The Annals of Statistics
2010, Vol. 38, No. 3, 1433–1459
DOI: 10.1214/09-AOS755

OPTIONAL PÓLYA TREE AND BAYESIAN INFERENCE

BY WING H. WONG[1] AND LI MA[2]

Stanford University

We introduce an extension of the Pólya tree approach for constructing distributions on the space of probability measures. By using optional stopping and optional choice of splitting variables, the construction gives rise to random measures that are absolutely continuous with piecewise smooth densities on partitions that can adapt to fit the data. The resulting "optional Pólya tree" distribution has large support in total variation topology and yields posterior distributions that are also optional Pólya trees with computable parameter values.

1. Introduction. Ferguson [7] formulated two criteria for desirable prior distributions on the space of probability measures: (i) The support of the prior should be large with respect to a suitable topology, and (ii) the corresponding posterior distribution should be analytically manageable. Extending the work by Freedman [9] and Fabius [6], he introduced the Dirichlet process as a prior that satisfies these criteria. Specifically, assuming for simplicity that the parameter space Ω is a bounded interval of real numbers, and the base measure in the Dirichlet process prior is the Lebesgue measure, then the prior will have positive probability in all weak neighborhoods of any absolutely continuous probability measure, and given i.i.d. observations, the posterior distribution is also a Dirichlet process with its base measure obtainable from that of the prior by the addition of delta masses at the observed data points.

While these properties made it an attractive prior in many Bayesian nonparametric problems, the use of the Dirichlet process prior is limited by its inability to generate absolutely continuous distributions; that is, a random probability measure sampled from the Dirichlet process prior is almost surely a discrete measure [1, 2, 7]. Thus in applications that require the existence of densities under the prior, such as the estimation of a density from a sample [16] or the modeling of error distributions in location or regression problems [5], there is a need for alternative ways to specify the prior. Lo [16] proposed an elegant prior in the space of densities by assuming the density is a mixture of kernel functions where the mixing

Received May 2009; revised September 2009.

[1] Supported in part by NSF Grants DMS-05-05732 and DMS-09-06044.

[2] Supported by a Gerhard Casper Stanford Graduate Fellowship.

AMS 2000 subject classifications. Primary 62F15, 62G99; secondary 62G07.

Key words and phrases. Pólya tree, Bayesian inference, nonparametric, recursive partition, density estimation.

distribution is modeled by a Dirichlet process. Under Lo's model, the random distributions are guaranteed to have smooth densities and the predictive density is still analytically tractable. However the degree of smoothness is not adaptive.

Another approach to deal with the discreteness problem is to use Pólya tree priors [8]. This class of random probability measures includes the Dirichlet process as a special case and yet is itself a special case of the more general class of "tail free" processes previously studied by Freedman [9]. Pólya tree prior satisfies Ferguson's two criteria. First, it is possible to construct Pólya tree priors with positive probability in neighborhoods around arbitrary positive densities [14]. Second, the posterior distribution arising from a Pólya tree prior is available in close form [8]. Further properties and applications of Pólya tree priors are found in [10–12, 15] and [17].

In this paper we study the extension of the Pólya tree prior construction by allowing optional stopping and randomized partitioning schemes. To motivate optional stopping, consider the standard construction of the Pólya tree prior for probability measures in an interval Ω. The interval is recursively bisected into subintervals. At each stage, the probability mass already assigned to an interval is randomly divided and assigned into its subintervals according to the independent draw of a Beta variable. However, in order for the prior to generate absolutely continuous measures, it is necessary for the parameters in the Beta distribution to increase rapidly as the depth of the bisection increases, that is, as we move into more and more refined levels of partitioning [13].

In any case, even when the construction yields a random distribution with density, with probability 1 the density will have discontinuity almost everywhere. The use of Beta variables with large magnitudes for its parameters, although useful in forcing the random distribution to be absolutely continuous, has the effect of severely constraining our ability to allocate conditional probability to represent faithfully the data distributions within small intervals. To resolve this conflict between smoothness and faithfulness to the data distribution, one can introduce an optional stopping variable for each subregion obtained in the partitioning process [12]. By putting uniform distributions within each stopped subregion, we can achieve the goal of generating absolutely continuous distributions without having to force the Beta parameters to increase rapidly. In fact, we will be able to use Jeffrey's rule of Beta $(\frac{1}{2}, \frac{1}{2})$ in the inference of conditional probabilities, regardless of the depth of the subregion in the partition tree. We believe this is a desirable consequence of optional stopping.

Our second extension is to allow randomized partitioning. Standard Pólya tree construction relies on a fixed scheme for partitioning. For example in [11] a k-dimensional rectangle is recursively partitioned where in each stage of the recursion the subregions are further divided into 2^k quadrants by bisecting each of the k coordinate variables. In contrast, when recursive partitioning is used in other statistical problems, it is customary to allow flexible choices of the variables to use to further divide a subregion. This allows the subregion to take very different shapes depending on the information in the data. The data-adaptive nature of the recursive partitioning is a reason for the success of tree-based learning methodologies such as CART [3]. Thus it is desirable to allow Pólya tree priors to use partitions that are the result of randomized choices of divisions in each of the subregions at each stage of the recursion. Once the partitioning is randomized in the prior, the posterior distribution will give more weights on those partitions that provide better fits to the data. In this way the data is allowed to influence the choice of the partitioning. This will be especially useful in high-dimensional applications.

In Section 2 we introduce the construction of "Optional Pólya trees" that allow optional stopping and randomized partitioning. It is shown that this construction leads to priors that give absolutely continuous distributions almost surely. We also show how to specify the prior so that it has positive probability in all total variation neighborhoods in the space of absolutely continuous distributions on Ω. In Section 3 we show that the use of optional Pólya tree priors will lead to posterior distributions that are also optional Pólya trees. We present a recursive algorithm for the computation of the parameters governing the posterior optional Pólya tree. These results ensure that Ferguson's two criteria are satisfied by optional Pólya tree priors, but now on the space of absolutely continuous probability measures. In this section, we also show that the posterior Pólya tree is weakly consistent in the sense that asymptotically it concentrates all its probability in any weak neighborhood of a true distribution whose density is bounded. In Section 4, we develop and test the optional Pólya tree approach to density estimation in Euclidean space. Concluding remarks are given in Section 5.

We end this introduction with brief remarks on related works. The important idea of early stopping was first introduced by Hutter [12]. Ways to attenuate the dependency of Pólya trees on the partition include mixing the base measure used to define the tree [11, 14, 15], random perturbation of the dividing boundary in the partition of intervals [19] and the use of positively correlated variables for the conditional probabilities at each level of the tree definition (Nieto-Barajas and Müller [18]). Compared to these works, our approach allows not only early stopping but also randomized choices of the splitting variables. This provides a much richer class of partitions than previous models and raises the new challenge of learning the partition based on the observed data. We show that under mild conditions such learning is achievable by finite computation. We also provide a relatively complete mathematical foundation which represents the first theory for Bayesian density estimation based on recursive partitioning. Although a Bayesian version of recursive partitioning has been proposed previously (Bayesian CART, [4]), it was formulated for a different problem (classification instead of density estimation). Furthermore, it studied mainly model specification and computational algorithm, and did not discuss the mathematical and asymptotic properties of the method.

2. Optional Pólya tree. We are interested in constructing random probability measures on a space (Ω, μ). Ω is either finite or a bounded rectangle in $\mathbb{R}^p$. In this paper we assume for simplicity that μ is the counting measure in the finite case and the Lebesgue measure in the continuous case. Suppose that Ω can be partitioned in M different ways; that is, for $j = 1, 2, \dots, M$,

$$\Omega = \bigcup_{k=1}^{K^j} \Omega_k^j \qquad \text{where } \Omega_k^j\text{'s are disjoint.}$$

Each Ω_k^j, called a level-1 elementary region, can in turn be divided into level-2 elementary regions. Assume there are $M_{k_1}^{j_1}$ ways to divide $\Omega_{k_1}^{j_1}$; then for $j_2 = 1, \dots, M_{k_1}^{j_1}$, we have

$$\Omega_{k_1}^{j_1} = \bigcup_{k_2=1}^{K_{k_1}^{j_1 j_2}} \Omega_{k_1 k_2}^{j_1 j_2}.$$

In general, for any level-k elementary region A, we assume there are $M(A)$ ways to partition it; that is, for $j = 1, 2, \dots, M(A)$,

$$A = \bigcup_{k=1}^{K^j(A)} A_k^j.$$

Let $\mathcal{A}^k$ be the set of all possible level-k elementary regions, and $\mathcal{A}^{(k)} = \bigcup_{l=1}^k \mathcal{A}^l$. If Ω is finite, we assume that $\mathcal{A}^k$ separates points in Ω if k is large enough. If Ω is a rectangle in $\mathbb{R}^p$, we assume that every open set $B \subset \Omega$ is approximated by unions of sets in $\mathcal{A}^{(n)}$, that is, $\exists B_n \uparrow B$ where B_n is a finite union of disjoint regions in $\mathcal{A}^{(n)}$.

EXAMPLE 1.

$$\begin{aligned} \Omega &= \{x = (x_1, \dots, x_p) : x_i \in \{1, 2\}\}, \\ \Omega_k^j &= \{x : x_j = k\}, \qquad k = 1 \text{ or } 2, \\ \Omega_{k_1 k_2}^{j_1 j_2} &= \{x : x_{j_1} = k_1, x_{j_2} = k_2\}, \qquad \text{etc.} \end{aligned}$$

In this example, the number of ways to partition a level-k elementary region decreases as k increases.

