary accuracy by taking m sufficiently large. Also, $\int Z\,dQ|\Theta_1, \ldots, \Theta_k$ is computable, but if $\lim_{m\to\infty} t(m) \neq 0$, then even when $\int Z\,dQ$ exists it cannot be evaluated by this method.

Often the Polya tree is constructed so that $\int Z\,dQ$ is computable. For example, when $Z(\theta) = \theta$ then $\int Z\,dQ = E[\Theta_1]$, which may be known if the Polya tree was constructed to have a given predictive distribution. However,

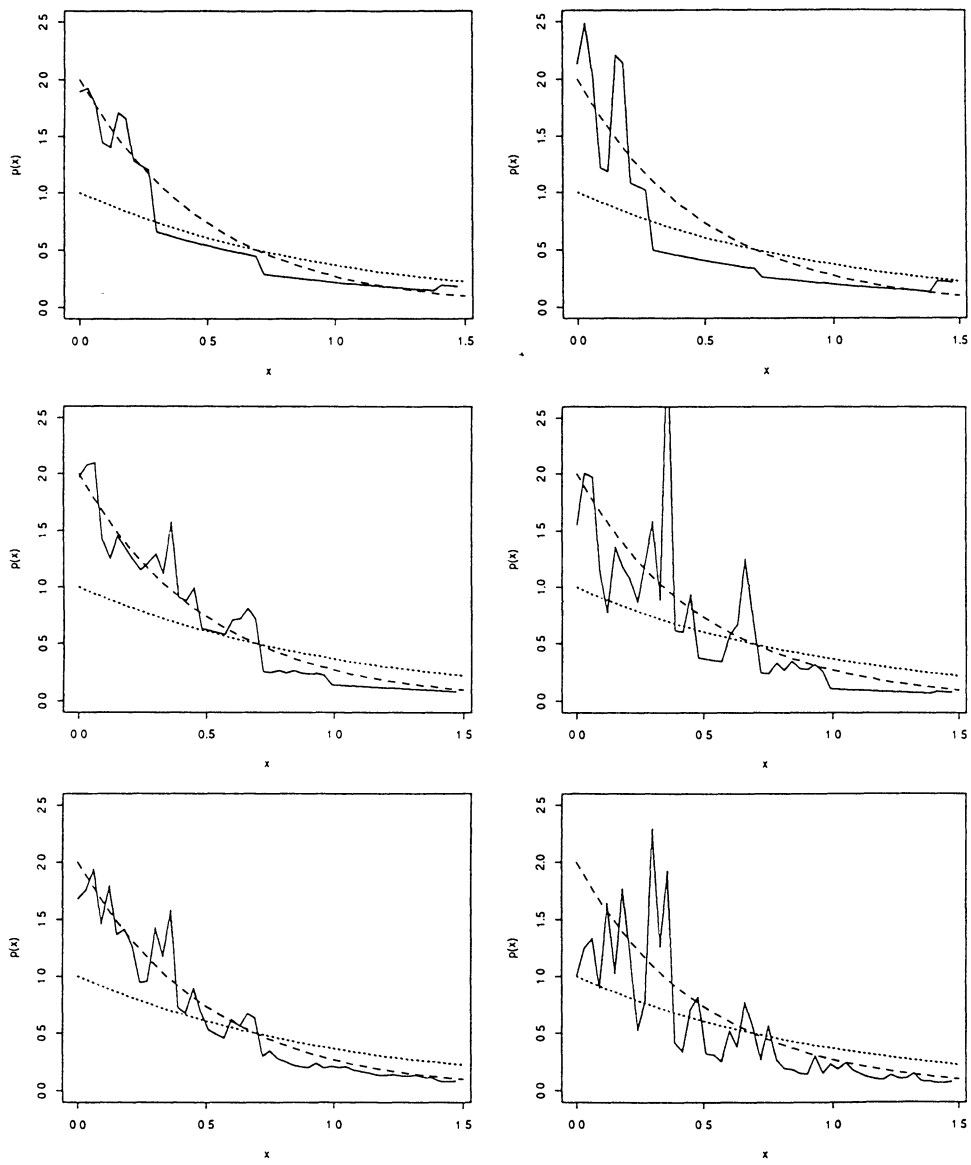

FIG. 1.

