3

Priors on Spaces of Probability Measures

In the nonparametric setting it is natural to place a prior distribution directly on the law of the data. After presenting a general background on priors on spaces of measures, in this chapter we introduce several methods of constructing priors: stick breaking, successive partitioning, random distribution functions, etc. In particular, we discuss the rich class of tail-free processes and their properties which includes the important special case of a Pólya tree process.

3.1 Random Measures

Constructing a prior distribution on a space of probability measures comes with some technical complications. To limit these as much as possible, we assume that the sample space $(\mathfrak{X}, \mathscr{X})$ is a Polish space, and consider priors on the collection $\mathfrak{M} = \mathfrak{M}(\mathfrak{X})$ of all probability measures on $(\mathfrak{X}, \mathscr{X})$.

A prior Π on $\mathfrak{M}$ can be viewed as the law of a *random measure* P (a map from some probability space into $\mathfrak{M}$), and can be identified with the collection of "random probabilities" $P(A)$ of sets $A \in \mathscr{X}$. It is natural to choose the measurability structure on $\mathfrak{M}$ so that at least each $P(A)$ is a random variable; in other words, $(P(A): A \in \mathscr{X})$ is a *stochastic process* on the underlying probability space. In this chapter we choose the σ-field $\mathscr{M}$ on $\mathfrak{M}$, the minimal one to make this true: we set $\mathscr{M}$ equal to the smallest σ-field that makes all maps $M \mapsto M(A)$ from $\mathfrak{M}$ to $\mathbb{R}$ measurable, for $A \in \mathscr{X}$, and consider priors Π that are measures on $(\mathfrak{M}, \mathscr{M})$. Although other measurability structures are possible, the σ-field $\mathscr{M}$ is attractive for two reasons.

First, it is identical to the Borel σ-field for the weak topology on $\mathfrak{M}$ (the topology of convergence in distribution in this space, see Proposition A.5). As $\mathfrak{M}$ is Polish under the weak topology (see Theorem A.3), this means that $(\mathfrak{M}, \mathscr{M})$ is a standard Borel space. As is noted in Section 1.3, this is desirable for the definition of posterior distributions and also permits us to speak of the support of a prior, called the *weak support* in this situation. Furthermore, the parameter θ from Section 1.3 that indexes the statistical model $(P_\theta: \theta \in \Theta)$, can be taken equal to the distribution P itself, with $\mathfrak{M}$ (or a subset) as the parameter set, giving a model of the form $(P: P \in \mathfrak{M})$. With respect to the σ-field $\mathscr{M}$ on the parameter set $\mathfrak{M}$, the data distributions are trivially "regular conditional probabilities":

(i) $P \mapsto P(A)$ is $\mathscr{M}$-measurable for every $A \in \mathscr{X}$,
(ii) $A \mapsto P(A)$ is a probability measure for every realization of P.

This mathematically justifies speaking of "drawing a measure P from the prior Π and next sampling observations X from P."

Second, the fact that $\mathscr{M}$ is generated by all maps $M \mapsto M(A)$, for $A \in \mathscr{X}$, implies that a map $P\colon (\Omega, \mathscr{U}, \mathrm{P}) \to (\mathfrak{M}, \mathscr{M})$ defined on some probability space is measurable precisely if the induced measure $P(A)$ of every set $A \in \mathscr{X}$ is a random variable. Thus, as far as measurability goes, a random probability measure can be identified with a random element $(P(A)\colon A \in \mathscr{X})$ in the product space $\mathbb{R}^{\mathscr{X}}$ (or $[0, 1]^{\mathscr{X}}$).

3.1.1 Other Topologies

In the preceding we work with a measurability structure that is linked to the weak topology (or topology of convergence in distribution). Other natural topologies lead to different measurability structures. Two topologies are common.

A stronger topology is induced by the *total variation distance*

$$d_{TV}(P, Q) = \sup_{A \in \mathscr{X}} |P(A) - Q(A)|.$$

In general, the σ-field $\mathscr{M}$ is smaller than the Borel σ-field arising from the total variation distance on $\mathfrak{M}$. This difference disappears if the set of measures of interest is dominated: the traces of the two σ-fields on a dominated subset $\mathfrak{M}_0 \subset \mathfrak{M}$ are equal (see Proposition A.10; $\mathscr{M}$ is always equal to the ball σ-field). However, this equivalence does not extend to the topologies and the supports of priors: in general the weak support is bigger than the support relative to the total variation distance, even on dominated sets of measures.

The topology of uniform convergence of cumulative distribution functions, in the case that $\mathfrak{X} = \mathbb{R}^k$, generated by the *Kolmogorov-Smirnov distance* defined in (A.5), is intermediate between the weak and total variation topologies. Thus the Kolmogorov-Smirnov support is also smaller than the weak support. However distributions with a *continuous* distribution function are in the Kolmogorov-Smirnov support as soon as they are in the weak support. This is a consequence of Pólya's theorem (see Proposition A.11), which asserts that weak convergence to a continuous distribution function implies uniform convergence.

Appendices A and B contain extended discussions of these issues.

3.2 Construction through a Stochastic Process

One general method of constructing a random measure is to start with the stochastic process $(P(A)\colon A \in \mathscr{X})$, constructed using Kolmogorov's consistency theorem, and next show that this process can be realized within $\mathfrak{M}$, viewed as a subset of $\mathbb{R}^{\mathscr{X}}$. As measures have much richer properties than can be described by the finite-dimensional distributions involved in Kolmogorov's theorem, this approach is nontrivial, but it can be pushed through by standard arguments. The details are as follows.

For every finite collection $A_1, \ldots, A_k$ of Borel sets in $\mathfrak{X}$, the vector $(P(A_1), \ldots, P(A_k))$ of probabilities obtained from a random measure P is an ordinary random vector in $\mathbb{R}^k$. The construction of P may start with the specification of the distributions of all vectors of this type. A simple, important example would be to specify these as Dirichlet distributions

with parameter vector $(\alpha(A_1), \ldots, \alpha(A_k))$, for a given Borel measure α. For any *consistent* specification of the distributions, Kolmogorov's theorem allows us to construct on a suitable probability space $(\Omega, \mathscr{U}, \mathrm{P})$ a stochastic process $(P(A): A \in \mathscr{X})$ with the given finite-dimensional distributions. If the marginal distributions correspond to those of a random measure, then it will be true that

(i) $P(\varnothing) = 0$, $P(\mathfrak{X}) = 1$, a.s.
(ii) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, a.s., for any disjoint A_1, A_2.

Assertion (i) follows, because the distributions of $P(\varnothing)$ and $P(\mathfrak{X})$ will be specified to be degenerate at 0 and 1, respectively, while (ii) can be read off from the degeneracy of the joint distribution of the three variables $P(A_1)$, $P(A_2)$ and $P(A_1 \cup A_2)$. Thus the process $(P(A): A \in \mathscr{X})$ will automatically define a *finitely additive* measure on $(\mathfrak{X}, \mathscr{X})$.

A problem is that the exceptional null sets in (ii) might depend on the pair (A_1, A_2). If restricted to a countable subcollection $\mathscr{X}_0 \subset \mathscr{X}$, there would only be countably many pairs and the null sets could be gathered in a single null set. Then still when extending (ii) to σ-additivity, which is typically possible by similar distributional arguments, there would be uncountably many sequences of sets. This problem can be overcome through existence of a *mean measure*

$$\mu(A) = \mathrm{E}[P(A)].$$

For a valid random measure P, this necessarily defines a Borel measure on $\mathfrak{X}$. Existence of a mean measure is also sufficient for existence of a version of $(P(A): A \in \mathscr{X})$ that is a measure on $(\mathfrak{X}, \mathscr{X})$.

Theorem 3.1 (Random measure) *Suppose that $(P(A): A \in \mathscr{X})$ is a stochastic process that satisfies (i) and (ii) and whose mean $A \mapsto \mathrm{E}[P(A)]$ is a Borel measure on $\mathfrak{X}$. Then there exists a version of P that is a random measure on $(\mathfrak{X}, \mathscr{X})$. More precisely, if P is defined on the complete probability space $(\Omega, \mathscr{U}, \mathrm{P})$, then there exists a measurable map $\tilde{P}: (\Omega, \mathscr{U}, \mathrm{P}) \to (\mathfrak{M}, \mathscr{M})$ such that $P(A) = \tilde{P}(A)$ almost surely, for every $A \in \mathscr{X}$.*

Proof Let $\mathscr{X}_0$ be a countable field that generates the Borel σ-field $\mathscr{X}$, enumerated arbitrarily as $A_1, A_2, \ldots$. Because the mean measure $\mu(A) := \mathrm{E}[P(A)]$ is regular, there exists for every $i, m \in \mathbb{N}$ a compact set $K_{i,m} \subset A_i$ with $\mu(A_i \setminus K_{i,m}) < 2^{-2i-2m}$. By Markov's inequality

$$\mathrm{P}(P(A_i \setminus K_{i,m}) > 2^{-i-m}) \leq 2^{i+m}\mathrm{E}P(A_i \setminus K_{i,m}) \leq 2^{-i-m}.$$

Consequently, the event $\Omega_m = \cap_i \{P(A_i \setminus K_{i,m}) \leq 2^{-i-m}\}$ possesses probability at least $1 - 2^{-m}$, and $\liminf \Omega_m$ possesses probability 1, by the Borel-Cantelli lemma.

Because $\mathscr{X}_0$ is countable, the null sets involved in (i) and (ii) with $A_1, A_2 \in \mathscr{X}_0$ can be aggregated into a single null set N. For every $\omega \notin N$, the process P is a finitely additive measure on $\mathscr{X}_0$, with the resulting usual properties of monotonicity and subadditivity. By increasing N, if necessary, we can also ensure that it is subadditive on all finite unions of sets $A_i \setminus K_{i,m}$.

Let $A_{i_1} \supset A_{i_2} \supset \cdots$ be an arbitrary decreasing sequence of sets in $\mathscr{X}_0$ with empty intersection. Then, for every fixed m, the corresponding compacts $K_{i_j,m}$ also possess empty intersection, whence there exists a finite J_m such that $\cap_{j \le J_m} K_{i_j,m} = \varnothing$. This implies that

$$A_{i_{J_m}} = \cap_{j=1}^{J_m} A_{i_j} \setminus \cap_{j=1}^{J_m} K_{i_j,m} \subset \cup_{j=1}^{J_m} (A_{i_j} \setminus K_{i_j,m}).$$

Consequently, on the event $\Omega_m \setminus N$,

$$\limsup_j P(A_{i_j}) \le P(A_{i_{J_m}}) \le \sum_{j=1}^{J_m} P(A_{i_j} \setminus K_{i_j,m}) \le 2^{-m}.$$

Thus on the event $\Omega_0 := \liminf \Omega_m \setminus N$, the limit is zero. We conclude that for every $\omega \in \Omega_0$, the restriction of $A \mapsto P(A)$ to $\mathscr{X}_0$ is countably additive. By Carathéodory's theorem, it extends to a measure $\tilde{P}$ on $\mathscr{X}$.

By construction $\tilde{P}(A) = P(A)$, almost surely, for every $A \in \mathscr{X}_0$. In particular, $\mathrm{E}[\tilde{P}(A)] = \mathrm{E}[P(A)] = \mu(A)$, for every A in the field $\mathscr{X}_0$, whence by uniqueness of extension the mean measure of $\tilde{P}$ coincides with the original mean measure μ on $\mathscr{X}$. For every $A \in \mathscr{X}$, there exists a sequence $\{A_m\} \subset \mathscr{X}_0$ such that $\mu(A \,\Delta\, A_m) \to 0$. Then both $P(A_m \,\Delta\, A)$ and $\tilde{P}(A_m \,\Delta\, A)$ tend to zero in mean. Finite-additivity of P gives that $|P(A_m) - P(A)| \le P(A_m \,\Delta\, A)$, almost surely, and by σ-additivity the same is true for $\tilde{P}$. This shows that $\tilde{P}(A) = P(A)$, almost surely, for every $A \in \mathscr{X}$.

This also proves that $\tilde{P}(A)$ is a random variable for every $A \in \mathscr{X}$, whence $\tilde{P}$ is a measurable map in $(\mathfrak{M}, \mathscr{M})$. □

Rather than starting from the process $(P(A): A \in \mathscr{X})$ indexed by all Borel sets, we may wish to start from a smaller set $(P(A): A \in \mathscr{X}_0)$ of variables, for some $\mathscr{X}_0 \subset \mathscr{X}$. As shown in the proof of the preceding theorem, a countable collection $\mathscr{X}_0$ suffices, but compact sets play a special role.

Theorem 3.2 (Random measure) *Suppose that $(P(A): A \in \mathscr{X}_0)$ is a stochastic process that satisfies (i) and (ii) for a countable field $\mathscr{X}_0$ that generates $\mathscr{X}$ and is such that for every $A \in \mathscr{X}_0$ and $\epsilon > 0$ there exists a compact $K_\epsilon \subset \mathfrak{X}$ and $A_\epsilon \in \mathscr{X}_0$ such that $A_\epsilon \subset K_\epsilon \subset A$ and $\mu(A \setminus A_\epsilon) < \epsilon$, where μ is the mean $\mu(A) = \mathrm{E}[P(A)]$. Then there exists a random measure that extends P to $\mathscr{X}$.*

The proof of the theorem follows the same lines, except that, if the compacts K_ϵ are not elements of $\mathscr{X}_0$, the bigger sets $A \setminus A_\epsilon$ must be substituted for $A \setminus K_\epsilon$ when bounding the P-measure of this set.

For instance, for $\mathfrak{X} = \mathbb{R}^k$ we can choose $\mathscr{X}_0$ equal to the finite unions of cells $(a, b]$, with the compacts and A_ϵ equal to the corresponding finite unions of the intervals $[a_\epsilon, b]$ and $(a_\epsilon, b]$ for a_ϵ descending to a. By restricting to rational endpoints we obtain a countable collection.

3.3 Countable Sample Spaces

A probability distribution on a countable sample space (equipped with the σ-field of all its subsets) can be represented as an infinite-length probability vector $s = (s_1, s_2, \ldots)$. A prior on the set $\mathfrak{M}$ of all probability measures on a countable sample space can therefore be identified with the distribution of a random element with values in the countable-dimensional unit simplex

$$\mathbb{S}_\infty = \Big\{ s = (s_1, s_2, \ldots) \colon s_j \geq 0,\ j \in \mathbb{N},\ \sum_{j=1}^{\infty} s_j = 1 \Big\}.$$

The usual σ-field $\mathscr{M}$ on $\mathbb{S}_\infty \equiv \mathfrak{M}$ can be characterized in various ways, the simplest being that it is generated by the coordinate maps $s \mapsto s_i$, for $i \in \mathbb{N}$. This shows that a map p from some probability space into $\mathbb{S}_\infty$ is a random element if and only if every coordinate variable p_i is a random variable. Hence a prior simply corresponds to an infinite sequence of nonnegative random variables $p_1, p_2, \ldots$ that add up to 1.