EXAMPLE 2.

$$\Omega = \{(x_1, x_2, \dots, x_p) : x_i \in [0, 1]\} \subset \mathbb{R}^p.$$

If A is a level-k elementary region (a rectangle), and $m_j(A)$ is the midpoint of the range of x_j for A, we set $A_1^j = \{x \in A : x_j \le m_j(A)\}$ and $A_2^j = A \setminus A_1^j$. There are exactly $M(A) = p$ ways to partition each A, regardless of its level.

Once a system to generate partitions has been specified as above, we can formally define recursive partitions as follows. A recursive partition of depth k is a series of decisions $J^{(k)} = (J_1, J_2, \dots, J_k)$ where J_l represents all the decisions made at level l to decide, for each region produced at the previous level, whether or not to stop partitioning it further and if not, which way to use to partition it. Once we have decided not to partition a region, then it will remain intact at all subsequent levels. Thus each $J^{(k)}$ specifies a partition of Ω into a subset of regions in $\mathcal{A}^{(k)}$.

We use a recursive procedure to produce a random recursive partition of Ω and a random probability measure Q that is uniformly distributed within each part of the partition. Suppose after k steps of the recursion, we have obtained a random recursive partition $\mathbf{J}^{(k)}$ and we write

$$\Omega = T_0^k \cup T_1^k,$$

where

$$T_0^k = \bigcup_{i=1}^{I} A_i \qquad \text{is a union of disjoint } A_i \in \mathcal{A}^{(k-1)},$$

$$T_1^k = \bigcup_{i=1}^{I'} A_i' \qquad \text{is a union of disjoint } A_i' \in \mathcal{A}^{k}.$$

The set T_0 represents the part of Ω where the partitioning has already been stopped and T_1 represents the complement. In addition, we have also obtained a random probability measure $Q^{(k)}$ on Ω which is uniformly distributed within each region in T_0^k and T_1^k.

In the $(k+1)$th step, we define $Q^{(k+1)}$ by further partitioning of the regions in T_1^k as follows. For each elementary region A in the above decomposition of T_1^k, generate an independent random variable,

$$S \sim \text{Bernoulli}(\rho).$$

If $S = 1$, stop further partitioning of A and add it to the set of stopped regions. If $S = 0$, draw $J \in \{1, 2, \dots, M(A)\}$ according to a nonrandom vector $\boldsymbol{\lambda}(A) = (\lambda_1, \dots, \lambda_{M(A)})$, called the selection probability vector, that is, $P(J = j) = \lambda_j$ and $\sum_{l=1}^{M(A)} \lambda_l = 1$. If $J = j$, apply the jth way of partitioning A,

$$A = \bigcup_{l=1}^{K} A_l^j \qquad \text{(here } K \text{ depends on } A \text{ and } j)$$

and set $Q^{(k+1)}(A_l^j) = Q^{(k)}(A)\theta_l^j$ where $\boldsymbol{\theta}^j = (\theta_1^j, \dots, \theta_K^j)$ is generated from a Dirichlet distribution with parameter $(\alpha_1^j, \dots, \alpha_K^j)$. The nonrandom vector $\boldsymbol{\alpha}^j = \boldsymbol{\alpha}^j(A)$ is referred to as the assignment weight vector.

After this step, we have obtained T_0^{k+1} and T_1^{k+1}, the respective unions of the stopped and continuing regions. Clearly

$$\Omega = T_0^{k+1} \cup T_1^{k+1},$$
$$T_0^{k+1} \supset T_0^k, \qquad T_1^{k+1} \subset T_1^k.$$

The new measure $Q^{(k+1)}$ is then defined as a refinement of $Q^{(k)}$. For $B \subset T_0^{(k+1)}$, we set

$$Q^{(k+1)}(B) = Q^{(k)}(B).$$

For $B \subset T_1^{(k+1)}$ where T_1^{k+1} is partitioned as

$$T_1^{k+1} = \bigcup_{i=1}^{J} A_i, \qquad A_i \in \mathcal{A}^{k+1},$$

we set

$$Q^{(k+1)}(B) = \sum_{i=1}^{J} Q^{(k+1)}(A_i)\left(\frac{\mu(A_i \cap B)}{\mu(A_i)}\right).$$

Recall that for each A_i in the partition of T_1^{k+1}, we have already generated its $Q^{(k+1)}$ probability.

Let $\mathcal{F}^{(k)}$ be the σ-field of events generated by all random variables used in the first k steps; the stopping probability $\rho = \rho(A)$ is required to be measurable with respect to $\mathcal{F}^{(k)}$. The specification of $\rho(\cdot)$ is called the stopping rule. In this paper we are mostly interested in the case when $\rho(\cdot)$ is an "independent stopping rule;" that is, $\rho(A)$ is a pre-specified constant for each possible elementary region A. However in some applications it is useful to let $\rho(A)$ depend on $Q^{(k)}(A)$.

Let $\mathcal{A}^{(\infty)} = \bigcup_{k=1}^{\infty} \mathcal{A}^k$ be the set of all possible elementary regions.

THEOREM 1. *Suppose there is a* $\delta > 0$ *such that with probability* 1, $1 - \delta > \rho(A) > \delta$ *for any region* A *generated during any step in the recursive partitioning process. Then with probability* 1, $Q^{(k)}$ *converges in variational distance to a probability measure* Q *that is absolutely continuous with respect to* μ.

DEFINITION 1. The random probability measure Q defined in Theorem 1 is said to have an optional Pólya tree distribution with parameters $\boldsymbol{\lambda}, \boldsymbol{\alpha}$ and stopping rule ρ.

PROOF OF THEOREM 1. We only need to prove this for the case when Ω is a bounded rectangle. We can think of $Q^{(k)}$'s as being generated in two steps.

1. Generate the nonstopped version $Q^{*(k)}$ by recursively choosing the ways of partitioning each level of regions but without stopping in any of the regions. Let $J^{*(k)}$ denote the decision made during this process in the first k levels of the recursion. Each realization of $J^{*(k)}$ determines a partition of Ω consisting of regions $A \in \mathcal{A}^k$ (not $\mathcal{A}^{(k)}$ as in the case of optional stopping). Let $\mathcal{A}^k(J^{*(k)}) = \{A \in \mathcal{A}^k : A$ is a region in the partition induced by $J^{*(k)}\}$. If $A \in \mathcal{A}^k(J^{*(k)})$, then it can be written as
$$A = \Omega_{l_1 l_2 \cdots l_k}^{j_1 j_2 \cdots j_k}.$$
We set
$$Q^{*(k)}(A) = \theta_{l_1}^{j_1} \cdot \theta_{l_1 l_2}^{j_1 j_2} \cdots \theta_{l_1 \cdots l_k}^{j_1 \cdots j_k} \quad \text{and} \quad Q^{*(k)}(\cdot|A) = \mu(\cdot|A).$$
This defines $Q^{*(k)}$ as a random measure.
2. Given the results in Step 1, generate the optional stopping variables $S = S(A)$ for each region $A \in \mathcal{A}^k(J^{*(k)})$, successively for each level $k = 1, 2, 3, \ldots$. Then for each k, modify $Q^{*(k)}$ to get $Q^{(k)}$ by replacing $Q^{*(k)}(\cdot|A)$ with $\mu(\cdot|A)$ for any stopped region A up to level k.

For each $A \in \mathcal{A}^k(J^{*(k)})$, let $I^k(A) =$ indicator of the event that A has not been stopped during the first k levels of the recursion:
$$\begin{aligned}
E(Q^{(k)}(T_1^k)|J^{*(k)}) &= E\Bigl(\sum_{A \in \mathcal{A}^k(J^{*(k)})} Q^{*(k)}(A) I^k(A)|J^{*(k)}\Bigr) \\
&= \sum_{A \in \mathcal{A}^k(J^{*(k)})} E(Q^{*(k)}(A)|J^{*(k)}) E(I^k(A)|J^{*(k)}) \\
&\le (1-\delta)^k \sum_{A \in \mathcal{A}^k(J^{*(k)})} E(Q^{*(k)}(A)|J^{*(k)}) \\
&= (1-\delta)^k.
\end{aligned}$$
Thus $E(Q^{(k)}(T_1^k)) \to 0$ geometrically and hence $Q^{(k)}(T_1^k) \to 0$ with probability 1. Similarly, $\mu(T_1^k) \to 0$ with probability 1.

For any Borel set $B \subset \Omega$, we claim that $\lim Q^{(k)}(B)$ exists with probability 1. To see this, write
$$\begin{aligned}
Q^{(k)}(B) &= Q^{(k)}(B \cap T_0^k) + Q^{(k)}(B \cap T_1^k) \\
&= a_k + b_k;
\end{aligned}$$
a_k is increasing since
$$\begin{aligned}
Q^{(k+1)}(B \cap T_0^{k+1}) &\ge Q^{(k+1)}(B \cap T_0^k) \\
&= Q^{(k)}(B \cap T_0^k),
\end{aligned}$$

and $b_k \to 0$ since $Q^{(k)}(T_1^k) \to 0$ with probability 1.

Since the Borel σ-field $\mathcal{B}$ is generated by countably many rectangles, we have with probability 1 that $\lim Q^{(k)}(B)$ exists for all $B \in \mathcal{B}$. Define $Q(B)$ as this limit. If $Q(B) > 0$ then $Q^{(k)}(B) > 0$ for some k. Since $Q^{(k)} \ll \mu$ by construction, we must also have $\mu(B) > 0$. Thus Q is absolutely continuous.