even assuming $\int Z\,dQ$ is known, we still want to evaluate $\int Z\,dQ|\Theta_1,\ldots,\Theta_k$. If there is a set $C \subset \Omega$ such that each of the first k observations falls in C, C is the union of finitely many B_ε's and $\lim_{m\to\infty} t(m) = 0$ for the restriction of Z to C, then the method in the previous paragraph can be used to evaluate $E[\int Z\,dQ|\Theta_1,\ldots,\Theta_k;\ \Theta \in C]$. If we assume further that $E[\int Z\,dQ|\Theta_1,\ldots,\Theta_k;\ \Theta \notin C] = E[\int Z\,dQ|\Theta \notin C]$ is evaluable then, because $\Pr[\Theta \in C|\Theta_1,\ldots,\Theta_k]$ is

computable from the updated Polya tree, the mean of $Z|\Theta_1, \ldots, \Theta_k$ can still be estimated.

For example, suppose that a Polya tree is constructed on $(0, \infty)$ with canonical partitions, with $\alpha_\varepsilon = 1$ for every ε, and so that the initial predictive density is $g(\theta_1) = \exp(-\theta_1)$. After $\Theta_1 = 0.5$ has been observed the method of the preceding paragraphs, with $m = 2$, can be used to bound the predictive mean $E[\Theta_2|\Theta_1 = 0.5]$. Note that $B_{00} = (0, \ln(4/3))$, $B_{01} = (\ln(4/3), \ln(2))$ and $B_1 = (\ln(2), \infty)$. The posterior probabilities are $Q[B_{00}|\Theta_1 = 0.5] = 2/9$, $Q[B_{01}|\Theta_1 = 0.5] = 4/9$ and $Q[B_1|\Theta_1 = 0.5] = 3/9$. The predictive density is $g_{\Theta_2|\Theta_1}(\theta|0.5) = (8/9)\exp(-\theta)$ on B_{00} and $g_{\Theta_2|\Theta_1}(\theta|0.5) = (2/3)\exp(-\theta)$ on B_1. Therefore

$$
\begin{aligned}
E[\Theta_2|\Theta_1 = 0.5] &= \int_{B_{00}} \theta g_{\Theta_2|\Theta_1}(\theta|0.5)\, d\theta + \int_{B_{01}} \theta g_{\Theta_2|\Theta_1}(\theta|0.5)\, d\theta \\
&\quad + \int_{B_1} \theta g_{\Theta_2|\Theta_1}(\theta|0.5)\, d\theta \\
&= \tfrac{8}{9}\int_0^{\ln(4/3)} \theta e^{-\theta}\, d\theta + \int_{\ln(4/3)}^{\ln 2} \theta g_{\Theta_2|\Theta_1}(\theta|0.5)\, d\theta + \tfrac{2}{3}\int_{\ln 2}^{\infty} \theta e^{-\theta}\, d\theta \\
&= 0.595 + \int_{\ln(4/3)}^{\ln 2} \theta g_{\Theta_2|\Theta_1}(\theta|0.5)\, d\theta \\
&\in \left(0.595 + \tfrac{4}{9}\ln\tfrac{4}{3}, 0.595 + \tfrac{4}{9}\ln 2\right) \\
&= (0.723, 0.903).
\end{aligned}
$$

3.3. *Lifetimes of spherical pressure vessels.* This example illustrates the use of mixtures of Polya trees and also shows the influence of the value of α_ε on the predictive density. The data are time-to-failure of Kevlar 49/epoxy spherical vessels under pressure and come from [Andrews and Herzberg (1985), page 185], they have been divided by 300 for this example. According to Andrews and Herzberg (1985), "The NASA space shuttle uses Kevlar/epoxy spherical pressure vessels in a sustained pressure mode throughout the usage life of the vessel, and several commercial applications, such as fire-fighters' air-breathing apparatus, are also subject to this service condition. The study was done to generate baseline data on vessel life under pressure and to predict vessel life and design reliability."