We can also embed $\mathbb{S}_\infty$ in $\mathbb{R}^\infty$, from which it then inherits its measurable structure (the projection σ-field) and a topology: the topology of coordinatewise convergence. By Scheffé's theorem the restriction of this topology to $\mathbb{S}_\infty$ is equivalent to the topology derived from the norm $\|s\|_1 = \sum_{j=1}^{\infty} |s_j|$ of the space $\ell_1 := \{s = (s_1, s_2, \ldots) \in \mathbb{R}^\infty \colon \sum_{j=1}^{\infty} |s_j| < \infty\} \supset \mathbb{S}_\infty$. The weak topology on $\mathbb{S}_\infty$, viewed as probability measures, coincides with these topologies as well. Thus the projection σ-field $\mathscr{M}$ is also the Borel σ-field for these topologies and $(\mathfrak{M}, \mathscr{M})$ is a Polish space.

The general approach of Section 3.2 of constructing priors using Kolmogorov's theorem applies, but can be simplified by ordering the coordinates: it suffices to construct consistent marginal distributions for $(p_1, \ldots, p_k)$, for every $k = 1, 2, \ldots$ In the next sections, we also present two structural methods.

The support of a prior Π relative to the ℓ_1-norm is the set of all p_0, such that $\Pi(p \colon \|p - p_0\|_1 < \epsilon) > 0$, for every $\epsilon > 0$. Since the norm topology coincides with the topology of coordinatewise convergence (and hence finite intersections of sets of the form $\{p \colon |p_j - p_{0j}| < \epsilon\}$ form a base of the topology), it can also be described as the set of all p_0 such that $\Pi(p \colon |p_j - p_{0j}| < \epsilon, 1 \leq j \leq k) > 0$, for every $\epsilon > 0$ and $k \in \mathbb{N}$. This reduction to the finite-dimensional marginal distributions makes it easy to construct priors with large support.

3.3.1 Construction through Normalization

Given nonnegative random variables $Y_1, Y_2, \ldots$ such that $\sum_{j=1}^{\infty} Y_j$ is positive and converges a.s., we can define a prior on $\mathbb{S}_\infty$ by putting

$$p_k = \frac{Y_k}{\sum_{j=1}^{\infty} Y_j}, \qquad k \in \mathbb{N}. \tag{3.1}$$

A simple, sufficient condition for the convergence of the random series is that $\sum_{j=1}^{\infty} \mathrm{E}(Y_j) < \infty$. It is convenient to use independent random variables.

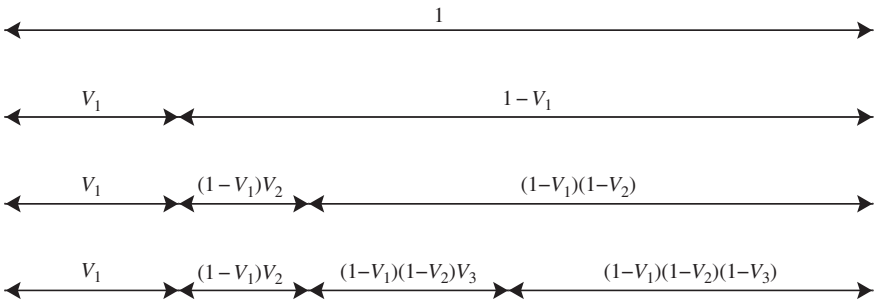

Figure 3.1 Stick breaking. A stick of length 1 is sequentially broken in smaller sticks in proportions given by a random sequence $V_1, V_2, \ldots$

Lemma 3.3 *If $Y_1, Y_2, \ldots$ are independent, positive random variables with $\sum_{j=1}^{\infty} Y_j < \infty$ a.s. and marginal densities that are positive everywhere in $(0, \infty)$, then the support of the prior defined by* (3.1) *is the full space $\mathbb{S}_\infty$.*

Proof Fix an arbitrary point $p_0 \in \mathbb{S}$, $\epsilon > 0$ and $k \in \mathbb{N}$. Since $\sum_{j=1}^{\infty} Y_j$ converges a.s., there exists $K \geq k$ such that the event $\{\sum_{j=K+1}^{\infty} Y_j < \epsilon/2\}$ has positive probability. By choosing K larger, if necessary, we can also ensure that $\sum_{j=K+1}^{\infty} p_{0j} < \epsilon/2$. By continuity of the maps $(y_1, \ldots, y_K) \mapsto y_j/(\sum_{j=1}^{K} y_j + c)$ from $\mathbb{R}^K \to \mathbb{R}$, the set

$$\mathcal{N}_K = \left\{(y_1, \ldots, y_K): \max_{1 \leq j \leq k} \left| \frac{y_j}{\sum_{l=1}^{K} y_l} - p_{0j} \right| \vee \left| \frac{y_j}{\sum_{l=1}^{K} y_l + \epsilon/2} - p_{0j} \right| < \epsilon \right\}$$

is open in $(0, \infty)^K$. Also $\mathcal{N}_K$ is nonempty, since $(p_{01}, \ldots, p_{0K}) \in \mathcal{N}_K$, as is easy to verify. Thus the event $\{(Y_1, \ldots, Y_K) \in \mathcal{N}_K\}$ has positive probability by the assumption of positive densities.

On the intersection of the events $\{\sum_{j=K+1}^{\infty} Y_j < \epsilon/2\}$ and $\{(Y_1, \ldots, Y_K) \in \mathcal{N}_K\}$, we have that $|Y_j/(\sum_{l=1}^{\infty} Y_l) - p_{0j}| < \epsilon$ for all $1 \leq j \leq k$. Because the events are independent and both have positive probability, the intersection has positive probability. □

3.3.2 Construction through Stick Breaking

Stick-breaking is a technique to construct a prior directly on $\mathbb{S}_\infty$. The problem at hand is to distribute the total mass 1, which we identify with a stick of length 1, randomly to each element of $\mathbb{N}$. We first break the stick at a point given by the realization of a random variable $0 \leq V_1 \leq 1$ and assign mass V_1 to $1 \in \mathbb{N}$. We think of the remaining mass $1 - V_1$ as a new stick, and break it into two pieces of relative lengths V_2 and $1 - V_2$ according to the realized value of another random variable $0 \leq V_2 \leq 1$. We assign mass $(1 - V_1)V_2$ to the point 2, and are left with a new stick of length $(1 - V_1)(1 - V_2)$. Continuing in this way, we assign mass to the point j equal to

$$p_j = \Big(\prod_{l=1}^{j-1}(1 - V_l)\Big)V_j. \tag{3.2}$$

Clearly, by continuing to infinity, this scheme will attach a random subprobability distribution to $\mathbb{N}$ for any sequence of random variables $V_1, V_2, \ldots$ with values in $[0, 1]$. Under mild conditions, the probabilities p_j will sum to one.

Lemma 3.4 (Stick-breaking) *The random subprobability distribution* $(p_1, p_2, \ldots)$ *lies in* $\mathbb{S}_\infty$ *almost surely if and only if* $\mathrm{E}[\prod_{l=1}^{j}(1-V_l)] \to 0$ *as* $j \to \infty$. *For independent variables* $V_1, V_2, \ldots$, *this condition is equivalent to* $\sum_{l=1}^{\infty} \log \mathrm{E}(1 - V_l) = -\infty$. *In particular, for i.i.d. variables* $V_1, V_2, \ldots$, *it suffices that* $\mathrm{P}(V_1 > 0) > 0$. *If for every* $k \in \mathbb{N}$ *the support of* $(V_1, \ldots, V_k)$ *is* $[0, 1]^k$, *then the support of* $(p_1, p_2, \ldots)$ *is the whole space* $\mathbb{S}_\infty$.

Proof By induction, it easily follows that the leftover mass at stage j is equal to $1 - \sum_{l=1}^{j} p_l = \prod_{l=1}^{j}(1 - V_l)$. Hence the random subprobability distribution will lie in $\mathbb{S}_\infty$ a.s. if and only if $\prod_{l=1}^{j}(1 - V_l) \to 0$ a.s. Since the leftover sequence is decreasing, nonnegative and bounded by 1, the almost sure convergence is equivalent to convergence in mean. If the V_js are independent, then this condition becomes $\prod_{l=1}^{j}(1 - \mathrm{E}(V_l)) \to 0$ as $j \to \infty$, which is equivalent to the condition $\sum_{l=1}^{\infty} \log \mathrm{E}(1 - V_l) = -\infty$.

The last assertion follows, because the probability vector $(p_1, \ldots, p_k)$ is a continuous function of $(V_1, \ldots, V_k)$, for every k. □

In the stick-breaking construction, if X is a random variable distributed according to probability mass function $p = (p_1, p_2, \ldots)$,

$$V_j = \frac{p_j}{1 - \sum_{l=1}^{j-1} p_l} = \mathrm{P}(X = j \,|\, X \geq j). \tag{3.3}$$

This is known as the *discrete hazard rate* for X. Thus the construction may be interpreted in terms of hazard rates, or equivalently in terms of the negative log-hazard $h_j = -\log V_j$. Any consistent system of joint distributions of h_js lying in $[0, \infty)$ (in particular, independent) gives rise to a possibly defective prior on p. For independent h_js, the necessary and sufficient condition for this to be proper prior is given by $\sum_{j=1}^{\infty} \log(1 - \mathrm{E}(e^{-h_j})) = -\infty$.

3.3.3 Countable Dirichlet Process

The *countable Dirichlet distribution* can be obtained both through normalization and stick breaking.

In the construction by normalization, we choose the variables $Y_1, Y_2, \ldots$ to be independent with $Y_j \sim \mathrm{Ga}(\alpha_j, 1)$, for $j \in \mathbb{N}$. Then, by Proposition G.2, the vector $(p_1, p_2, \ldots)$ defined in (3.1) satisfies, for any $k \in \mathbb{N}$,

$$\Big(p_1, \ldots, p_k, 1 - \sum_{j=1}^{k} p_j\Big) \sim \mathrm{Dir}\Big(k+1; \alpha_1, \ldots, \alpha_k, \sum_{j=k+1}^{\infty} \alpha_j\Big). \tag{3.4}$$

In particular, every p_j is $\mathrm{Be}(\alpha_j, \sum_{l \neq j} \alpha_l)$-distributed.

In the construction by stick breaking, we choose the variables $V_1, V_2, \ldots$ independent with $V_j \sim \mathrm{Be}(\alpha_j, \sum_{l=j+1}^{\infty} \alpha_l)$ and define $(p_1, p_2, \ldots)$ by (3.2). To see that this yields the same distribution, it suffices to check that the joint distribution of $(p_1, \ldots, p_k)$

is the same in the two cases, for every $k \in \mathbb{N}$. Equivalently, it suffices to show that $V_1, V_2, \ldots$, defined by (3.3) from $p_1, p_2, \ldots$ satisfying (3.4), are independent variables with $V_j \sim \mathrm{Be}(\alpha_j, \sum_{l=j+1}^{\infty} \alpha_l)$. This is a consequence of the aggregation properties of the finite-dimensional Dirichlet distribution described in Proposition G.3, as noted in Corollary G.5.

Explicit calculation of the posterior distribution given a sample of i.i.d. observations $X_1, \ldots, X_n$ from p is possible in this case. The likelihood can be written as $p \mapsto \prod_j p_j^{N_j}$, for N_j the number of observations equal to j. The vector $N = (N_1, N_2, \ldots)$ is a sufficient statistic, and hence the posterior given N is the same as the posterior given the original observations. For any $l \in \mathbb{N}$,

$$\Big(N_1, \ldots, N_l, n - \sum_{j=1}^{l} N_j\Big) \sim \mathrm{MN}_{l+1}\Big(n; p_1, \ldots, p_l, 1 - \sum_{j=1}^{l} p_j\Big).$$

Therefore, by (3.4) with l replacing k and the conjugacy of the finite Dirichlet distribution with the multinomial likelihood, expressed in Proposition G.8, it follows that the posterior density of $(p_1, \ldots, p_l, 1 - \sum_{j=1}^{l} p_j)$ given $N_1, \ldots, N_l$ is given by

$$\mathrm{Dir}\Big(l+1; \alpha_1 + N_1, \ldots, \alpha_l + N_l, \sum_{j=l+1}^{\infty} \alpha_j + n - \sum_{j=1}^{l} N_j\Big). \tag{3.5}$$

Marginalizing to the first $k \leq l$ cells, it follows from Proposition G.3 that the posterior density of $(p_1, \ldots, p_k, 1 - \sum_{j=l}^{k} p_j)$ given $N_1, \ldots, N_l$ is given by

$$\mathrm{Dir}\Big(k+1; \alpha_1 + N_1, \ldots, \alpha_k + N_k, \sum_{j=k+1}^{\infty} \alpha_j + n - \sum_{j=1}^{k} N_j\Big). \tag{3.6}$$

Because this depends only on $(N_1, \ldots, N_k)$, the posterior density of $(p_1, \ldots, p_k, 1 - \sum_{j=l}^{k} p_j)$ given $N_1, \ldots, N_l$ is the same for every $l \geq k$. Because $\sigma\langle N_1, \ldots, N_l\rangle$ increases to $\sigma\langle N_1, N_2, \ldots\rangle$ as $l \to \infty$, it follows by the martingale convergence theorem that the preceding display also gives the posterior density of $(p_1, \ldots, p_k, 1 - \sum_{j=l}^{k} p_j)$ given N.

The distribution (3.6) has the same form as the distribution in the right side of (3.4), whence the posterior distribution is also a countable Dirichlet distribution. The posterior parameters depend on the prior parameters and the cell counts and can be found by the straightforward updating rule $\alpha \mapsto \alpha + N$.

In analogy with the finite dimensional Dirichlet distribution, it is natural to call the prior the *Dirichlet process* on $\mathbb{N}$ or the *countable Dirichlet process*. We shall write $(p_1, p_2, \ldots) \sim \mathrm{DP}((\alpha_1, \alpha_2, \ldots))$. This Dirichlet process will be generalized to arbitrary spaces in the next chapter and will be seen to admit similar explicit expressions. It is a central object in Bayesian nonparametrics.

From (3.6) and the properties of the finite-dimensional Dirichlet distribution, the posterior mean, variance and covariances are seen to be

$$\mathrm{E}(p_j \,|\, X_1, \ldots, X_n) = \frac{\alpha_j + N_j}{\sum_{l=1}^{\infty} \alpha_l + n}, \tag{3.7}$$

$$\operatorname{var}(p_j \mid X_1, \ldots, X_n) = \frac{(\alpha_j + N_j)(\sum_{l \neq j} \alpha_l + n - N_j)}{(\sum_{l=1}^{\infty} \alpha_l + n)^2 (\sum_{l=1}^{\infty} \alpha_l + n + 1)}, \tag{3.8}$$

$$\operatorname{cov}(p_j, p_{j'} \mid X_1, \ldots, X_n) = -\frac{(\alpha_j + N_j)(\alpha_{j'} + N_{j'})}{(\sum_{l=1}^{\infty} \alpha_l + n)^2 (\sum_{l=1}^{\infty} \alpha_l + n + 1)}. \tag{3.9}$$

3.4 Construction through Structural Definitions

In this section we collect priors on measures on a general Polish space that are defined explicitly or through an iterative algorithm.

3.4.1 Construction through a Distribution on a Dense Subset

An easy method to obtain a prior with full support is to assign positive prior mass to every point in a given countable dense subset. Such a prior distribution can be considered a default prior if the point masses are constructed by a default mechanism. The set $\mathfrak{M}$ of Borel measures on a Polish space $\mathfrak{X}$ is also Polish by Theorem A.3, and hence has a countable dense subset to which this construction can be applied.