For any $B \in \mathcal{B}$, $Q^{(k)}(B \cap T_0^k) = Q(B \cap T_0^k)$, and hence

$$\begin{aligned}|Q^{(k)}(B) - Q(B)| &= |Q^{(k)}(B \cap T_1^k) - Q(B \cap T_1^k)| \\ &< 2Q^{(k)}(T_1^k) \longrightarrow 0.\end{aligned}$$

Thus the convergence of $Q^{(k)}$ to Q is in variational distance. □

The next result shows that it is possible to construct optional Pólya tree distribution with positive probability on all L_1 neighborhoods of densities.

THEOREM 2. *Let Ω be a bounded rectangle in $\mathbb{R}^p$. Suppose that the condition of Theorem 1 holds and that the selection probabilities $\lambda_i(A)$, the assignment probabilities $\alpha_i^j(A)/(\sum_l \alpha_l^j(A))$ for all i, j and $A \in \mathcal{A}^{(\infty)}$ are uniformly bounded away from 0 and 1. Let $q = dQ/d\mu$; then for any density f and any $\tau > 0$, we have*

$$P\left(\int |q(x) - f(x)|\, d\mu < \tau\right) > 0.$$

PROOF. First assume that f is uniformly continuous. Let

$$\delta(\varepsilon) = \sup_{|x-y|<\varepsilon} |f(x) - f(y)|;$$

then $\delta(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$. For any k large enough, we can find a partitioning $\Omega = \bigcup_{i=1}^I A_i$ where $A_i \in \mathcal{A}^k$ is arrived at by k steps of recursive partitioning (deterministic and without stopping) and that each A_i has diameter $< \varepsilon$.

Approximate f by a step function $f^*(x) = \sum_i f_i^* I_{A_i}(x)$, $f_i^* = \int_{A_i} f\, d\mu / \mu(A_i)$. Let $D_\varepsilon(f)$ be the set of step functions $g(\cdot) = \sum g_i I_{A_i}(\cdot)$ satisfying

$$\sup_i |g_i - f_i^*| < \delta(\varepsilon).$$

Suppose $g \in D_\varepsilon(f)$; then for any B we have $B = \bigcup_{i=1}^I (B \cap A_i) = \bigcup_{i=1}^I B_i$ and

$$\begin{aligned}\left|\int_B (g - f)\, d\mu\right| &\le \sum_i |g_i - f_i^*| \mu(B_i) + \sum_i \left| f_i^* \mu(B_i) - \int_{B_i} f\, d\mu\right| \\ &\le \sum_i \delta(\varepsilon)\mu(B_i) + \sum_i r_i,\end{aligned}$$

where

$$r_i = \mu(B_i)\left|\frac{\int_{A_i} f\,d\mu}{\mu(A_i)} - \frac{\int_{B_i} f\,d\mu}{\mu(B_i)}\right|$$
$$= \mu(B_i)\left|\frac{\int_{A_i}(f(x) - f(x_k))\,d\mu}{\mu(A_i)} - \frac{\int_{B_i}(f(x) - f(x_k))\,d\mu}{\mu(B_i)}\right|,$$

where $x_i \in B_i$. Since

$$|f(x) - f(x_i)| < \delta(\varepsilon) \qquad \text{for } x \in A_i,$$

we have

$$|r_i| < 2\delta(\varepsilon)\mu(B_i).$$

Hence

$$\left|\int_B (g - f)\,d\mu\right| < 3\delta(\varepsilon)\mu(B) \qquad \forall B,$$

and thus

$$\int |g - f|\,d\mu < 3\delta(\varepsilon)\mu(\Omega) = 3\delta'(\varepsilon),$$

where $\delta'(\varepsilon) = \delta(\varepsilon)\mu(\Omega)$. Since all probabilities in the construction of $q^k = \frac{dQ^{(k)}}{d\mu}$ are bounded away from 0 and 1, we have

$$P\bigl(q^k \in D_\varepsilon(f) \text{ for all large } k\bigr) > 0.$$

Hence

$$P\biggl(\int |q^k - f|\,d\mu < 3\delta'(\varepsilon) \text{ for all large } k\biggr) > 0.$$

On the other hand, by Theorem 1, we have

$$P\biggl(\int |q^k - q|\,d\mu \to 0\biggr) = 1.$$

Thus

$$P\biggl(\int |q - f|\,d\mu < 4\delta'(\varepsilon)\biggr) > 0.$$

Finally, the result also holds for a discontinuous f since we can approximate it arbitrarily closely in L_1 distance by a uniformly continuous one. □

It is not difficult to specify $\alpha_i^j(A)$ to satisfy the assumption of Theorem 2. A useful choice is

$$\alpha_i^j(A) = \tau^k \mu(A_i^j)/\mu(\Omega) \qquad \text{for } A \in \mathcal{A}^k,$$

where $\tau > 0$ is a suitable constant.

The reason for including the factor τ^k when $A \in \mathcal{A}^k$ is to ensure that the strength of information we specified for the conditional probabilities within A is not diminishing as the depth of partition k increases. For example, in Example 2 each A is partitioned into two parts of equal volumes; that is,

$$A = A_1^j \cup A_2^j, \qquad \mu(A_1^j) = \mu(A_2^j) = \tfrac{1}{2}\mu(A).$$

Thus $A \in \mathcal{A}^k \Rightarrow \mu(A_i^j) = 2^{-(k+1)}\mu(\Omega)$, and

$$\alpha_i^j(A) = 2^k \frac{\mu(A_i^j)}{\mu(\Omega)} = \frac{1}{2} \qquad \text{for all } k.$$

In this case, by choosing $\tau = 2$ we have obtained a nice "self-similarity" property for the optional Pólya tree, in the sense that the conditional probability measure $Q(\cdot|A)$ will have an optional Pólya tree distribution with the same specification for α_i^j's as in the original optional Pólya tree distribution for Q.

Furthermore, in this example if we use $\tau = 2$ to specify a prior distribution for Bayesian inference of Q, then for any $A \in \mathcal{A}^k$, the inference for the conditional probability $\theta_1^j(A)$ will follow a classical binomial Bayesian inference with the Jeffrey's prior Beta $(\frac{1}{2}, \frac{1}{2})$.

3. Bayesian inference with an optional Pólya tree prior. Suppose we have observed $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ where x_i's are independent draws from a probability measure Q, where Q is assumed to have an optional Pólya tree as a prior distribution. In this section we show that the posterior distribution of Q given $\mathbf{x}$ also follows an optional Pólya tree distribution.

We denote the prior distribution for $q = \frac{dQ}{d\mu}$ by $\pi(\cdot)$. For any $A \subset \Omega$, we define $\mathbf{x}(A) = \{x_i \in \mathbf{x} : x_i \in A\}$ and $n(A) = \#(\mathbf{x}(A)) =$ cardinality of the set $\mathbf{x}(A)$. Let

$$q(x) = \frac{dQ}{d\mu}(x) \qquad \text{for } x \in \Omega$$

and

$$q(x|A) = \frac{q(x)}{Q(A)} \qquad \text{for } x \in A;$$

then the likelihood for $\mathbf{x}$ and the marginal density for $\mathbf{x}$ can be written, respectively, as

$$P(\mathbf{x}|Q) = \prod_{i=1}^{n} q(x_i) = q(\mathbf{x}),$$

$$P(\mathbf{x}) = \int q(\mathbf{x})\, d\pi(q).$$

The variable q (or Q) represents the whole set of random variables, that is, the stopping variable $S(A)$, the selection variable $J(A)$ and the condition probability allocation $\theta_i^j(A)$, etc., for all regions A generated during the generation of the random probability measure Q.

In what follows, we assume that the stopping rule needed for Q is an independent stopping rule. By considering how Ω is partitioned and how probabilities are assigned to the parts of this partition, we have

$$q(\mathbf{x}) = Su(\mathbf{x}) + (1-S)\left(\prod_{i=1}^{K^J}(\theta_i^J)^{n_i^J}\right)q(\mathbf{x}|\mathbf{N}^J = \mathbf{n}^J). \tag{3.1}$$

In this expression:

(i) $u(\mathbf{x}) = \prod_{i=1}^n u(x_i)$ where $u(x) = \frac{1}{\mu(\Omega)}$ is the uniform density on Ω.
(ii) $S = S(\Omega)$ is the stopping variable for Ω.
(iii) J is the choice of partitioning to use on Ω.
(iv) $\mathbf{N}^J = (n(\Omega_1^J), \dots, n(\Omega_{K^J}^J))$ is the counts of observations in $\mathbf{x}$ falling into each part of the partition J.

To understand $q(\mathbf{x}|\mathbf{N}^J = \mathbf{n}^j)$, suppose $J = j$ specifies a partition $\Omega = \Omega_1^j \cup \Omega_2^j \cup \cdots \cup \Omega_{K^j}^j$; then the sample $\mathbf{x}$ is partitioned accordingly into subsamples,

$$\mathbf{x} = \mathbf{x}(\Omega_1^j) \cup \cdots \cup \mathbf{x}(\Omega_{K^j}^j).$$

Under Q, if the subsample sizes $n_1^j, \dots, n_{K^j}^j$ are given, then the positions of points in $\mathbf{x}(\Omega_i^j)$ within Ω_i^j are generated independently of those in the other subregions. Thus

$$q(\mathbf{x}|\mathbf{N}^J = n^j) = \prod_{i=1}^{K^j} q(\mathbf{x}(\Omega_i^j)|\Omega_i^j),$$

where

$$q(\mathbf{x}(\Omega_i^j)|\Omega_i^j) = \prod_{x\in\mathbf{x}(\Omega_i^j)} q(x|\Omega_i^j).$$

Note that once $J = j$ is given, $q(\cdot|\Omega_i^j)$ is generated independently as an optional Pólya tree according to the parameters $\boldsymbol{\rho}, \boldsymbol{\lambda}, \boldsymbol{\alpha}$ that are relevant within Ω_i^j. We denote by $\Phi(\Omega_i^j)$ the expectation of $q(\mathbf{x}(\Omega_i^j)|\Omega_i^j)$ under this induced optional Pólya tree within Ω_i^j.