Four mixtures of Polya trees were created as models for the data. Each mixture was created with $h(u) = \exp(-u)$, $g(\theta|u) = u\exp(-u\theta)$ and the canonical partitions. The value of $\alpha_{\varepsilon_1 \cdots \varepsilon_m}$ depends only on m and on the mixture, not on u. In the first mixture, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 20$; in the second mixture, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 2$; in the third mixture, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 2^m$; in the fourth mixture, $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 10^6$.

There are 39 observations; the quartiles are 0.03, 0.18 and 1.32. The data are plotted as points on the horizontal axis in Figure 2. For each mixture of Polya trees, the predictive density $g_{\Theta_{40}|\Theta_1, \ldots, \Theta_{39}}(\theta|\theta_1, \ldots, \theta_{39})$ is plotted in

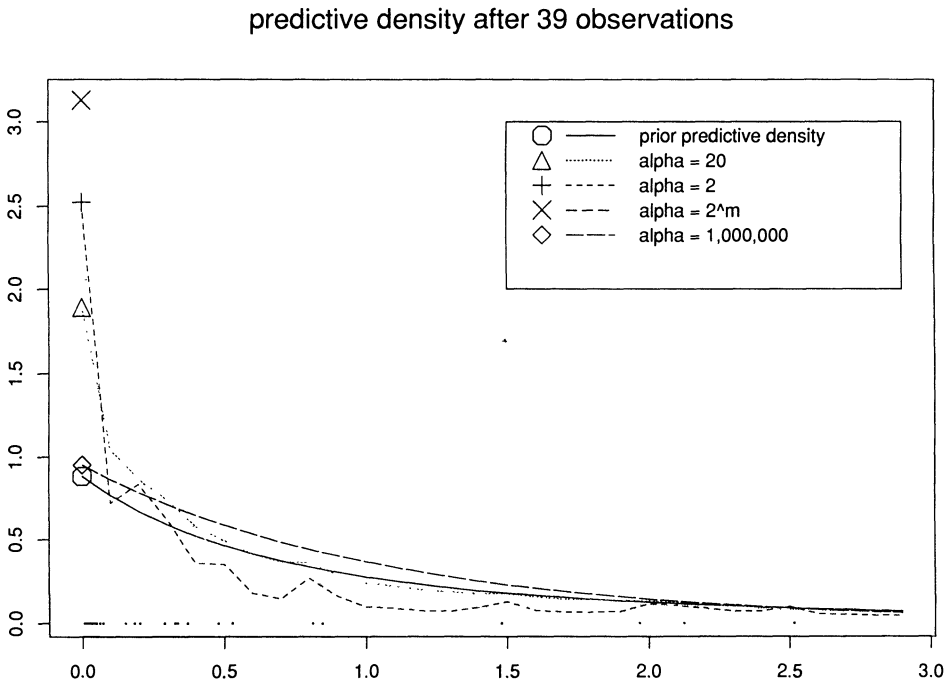

FIG. 2.

Figure 2. When $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 10^6$, each Polya tree in the mixture changes very little after only 39 observations, so this mixture mimics the usual parametric Bayesian analysis. The other Polya trees are flexible enough to model the data more closely, allowing for different degrees of smoothness and compromise between the data and the original predictive distribution.

REFERENCES

ANDREWS, D. F. and HERZBERG, A. M. (1985). *Data*. Springer, New York.

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

BILLINGSLEY, P. (1986). *Probability and Measure*, 2nd ed. Wiley, New York.

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

DOKSUM, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183–201.

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

FABIUS, J. (1973). Neutrality and Dirichlet distributions. In *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes* 175–181.

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

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

FREEDMAN, D. A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403.

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

LAVINE, M. (1988). Prior influence in Bayesian statistics. *J. Amer. Statist. Assoc.* To appear.

LAVINE, M., WASSERMAN, L. and WOLPERT, R. (1991). Bayesian influence with specified prior marginals. *J. Amer. Statist. Assoc.* **86** 964–971.

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

MÉTIVIER, M. (1971). Sur la construction de mesures aléatoires presque sûrement absolument continues par rapport à une mesure donnée. *Z. Wahrsch. Verw. Gebiete* **20** 332–344.

INSTITUTE OF STATISTICS
AND DECISION SCIENCES
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27706