Often it is meaningful to construct the prior using a sequence of finite subsets that gradually improve the approximation. For instance, at stage m we choose a finite subset $\mathfrak{M}_m$ which approximates the elements of $\mathfrak{M}$ within a distance ϵ_m, and put the discrete uniform distribution on $\mathfrak{M}_m$. A convex combination over m (or the sequence of discrete priors) may be regarded as a default prior. If $\epsilon_m \downarrow 0$, then the weak support of a prior of this type is the whole of $\mathfrak{M}$. It will be seen in Chapters 6 and 8 that such a prior, with carefully chosen support points guided by covering numbers, results in posterior distributions with good large-sample properties.

Computation of the resulting posterior distribution may be difficult to carry out.

3.4.2 Construction through a Randomly Selected Discrete Set

Given an integer $N \in \mathbb{N} \cup \{\infty\}$, nonnegative random variables $W_{1,N}, \ldots, W_{N,N}$ with $\sum_{i=1}^{N} W_{i,N} = 1$ and random variables $\theta_{1,N}, \ldots, \theta_{N,N}$ taking their values in $(\mathfrak{X}, \mathscr{X})$, we can define a random probability measure by

$$P = \sum_{i=1}^{N} W_{i,N} \delta_{\theta_{i,N}}.$$

The realizations of this prior are discrete with finitely or countably many support points, which may be different for each realization. Given the number N of support points, their "weights" $W_{1,N}, \ldots, W_{N,N}$ and "locations" $\theta_{1,N}, \ldots, \theta_{N,N}$ are often chosen independent. Assume that $\mathfrak{X}$ is a separable metric space with a metric d.

Lemma 3.5 *If the support of N is unbounded and given $N = n$, the weights and locations are independent with full supports $\mathbb{S}_n$ and $\mathfrak{X}^n$, respectively, for every n, then P has full support $\mathfrak{M}$.*

Proof Because the finitely discrete distributions are weakly dense in $\mathfrak{M}$, it suffices to show that P gives positive probability to any weak neighborhood of a distribution $P^* = \sum_{i=1}^k w_i^* \delta_{\theta_i^*}$ with finite support. All distributions $P' := \sum_{i=1}^k w_i \delta_{\theta_i}$ with $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to $(w_1^*, \ldots, w_k^*)$ and $(\theta_1^*, \ldots, \theta_k^*)$ are in such a weak neighborhood. So are the measures $P' = \sum_{i=1}^\infty w_i \delta_{\theta_i}$ with $\sum_{i>k} w_i$ sufficiently small and $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to their targets, as before.

If $\mathrm{P}(N = \infty) > 0$, then the assertion follows upon considering the events $\{\sum_{i>k} W_{i,\infty} < \epsilon, \max_{i \le k} |W_{i,k'} - w_i^*| \vee d(\theta_{i,k}, \theta_i^*) < \epsilon\}$. These events have positive probability, as they refer to an open subset of $\mathbb{S}_\infty \times \mathfrak{X}^k$.

If N is finite almost surely, then the assertion follows from the assumed positive probability of the events $\{N = k', \max_{i \le k'} |W_{i,k'} - w_i^*| \vee d(\theta_{i,k'}, \theta_i^*) < \epsilon\}$ for every $\epsilon > 0$ and some $k' > k$, where we define $w_i^* = 0$ and θ_i^* arbitrarily for $k < i \le k'$. □

The prior is computationally more tractable if N is finite and bounded, but such a prior does not have full support on an infinite sample space. To achieve reasonable large sample properties, N must either depend on the sample size n, or be given a prior with infinite support.

An important special case is obtained by choosing $N \equiv \infty$, yielding a prior of the form

$$P = \sum_{i=1}^\infty W_i \delta_{\theta_i}. \tag{3.10}$$

Further specializations are to choose $\theta_1, \theta_2, \ldots$ an i.i.d. sequence in $\mathfrak{X}$, and independently to choose the weights $W_1, W_2, \ldots$ by the stick-breaking algorithm of Section 3.3.2. If the common distribution of the θ_i has support equal to the full space $\mathfrak{X}$ and the stick-breaking weights are as in Lemma 3.4, then this prior has full support. The assumed independence of locations and weights gives that the mean measure of P is equal to the distribution of the θ_i:

$$\mathrm{E}P(A) = \mathrm{P}(\theta_i \in A), \qquad A \in \mathscr{X}.$$

Because the realizations $\{\theta_1, \theta_2, \ldots\}$ will be dense in $\mathfrak{X}$ a.s., it is then also true that $P(U) = \sum_{i:\theta_i \in U} W_i > 0$ a.s. for every open U, as soon as the stick-breaking weights are strictly positive. In Section 4.2.5 the Dirichlet process prior will be seen to take this form.

Lemma 3.6 *If $(W_1, W_2, \ldots)$ are stick-breaking weights based on stick lengths $V_i \overset{iid}{\sim} H$ for a fully supported measure H on $[0, 1]$, independent of $\theta_i \overset{iid}{\sim} G$ with full support $\mathfrak{X}$, then the random measure* (3.10) *has full support $\mathfrak{X}$. Furthermore, the mean measure of P is G.*

3.4.3 Construction through Random Rectangular Partitions

Consider the sample space $\mathfrak{X} = [0, 1]$ (other domains like $\mathbb{R}$ may be handled through appropriate transformations), and fix a probability measure μ on the unit square $[0, 1]^2$, not entirely concentrated on the boundary. Draw a point randomly from the unit square according to μ. This divides the unit square into four rectangles: lower left (LL), upper left (UL), lower right (LR) and upper right (UR). Keep LL and UR, including their boundaries, and

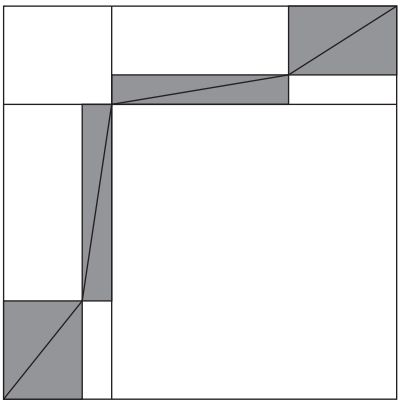

Figure 3.2 Random rectangular partitions. The shaded rectangles are the ones kept after two rounds of choosing splitting points. The curve following their diagonals is the stage 2 approximation of the random distribution function.

discard the other two. Construct the probability measures induced by μ on LL and UR by affinely mapping the unit square onto these rectangles. Draw a point randomly from LL and discard its UL and LR. Similarly, draw a point randomly from UR and discard its UL and LR. Continue this process indefinitely (see Figure 3.2).

The union of the rectangles kept by this algorithm forms an infinite sequence of decreasing compact sets. Their intersection is the graph of a continuous increasing function from $[0, 1]$ onto $[0, 1]$ and hence defines a random cumulative distribution function on $[0, 1]$. To verify this claim, observe that each vertical cross-section of the compact sets forms a sequence of nested closed *intervals* whose lengths converge to zero a.s., and hence the intersection is a singleton a.s. Alternatively, it can be seen that the function is given by the limit of the sequence of functions obtained by joining the split points at a given stage (see Figure 3.2). This is a sequence of increasing continuous functions converging uniformly a.s., and hence the limit is also increasing and continuous a.s.

This construction is similar to the one of Section 3.5 with the difference that presently both probabilities and partitions are generated randomly. Some examples of measures μ of particular interest are the uniform distributions on the sets:

(a) the vertical line segment $x = 1/2, 0 < y < 1$;
(b) the horizontal line segment $0 < x < 1, y = 1/2$;
(c) the unit square.

Case (a) is equivalent to each time breaking the horizontal line segment evenly into two subintervals and assigning probabilities according to the uniform distribution; it is the special case of the tree construction of Section 3.5, where the $V_{\varepsilon 0}$s are i.i.d. uniformly distributed. Case (b) is equivalent to breaking the horizontal axis randomly into two and assigning probability half to each segment; this generates binary quantiles of a distribution. Case (c) seems to be a natural default choice.

If μ is concentrated on the diagonal, then the procedure always leads to the uniform distribution, and hence the prior is degenerate.

3.4.4 Construction through Moments

If the domain is a bounded interval in $\mathbb{R}$, then the sequence of moments uniquely determines the probability measure. Hence a prior on the space of probability measures can be induced from one on the sequence of moments. One may control the location, scale, skewness and kurtosis of the random probability by using subjective priors on the first four moments. Priors for the higher-order moments are difficult to elicit, and some default method should be used. Maintaining the necessary constraints in the prior specification linking various moments is difficult; hence, the approach may be hard to implement.

3.4.5 Construction through Quantiles

A prior for quantiles is much easier to elicit than for moments. One may put priors on all dyadic quantiles honoring the order restrictions. Conceptually, this operation is the opposite of specifying a tree-based prior as considered in Section 3.5. For quantile priors the masses are predetermined and the partitions are random. In practice, one may put priors only for a finite number of quantiles, and then distribute the remaining masses uniformly over the corresponding interval.

3.4.6 Construction by Normalization

A prior distribution on a probability measure P needs to honor not only the countable additivity of P, but also the normalization condition $P(\mathfrak{X}) = 1$. The additional restriction typically rules out assignments such as independence of $P(A)$ and $P(B)$, when A and B are disjoint. One possible approach is to disregard this restriction and construct a prior distribution Π_∞ on $\mathfrak{M}_\infty(\mathfrak{X})$, the space of positive finite measures, and apply a renormalization step $\mu \mapsto \mu/\mu(\mathfrak{X})$ in the end. If the prior Π_∞ has full support $\mathfrak{M}_\infty(\mathfrak{X})$ with respect to the weak topology, then the prior Π on $\mathfrak{X}$ obtained by renormalization will also have full support $\mathfrak{M}(\mathfrak{X})$ with respect to the weak topology, as the following result shows.

Proposition 3.7 *Let Π_∞ be a probability distribution on $\mathfrak{M}_\infty(\mathfrak{X})$ and $\mu_0 \in \mathfrak{M}_\infty(\mathfrak{X})$ be such that, for any collection of bounded continuous functions $g_1, \ldots, g_k$ and every $\epsilon > 0$,*

$$\Pi_\infty\Big(\mu : \Big|\int g_j\, d\mu - \int g_j\, d\mu_0\Big| < \epsilon\Big) > 0.$$

Then the probability measure $\mu_0/\mu_0(\mathfrak{X})$ belongs to the weak support of the probability measure Π on $\mathfrak{M}(\mathfrak{X})$ induced by the map $\mu \mapsto \mu/\mu(\mathfrak{X})$ on $\mathfrak{M}_\infty(\mathfrak{X})$ equipped with Π_∞.

Proof A typical neighborhood for the weak topology takes the form $\{P : |\int g_j dP - \int g_j dP_0| < \epsilon\}$, for given bounded continuous functions $g_1, \ldots, g_k$ and $\epsilon > 0$. If $|\int g_j\, d\mu - \int g_j\, d\mu_0| < \delta$ and $|\mu(\mathfrak{X}) - \mu_0(\mathfrak{X})| = |\int 1\, d\mu - \int 1\, d\mu_0| < \delta$, then

$$\begin{aligned}\left|\frac{\int g_j\, d\mu}{\int g_j\, d\mu_0} - \frac{\int g_j\, d\mu}{\mu(\mathfrak{X})}\right| &\le \frac{1}{\mu(\mathfrak{X})}\left|\int g_j\, d\mu - \int g_j\, d\mu_0\right| + \int |g_j|\, d\mu \frac{|\mu(\mathfrak{X}) - \mu_0(\mathfrak{X})|}{\mu(\mathfrak{X})\mu_0(\mathfrak{X})} \\ &\le \frac{1}{\mu_0(\mathfrak{X}) - \delta}\Big(\delta + \frac{\max_j \|g_j\|_\infty \delta}{\mu_0(\mathfrak{X})}\Big) < \epsilon.\end{aligned}$$

Thus, for sufficiently small δ, the measure $\mu/\mu(\mathfrak{X})$ belongs to the given neighborhood. □

A random finite measure μ on $\mathfrak{M}_\infty(\mathfrak{X})$ such that the variables $\mu(A)$ and $\mu(B)$ are independent, for every disjoint sets A and B, is called a *completely random measure*, as discussed in Appendix J. The probability measure obtained by normalization of such a measure is a *normalized completely random measure*, and is studied in Section 14.7.

3.5 Construction through a Tree

Consider a sequence $\mathcal{T}_0 = \{\mathfrak{X}\}$, $\mathcal{T}_1 = \{A_0, A_1\}$, $\mathcal{T}_2 = \{A_{00}, A_{01}, A_{10}, A_{11}\}$, and so on, of measurable partitions of the sample space $\mathfrak{X}$, obtained by splitting every set in the preceding partition into two new sets. With $\mathcal{E} = \{0, 1\}$ and $\mathcal{E}^* = \cup_{m=0}^{\infty} \mathcal{E}^m$, the set of all finite strings $\varepsilon_1 \cdots \varepsilon_m$ of 0s and 1s, we can index the 2^m sets in the mth partition $\mathcal{T}_m$ by $\varepsilon \in \mathcal{E}^m$, in such a way that $A_\varepsilon = A_{\varepsilon 0} \cup A_{\varepsilon 1}$ for every $\varepsilon \in \mathcal{E}^*$. Here $\varepsilon 0$ and $\varepsilon 1$ are the extensions of the string ε with a single symbol 0 or 1; the empty string indexes $\mathcal{T}_0$ (see Figure 3.3). Let $|\varepsilon|$ stand for the length of a string ε and let $\varepsilon\delta$ be the concatenation of two strings $\varepsilon, \delta \in \mathcal{E}^*$. The set of all finite unions of sets A_ε, for $\varepsilon \in \mathcal{E}^*$, forms a subfield of the Borel sets. We assume throughout that the splits are chosen rich enough that this generates the Borel σ-field.

Because the probability of any A_ε must be distributed to its "offspring" $A_{\varepsilon 0}$ and $A_{\varepsilon 1}$, a probability measure P must satisfy the *tree additivity* requirement $P(A_\varepsilon) = P(A_{\varepsilon 0}) + P(A_{\varepsilon 1})$. The relative weights of the offspring sets are the conditional probabilities

$$V_{\varepsilon 0} = \mathrm{P}(A_{\varepsilon 0} | A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = \mathrm{P}(A_{\varepsilon 1} | A_\varepsilon). \tag{3.11}$$

This motivates to define, for a given specification of a set $(V_\varepsilon : \varepsilon \in \mathcal{E}^*)$ of $[0, 1]$-valued random variables,

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^m. \tag{3.12}$$

If $V_{\varepsilon 0} + V_{\varepsilon 1} = 1$ for every ε, then the stochastic process $(P(A_\varepsilon) : \varepsilon \in \mathcal{E}^*)$ will satisfy the tree-additivity condition and define a finitely additive measure on the field of all finite unions of sets A_ε, for $\varepsilon \in \mathcal{E}^*$.

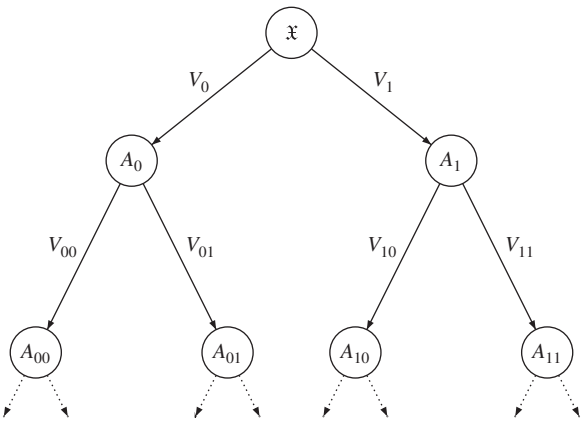

Figure 3.3 Tree diagram showing the distribution of mass over the first two partitions $\mathfrak{X} = A_0 \cup A_1 = (A_{00} \cup A_{01}) \cup (A_{10} \cup A_{11})$ of the sample space. Mass at a given node is distributed to its two children proportionally to the weights on the arrows. Every pair of Vs on arrows originating from the same node add to 1.