In fact, for any $A \subset \bigcup_{k=1}^\infty \mathcal{A}^k$, we have an induced optional Pólya tree distribution $\pi_A(q)$ for the conditional density $q(\cdot|A)$, and we define

$$\Phi(A) = \int q(\mathbf{x}(A)|A)\, d\pi_A(q),$$

if $\mathbf{x}(A) \neq \varnothing$ and $\Phi(A) = 1$ if $\mathbf{x}(A) = \varnothing$. Similarly, we define

$$\Phi_0(A) = u(\mathbf{x}(A)|A) = \prod_{x \in \mathbf{x}(A)} u(x|A)$$

and $\Phi_0(A) = 1$ if $\mathbf{x}(A) = \varnothing$. Note that $P(\mathbf{x}) = \Phi(\Omega)$ and $u(\mathbf{x}) = \Phi_0(\Omega)$.

Next, we successively integrate out [w.r.t. $\pi(\cdot)$] the random variables in the right-hand side of (3.1) according to the order $q(\mathbf{x}|\mathbf{n}^J), \boldsymbol{\theta}^J$, J and S (last). This gives us

$$\Phi(\Omega) = \rho\Phi_0(\Omega) + (1-\rho)\sum_{j=1}^{M} \lambda_j \frac{D(\mathbf{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} \Phi(\Omega_i^j), \tag{3.2}$$

where $D(\mathbf{t}) = \Gamma(t_1)\cdots\Gamma(t_k)/\Gamma(t_1 + \cdots + t_k)$.

Similarly, for any $A \in \bigcup_{k=1}^{\infty} \mathcal{A}^k$ with $\mathbf{x}(A) \neq \varnothing$, we have

$$\Phi(A) = \rho\Phi_0(A) + (1-\rho)\sum_{j=1}^{M} \lambda_j \frac{D(\mathbf{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} \Phi(A_i^j), \tag{3.3}$$

where $\mathbf{n}^j$ is the vector of counts in the partition $A = \bigcup_{i=1}^{K^j} A_i^j$, and $M, K^j, \rho, \boldsymbol{\lambda}^j$, $\boldsymbol{\alpha}^j$, etc., all depend on A. We note that in the special case when the choice of splitting variables are nonrandom, a similar recursion was given in [12].

We can now read off the posterior distribution of $S = S(\Omega)$ from equation (3.2) by noting that the first term $\rho\Phi_0(\Omega)$ and the remainder in the right-hand side of (3.2) are, respectively, the probabilities of the events

$$\{\text{stopped at } \Omega, \text{ generate } \mathbf{x} \text{ from } u(\cdot)\}$$

and

$$\{\text{not stopped at } \Omega, \text{ generate } \mathbf{x} \text{ by one of the } M \text{ partitions}\}.$$

Thus $S \sim$ Bernoulli with probability $\rho\Phi_0(\Omega)/\Phi(\Omega)$. Similarly, the jth term in the sum (over j) appearing in the right-hand side of (3.2) is the probability of the event

$$\{\text{not stopped at } \Omega, \text{ generate } \mathbf{x} \text{ by using the } j\text{th way to partition } \Omega\}.$$

Hence, conditioning on not stopping at Ω, J takes value j with probability proportional to

$$\lambda_j \frac{D(\mathbf{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} \Phi(\Omega_i^j).$$

Finally, given $J = j$, the probabilities assigned to the parts of this partition are $\boldsymbol{\theta}^j$ whose posterior distribution is Dirichlet $(\mathbf{n}^j + \boldsymbol{\alpha}^j)$.

By similar reasoning, we can also read off the posterior distribution of $S = S(A), J = J(A), \boldsymbol{\theta}^j = \boldsymbol{\theta}^j(A)$ from (3.3) for any $A \subset \mathcal{A}^k$. Thus we have proven the following.

THEOREM 3. *Suppose* $\mathbf{x} = (x_1, \dots, x_n)$ *are independent observations from* Q *where* Q *has a prior distribution* $\pi(\cdot)$ *that is an optional Pólya tree with independent stopping rule, and satisfying the condition of Theorem 2, the conditional distribution of* Q *given* $\mathbf{X} = \mathbf{x}$ *is also an optional Pólya tree where, for each* $A \subset A^\infty$, *the parameters are given as follows*:

1. *Stopping probability*:

$$\rho(A|\mathbf{x}) = \rho(A)\Phi_0(A)/\Phi(A).$$

2. *Selection probabilities*:

$$P(J = j|\mathbf{x}) \propto \lambda_j \frac{D(\mathbf{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} \Phi(A_i^j), \qquad j = 1, \dots, M.$$

3. *Allocation of probability to subregions*: *the probabilities* θ_i^j *for subregion* A_i^j, $i = 1, \dots, K^j$ *are drawn from Dirichlet* $(\mathbf{n}^j + \boldsymbol{\alpha}^j)$.

In the above, it is understood that M, K^j, λ_j, $\mathbf{n}^j$, $\boldsymbol{\alpha}^j$ *all depend on* A.

We use the notation $\pi(\cdot|x_1, x_2, \dots, x_n)$ to denote this posterior distribution for Q.

To use Theorem 3, we need to compute $\Phi(A)$ for $A \in \mathcal{A}^\infty$. This is done by using the recursion (3.3), which says that $\Phi(\cdot)$ is determined for a region A if it is first determined for all subregions A_i^j. By going into subregions of increasing levels of depth, we will eventually arrive at some regions having certain simple relations with the sample $\mathbf{x}$. We can often derive close form solutions for $\Phi(\cdot)$ for such "terminal regions" and hence determine all the parameters in the specifications of the posterior optional Pólya tree by a finite computation. We give two examples.

EXAMPLE 3 (2^p contingency table). Let $\Omega = \{1, 2\} \times \{1, 2\} \times \cdots \times \{1, 2\}$ be a table with 2^p cells. Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be n independent observations where each x_i falls into one of the 2^p cells according to the cell probabilities $\{q(y) : y \in \Omega\}$. Assume that q has an optional Pólya tree distribution according to the partitioning scheme in Example 1 where $\lambda_j = \frac{1}{M}$ if there are M variables still available for further splitting of a region A, and $\alpha_i^j = \frac{1}{2}$, $i = 1, 2$. Finally, assume that $\rho(A) \equiv \rho$ where $\rho \in (0, 1)$ is a constant.

In this example, there are three types of terminal regions.

1. A contains no observation. In this case, $\Phi(A) = 1$.
2. A is a single cell (in the 2^p table) containing any number of observations. In this case, $\Phi(A) = 1$.

3. A contains exactly one observation, and A is a region where M of the p variables are still available for splitting. In this case,

$$\Phi(A) = r_M = \int q(x)\, d\pi_M(Q),$$

where $\pi_M(\cdot)$ is the optional Pólya tree on a 2^M table. By recursion (3.3) we have

$$\begin{aligned} r_M &= \rho 2^{-M} + (1-\rho)\left(\frac{1}{M}\sum_{j=1}^{M}\frac{B(3/2,1/2)}{B(1/2,1/2)}\right)\cdot r_{M-1} \\ &= \rho 2^{-M} + (1-\rho)\frac{1}{2} r_{M-1} \\ &= \rho 2^{-M}\frac{(1-(1-\rho)^M)}{1-(1-\rho)} + \left(\frac{1-\rho}{2}\right)^M \\ &= 2^{-M}. \end{aligned}$$

EXAMPLE 4. Ω is a bounded rectangle in $\mathbb{R}^p$ with a partitioning scheme as in Example 2. Assume that for each region, one of the p variables is chosen to split it ($\lambda_j \equiv \frac{1}{p}$), and that $\alpha_i^j = \frac{1}{2}, i = 1, 2$. Assume $\rho(A)$ is a constant, $\rho \in (0, 1)$. In this case, a terminal region A contains either no observations [then $\Phi(A) = 1$] or a single observation $x \in A$. In the latter case,

$$\Phi(A) = r_A(x) = \int_A q(x|A)\, d\pi_A(Q)$$

and

$$\begin{aligned} r_A(x) &= \frac{\rho}{\mu(A)} + (1-\rho)\frac{1}{p}\sum_{j=1}^{p}\frac{B(3/2,1/2)}{B(1/2,1/2)}\cdot r_{A^j_{i(x)}}(x) \\ &= \frac{\rho}{\mu(A)} + (1-\rho)\frac{1}{2} r_{A^j_{i(x)}}(x), \end{aligned}$$

where $i(x) = 1$ or 2 according to whether $x \in A_1^j$ or A_2^j. Since $\mu(A_1^j) = \mu(A_2^j) = \frac{1}{2}\mu(A)$ for the Lebesgue measure, we have

$$\begin{aligned} r_A(x) &= \frac{\rho}{\mu(A)} + (1-\rho)\frac{1}{2}\left[\frac{\rho}{\mu(A)\cdot 1/2} + (1-\rho)\frac{1}{2}[\cdots]\right] \\ &= \frac{\rho}{\mu(A)}[1 + (1-\rho) + (1-\rho)^2 + \cdots] \\ &= \frac{1}{\mu(A)}. \end{aligned}$$

EXAMPLE 5. Ω is a bounded rectangle in $\mathbb{R}^p$. At each level, we split the regions according to just one coordinate variable, according to a predetermined order; for example, coordinate variable x_i is used to split all regions at the kth step whenever $k \equiv i \pmod{p}$. In this case, $\Phi(A)$ for terminal regions are determined exactly as in Example 4. By allowing only one way to split a region, we sacrifice some flexibility in the resulting partition in exchange for a great reduction of computational complexity.

Our final result in this section shows that optional Pólya tree priors lead to posterior distributions that are consistent in the weak topology. For any probability measure Q_0 on Ω, a weak neighborhood U of Q_0 is a set of probability measures of the form

$$U = \left\{ Q : \left| \int g_i(\cdot)\, dQ - \int g_i(\cdot)\, dQ_0 \right| < \varepsilon_i, i = 1, 2, \ldots, K \right\},$$

where $g_i(\cdot)$ is a bounded continuous function on Ω.

THEOREM 4. *Let $x_1, x_2, \ldots$ be independent, identically distributed variables from a probability measure Q, $\pi(\cdot)$ and $\pi(\cdot|x_1, \ldots, x_n)$ be the prior and posterior distributions for Q as defined in Theorem 3. Then, for any Q_0 with a bounded density, it holds with $Q_0^{(\infty)}$ probability equal to 1 that*

$$\pi(U|x_1, \ldots, x_n) \longrightarrow 1$$

for all weak neighborhoods U of Q_0.

PROOF. It is a consequence of Schwarz's theorem [20] that the posterior is weakly consistent if the prior has positive probability in Kullback–Leibler neighborhoods of the true density [10], Theorem 4.4.2. Thus, by the same argument as in Theorem 2, we only need to show that it is possible to approximate a bounded density in Kullback–Leibler distance by step functions on a suitably refined partition.

Let f be a density satisfying $\sup_{x \in \Omega} f(x) \le M < \infty$. First assume that f is continuous with modulus of continuity $\delta(\varepsilon)$. Let $\bigcup_{i=1}^{I} A_i$ be a recursive partition of Ω satisfying $A_i \in \mathcal{A}^k$ and diameter $(A_i) \le \varepsilon$. Let

$$g_i = \sup_{x \in A_i} f(x), \qquad g(x) = \sum_{i=1}^{I} g_i I_{A_i}(x)$$

and $G = \int g(x)\, d\mu$. We claim that as $\varepsilon \to 0$, the density g/G approximates f arbitrarily well in Kullback–Leibler distance. To see this, note that

$$0 \le G - 1 = \int (g - f)\, d\mu = \sum_i \int_{A_i} \bigl(g(x) - f(x)\bigr)\, d\mu$$

$$\le \sum_i \int_{A_i} \delta(\varepsilon)\, d\mu = \delta(\varepsilon)\mu(\Omega).$$

Hence

$$\begin{aligned}0 &\le \int f \log(f/(g/G))\,d\mu\\ &= \int f \log(f/g)\,d\mu + \int f \log G\,d\mu\\ &\le \log(G) \le \log(1+\delta(\varepsilon)\mu(\Omega)).\end{aligned}$$

Finally, if f is not continuous, we can find a set $B \subset \Omega$ with $\mu(B^c) < \varepsilon'$ such that f is uniformly continuous on B. Then

$$\begin{aligned}\int (g-f)\,d\mu &= \int_B (g-f)\,d\mu + \int_{B^c} (g-f)\,d\mu\\ &\le \delta(\varepsilon)\mu(\Omega) + M\varepsilon'\end{aligned}$$

and the result still holds. □

4. Density estimation using an optional Pólya tree prior.

In this section we develop and test the methods for density estimation using an optional Pólya tree prior. Two different strategies are considered. The first is through computing the posterior mean density. The other is a two-stage approach—first learn a fixed tree topology that is representative of the underlying structure of the distribution, and then compute a piecewise constant estimate *conditional* on this tree topology. Our numerical examples start with the one-dimensional setting to demonstrate some of the basic properties of optional Pólya trees. We then move onto the two-dimensional setting to provide a flavor of what happens when the dimensionality of the distribution increases.

4.1. *Computing the mean.* For the purpose of demonstration, we first consider the situation described in Example 2 with $p = 1$ where the state space is the unit interval and the splitting point of each elementary region (or tree node) is the middle point of its range. In this simple scenario, each node has only one way to divide, so the only decision to make is whether to stop or not. Each point x in the state space Ω belongs to one and only one elementary region in A^k for each k. In this case, the posterior mean density function can be computed very efficiently using an inductive procedure. (See the Appendix for details.)

In a multi-dimensional setting with multiple ways to split at each node, the sets in each A^k could overlap, and so the computation of the posterior mean is more difficult. One way to get around this problem is to place some restriction on how the elementary regions can split. For example, an alternate splitting rule requires that each dimension is split in turn (Example 5). This limits the number of choices to split for each elementary region to one and effectively reduces the dimensionality of the problem to one. However, in restricting the ways to divide, one wastes a lot of computation on cutting dimensions that need not be cut which affects the variability of the estimate significantly. We demonstrate this phenomenon in our later examples.

Another way to compute (or at least approximate) the posterior mean density is first explored by Hutter [12]. For any point $x \in \Omega$, Hutter proposed computing $\Phi(\Omega|x, D)$ and using $\Phi(\Omega|x, D)/\Phi(\Omega|D)$ as an estimate of the posterior mean density at x. [Here D represents the observed data; $\Phi(\Omega|D)$ denotes the Φ computed for the root node given the observed data points and $\Phi(\Omega|x, D)$ is computed treating x as an extra data point observed.] This method is general but computationally intensive, especially when there are multiple ways to divide each node. Also, because this method is for estimating the density at a specific point, to investigate the entire function one must evaluate $\Phi(\Omega|x, D)$ on a grid of x values which makes it even more unattractive computationally. For this reason, in our later two-dimensional examples we only use the restriction method discussed above to compute the posterior mean.

4.2. *The hierarchical MAP method.* Another approach for density estimation using an optional Pólya tree prior is to proceed in two steps—first learn a "good" partition or tree topology over the state space, and then estimate the density conditional on this tree topology. The first step reduces the prior process from an infinite mixture of infinite trees to a fixed finite tree. Given such a fixed tree topology (i.e., whether to stop or not at each step, and if not, which way to divide), we can easily compute the (conditional) mean density function. The posterior probability mass over each node is simply a product of Beta means, and the distribution within those stopped regions is uniform by construction. So the key lies in learning a reliable tree structure. In fact, learning the tree topology is useful beyond facilitating density estimation. A representative partition over the state space by itself sheds light on the underlying structure of the distribution. Such information is particularly valuable in high-dimensional problems where direct visualization of the data is difficult.

Because a tree topology depends only on the decisions to stop and the ways to split, its posterior probability is determined by the posterior ρ's and λ's. The likelihood of each fixed tree topology is the product of a sequence of terms in the form, ρ, $1-\rho$, λ_k, depending on the stopping and splitting decisions at each node. One seemingly obvious candidate tree topology for representing the data structure is the maximum a posteriori (MAP) topology, that is, the topology with the highest posterior probability. However, in this setting the MAP topology often does not produce the most descriptive partition for the distribution. It biases toward shorter tree branches in that deeper tree structures simply have more terms less than 1 to multiply into their posterior probability. While the data typically provide strong evidence for the stopping decisions (and so the posterior ρ's for all but the very deep nodes are either very close to 1 or very close to 0), this is not the case for the λ's. It occurs often that for an elementary region the data points are distributed relatively symmetrically in two or more directions, and thus the posterior λ's for those directions will be much less than 1. As a consequence, deep tree topologies, even if they reflect the actual underlying data structure, often have lower posterior probabilities than shallow trees do. (This failure of the MAP estimate relates more generally to the multi-modality of the posterior distribution as well as the self-similarity of the prior process and deserves more studies in its own right.)

We propose the construction of the representative tree topology through a simple top-down sequential procedure. Starting from the root node, if the posterior $\rho > 0.5$ then we stop the tree; otherwise we divide the tree in the direction k that has the highest λ_k. (When there is more than one direction with the same highest λ_k, the choice among them is arbitrary.) Then we repeat this procedure for each A_k^j until all branches of the tree have been stopped. This can be viewed as a hierarchical MAP decision procedure—with each MAP decision being made based on those made in the previous steps. In the context of building trees, this approach is natural in that it exploits the hierarchy inherent in the problem.

4.3. *Numerical examples.* Next we apply the optional Pólya tree prior to several examples of density estimation in one and two dimensions. We consider the situation described in Example 2 with $p = 1$ and 2 where the state space is the unit interval [0, 1] and the unit square [0, 1] × [0, 1], respectively. The cutting point of each coordinate is the middle point of its range for the corresponding elementary region. For all the optional Pólya tree priors used in the following examples, the prior stopping probability $\rho = 0.5$ and the prior pseudo-count $\alpha = 0.5$ for all elementary regions. The standard Pólya tree priors examined (as a comparison) have quadratically increasing pseudo-counts $\alpha = \text{depth}^2$ (see [8] and [13]). For numerical purpose, we stop dividing the nodes if their support is under a certain threshold which we refer to as the precision threshold. We used 10^{-6} as the precision threshold in the one-dimensional examples and 10^{-4} in the two-dimensional examples. Note that in the 1D examples, each node has only one way to divide, and so we can use the inductive procedure described in the Appendix to compute the posterior mean density function. For the 2D examples, we implemented and tested the full optional tree as well as a restricted version based on "alternate cutting" (see Example 5).