Countable additivity is not immediate, but may be established using a mean measure by the approach of Theorem 3.2.

Theorem 3.8 *Consider a sequence of partitions $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets in $(\mathfrak{X}, \mathscr{X})$ and is such that every A_ε is the union of all $A_{\varepsilon\delta}$ whose closure is compact and satisfies $\overline{A}_{\varepsilon\delta} \subset A_\varepsilon$, where $\delta \in \mathcal{E}^*$. If $(V_\varepsilon : \varepsilon \in \mathcal{E}^*)$ is a stochastic process with $0 \le V_\varepsilon \le 1$ and $V_{\varepsilon 0} + V_{\varepsilon 1} = 1$ for all $\varepsilon \in \mathcal{E}^*$, and there exists a Borel measure μ such that $\mu(A_\varepsilon) = \mathrm{E}[V_{\varepsilon_1} V_{\varepsilon_1\varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}]$, for every $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^*$, then there exists a random Borel measure P satisfying* (3.12).

Proof For fixed $\varepsilon \in \mathcal{E}^*$ there are at most countably many $A_{\varepsilon\delta}$ as stated, and their union is A_ε. Thus for any given $\eta > 0$ there exists a finite subcollection whose union $B_{\varepsilon,\eta}$ satisfies $\mu(A_\varepsilon \setminus B_{\varepsilon,\eta}) < \eta$. The corresponding union $K_{\varepsilon,\eta}$ of the closures $\bar{A}_{\varepsilon\delta}$ is compact and satisfies $B_{\varepsilon,\eta} \subset K_{\varepsilon,\eta} \subset A_\varepsilon$. Thus we are in the situation of Theorem 3.2, with P defined by (3.12) as a finitely additive measure on the field consisting of all finite unions of A_ε. □

In the case of $\mathfrak{X} = \mathbb{R}$ splits in intervals are natural. The condition of the preceding theorem is then met as soon as a mean measure exists. Alternatively, an explicit condition for countable additivity can be given as follows.

Suppose that we use left-open and right-closed cells $(a, b]$, except for a single cell of the form (a, ∞) that is unbounded to the right at every level, and choose the index such that $A_{\varepsilon 0}$ lies below $A_{\varepsilon 1}$. (Thus the sets A_ε with ε a string of only 0s or 1s are unbounded to the left and right, respectively; the other sets are bounded.) Furthermore, suppose that the split points are dense in $\mathbb{R}$, so that the partitions generate the Borel σ-field.

To check in this special case that P defined by (3.12) extends to a countably additive random measure on the Borel sets, it suffices to check that its induced distribution function is right continuous at every of the boundary points of the partitions and tends to 0 and 1 as $t \to -\infty$ or $+\infty$. This can be easily translated in the conditional probabilities V_ε.

Theorem 3.9 *Consider a sequence of partitions $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m\}$ of $\mathbb{R}$ such that $A_{\varepsilon 0} = \{x \in A_\varepsilon : x \le a_\varepsilon\}$ and $A_{\varepsilon 1} = \{x \in A_\varepsilon : x > a_\varepsilon\}$, where $a_\varepsilon \in \operatorname{int} A_\varepsilon$ and $\{a_\varepsilon : \varepsilon \in \mathcal{E}^*\}$ is dense in $\mathbb{R}$. If $(V_\varepsilon : \varepsilon \in \mathcal{E}^*)$ is a stochastic process with $0 \le V_\varepsilon \le 1$ and $V_{\varepsilon 0} + V_{\varepsilon 1} = 1$ for all $\varepsilon \in \mathcal{E}^*$, then P defined by* (3.12) *extends to a random measure on the Borel sets in $\mathbb{R}$ a.s. if and only if*

$$\mathrm{E}[V_\varepsilon V_{\varepsilon 0} V_{\varepsilon 00} \cdots] = 0 \text{ for all } \varepsilon \in \mathcal{E}^*, \text{ and } \mathrm{E}[V_1 V_{11} V_{111} \cdots] = 0. \tag{3.13}$$

Proof The probabilities (3.12) define a "distribution function" on the collection $A = \{a_\varepsilon : \varepsilon \in \mathcal{E}^*\}$ of endpoints of the partitions. If this function is right continuous and has the appropriate limits at $\pm\infty$, then it is a valid distribution function on A, which extends (uniquely) to a distribution function on $\mathbb{R}$.

The left endpoint of A_ε (for $\varepsilon \in \mathcal{E}^*$, $\varepsilon \ne 000\cdots$) is also the left endpoint of the cells $A_{\varepsilon 0}, A_{\varepsilon 00}, \ldots$ Therefore right continuity of the distribution function of P at this point holds if and only if $V_{\varepsilon 0} V_{\varepsilon 00 \ldots} = 0$. Similarly convergence to 0 and 1 in the tails correspond to $V_0 V_{00} \cdots = 0$ and $V_1 V_{11} \cdots = 0$. By assumption (3.13) these sequences tend to zero in

mean. Because they are monotone and bounded, the convergence is also a.s. The exceptional null sets add to a single null set, as $\mathcal{E}^*$ has only countably many elements. □

Rather than in two subsets, the sets may be split in different (even varying, including infinite) numbers of subsets, as long as the V-variables that regulate the mass distribution satisfy the appropriate restriction. In particular, in the case that $\mathfrak{X} = \mathbb{R}^k$, splitting in 2^k sets would be natural. (These splits could also be incorporated in a longer, binary tree with sequential splits along the coordinate axes.) The preceding theorem can be extended to this case, although the condition for countable additivity (3.13) will become more complicated.

To study properties of tree-based priors, we consider a given partitioning tree $\mathcal{T}_1, \mathcal{T}_2, \ldots$ and a random measure P on the Borel sets. The measure may have been constructed through the partitioning tree, but this is not assumed in the following. Given a tree and a random measure P, we define *splitting variables* $(V_\varepsilon, \varepsilon \in \mathcal{E}^*)$ through (3.11) and then the representation (3.12) holds.

Tree-based priors have full weak support under weak conditions. Call a collection of sets $\mathcal{X}_0 \subset \mathcal{X}$ *convergence-forcing* for weak convergence if $P_m(A) \to P(A)$ for every $A \in \mathcal{X}_0$, for probability measures $\{P_m\}$ and P, implies that $P_m \rightsquigarrow P$. An example is the set of all cells with endpoints in a dense subset of $\mathbb{R}^k$.

Theorem 3.10 (Weak support) *Let $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m)$ be a sequence of successive binary partitions that generates the Borel sets and such that $\cup_m \mathcal{T}_m$ is convergence-forcing for weak convergence. If P is a random probability measure with splitting variables $(V_{\varepsilon 0} : \varepsilon \in \mathcal{E}^*)$ of which every finite-dimensional subvector has full support (a unit cube), then the weak support of P is the full space $\mathfrak{M}$.*

Proof Because $\cup_m \mathcal{T}_m$ is convergence-forcing for weak convergence, the topology generated by the metric $d(P, Q) = \sum_{\varepsilon \in \mathcal{E}^*} |P(A_\varepsilon) - Q(A_\varepsilon)| 2^{-2|\varepsilon|}$, is at least as strong as the weak topology. Therefore, it suffices to show that every open d-ball receives positive prior mass. For given $\eta > 0$, we can find m such that $\sum_{j>m} 2^{-j} < \eta$. Hence all probability measures P with

$$\sum_{\varepsilon \in \mathcal{E}^m} |P(A_\varepsilon) - P_0(A_\varepsilon)| < \eta$$

are in a d-ball of radius 2η around a given P_0. The latter event can be described in terms of a continuous function of the vectors $(V_\varepsilon : \varepsilon \in \mathcal{E}^j)$, for $j = 1, \ldots, m$. The inverse image of the set under this continuous map is open and has positive mass under the joint distribution of these vectors, as this has full support by assumption. □

3.6 Tail-Free Processes

It is simple and appealing to choose the splitting variables in a tree-based prior independent across the levels of the hierarchy. This leads to a "tail-free" random measure. In this section we study this property in general.

Consider a given partitioning tree $\mathcal{T}_1, \mathcal{T}_2, \ldots$ and a random measure P on the Borel sets, and define the splitting variables $(V_\varepsilon, \varepsilon \in \mathcal{E}^*)$ as in (3.11).

Definition 3.11 (Tail-free) The random measure P is a *tail-free process* with respect to the sequence of partitions $\mathcal{T}_m$ if $\{V_0\} \perp\!\!\!\perp \{V_{00}, V_{10}\} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \{V_{\varepsilon 0} : \varepsilon \in \mathcal{E}^m\} \perp\!\!\!\perp \cdots$.

A degenerate prior is certainly tail-free according to this definition (with respect to any sequence of partitions), since all its V-variables are degenerate at appropriate values. Nontrivial and important examples are the Pólya-tree and Dirichlet processes, which are discussed respectively in Section 3.7 and Chapter 4.

Tail-free processes enjoy an obvious *equivariance* property under transformations: if P is a tail-free process with respect to a given sequence of partitions $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m\}$ and g is a measurable isomorphism, then the induced random measure $P \circ g^{-1}$ is tail-free with respect to the partitions $\{g^{-1}(A_\varepsilon) : \varepsilon \in \mathcal{E}^m\}$. In the case that $\mathfrak{X} = \mathbb{R}$, where ordered partitions in intervals, as in Theorem 3.9, are natural, the transformed partitions will again consist of ordered intervals if g is monotonically increasing.

On the other hand, the tail-free property is not preserved under the formation of mixtures. For instance, a mixture of a nondegenerate and a degenerate tail-free process fails to be tail-free. In statistical applications this means that if the partitions and/or the V-variables involve hyperparameters, which are given a nondegenerate distribution, then the resulting mixture prior will not be tail-free, except in trivial cases. In Chapter 6 this will be seen to have important negative implications for posterior consistency.

The factorization (3.12) of the probabilities $P(A_\varepsilon)$ of a tail-free process and independence of the V-variables make it easy to compute means and variances of probabilities and log-probabilities of the sets in the partitions. Part (iv) of the following proposition shows that the probabilities of sets appearing later in the hierarchy have more variation in the logarithmic scale.

Proposition 3.12 (Moments) *For every tail-free process P and $\varepsilon = \varepsilon_1 \varepsilon_2 \cdots \in \mathcal{E}^*$,*

(i) $\mathrm{E}(P(A_\varepsilon)) = \prod_{j=1}^{|\varepsilon|} \mathrm{E}(V_{\varepsilon_1 \cdots \varepsilon_j})$.

(ii) $\mathrm{var}(P(A_\varepsilon)) = \prod_{j=1}^{|\varepsilon|} \mathrm{E}(V_{\varepsilon_1 \cdots \varepsilon_j}^2) - (\mathrm{E}(P(A_\varepsilon)))^2$.

(iii) $\mathrm{E}(\log P(A_\varepsilon)) = \sum_{j=1}^{|\varepsilon|} \mathrm{E}(\log V_{\varepsilon_1 \cdots \varepsilon_j})$.

(iv) $\mathrm{var}(\log P(A_\varepsilon)) = \sum_{j=1}^{|\varepsilon|} \mathrm{var}(\log V_{\varepsilon_1 \cdots \varepsilon_j})$.

Consequently, if $\sup_{\varepsilon \in \mathcal{E}^*} \mathrm{E}(V_\varepsilon^2) < 1/2$, then $\max\{P(A_\varepsilon) : \varepsilon \in \mathcal{E}^m\} \to 0$ a.s. at an exponential rate.

Proof Parts (i)–(iv) are immediate from (3.12) and the independence of the splitting variables. For the final assertion we bound the maximum by the sum and apply (i) and (ii) to see that, if $r < 1$ is a uniform upper bound on $2\mathrm{E}(V_\varepsilon^2)$,

$$\mathrm{E}[\max_{\varepsilon \in \mathcal{E}^m} P(A_\varepsilon)]^2 \leq \sum_{\varepsilon \in \mathcal{E}^m} \prod_{j=1}^{m} \mathrm{E}(V_{\varepsilon_1 \cdots \varepsilon_j}^2) \leq 2^m (r/2)^m = r^m. \tag{3.14}$$

The result is now immediate from the Borel-Cantelli lemma. □

In applications, we may wish to choose the partitions and the V-variables such that the tail-free process possesses a given mean measure μ. Part (i) of the preceding proposition

shows that this can always be achieved by combining equal probability splits at every level (i.e. $\mu(A_{\varepsilon 0}) = \mu(A_{\varepsilon 1})$ for every $\varepsilon \in \mathcal{E}^*$) with V-variables with mean $\mathrm{E}(V_\varepsilon) = 1/2$. For a continuous measure μ on $\mathfrak{X} = \mathbb{R}$ this is always possible: the first split is at the median of μ, the second at the quartiles, the third at the octiles, etc.; the mth partition will have the $(k/2^m)$th quantiles of μ as its boundary points and consist of sets of μ-probability 2^{-m}. We refer to a partition with $\mu(A_\varepsilon) = 2^{-|\varepsilon|}$, for every $\varepsilon \in \mathcal{E}^*$, as a *canonical partition* relative to μ. (On $\mathbb{R}$ with interval splits this is uniquely determined by μ, whereas on a general space μ is determined by the partition, but not the other way around.)

The uniform distribution over the unit interval may seem a natural default choice for meeting the condition $\mathrm{E}(V_\varepsilon) = 1/2$. However, the resulting prior is peculiar in that its realizations are continuous, but a.s. singular with respect to the Lebesgue measure (see Lemma 3.17).

The mass $P(A_\varepsilon)$ of a partitioning set at level m can be expressed in the V-variables up to level m (see (3.12)), while, by their definition (3.11), the V-variables at higher levels control conditional probabilities. Therefore, tail-freeness makes the distribution of mass *within* every partitioning set in $\mathcal{T}_m$ independent of the distribution of the total mass one *among* the sets in $\mathcal{T}_m$. Definition 3.11 refers only to masses of partitioning sets, but under the assumption that the partitions generate the Borel sets, the independence extends to all Borel sets.

Lemma 3.13 *If P is a random measure that is tail-free relative to a sequence of partitions $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets $\mathscr{X}$ in $\mathfrak{X}$, then for every $m \in \mathbb{N}$ the process $(P(A|\, A_\varepsilon) : A \in \mathscr{X}, \varepsilon \in \mathcal{E}^m)$ is independent of the random vector $(P(A_\varepsilon) : \varepsilon \in \mathcal{E}^m)$.*

Proof Because P is a random measure, its mean measure $\mu(A) = \mathrm{E}P(A)$ is a well defined Borel probability measure. As $\mathcal{T} := \cup_m \mathcal{T}_m$ is a field, which generates the Borel σ-field by assumption, there exists for every $A \in \mathscr{X}$ a sequence A_n in $\mathcal{T}$ such that $\mu(A_n \,\Delta\, A) \to 0$. Because P is a random measure, $P(A_n|\, A_\varepsilon) \to P(A|\, A_\varepsilon)$ in mean and hence a.s. along a subsequence. It follows that the random variable $P(A|\, A_\varepsilon)$ is measurable relative to the completion of the σ-field generated by the variables $P(C|\, A_\varepsilon)$, for $C \in \mathcal{T}$. Every of these conditional probabilities is a finite sum of probabilities of the form $P(A_{\varepsilon\delta}|\, A_\varepsilon) = V_{\varepsilon\delta_1} \cdots V_{\varepsilon\delta_1 \cdots \delta_k}$, for $\delta = \delta_1 \cdots \delta_k \in \mathcal{E}^k$ and $k \in \mathbb{N}$. Therefore, by tail-freeness this σ-field is independent of the σ-field generated by the variables $P(A_\varepsilon) = V_{\varepsilon_1} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}$, for $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^m$. □

Relative to the σ-field $\mathscr{M}$ generated by all maps $M \mapsto M(A)$, the process $(P(A|\, A_\varepsilon) : A \in \mathscr{X})$ contains all information about the conditional random measure $P(\cdot|\, A_\varepsilon)$. Thus the preceding theorem truly expresses that the "conditional measure within partitioning sets is independent of the distribution of mass among them."

Suppose that the data consist of an i.i.d. sample $X_1, \ldots, X_n$ from a distribution P, which is a priori modeled as a tail-free process. For each $\varepsilon \in \mathcal{E}^*$, denote the number of observations falling in A_ε by

$$N_\varepsilon := \#\{1 \le i \le n : X_i \in A_\varepsilon\}. \tag{3.15}$$

For each m the vector $(N_\varepsilon : \varepsilon \in \mathcal{E}^m)$ collects the counts of all partitioning sets at level m. The following theorem shows that this vector contains all information (in the Bayesian sense)

about the probabilities $(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m)$ of these sets: the additional information about the precise positions of the X_i within the partitioning sets is irrelevant.

Theorem 3.14 *A random measure P is tail-free relative to a given sequence of partitions $\mathcal{T}_m = \{A_\varepsilon: \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets if and only if for every m and n the posterior distribution of $(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m)$ given an i.i.d. sample $X_1, \ldots, X_n$ from P is the same as the posterior distribution of this vector given $(N_\varepsilon: \varepsilon \in \mathcal{E}^m)$ defined in* (3.15)*, a.s.*

Proof Fix m. Given P the data $X = (X_1, \ldots, X_n)$ can be generated in two steps. First we generate a multinomial vector $N = (N_\varepsilon: \varepsilon \in \mathcal{E}^m)$ with parameters n and $(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m)$. Next, given N, we generate an i.i.d. sample of size N_ε from the measure $P(\cdot\,|\,A_\varepsilon)$, independently for every $\varepsilon \in \mathcal{E}^m$, and randomly order the n values so obtained.

The random measure P can also be generated in two steps. First we generate the vector $\theta := (P(A_\varepsilon): \varepsilon \in \mathcal{E}^m)$, and second the process $\eta := (P(A\,|\,A_\varepsilon): A \in \mathcal{X}, \varepsilon \in \mathcal{E}^m)$. For a tail-free measure P these steps are independent.

The first step of the generation of the data depends on θ only, and the second only on (N, η). The set of (conditional) independencies $\eta \perp\!\!\!\perp (N, \theta)$, if P is tail-free, and $X \perp\!\!\!\perp \theta\,|\,(N, \eta)$ implies that $\theta \perp\!\!\!\perp X\,|\,N$. Indeed, these assumption imply that $\mathrm{E}(f(X)g(\theta)\,|\,N, \eta) = \mathrm{E}(f(X)\,|\,N, \eta)\mathrm{E}(g(\theta)\,|\,N)$, for any bounded, measurable functions f and g, from which the assertion follows by taking the conditional expectation given N left and right. Now $\theta \perp\!\!\!\perp X\,|\,N$ is equivalent to the "only if" part of the theorem for this special representation of prior and data. Because the assertion depends on the joint distribution of (P, X, N) only, it is true in general.

For the proof that dependence on the cell counts only implies tail-freeness, let $N' = (N_\varepsilon: \varepsilon \in \mathcal{E}^{m+1})$. First note that $\theta \perp\!\!\!\perp X\,|\,N$ implies $\theta \perp\!\!\!\perp N'\,|\,N$, which can also be represented as $p(\theta\,|\,N') = p(\theta\,|\,N)$. Thus

$$\mathrm{P}(\theta \in C, N' = y')\mathrm{P}(N = y) = \mathrm{P}(\theta \in C, N = y)\mathrm{P}(N' = y'),$$

for every measurable set C and every pair $y' = (y'_\varepsilon: \varepsilon \in \mathcal{E}^{m+1})$ and $y = (y_\varepsilon: \varepsilon \in \mathcal{E}^m)$ that are compatible (i.e. $y_\varepsilon = \sum_{\delta \in \mathcal{E}} y'_{\varepsilon\delta}$ for all ε). Because given P, the vectors N' and N possess multinomial distributions and θ is a function of P, this can be further rewritten as

$$\mathrm{E}\mathbb{1}_{\theta \in C}\binom{n}{y'}\prod_{\varepsilon \in \mathcal{E}^{m+1}} P(A_\varepsilon)^{y'_\varepsilon}\mathrm{P}(N = y) = \mathrm{E}\mathbb{1}_{\theta \in C}\binom{n}{y}\prod_{\varepsilon \in \mathcal{E}^m} P(A_\varepsilon)^{y_\varepsilon}\mathrm{P}(N' = y').$$

In view of the definition of θ we conclude that, provided that $\mathrm{P}(N = y) > 0$,

$$\mathrm{E}\Big(\prod_{\varepsilon \in \mathcal{E}^{m+1}} P(A_\varepsilon)^{y'_\varepsilon}\,|\,\theta\Big) = \prod_{\varepsilon \in \mathcal{E}^m} P(A_\varepsilon)^{y_\varepsilon}\, c(y'),$$

for $c(y') = \binom{n}{y}\mathrm{P}(N' = y')/\mathrm{P}(N = y)/\binom{n}{y'}$. If $\mathrm{P}(N = y) = \mathrm{E}\binom{n}{y}\prod_{\varepsilon \in \mathcal{E}^m} P(A_\varepsilon)^{y_\varepsilon} = 0$, then the variable in the left side is zero and the identity is true with $c(y') = 0$. In other words, the conditional mixed moment of degree y' of the vector $V = (P(A_{\varepsilon\delta})/P(A_\varepsilon): \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E})$ given $(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m)$ is equal to a deterministic number $c(y')$. This being true for every y' in n times the unit simplex, for every n, implies independence of V and $(P(A_\varepsilon): \varepsilon \in \mathcal{E}^m)$. □

Given P the vector $(N_\varepsilon\colon \varepsilon \in \mathcal{E}^m)$ possesses a multinomial distribution with parameters n and $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$. Finding the posterior distribution of the latter vector of cell probabilities therefore reduces to the finite dimensional problem of multinomial probabilities. This not only makes computations easy, but also means that asymptotic properties of the posterior distribution follow those of parametric problems, for instance easily leading to consistency in an appropriate sense, as discussed in Chapter 6. The result also justifies the term "tail-free" in that posterior computation can be carried out without looking at the tail of the prior.

Tail-free processes form a conjugate class of priors, in the sense that the posterior process is again tail-free.

Theorem 3.15 (Conjugacy) *The posterior process corresponding to observing an i.i.d. sample $X_1, \ldots, X_n$ from a distribution P that is a priori modeled by a tail-free process is tail-free with respect to the same sequence of partitions as in the definition of the prior.*

Proof We must show that the vectors $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ defined in (3.11) are mutually conditionally independent across levels m, given the data. It suffices to show sequentially for every m that this vector is conditionally independent of the vectors corresponding to lower levels. Because the vectors $(V_\varepsilon\colon \varepsilon \in \cup_{k\le m}\mathcal{E}^k)$ and $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$ generate the same σ-field, it suffices to show that $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ is conditionally independent of $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$, for every fixed m.

Together these vectors are equivalent to the vector $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^{m+1})$. Therefore, by Theorem 3.14 the joint posterior distribution of the latter vectors depends only on the cell counts, $N = (N_{\varepsilon\delta}\colon \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E})$, and "conditionally given the data" can be interpreted as "given this vector N." Writing $V = (V_{\varepsilon\delta}\colon \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E})$ and $\theta = (\theta_\varepsilon\colon \varepsilon \in \mathcal{E}^m)$, for $\theta_\varepsilon = P(A_\varepsilon)$, we can write the likelihood for (V, θ, N) as

$$\binom{n}{N} \prod_{\varepsilon\in\mathcal{E}^m,\delta\in\mathcal{E}} (\theta_\varepsilon V_{\varepsilon\delta})^{N_{\varepsilon\delta}}\, d\Pi_1(V)\, d\Pi_2(\theta).$$

Here Π_1 and Π_2 are the marginal (prior) distributions of V and θ, and we have used that these vectors are independent under the assumption that P is tail-free. Clearly the likelihood factorizes in parts involving (V, N) and involving (θ, N). This shows that V and θ are conditionally independent given N. □

The atoms in the σ-field generated by $\mathcal{T}_0, \mathcal{T}_1, \ldots$ are the sets $\cap_{m=1}^\infty A_{\varepsilon_1\cdots\varepsilon_m}$, for $\varepsilon_1\varepsilon_2\cdots$ an infinite string of 0s and 1s. If the sequence of partitions generates the Borel sets, then these atoms are single points (or empty). Their probabilities under a tree-based random measure are $\prod_{m=1}^\infty V_{\varepsilon_1\cdots\varepsilon_m}$, and hence $P\{x\} = 0$ a.s. for every x as soon as these infinite products are zero. For tail-free random measures this is typical, as the means $\mathrm{E}[V_\varepsilon]$ will usually be smaller and bounded away from 1.

However, this does not imply that P is a.s. continuous (atomless), but only says that P has no *fixed atoms*. Because there are uncountably many possible atoms (or infinitely long sequences of 0s and 1s), every realization of P may well have some atom. In fact, every realization of P can be discrete, as is illustrated by the Dirichlet process priors discussed in Chapter 4.

In practice, an atomless, or even absolutely continuous, random measure may be preferable. This may be constructed by choosing splitting variables that are "close" to the splits of a deterministic absolutely continuous measure. For instance, to make the random measure resemble the canonical measure for the partition, we choose the splitting variables close to 1/2. In the other case, when the mass is frequently divided unevenly between two offspring sets in the splitting tree, then the realizations of the resulting random measure will tend to concentrate more around some points and fail to have a density. The following theorem makes "close to $1/2$" precise.

Theorem 3.16 (Absolute continuity) *Let $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m)$ be a sequence of successive binary partitions that generates the Borel sets. If P is a random measure with splitting variables $(V_\varepsilon : \varepsilon \in \mathcal{E}^*)$ that satisfy, for an arbitrary probability measure μ,*

$$\sup_{m\in\mathbb{N}} \max_{\varepsilon\in\mathcal{E}^m} \frac{\mathrm{E}(\prod_{j=1}^m V_{\varepsilon_1\cdots\varepsilon_j}^2)}{\mu^2(A_{\varepsilon_1\cdots\varepsilon_m})} < \infty, \tag{3.16}$$

then almost all realizations of P are absolutely continuous with respect to μ, i.e. $\Pi(P \ll \mu) = 1$. In particular, this condition is satisfied if P is tail-free, the partitions are canonical relative to μ and the following two conditions hold:

$$\sum_{m=1}^\infty \max_{\varepsilon\in\mathcal{E}^m} \left|\mathrm{E}(V_\varepsilon) - \frac{1}{2}\right| < \infty, \qquad \sum_{m=1}^\infty \max_{\varepsilon\in\mathcal{E}^m} \mathrm{var}(V_\varepsilon) < \infty. \tag{3.17}$$

In this case a density process $x \mapsto p(x)$ is given by, for $x \in \cap_{m=1}^\infty A_{x_1\cdots x_m}$[1]*,*

$$p(x) = \prod_{j=1}^\infty (2V_{x_1x_2\cdots x_j}). \tag{3.18}$$

For the canonical partitions relative to the Lebesgue measure on $\mathfrak{X} = (0,1)$ and fully supported and independent variables $V_{\varepsilon 0}$, for $\varepsilon \in \mathcal{E}^*$, any version of this process is discontinuous almost surely at every boundary point of the partitions.

Proof The σ-fields $\sigma(\mathcal{T}_1) \subset \sigma(\mathcal{T}_2) \subset \cdots$ are increasing and generate the Borel σ-field. If the restrictions P_m and μ_m of two fixed (nonrandom) Borel probability measures P and μ to $\sigma(\mathcal{T}_m)$ satisfy $P_m \ll \mu_m$, then by Lemma L.7, the sequence of densities $p_m = dP_m/d\mu_m$ tends μ almost surely to the density of the absolutely continuous part of P with respect to μ. In particular, if p_m tends to p, and $\int p\,d\mu = 1$, then P is absolutely continuous with respect to μ, with density p. We shall apply this to the present *random* measures.

Condition (3.16) implies that $\mu(A_\varepsilon) > 0$, for every $\varepsilon \in \mathcal{E}^m$, and hence μ_m dominates any measure on $\sigma(\mathcal{T}_m)$. The density of P_m is given by

$$p_m = \sum_{\varepsilon\in\mathcal{E}^m} \frac{P(A_\varepsilon)}{\mu(A_\varepsilon)} \mathbb{1}_{A_\varepsilon}, \tag{3.19}$$

[1] This condition defines a binary coding $x_1x_2\cdots$ for any $x \in \mathfrak{X}$. In the case of the canonical partitions of the unit interval in cells, it coincides with the infinite binary expansion of x (not ending in $000\cdots$).

and satisfies, for Π the law of P,

$$\int\int p_m^2 \, d\mu \, d\Pi(p) = \sum_{\varepsilon \in \mathcal{E}^m} \frac{\mathrm{E}(\prod_{j=1}^m V_{\varepsilon_1 \cdots \varepsilon_j}^2)}{\mu(A_{\varepsilon_1 \cdots \varepsilon_m})} \leq \max_{\varepsilon \in \mathcal{E}^m} \frac{\mathrm{E}(\prod_{j=1}^m V_{\varepsilon_1 \cdots \varepsilon_j}^2)}{\mu^2(A_{\varepsilon_1 \cdots \varepsilon_m})}.$$

By assumption, the right side is bounded in m, which implies that the sequence p_m is uniformly $(\mu \times \Pi)$-integrable. It follows that $\int\int p \, d\mu \, d\Pi = \lim_{m \to \infty} \int\int p_m \, d\mu \, d\Pi = 1$, which implies that $\int p \, d\mu = 1$ a.s. $[\Pi]$.

For a canonical partition the denominator of the quotient in (3.16) is 2^{-2m} and the quotient can be written in the form $\mathrm{E} \prod_{j=1}^m (2V_{\varepsilon_1 \cdots \varepsilon_j})^2 = \prod_{j=1}^m \mathrm{E}(2V_{\varepsilon_1 \cdots \varepsilon_j})^2$, if P is tail-free. If $|\mathrm{E}(V_\varepsilon) - 1/2| \leq \delta_j$ and $var(V_\varepsilon) \leq \gamma_j$ for $\varepsilon \in \mathcal{E}^j$, then

$$\mathrm{E}(V_{\varepsilon_1 \cdots \varepsilon_j}^2) = var(V_{\varepsilon_1 \cdots \varepsilon_j}) + (\mathrm{E}(V_{\varepsilon_1 \cdots \varepsilon_j}))^2 \leq (1 + 4\gamma_j + 4\delta_j + 4\delta_j^2)/4.$$

Therefore, the expression in (3.16) is bounded by $\prod_{j=1}^\infty (1 + 4\gamma_j + 4\delta_j + 4\delta_j^2)$, which is finite if $\sum_j (\delta_j + \gamma_j) < \infty$.