EXAMPLE 6 (Mixture of two close spiky uniforms). We simulate data from the following mixture of uniforms:

$$0.5U(0.23, 0.232) + 0.5U(0.233, 0.235)$$

and we apply three methods to estimate the density function. The first is to compute the posterior mean density using an optional Pólya tree prior. The second is to apply the hierarchical MAP method using an optional Pólya tree prior. The third is to compute the posterior mean using a standard Pólya tree prior. The results are presented in Figure 1. Several points can be made from this figure. (1) A sample

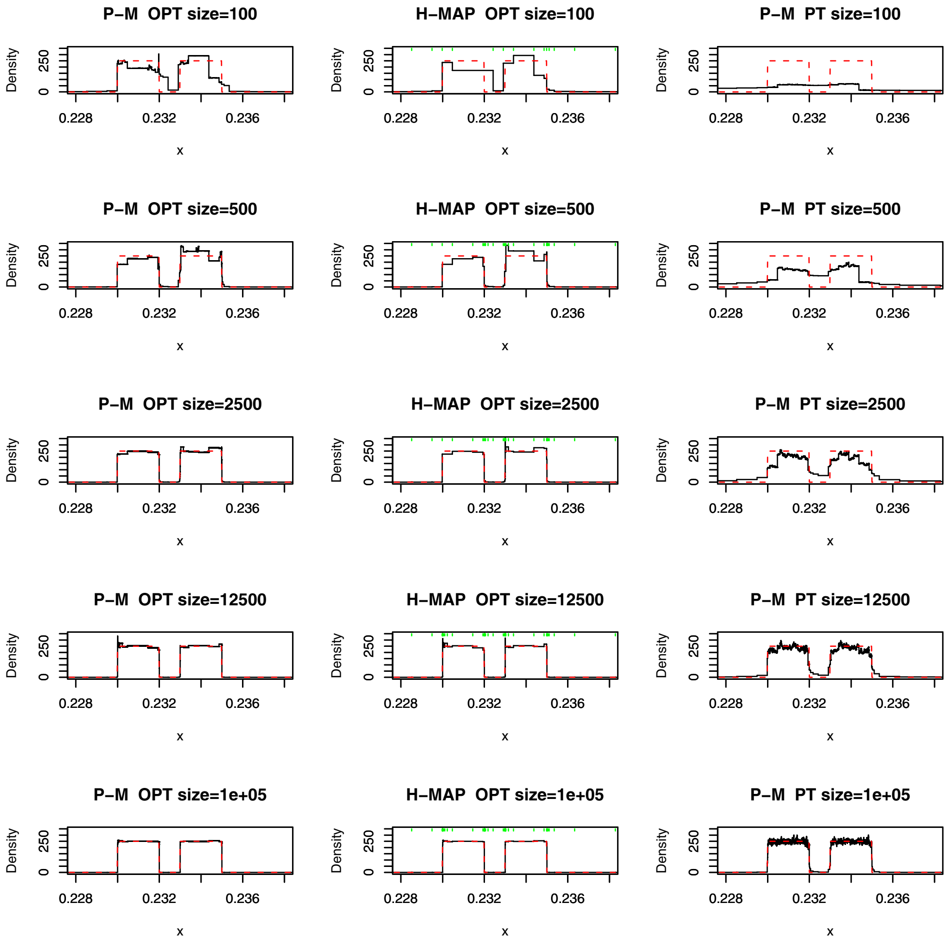


FIG. 1. *Density estimation for* $0.5U(0.23, 0.232) + 0.5U(0.233, 0.235)$. *The five rows represent five different sample sizes* $n = 100$, 500, 2500, 12,500 *and* 100,000. *The first column corresponds to the posterior mean approach using an optional Pólya tree prior. The second column corresponds to the hierarchical MAP method using an optional Pólya tree prior. The green ticks along the top margins of this column indicate the partition learned from this method. The third column corresponds to the posterior mean approach using a standard Pólya tree prior with* $\alpha = \text{depth}^2$. *The red dashed lines in all plots represent the true density function.*

size of 500 is sufficient for the optional tree methods to capture the boundaries as well as the modes of the uniform distributions whereas the Pólya tree prior with quadratic pseudo-counts requires thousands of data points to achieve this. (2) With increasing sample size, the estimates from the optional Pólya tree methods become smoother, while the estimate from the standard Pólya tree with quadratic pseudo-counts is still “locally spiky” even for a sample size of 10^5. (This problem can be remedied by increasing the prior pseudo-counts faster than the quadratic rate at the price of further loss of flexibility.) (3) The hierarchical MAP method performs just as well as the posterior mean approach even though it requires much less computation and memory. (4) The partition learned in the hierarchical MAP approach reflects the structure of the distribution.

EXAMPLE 7 (Mixture of two Betas). Next we apply the same three methods to simulated samples from a mixture of two Beta distributions,

$$0.7\,\mathrm{Beta}(40, 60) + 0.3\,\mathrm{Beta}(2000, 1000).$$

The results are given in Figure 2. Both the optional and the standard Pólya tree methods do a decent job in capturing the locations of the two mixture components (with smooth boundaries). The optional Pólya tree does quite well with just 100 data points.

EXAMPLE 8 (Mixture of Uniform and "semi-Beta" in the unit square). In this example, we consider a mixture distribution over the unit square $[0, 1] \times [0, 1]$. The first component is a uniform distribution over $[0.78, 0.80] \times [0.2, 0.8]$. The second component has support $[0.25, 0.4] \times [0, 1]$ with X being uniform over $[0.25, 0.4]$ and Y being Beta(100, 120), independent of each other. The mixture probability for the two components is $(0.35, 0.65)$. Therefore, the actual density function of the distribution is

$$\frac{0.35}{0.012} \times \mathbf{1}_{[0.78,0.80]\times[0.2,0.8]} + \frac{0.65}{0.15} \times \frac{\Gamma(220)}{\Gamma(120)\Gamma(100)} y^{99}(1-y)^{119}\mathbf{1}_{[0.25,0.4]\times[0,1]}.$$

We apply the following methods to estimate this density—(1) the posterior mean approach using an optional Pólya tree prior with the alternate cutting restriction (Figure 3); (2) the hierarchical MAP method using an optional Pólya tree prior with the alternate cutting restriction (Figure 4); and (3) the hierarchical MAP method using an optional Pólya tree prior without any restriction on division (Figure 5). The last method does a much better job in capturing the underlying structure of the data, and thus requires a much smaller sample size to achieve decent estimates of the density.

EXAMPLE 9 (Bivariate normal). In our last example, we apply the hierarchical MAP method using an optional Pólya tree prior to samples from a bivariate normal distribution,

$$\mathrm{BN}\left(\begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.1^2 & 0 \\ 0 & 0.1^2 \end{pmatrix}\right).$$

This example demonstrates how the posterior optional Pólya tree behaves in a multi-dimensional setting when the underlying distribution has smooth boundary (Figure 6). Not surprisingly, the gradient or change in density is best captured when its direction is perpendicular to one of the coordinates (and thus is parallel to the other in the 2D case).