If $x \in A_{x_1 \cdots x_m}$ and $\mu(A_\varepsilon) = 2^{-|\varepsilon|}$, then the right-hand side of (3.19) evaluated at x reduces to $2^{-m} P(A_{x_1 \cdots x_m})$. Combined with (3.12) this gives formula (3.18).

For the proof of the last assertion consider an arbitrary point $x = x_1 x_2 \cdots \in (0, 1)$ in its infinite binary expansion (so not ending in $000 \cdots$), so that $A_{x_1 \cdots x_m} = (x_m^-, x_m^+]$, for $x_m^- = x_1 \cdots x_m 000 \cdots < x \leq x_m^+ = x_m^- + 2^{-m} = x_1 \cdots x_m 111 \cdots$. The finite binary expansion of x_m^+ is given by $x_m^+ = x_1 \cdots x_{j_m - 1} 1 \cdots 1000 \cdots$, where j_m is the biggest integer $j \leq m$ with $x_j = 0$ and the zeros start at coordinate $m + 1$; hence, the partitioning set immediately to the right of $A_{x_1 \cdots x_m}$ is $A_{x_1 \cdots x_{j_m - 1} 1 \cdots 1}$. If F is the cumulative distribution function of P, then

$$F(x_m^- + 2^{-m}) - F(x_m^-) = P(A_{x_1 \cdots x_m}) = \prod_{j=1}^m V_{x_1 \cdots x_j},$$

$$F(x_m^+ + 2^{-m}) - F(x_m^+) = P(A_{x_1 \cdots x_{j_m - 1} 1 \cdots 1}) = \prod_{j=1}^{j_m - 1} V_{x_1 \cdots x_j} \prod_{j=j_m}^m V_{x_1 \cdots x_{j_m - 1} 1 \cdots 1}.$$

If F is absolutely continuous with density f that is continuous at x, then $2^m (F(y_m + 2^{-m}) - F(x_m)) = 2^m \int_{y_m}^{y_m + 2^{-m}} f(s) \, ds \to f(x)$ for any $y_m \to x$. In particular the quotient of the right sides of the preceding display tends to 1 as $m \to \infty$.

If x is a boundary point of the $(m_0 + 1)$th partition, then $x = x_1 \cdots x_{m_0} 0111 \cdots$ for some m_0, and $j_m = m_0 + 1$ for $m > m_0$ and hence the quotient tends to the infinite product $\prod_{j > m_0} (V_{x_1 \cdots x_j} / V_{x_1 \cdots x_{m_0} 1 \cdots 1})$. The logarithm of this expression is a sum of independent, continuous variables and hence possesses a continuous (atomless) distribution; in particular the probability that it is $\log 1 = 0$ is zero. □

It follows that the realizations of a tail-free measure with splitting variables such that $\mathrm{E}|V_\varepsilon - 1/2|^2 \to 0$ as $|\varepsilon| \to \infty$ at a fast enough speed possess densities relative to the canonical measure for the partition. On the other hand, fixed (for instance, uniform) splitting variables give widely unbalanced divisions with appreciable probabilities and lead to singular measures.

Lemma 3.17 *Let P be a tail-free process with respect to a canonical partition relative to a probability measure μ and with i.i.d. splitting variables $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^*)$.*

(a) *If $\mathrm{P}(V_{\varepsilon 0} \neq 1/2) > 0$, then P is a.s. singular with respect to μ.*
(b) *If $\mathrm{E}(V_{\varepsilon 0}) = 1/2$ and $\mathrm{P}(0 < V_{\varepsilon 0} < 1) > 0$, then P is a.s. nonatomic.*

Proof As in the proof of Theorem 3.16, the density of the absolutely continuous part of P with respect to μ can be computed as the limit of the sequence p_m given in (3.19), where presently we substitute $\mu(A_\varepsilon) = 2^{-|\varepsilon|}$. For (a) it suffices to prove that this limit is zero a.s.

Define a map from $\mathfrak{X}$ to $\{0, 1\}^\infty$ by $x \mapsto \varepsilon_1\varepsilon_2\cdots$ if $x \in \cap_{j=1}^\infty A_{\varepsilon_1\varepsilon_2\cdots\varepsilon_j}$. Write the map as $x \mapsto x_1x_2\cdots$, so that $x \in A_{x_1\cdots x_m}$, for every m. Formula (3.19) can then be written in the form

$$\frac{1}{m}\log p_m(x) = \frac{1}{m}\log\left(2^m P(A_{x_1\cdots x_m})\right) = \frac{1}{m}\sum_{j=1}^m \log(2V_{x_1\cdots x_j}).$$

If X is a random element in $\mathfrak{X}$ with law μ and $X_1X_2\ldots$ is its image under the map $x \mapsto x_1x_2\cdots$ defined previously, then $\mathrm{P}(X_1 = \varepsilon_1, \ldots, X_m = \varepsilon_m) = \mu(A_\varepsilon) = 2^{-m}$, for every $\varepsilon = \varepsilon_1\cdots\varepsilon_m \in \mathcal{E}^m$, and hence $X_1, X_2, \ldots$ are i.i.d. Bernoulli variables.

The variables $V_{x_1\cdots x_j}$ are independent and distributed as V if $x_j = 0$ and as $1 - V$ if $x_j = 1$, for V a variable with a given distribution with $\mathrm{P}(V \neq 1/2) > 0$. By the strong law of large numbers μ-almost every sequence $X_1, X_2, \ldots$ has equal limiting frequencies of 0s and 1s. For any such realization $x_1x_2\cdots$, another application of the strong law of large numbers gives that

$$\frac{1}{m}\log p_m(x) \to \frac{1}{2}\mathrm{E}(\log(2V)) + \frac{1}{2}\mathrm{E}(\log(2(1-V))), \qquad \text{a.s.}$$

Because $4v(1 - v) < 1$ for all $v \neq 1/2$, under assumption (a) the right side of this display is negative, which implies that $p_m(x) \to 0$.

We have shown that the event $\{\lim p_m(x) = 0\}$ has probability 1 for every x in a set of μ-measure 1. Using Fubini's theorem we can also formulate this in the form that $\lim p_m(x) = 0$ for a. e. x $[\mu]$, with probability one.

To prove (b), observe that the probability of any atom is bounded by $\max_{\varepsilon\in\mathcal{E}^m} P(A_\varepsilon)$ for every m, and hence is 0 a.s., in view of Proposition 3.12, since $\mathrm{E}(V^2) < \mathrm{E}(V)$ whenever $\mathrm{P}(0 < V < 1) > 0$, and both $V_{\varepsilon 0}$ and $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$ have mean 1/2 by assumption. □

The last two theorems concern the event that a random measure P is singular or absolutely continuous. From Proposition A.7 it follows that the sets of singular probability measures and absolutely continuous probability measures are indeed measurable subsets of $\mathfrak{M}$, as are the sets of nonatomic and discrete probability measures.

The following theorem shows that it is no coincidence that these events were trivial in all cases: tail-free processes satisfy the following zero-one law.

Theorem 3.18 (Zero-one law) *A random measure that is tail-free with respect to a sequence of partitions that generates the Borel σ-field and with splitting variables such that $0 < V_\varepsilon < 1$ for all $\varepsilon \in \mathcal{E}^*$ is, with probability zero or one,*

(i) *absolutely continuous with respect to a given measure,*

(ii) *mutually singular with respect to a given measure,*
(iii) *atomless,*
(iv) *discrete.*

Proof We prove only (i); the proofs of the other parts are virtually identical. Fix $m \in \mathbb{N}$, and decompose P as $P(A) = \sum_{\varepsilon \in \mathcal{E}^m} P(A|\, A_\varepsilon) P(A_\varepsilon)$. Because $P(A_\varepsilon) > 0$ for all $\varepsilon \in \mathcal{E}^m$, the measure P is absolutely continuous with respect to a given measure λ if and only if $P(\cdot|\, A_\varepsilon)$ is so for all $\varepsilon \in \mathcal{E}^m$. Since every measure $P(\cdot|\, A_\varepsilon)$ is describable in terms of (only) $(V_{\varepsilon\delta}, \delta \in \mathcal{E}^*)$, so is the event that $P(\cdot|\, A_\varepsilon)$ is absolutely continuous with respect to λ. In other words, this event is tail-measurable for the sequence of independent random vectors $(V_\varepsilon\colon \varepsilon \in \mathcal{E}^1), (V_\varepsilon\colon \varepsilon \in \mathcal{E}^2), \ldots$ By Kolmogorov's zero-one law, it has probability zero or one. □

Under the sufficient condition for absolute continuity given in Theorem 3.16, a tail-free process automatically has all absolutely continuous measures in its total variation support.

Theorem 3.19 (Strong support) *Let P be a random measure that is tail-free relative to a μ-canonical sequence of partitions that generates the Borel σ-field. If the splitting variables V_ε satisfy* (3.17)*, and the random vectors $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ have full support $[0,1]^{2^m}$, then the total variation support of P consists of all probability measures that are absolutely continuous relative to μ.*

Proof By Theorem 3.16 P is μ-absolutely continuous with density p given by (3.18). As the partitions generate the Borel sets, the probability densities p_0 that are constant on every set in some partition $\mathcal{T}_m$, for some m, are dense in the set of all μ-probability densities relative to $\mathbb{L}_1(\mu)$-metric. For $p_m(x) = \prod_{j \le m}(2V_{x_1 x_2 \cdots x_j})$,

$$\int |p - p_0|\, d\mu \le \int \left| \frac{p}{p_m} - 1 \right| d\mu \, \|p_m\|_\infty + \|p_m - p_0\|_\infty.$$

Because under tail-freeness the process $x \mapsto (p/p_m)(x) = \prod_{j>m}(2V_{x_1 \cdots x_j})$ is independent of the process $x \mapsto p_m(x)$, the prior probability that the left side is smaller than ϵ is bigger than

$$\Pi\Big(\int \left| \frac{p}{p_m} - 1 \right| d\mu < \frac{\epsilon}{2\|p_0\|_\infty + \epsilon} \Big) \Pi(\|p_m - p_0\|_\infty < \epsilon/2). \tag{3.20}$$

The second probability refers to the event that the vector $(V_{\varepsilon 0}\colon \varepsilon \in \cup_{j \le m} \mathcal{E}^j)$ belongs to a certain nonempty open set, and hence is positive by assumption. The assumptions on the means and variances of the variables V_ε and the Cauchy-Schwarz inequality imply that $\sum_{j=1}^\infty |2V_{x_1 \cdots x_j} - 1| < \infty$ a.s. and hence, for every x and $m \to \infty$,

$$\frac{p}{p_m}(x) = \prod_{j>m} (2V_{x_1 \cdots x_j}) \sim e^{\sum_{j>m}(2V_{x_1 \cdots x_j} - 1)} \to 1, \qquad \text{a.s.}$$

Using Fubini's theorem we conclude that $p(x)/p_m(x) \to 1$ almost surely under the measure $\mu \times \Pi$, as $m \to \infty$. Furthermore, for $n > m$,

$$\int\int \Big(\frac{p_n}{p_m}\Big)^2 d\mu\, d\Pi(p) = \sum_{\varepsilon \in \mathcal{E}^n} \mathrm{E} \prod_{m<j\le n} (2V_{x_1\cdots x_j})^2 2^{-n} \le \max_{\varepsilon\in\mathcal{E}^n} \prod_{m<j\le n} \mathrm{E}(2V_{x_1\cdots x_j})^2.$$

With notation as in the proof of Theorem 3.16, the right-hand side can be bounded above by $\prod_{m<j\le n}(1+4\gamma_j+4\delta_j+4\delta_j^2)$, which remains finite in the limit $n \to \infty$. By Fatou's lemma $\int\int (p/p_m)^2\, d\mu\, d\Pi(p)$ is smaller than the limit and hence is bounded in m. This shows that p/p_m is uniformly integrable relative to $\mu \times \Pi$, and we can conclude that $\mathrm{E}\int |(p/p_m) - 1|\, d\mu \to 0$, as $m \to \infty$. This implies that the first probability in (3.20) tends to one, as $m \to \infty$, for any ϵ. Given a piecewise constant p_0 and $\epsilon > 0$, we choose m sufficiently large that the latter probability is positive and such that p_0 is piecewise constant on $\mathcal{T}_m$ to finish the proof. □

3.7 Pólya Tree Processes

Let $\mathcal{T}_m = \{A_\varepsilon : \varepsilon \in \mathcal{E}^m\}$ be a sequence of successive, binary, measurable partitions of the sample space, as before, and for a random measure P on $(\mathfrak{X}, \mathscr{X})$, define random variables $V_{\varepsilon 0}$ as in (3.11).

Definition 3.20 (Pólya tree) A random probability measure P is said to follow a *Pólya tree process* with parameters $(\alpha_\varepsilon : \varepsilon \in \mathcal{E}^*)$ with respect to the sequence $\{\mathcal{T}_m\}$ of partitions if the variables $V_{\varepsilon 0}$, for $\varepsilon \in \mathcal{E}^* \cup \{\varnothing\}$, are mutually independent and $V_{\varepsilon 0} \sim \mathrm{Be}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$.

Thus a Pólya tree process is a tail-free process with splitting variables $V_{\varepsilon 0}$ that are mutually independent not only between, but also *within* the levels, and possess beta-distributions. The parameters α_ε of the beta distributions must be nonnegative; we allow the value 0 provided that $\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1} > 0$, for every $\varepsilon \in \mathcal{E}^*$, with the beta-distributions $\mathrm{Be}(0, \alpha)$ and $\mathrm{Be}(\alpha, 0)$ interpreted as the degenerate distributions at 0 and 1, respectively, for any $\alpha > 0$.

We denote this process by $\mathrm{PT}(\mathcal{T}_m, \alpha_\varepsilon)$. If the partitions are canonical relative to a measure μ, then we also write $\mathrm{PT}(\mu, \alpha_\varepsilon)$. The choice $\mathrm{E}(V_{\varepsilon 0}) = 1/2$, which then renders the mean measure equal to μ, corresponds to setting the two parameters of each beta distribution equal: $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$. The parameter vector (α_ε), then, still has many degrees of freedom. A reasonable default choice is to use a single parameter per level, that is $\alpha_\varepsilon = a_m$ for all $\varepsilon \in \mathcal{E}^m$ and a sequence of numbers a_m. We refer to the resulting priors as a *canonical Pólya tree process* with center measure μ and rate sequence a_m, and denote it by $\mathrm{PT}^*(\mu, a_m)$. In practice it may be desirable to choose different parameters at the first few levels, since then the probability of choosing a density close to some target distribution may be raised without affecting the essential path properties of the random distribution.

Here are some easy facts or consequences of the preceding:

(i) The image $P \circ g^{-1}$ of a Pólya tree process under a measurable isomorphism g on the sample space is a Pólya tree process with respect to the transformed partitions $\{g(A_\varepsilon) : \varepsilon \in \mathcal{E}^m\}$, with the same set of V-variables.

(ii) The Pólya tree exists as a random measure on $\mathbb{R}$ if, for all $\varepsilon \in \mathcal{E}^*$ (see (3.13) and use that a $\mathrm{Be}(\alpha, \beta)$-variable has mean $\alpha/(\alpha+\beta)$),

$$\frac{\alpha_{\varepsilon 0}}{\alpha_{\varepsilon 0}+\alpha_{\varepsilon 1}} \times \frac{\alpha_{\varepsilon 00}}{\alpha_{\varepsilon 00}+\alpha_{\varepsilon 01}} \times \cdots = 0, \quad \frac{\alpha_1}{\alpha_0+\alpha_1} \times \frac{\alpha_{11}}{\alpha_{10}+\alpha_{11}} \times \cdots = 0.$$

(iii) The first two moments of probabilities of partitioning sets under a Pólya tree process distribution are (see Proposition 3.12)

$$\mathrm{E}(P(A_{\varepsilon_1 \cdots \varepsilon_m})) = \prod_{j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1}}, \tag{3.21}$$

$$\mathrm{E}(P(A_{\varepsilon_1 \cdots \varepsilon_m})^2) = \prod_{j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j}(\alpha_{\varepsilon_1 \cdots \varepsilon_j}+1)}{(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1})(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1}+1)}.$$

(iv) A Pólya tree process with $\alpha_\varepsilon > 0$, for every $\varepsilon \in \mathcal{E}^*$, has full weak support as soon as the sequence of partitions is convergence-forcing (see Theorem 3.10).

(v) A Pólya tree process with parameters such that $|\alpha_\varepsilon - a_m| \le K$ for every $\varepsilon \in \mathcal{E}^m$ and some constant K and numbers a_m with $\sum_{m=1}^{\infty} a_m^{-1} < \infty$, is absolutely continuous with respect to a canonical measure μ for the partition, with density $x \mapsto p(x)$ satisfying, for $x \in \cap_{m=1}^{\infty} A_{x_1 \cdots x_m}$,

$$\mathrm{E}(p(x)) = \prod_{j=1}^{\infty} \frac{2\alpha_{x_1 \cdots x_j}}{\alpha_{x_1 \cdots x_{j-1} 0} + \alpha_{x_1 \cdots x_{j-1} 1}}, \tag{3.22}$$

$$\mathrm{E}(p(x)^2) = \prod_{j=1}^{\infty} \frac{4\alpha_{x_1 \cdots x_j}(\alpha_{x_1 \cdots x_j}+1)}{(\alpha_{x_1 \cdots x_{j-1} 0} + \alpha_{x_1 \cdots x_{j-1} 1})(\alpha_{x_1 \cdots x_{j-1} 0} + \alpha_{x_1 \cdots x_{j-1} 1}+1)}.$$

Furthermore, its total variation support consists of all probability measures that are absolutely continuous relative to μ. (Observe that the assumption implies $|\mathrm{E}(V_\varepsilon) - 1/2| \le c/a_m$ and $varV_\varepsilon \le c/a_m$, for every $\varepsilon \in \mathcal{E}^m$, and some constant c, and apply Theorem 3.16.)

Like the tail-free class, the Pólya tree class of priors enjoys the conjugacy property.

Theorem 3.21 (Conjugacy) *The posterior process corresponding to observing an i.i.d. sample $X_1, \ldots, X_n$ from a distribution P that is a priori modeled as a Pólya tree process prior* $\mathrm{PT}(\mathcal{T}_m, \alpha_\varepsilon)$ *is a Pólya tree process* $\mathrm{PT}(\mathcal{T}_m, \alpha_\varepsilon^*)$*, where* $\alpha_\varepsilon^* := \alpha_\varepsilon + \sum_{i=1}^{n} \mathbb{1}(X_i \in A_\varepsilon)$*, for every $\varepsilon \in \mathcal{E}^*$.*

Proof Because the posterior process is tail-free by Theorem 3.15, it suffices to show that under the posterior distribution the variables $V_{\varepsilon 0} = P(A_{\varepsilon 0} | A_\varepsilon)$ *within* every given level are independent beta variables with the given parameters $(\alpha_{\varepsilon 0}^*, \alpha_{\varepsilon 1}^*)$. Fix $m \in \mathbb{N}$ and set $V = (V_\varepsilon : \varepsilon \in \cup_{k \le m} \mathcal{E}^k)$, where $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$. By Theorem 3.14 the (joint posterior) distribution of V given $X_1, \ldots, X_n$ is the same as their conditional distribution given the vector of cell counts $N = (N_\varepsilon : \varepsilon \in \mathcal{E}^m)$. The marginal likelihood of $V_{\varepsilon 0}$ is proportional to $V_{\varepsilon 0}^{\alpha_{\varepsilon 0}}(1 - V_{\varepsilon 0})^{\alpha_{\varepsilon 1}} = V_{\varepsilon 0}^{\alpha_{\varepsilon 0}} V_{\varepsilon 1}^{\alpha_{\varepsilon 1}}$, and these variables are independent. The conditional likelihood of N

given V is multinomial, with the probabilities $P(A_\varepsilon)$ determined by (3.12). Therefore, the joint likelihood of (N, V) is proportional to

$$\prod_{|\varepsilon|=m} (V_{\varepsilon_1} V_{\varepsilon_2\varepsilon_2} \cdots V_{\varepsilon_1\cdots\varepsilon_m})^{N_\varepsilon} \prod_{|\varepsilon|\le m} V_\varepsilon^{\alpha_\varepsilon} = \prod_{|\varepsilon|\le m} V_\varepsilon^{N_\varepsilon+\alpha_\varepsilon},$$

where $N_\varepsilon = \sum_{\delta\in\mathcal{E}^{m-|\varepsilon|}} N_{\varepsilon\delta}$ denotes the number of observations falling in A_ε, also for ε with $|\varepsilon| < m$. We recognize the right side as a product of beta-likelihoods with parameters α_ε^*. □

In conjunction with equations (3.21) and (3.22), the preceding result gives expressions for the posterior means and variances of probabilities and densities, simply by replacing α_ε with α_ε^*. For instance, for the canonical Pólya tree process, formula (3.22) shows that the mean posterior density, with $N_{x_1\cdots x_j}$ the number of observations falling in the set in the jth partition that contains x (i.e. $x \in \cap_{j=1}^\infty A_{x_1\cdots x_j}$ and N defined as in (3.15)),

$$\mathrm{E}(p(x)|\, X_1, \ldots, X_n) = \prod_{j=1}^\infty \frac{2a_j + 2N_{x_1\cdots x_j}}{2a_j + N_{x_1\cdots x_{j-1}}}. \tag{3.23}$$

If no observation falls in $A_{x_1\cdots x_j}$, then $N_{x_1\cdots x_j} = 0$, and the jth factor in the product is 1, as are all terms for higher j, as the sets are decreasing. The infinite product then reduces to a finite product. This happens for every $x \notin \{X_1, \ldots, X_n\}$ for sufficiently large $j = j(x)$. Furthermore, on every partitioning set A_ε that does not contain observations, the (in)finite product is constant, as all x in such a set share their membership in the partitioning sets at the coarser levels. This implies that the mean posterior density is very regular, in contrast to the sample paths of the posterior density, which are discontinuous at every boundary point by the last assertion of Theorem 3.16.

The condition $\sum_{m=1}^\infty a_m^{-1} < \infty$ ensures the absolute continuity of the realizations of a Pólya prior, and hence effectively reduces the model to a dominated set of measures. The posterior is then absolutely continuous with respect to the prior, by Bayes's rule, as noted in Lemma 1.2. It turns out that prior and posterior are mutually singular if the series diverges.

Theorem 3.22 *Consider the posterior distribution corresponding to observing an i.i.d. sample of finite size from a distribution P that is a priori modeled as a canonical Pólya tree process* $\mathrm{PT}^*(\lambda, a_m)$. *Then prior and posterior are mutually absolutely continuous or mutually singular, depending on whether $\sum_{m=1}^\infty a_m^{-1}$ converges or diverges.*

Proof The map $\phi\colon \mathfrak{M} \to [0,1]^{\mathcal{E}^*}$ given by $P \mapsto (P(A_\varepsilon), \varepsilon \in \mathcal{E}^*)$ is a measurable isomorphism, relative to $\mathscr{M}$ and the projection σ-field, and so is the further map into the vector $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^*)$ of conditional probabilities, also viewed as subset of $[0,1]^{\mathcal{E}^*}$. Therefore, prior and posterior are absolutely continuous or singular whenever the corresponding joint laws of $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^*)$ are so. Prior and posterior are both Pólya tree processes; hence, the variables $V_{\varepsilon 0}$ are independent beta variables. Therefore, by Kakutani's theorem (see Corollary B.9), the joint laws are absolutely continuous or singular depending on whether the product of the affinities of the marginal distributions of the $V_{\varepsilon 0}$ under prior and posterior is positive or zero.

The affinity of two beta distributions with parameters (α, β) and (α^*, β^*) is equal to

$$\frac{B((\alpha+\alpha^*)/2, (\beta+\beta^*)/2)}{\sqrt{B(\alpha,\beta)B(\alpha^*,\beta^*)}},$$

where B stands for the beta function $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. The beta distributions of the variables $V_{\varepsilon 0}$ at a given level $m = |\varepsilon| + 1$ have parameters (a_m, a_m) under prior and posterior, except for the indices ε of the partitioning cells A_ε that contain observations, where the parameters are $(a_m + N_{\varepsilon 0}, a_m + N_{\varepsilon 1})$ for the posterior. The product of the affinities for all variables $V_{\varepsilon 0}$ is therefore given by

$$\prod_{m=1}^{\infty} \prod_{\varepsilon \in \mathcal{E}^{m-1}} \frac{B(a_m + N_{\varepsilon 0}/2, a_m + N_{\varepsilon 1}/2)}{\sqrt{B(a_m, a_m)B(a_m + N_{\varepsilon 0}, a_m + N_{\varepsilon 1})}}.$$

With some effort, using Stirling's formula, it can be shown that, for fixed numbers $c, d \geq 0$, and $a \to \infty$,

$$B(a+c, a+d) = \sqrt{\frac{2\pi}{a}} 2^{-2a-c-d}\Big[1 + \frac{1}{a}\Big(c^2 + d^2 - \frac{c+d}{4} + \frac{1}{12}\Big) + O\Big(\frac{1}{a^2}\Big)\Big],$$

where the second order term is uniform in bounded c, d. Therefore, if $a_m \to \infty$, the product of affinities in the second last display can, again with some effort, be seen to be equal to

$$\prod_{m=1}^{\infty} \prod_{\varepsilon \in \mathcal{E}^{m-1}} \Big[1 - \frac{1}{a_m}\Big(\frac{N_{\varepsilon 0}^2 + N_{\varepsilon 1}^2}{4}\Big) + O\Big(\frac{1}{a_m^2}\Big)\Big].$$

Since $N_{\varepsilon 0}^2 + N_{\varepsilon 1}^2 > 0$ for at least one and at most n indices ε, for every m, the product is zero if $\sum_m a_m^{-1} = \infty$, and positive otherwise.

If a_m is bounded by a given constant for infinitely many m, then the corresponding terms of the double product are uniformly bounded by a constant strictly less than 1, and hence the product is zero. □

3.7.1 Relation with the Pólya Urn Scheme

In a *Pólya urn model* one repeatedly draws a random ball from an urn containing balls of finitely many types, replacing it with given numbers of balls of the various varieties before each following draw. In our situation there are two types of balls, marked "0" or "1," and every ball drawn from the urn is replaced by two balls of the same type (the chosen ball is replaced and an extra ball of the same type is added). There is an interesting link between Pólya tree processes and the Pólya urn scheme, which can be helpful for certain calculations.

The link employs infinitely many urns, arranged in a tree structure (see Figure 3.4). In the (binary) *Pólya tree urn model* there is a Pólya urn with balls marked "0" or "1" at every node of the (binary) tree, and balls are drawn sequentially from the urns that are situated on a random path of nodes through the tree. The path is determined, from the root down, by choosing at every node the left or right branch depending on the type of ball drawn from the urn at the node. After each draw the urn is updated, just as for an ordinary Pólya urn, thus creating a new tree of urns. Urns that are not on the path remain untouched and are copied unaltered to the new tree. The new tree is visited along a new, random path, starting

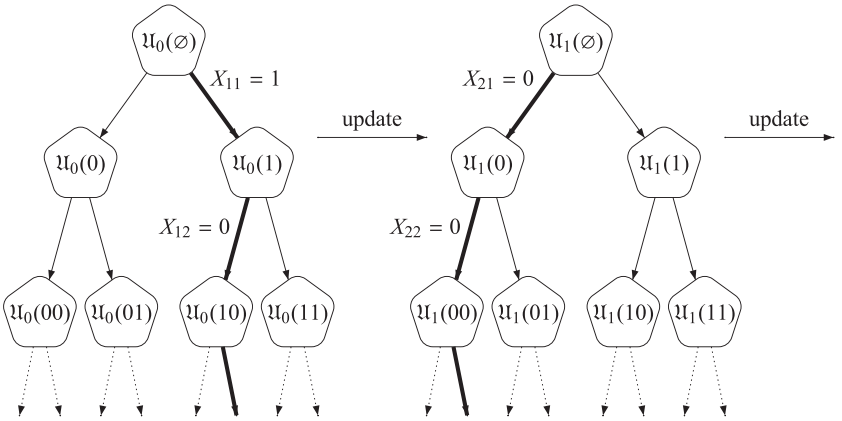

Figure 3.4 Pólya tree urn model. Shown are the urns at the first three levels of the initial tree (left) and its first update (right). Realizations of the paths chosen through the trees are indicated by bold arrows, the balls drawn from the urns on the paths are indicated on the outgoing arrows.

again at the root, and proceeding as before, albeit with different transition probabilities, as the contents of some urns have been updated.

We index the tree, as usual, by the elements of $\mathcal{E}^* \cup \varnothing$, with the empty string $\varnothing$ corresponding to the root and the strings of length m to the 2^m nodes at the mth level. The process starts with an initial tree of urns, which we denote by $\mathfrak{U}_0 = (\mathfrak{U}_0(\varepsilon)\colon \varepsilon \in \mathcal{E}^* \cup \varnothing)$, and this is sequentially replaced by urn trees $\mathfrak{U}_1, \mathfrak{U}_2, \ldots$ Every urn $\mathfrak{U}_m(\varepsilon)$ corresponds to a pair of nonnegative real numbers (r_0, r_1), which "physically" stand for the numbers of balls of colors 0 and 1. The numbers of balls in the initial urns $\mathfrak{U}_0(\varepsilon)$ are given; in the later urns they are random. Non-integer numbers (r_0, r_1) are permitted, in the understanding that a ball drawn from the urn is "0" with probability $r_0/(r_0 + r_1)$ and "1" with probability $r_1/(r_0 + r_1)$.

Every urn tree $\mathfrak{U}_i$ leads to an infinite sequence of draws, whose realizations we record as $X_i = X_{i1}, X_{i2}, X_{i3}, \ldots$, an infinite string of 0s and 1s. For instance, X_{i1} is the color of the ball drawn from urn $\mathfrak{U}_i(\varnothing)$, X_{i2} is the ball drawn from $\mathfrak{U}_i(X_{i1})$, X_{i3} is drawn from $\mathfrak{U}_i(X_{i1}X_{i2})$, etc. With this notation, the updating scheme can be described as $\mathfrak{U}_{i+1}(\varnothing) = \mathfrak{U}_i(\varnothing) \cup \{X_{i1}\}$, $\mathfrak{U}_{i+1}(X_{i1}) = \mathfrak{U}_i(X_{i1}) \cup \{X_{i2}\}$, etc.