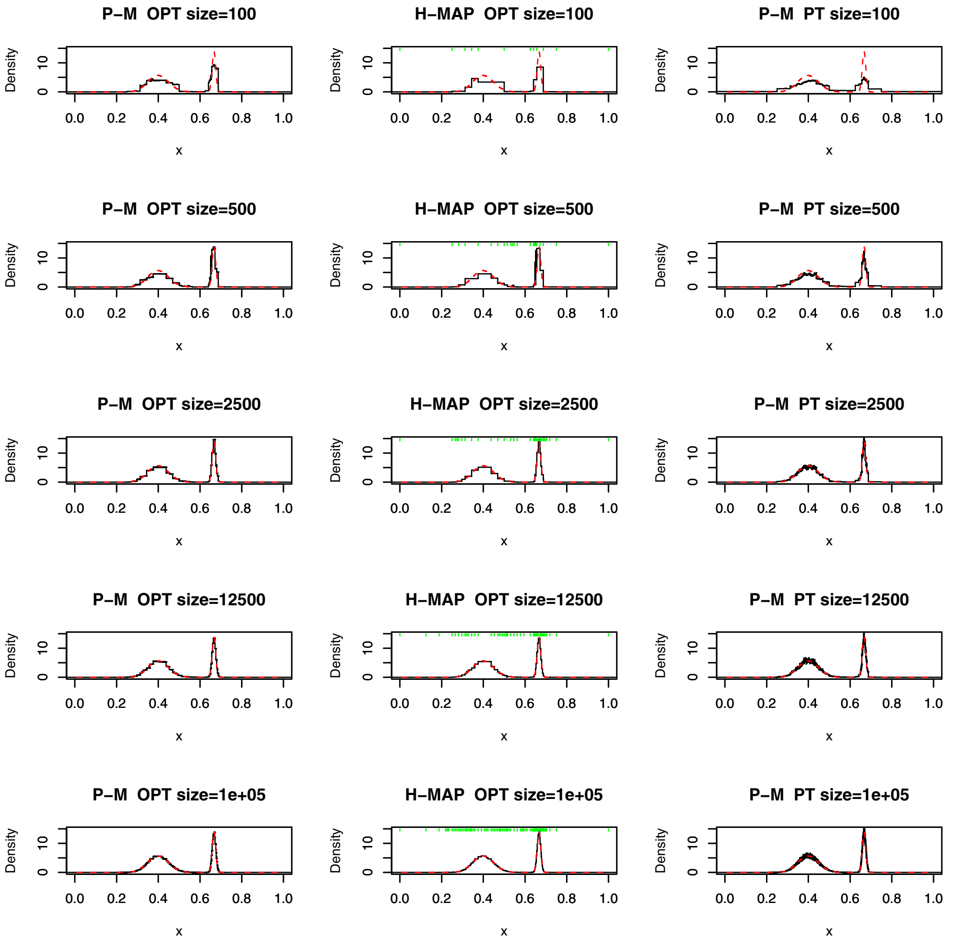
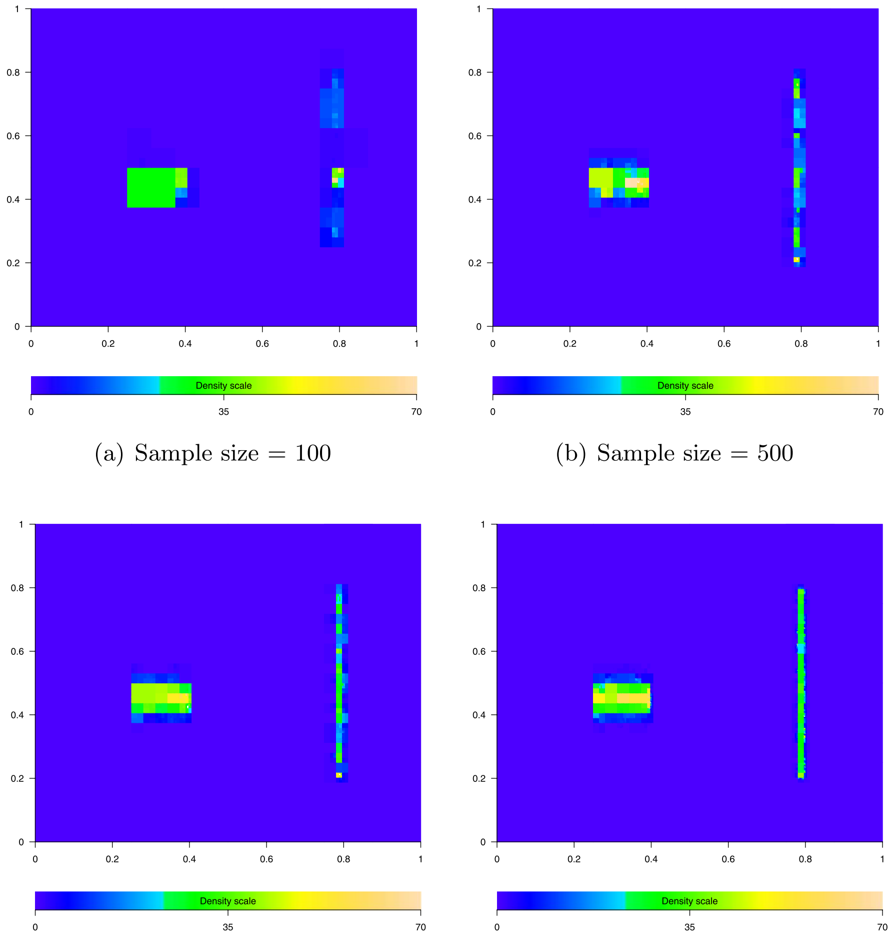


FIG. 2. *Density estimation for* $0.7\,\mathrm{Beta}(40, 60) + 0.3\,\mathrm{Beta}(2000, 1000)$. *The five rows represent five different sample sizes* $n = 100, 500, 2500, 12{,}500$ *and* $100{,}000$. *The first column corresponds to the posterior mean approach using an optional Pólya tree prior. The second column corresponds to the hierarchical MAP method using an optional Pólya tree prior. The green ticks along the top margins of this column indicate the partition learned from this method. The third column corresponds to the posterior mean approach using a standard Pólya tree prior with* $\alpha = \mathrm{depth}^2$. *The red dashed lines in all plots represent the true density function.*

5. Concluding remarks. In this paper we established the existence and the theoretical properties of absolutely continuous probability measures obtained through the Introduction of randomized splitting variables and early stopping rules into a Pólya tree construction. For low-dimensional densities, it is possible to carry out exact computation to obtain posterior inferences based on this "optional Pólya tree" prior. A conceptually important feature of this approach is the ability to learn the partition underlying a piecewise constant density in a principled manner. Al-



(a) Sample size = 100

(b) Sample size = 500

(c) Sample size = 1000

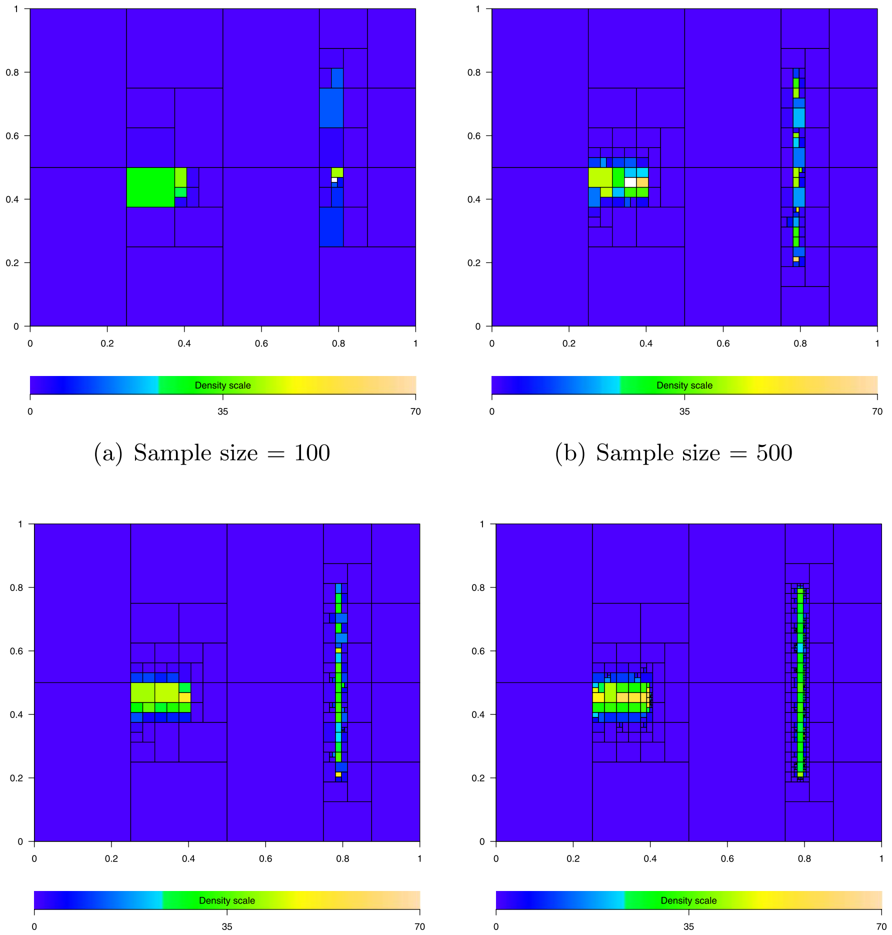
(d) Sample size = 5000

FIG. 3. *Density estimate for a mixture of uniform and "semi-Beta" using the posterior mean approach for an optional Pólya tree with the restriction of "alternate cutting." The white blocks represent the density estimates falling outside of the intensity range plotted.*

though the theory was motivated by applications in high-dimensional problems, at present exact computation is too demanding for such applications. The development of effective approximate computation should be a priority in future works.

APPENDIX

Here we describe an inductive procedure for computing the mean density function of an optional Pólya tree when the way to divide each elementary region is dichotomous and unique.



(a) Sample size = 100

(b) Sample size = 500

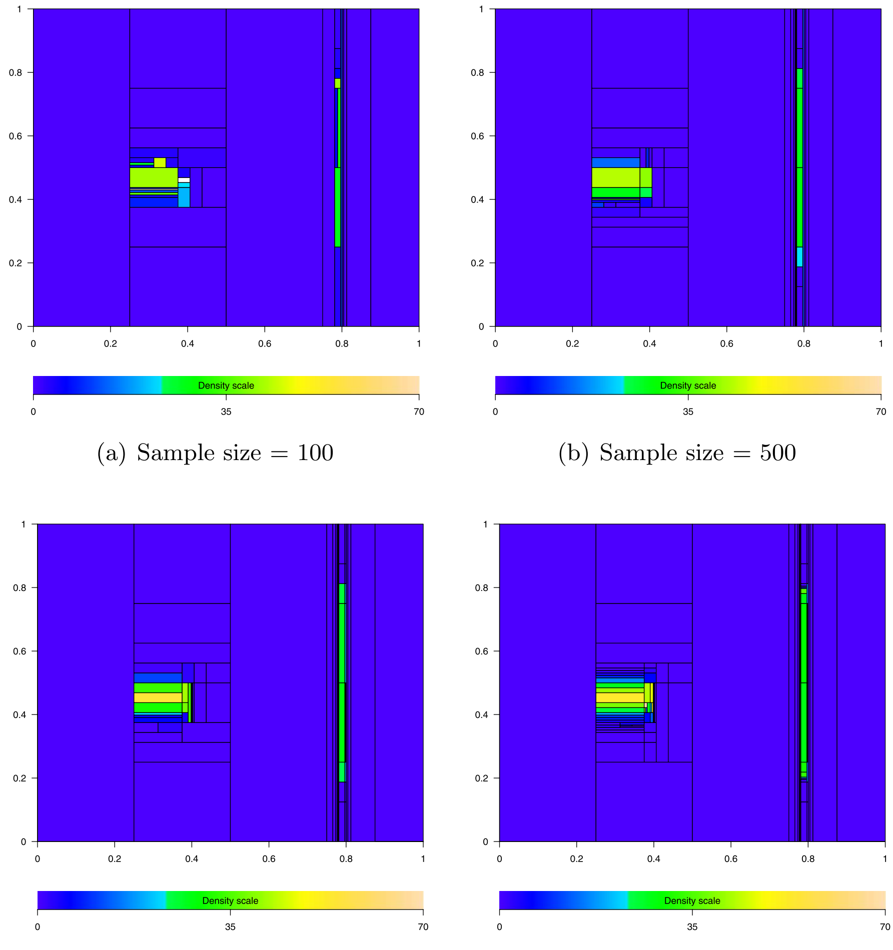
(c) Sample size = 1000

(d) Sample size = 5000

FIG. 4. *Density estimate for a mixture of uniform and "semi-Beta" by the hierarchical MAP method using an optional Pólya tree prior with the restriction of "alternate cutting." The dark lines mark the representative partition learned from the method. The white blocks represent the density estimates falling outside of the intensity range plotted.*

Let A_i denote a level-i elementary region and $(k_1, k_2, \ldots, k_i)$ the sequence of left and right decisions to reach A_i from the root node Ω. That is, $A_i = \Omega_{k_1 k_2 \cdots k_i}$, where the k's take values in $\{0, 1\}$ indicating left and right, respectively. For simplicity, we let $A_0 = \Omega$ represent the root node. Now for any point $x \in \Omega$, let $\{A_i\}$ be the sequence of nodes such that $x \in \bigcup_{i=0}^{\infty} A_i$. Assuming $\mu(A_i) \downarrow 0$, the density of the mean distribution at x is given by

$$\lim_{i \to \infty} EP(X \in A_i)/\mu(A_i).$$



(a) Sample size = 100

(b) Sample size = 500

(c) Sample size = 1000

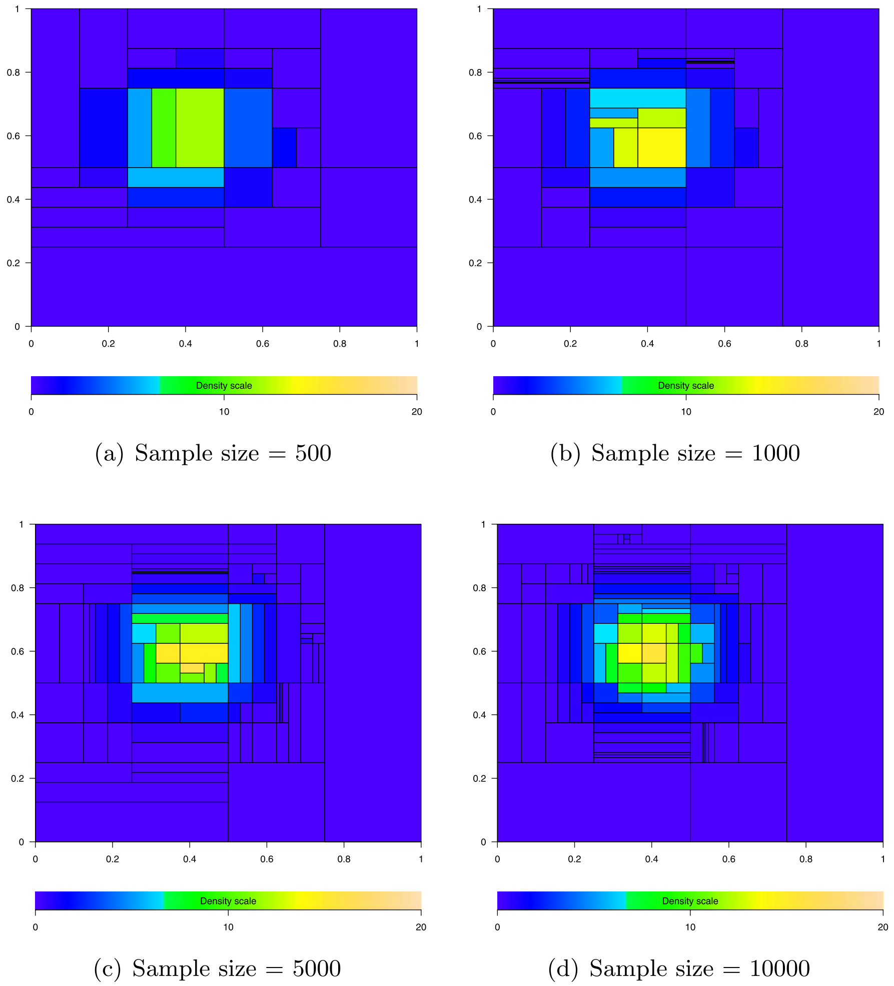
(d) Sample size = 5000

FIG. 5. *Density estimate for a mixture of uniform distribution and "semi-Beta" distribution by the hierarchical MAP method using an optional Pólya tree prior* (*with no restriction on division*). *The dark lines mark the representative partition learned from the method. The white blocks represent the density estimates falling outside of the intensity range plotted.*

Therefore, to compute the mean density we just need a recipe for computing $EP(X \in A_i)$ for any elementary region A_i. To achieve this goal, first let A'_i be the sibling of A_i for all $i \geq 1$. That is,

$$A'_i = \Omega_{k'_1 k'_2 \cdots k'_i} \qquad \text{where } k'_j = k_j \text{ for } j = 1, 2, \ldots, i-1 \text{ and } k'_i = 1 - k_i.$$

Next, for $i \geq 1$, let α_i and α'_i be the Beta parameters for node A_{i-1} associated with its two children A_i and A'_i. Also, for $i \geq 0$, let ρ_i be the stopping probability of A_i, and S_i the event that the tree has stopped growing on or before reaching node A_i.



(a) Sample size = 500

(b) Sample size = 1000

(c) Sample size = 5000

(d) Sample size = 10000

FIG. 6. *The hierarchical MAP method using an optional Pólya tree prior applied to samples from a bivariate normal distribution* $\mathrm{BN}((0.4, 0.6), 0.1^2 I)$.

With this notation, we have for all $i \geq 1$,

$$\begin{aligned}
&EP(X \in A_i)\mathbf{1}(S_i) \\
&\quad = EP(X \in A_i)\mathbf{1}(S_{i-1}) + EP(X \in A_i)\mathbf{1}(S_{i-1}^c)\mathbf{1}(S_i) \\
&\quad = \frac{\mu(A_i)}{\mu(A_{i-1})} EP(X \in A_{i-1})\mathbf{1}(S_{i-1}) \\
&\qquad + \frac{\alpha_i}{\alpha_i + \alpha_i'} \rho_i EP(X \in A_{i-1})\mathbf{1}(S_{i-1}^c)
\end{aligned}$$

and

$$\begin{aligned} EP(X \in A_i)\mathbf{1}(S_i^c) &= EP(X \in A^i)\mathbf{1}(S_i^c)\mathbf{1}(S_{i-1}^c) \\ &= \frac{\alpha_i}{\alpha_i + \alpha_i'}(1 - \rho_i)EP(X \in A_{i-1})\mathbf{1}(S_{i-1}^c). \end{aligned}$$

Now let $a_i = EP(X \in A_i)\mathbf{1}(S_i)$ and $b_i = EP(X \in A_i)\mathbf{1}(S_i^c)$, then the above equations can be rewritten as

$$\begin{cases} a_i = \dfrac{\mu(A_i)}{\mu(A_{i-1})}a_{i-1} + \dfrac{\alpha_i}{\alpha_i + \alpha_i'}\rho_i b_{i-1}, \\ b_i = \dfrac{\alpha_i}{\alpha_i + \alpha_i'}(1 - \rho_i)b_{i-1}, \end{cases} \tag{A.1}$$

for all $i \geq 1$. Because $a_o = EP(X \in \Omega)\mathbf{1}(S_0) = P(S_0) = \rho_0$, and $b_0 = 1 - a_0 = 1 - \rho_0$, we can apply (A.1) inductively to compute the a_i and b_i for all A_i's. Because $EP(X \in A_i) = a_i + b_i$, the mean density at x is given by

$$\lim_{i\to\infty} EP(X \in A_i)/\mu(A_i) = \lim_{i\to\infty}(a_i + b_i)/\mu(A_i).$$

Acknowledgments. The authors thank Persi Diaconis, Nicholas Johnson and Xiaotong Shen for helpful comments, and Cindy Kirby for help in typesetting.

REFERENCES

[1] Blackwell, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1** 356–358. MR0348905

[2] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

[3] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[4] Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. MR1649118

[5] Diaconis, P. and Freedman, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87. MR0829556

[6] Fabius, J. (1964). Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.* **35** 846–856. MR0162325

[7] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[8] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629. MR0438568

[9] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403. MR0158483

[10] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer, New York. MR1992245

[11] Hanson, T. E. (2006). Inference for mixtures of finite Pólya tree models. *J. Amer. Statist. Assoc.* **101** 1548–1565. MR2279479

[12] Hutter, M. (2009). Exact nonparametric Bayesian inference on infinite trees. Technical Report 0903.5342. Available at http://arxiv.org/abs/0903.5342.

[13] KRAFT, C. H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probab.* **1** 385–388. MR0171296

[14] LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235. MR1186248

[15] LAVINE, M. (1994). More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **22** 1161–1176. MR1311970

[16] LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

[17] MAULDIN, R. D., SUDDERTH, W. D. and WILLIAMS, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.* **20** 1203–1221. MR1186247

[18] NIETO-BARAJAS, L. E. and MÜLLER, P. (2009). Unpublished manuscript.

[19] PADDOCK, S. M., RUGGERI, F., LAVINE, M. and WEST, M. (2003). Randomized Polya tree models for nonparametric Bayesian inference. *Statist. Sinica* **13** 443–460. MR1977736

[20] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: whwong@stanford.edu
ma2@stanford.edu