As is well known, the consecutive draws from a Pólya urn, for instance the sequence $X_{11}, X_{21}, X_{31}, \ldots$ in our scheme, is exchangeable (see Feller 1968, page 120). It can be verified that, similarly, the sequence $X_1, X_2, \ldots$ is exchangeable, in the state space $\{0, 1\}^\infty$. Consequently $Y_1, Y_2, \ldots$, defined through the binary expansions $Y_i = \sum_{j=1}^{\infty} X_{ij} 2^{-j}$ based on X_i, are exchangeable random variables in $[0, 1]$. By de Finetti's theorem (e.g. Theorem 1.49 of Schervish 1995), there exists a unique measure Π on $\mathfrak{M} = \mathfrak{M}[0, 1]$ such that, for every Borel sets $A_1, \ldots, A_n$,

$$\mathrm{P}(Y_1 \in A_1, \ldots, Y_n \in A_n) = \int \prod_{i=1}^{n} P(A_i)\, d\Pi(P).$$

The de Finetti measure Π acts as a prior for P: it happens to be a Pólya tree process. For a precise proof of the following theorem, see Mauldin et al. (1992).

Theorem 3.23 *The de Finetti measure corresponding to the Pólya urn tree model is the Pólya tree process* $\mathrm{PT}(\lambda, \alpha_\varepsilon)$*, where* λ *is the Lebesgue measure on* $[0, 1]$ *and* $(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}) = \mathfrak{U}(\varepsilon)$ *is the configuration of the initial urn at node* $\varepsilon \in \mathcal{E}^* \cup \varnothing$.

Its connection with Pólya urn schemes justifies the name Pólya tree. Many properties of Pólya tree processes can be elegantly derived from the urn scheme formulations; see Problems 3.6 and 3.7.

3.7.2 Mixtures of Pólya Tree Processes

Although the support of a Pólya tree can be reasonably large (and it will be seen in Chapter 6 that the posterior distribution is consistent under certain conditions), practical implementation requires some meaningful elicitation of the mean measure. Often a class of target measures can be identified, like the normal family, but it is hard to single out one member from the family. Therefore it is natural to consider a mean measure that contains finitely many unspecified parameters and put a further prior on this set of hyperparameters. The resulting hierarchical prior is a mixture of Pólya tree processes.

Let the observations be a random sample $X_1, \ldots, X_n$ of real variables. Suppose that G_θ and g_θ are the distribution function and density of the elicited mean measure, and that the hyperparameter θ is given the prior density ρ. Given θ, let P follow the canonical Pólya tree process $\mathrm{PT}^*(G_\theta, a_m)$. Then the prior mean measure is $A \mapsto \mathrm{E}(P(A)) = \int G_\theta(A)\rho(\theta)\, d\theta$. Assume that $\sum_{m=1}^{\infty} a_m^{-1} < \infty$, so that the random measure P admits a density p a.s.

The canonical partitions can be constructed for all θ simultaneously as $(G_\theta^{-1}(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^*)$, for $(A_\varepsilon\colon \varepsilon \in \mathcal{E}^*)$, the canonical partition of the unit interval relative to the uniform measure. Given θ, the prior and posterior are ordinary Pólya tree processes, with the updating given in Theorem 3.21. In particular, (3.23) gives, with $N_{G_\theta(x)_1 \cdots G_\theta(x)_j}$ the number of transformed observations $G_\theta(X_i)$ falling in the same jth level partitioning set as $G_\theta(x)$,

$$\mathrm{E}(p(x)|\,\theta, X_1, \ldots, X_n) = g_\theta(x) \prod_{j=1}^{\infty} \frac{2a_j + 2N_{G_\theta(x)_1 \cdots G_\theta(x)_j}}{2a_j + N_{G_\theta(x)_1 \cdots G_\theta(x)_j}}. \tag{3.24}$$

The posterior density is obtained by integrating this relative to the posterior density of θ. By Bayes's theorem, the latter is given by

$$\rho(\theta|\, X_1, \ldots, X_n) \propto p(X_1, \ldots, X_n|\,\theta)\rho(\theta) = \int \prod_{i=1}^{n} p(X_i)\, d\Pi(P|\,\theta)\, \rho(\theta),$$

where $d\Pi(P|\,\theta)$ is the Pólya tree process $\mathrm{PT}^*(G_\theta, a_m)$. For a single observation X_1 this evaluates to $\rho(\theta)$ times the right side of (3.24) evaluated at $x = X_1$. The general formula can also be written explicitly, but depends on the pattern of joint memberships of the observations to the partitioning sets (at lower levels). The resulting expression can be evaluated numerically, and gives a method for density estimation.

This often has a desirable smoothness property. For instance, in the case that θ is a location parameter, the partitions defined by G_θ shift continuously as θ varies, resulting in a posterior expected mixture density that is continuous everywhere. The case that θ is a scale parameter is similar, except for a singularity at the point zero.

Theorem 3.24 *Let e be the function $x \mapsto \mathrm{E}(p(x)|\, X_1, \ldots, X_n)$ of pointwise posterior mean of $p(x)$, given a random sample $X_1, \ldots, X_n$ from a distribution P drawn from a canonical Pólya tree process $\mathrm{PT}^*(G_\theta, a_m)$ with $\sum_m a_m^{-1} < \infty$ with partitions as described. Let G_θ have Lebesgue density g_θ, and let g be a fixed bounded, continuous probability density with respect to Lebesgue measure on $\mathbb{R}$.*

(i) *If $g_\theta(x) = g(x - \theta)$, then e is continuous everywhere, a.s.*
(ii) *If $g_\theta(x) = \theta^{-1} g(x/\theta)$ and $\int_0^\infty \theta^{-1} \rho(\theta)\, d\theta < \infty$, then e is continuous everywhere except at zero, a.s.*

Proof The posterior mean density is the integral of (3.24) relative to the posterior distribution of θ. The latter is absolutely continuous and does not depend on the argument x. The former is an ordinary Pólya-tree mean density, and it is uniformly bounded, and continuous at every x that is not a boundary point of the sequence of partitions. To see this, note that the tail of the infinite product becomes uniformly in x close to 1, as $\sum_j a_j^{-1} < \infty$ and

$$\prod_{j>m} \Bigl(1 - \frac{n}{2a_j}\Bigr) \le \prod_{j>m} \frac{2a_j + 2N_{G_\theta(x)_1 \cdots G_\theta(x)_j}}{2a_j + N_{G_\theta(x)_1 \cdots G_\theta(x)_j}} \le \prod_{j>m} \Bigl(1 + \frac{n}{a_j}\Bigr). \tag{3.25}$$

Furthermore, for every given m, a non-boundary point x shares its mth level partitioning set with an open neighborhood surrounding it, rendering the first m terms of the product identical for every point in the neighborhood.

If D are the boundary points of the canonical uniform partition, then $G_\theta^{-1}(D)$ are the boundary points for the Pólya tree given θ. In the location case, this is the set $G^{-1}(D) + \theta$, whereas in the scale case this is $\theta G^{-1}(D)$.

It follows that in the location case the function (3.24) is continuous at x if $x \notin G^{-1}(D) + \theta$, or equivalently if $\theta \notin G^{-1}(D) - x$, which is a countable set of θ and hence a null set for the posterior distribution of θ. Because the function (3.24) is uniformly bounded in θ by $\|g\|_\infty$ times the right side of (3.25) for $m = 0$, the assertion follows from the dominated convergence theorem.

In the scale case the function (3.24) is continuous at x if $x \notin \theta G^{-1}(D)$; equivalently, $1/\theta \notin G^{-1}(D)/x$ if $x \neq 0$, which again is a countable set of θ. (If $x = 0$ and $0 \in G^{-1}(D)$, then $x \in \theta G^{-1}(D)$ for all θ.) In this case, (3.24) is bounded above by $\theta^{-1}\|g\|_\infty$ times the right side of (3.25) for $m = 0$, which is integrable relative to the posterior distribution of θ for almost every $X_1, \ldots, X_n$ as $\mathrm{E} \int \theta^{-1}\, \rho(\theta|\, X_1, \ldots, X_n)\, d\theta = \mathrm{E}\theta^{-1} < \infty$ by assumption. The assertion follows again by the dominated convergence theorem. □

An alternative way of mixing Pólya trees keeps the partitions intact, but varies the parameters α_ε with θ. A desired prior mean density g_θ can then be obtained from solving the relation

$$g_\theta(x) = \prod_{j=1}^{\infty} \frac{2\alpha_{x_1 \cdots x_j}(\theta)}{\alpha_{x_1 \cdots x_{j-1}0}(\theta) + \alpha_{x_1 \cdots x_{j-1}1}(\theta)}. \tag{3.26}$$

We shall refer to this type of mixtures as *Pólya tree mixtures of the second kind* and to the former mixtures as *Pólya tree mixtures of the first kind*. As the partitions do not vary, the density of a mixture of second kind is discontinuous at all boundary points of the partition

just like a usual Pólya tree. However, since the counts in the partitioning intervals are free of θ, the posterior expected density has a simpler expression than for a mixture of the first kind. A Pólya tree mixture of the second kind is useful in the context of testing a parametric family against a nonparametric alternative using Bayes factors.

3.7.3 Partially Specified Pólya Tree

Rather than continuing the splitting tree indefinitely, one may stop at a certain level m, and distribute the remaining probability masses uniformly over the partitioning set at the terminating level. On $[0, 1]$ with the standard binary intervals, this yields a prior on all regular histograms of width 2^{-m}. If the splitting variables are independent and possess beta distributions, then the resulting "partially specified Pólya tree" can formally be viewed as a regular Pólya tree with $\alpha_\varepsilon = \infty$ for all ε with $|\varepsilon| > m$ and $\mathrm{Be}(\infty, \infty)$ interpreted as the Dirac measure at $1/2$.

Computing the posterior distribution reduces to a multinomial problem where only the counts of the sets A_ε, for $\varepsilon \in \mathcal{E}^m$ are observed (consult Theorem 3.14 in this context). Furthermore, the density estimate given by (3.23) contains only m factors and is differentiable (with zero derivative) on the interior of all mth level partitioning sets.

One may similarly consider a mixture of these partially specified Pólya trees. Then, by a slight modification of the arguments given in the proof of Theorem 3.24, it follows that if g is bounded and continuous, then for a location mixture, the density estimate is differentiable everywhere except at the observations. For a scale mixture, if $\int \theta^{-1}\rho(\theta)\,d\theta < \infty$, the density estimate is differentiable everywhere except at zero and at the observations. The result can be extended to higher-order derivatives by appropriately strengthening the integrability condition.

Notwithstanding the differentiability of the density estimate, a partially specified Pólya tree process can support only histograms of some specified bandwidth, and hence not all densities can be consistently estimated by it as a prior. In practice, one considers a sequence of priors and works with a level fine enough for the accuracy associated with a given sample size, that is, one lets $m_n \to \infty$. Such a sequence of priors can maintain both consistency and differentiability properties.

3.7.4 Evenly Split Pólya Tree

A Pólya tree has full weak support, but rough sample paths. In contrast, a partially specified Pólya tree gives piecewise constant sample paths, but has limited support. The partial specification in a Pólya tree corresponds to half-half splitting of probabilities from a stage onwards, rather than continuing to split randomly, as in a Pólya tree. By introducing a random choice between even and random splitting, this makes it possible to design a process that combines the Pólya and partially specified Pólya trees and enjoys both large weak support and piecewise constant sample paths.

Suppose that the splitting variables are independent and distributed according to the mixture distribution, for some choice of numbers $\alpha_\varepsilon > 0$ and $\beta_\varepsilon \in [0, 1]$, for $\varepsilon \in \mathcal{E}^*$,

$$V_{\varepsilon_1\cdots\varepsilon_m 0} \sim (1 - \beta_{\varepsilon_1\cdots\varepsilon_m})\delta_{1/2} + \beta_{\varepsilon_1\cdots\varepsilon_m}\mathrm{Be}(\alpha_{\varepsilon_1\cdots\varepsilon_m 0}, \alpha_{\varepsilon_1\cdots\varepsilon_m 1}).$$

If $0 < \beta_\varepsilon < 1$ for all $\varepsilon \in \mathcal{E}^*$ and $\sum_{\varepsilon\in\mathcal{E}^*} \beta_\varepsilon < \infty$, then only finitely many splits can be random, almost surely, by the Borel-Cantelli lemma, and the other splits will be even. This ensures that the realizations of this process are almost surely piecewise constant functions with only finitely many points of discontinuity. As the nature of the splits is not predetermined, it is clear from Theorem 3.10 that the process has full weak support. In fact, Theorem 3.19 implies that every density is in the total variation support of the process. The process is not a Pólya tree, but it can be considered a mixture of partially specified Pólya trees. The mixing takes place over all splitting regimes that give even splits at all, but finitely many, occasions.

The parameters α_ε and β_ε may be chosen to depend only on the string lengths, i.e. $\alpha_\varepsilon = a_m$ and $\beta_\varepsilon = b_m$, for all $\varepsilon \in \mathcal{E}^m$. If $\sum_{m=1}^\infty 2^m b_m < \infty$, then for any choice of $a_m > 0$ the process is fully supported in both the weak and the total variation sense. In particular the parameters a_m need not grow with m. This is in sharp contrast with the usual Pólya tree, which requires $\sum_{m=1}^\infty a_m^{-1} < \infty$ already for the existence of a density.

3.8 Historical Notes

Priors for a probability distribution on a countable sample space were extensively discussed in Freedman (1963). Dubins and Freedman (1963) introduced the method of random rectangular partitions, discussed the singularity property and characterized the prior mean. Tail-free processes were studied by Freedman (1963, 1965) and Fabius (1964). The concept also appeared in Doksum (1974), who considered a very related concept, neutral to the right process, for priors on survival function. Theorem 3.16 is due to Kraft (1964). Theorem 3.14 was exploited by Freedman (1965) to show posterior consistency for tail-free process priors; its converse statement is due to Fabius (1964). Theorems 3.15 and 3.10 were obtained by Freedman (1965). Pólya trees were studied extensively by Lavine (1992, 1994), who showed conjugacy and derived formulas for density estimates, but many of the ideas actually appeared implicitly in the works of Freedman (1965) and Ferguson (1974). The connection with Pólya urn schemes was uncovered by Mauldin et al. (1992), based on earlier ideas of Blackwell and MacQueen (1973). Theorem 3.22 is due to Drăghici and Ramamoorthi (2000). Pólya tree mixtures of the first kind were studied by Hanson and Johnson (2002), who also gave a version of Theorem 3.24. Pólya tree mixtures of the second kind appeared in Berger and Guglielmi (2001) in the context of testing a parametric family against a nonparametric alternative.

Problems

3.1 Consider a stick-breaking scheme $p_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$, where the V_ks are independent with $\mathrm{P}(V_k = 0) = 1 - k^{-2}$ and $\mathrm{P}(V_k = 1 - e^{-k}) = k^{-2}$. Show that $\sum_{k=1}^\infty \mathrm{E}[\log(1 - V_k)] = -\infty$, yet $\sum_{k=1}^\infty p_k < 1$ a.s. [This contradicts the claim made in Ishwaran and James (2001) that the condition $\sum_{l=1}^\infty \mathrm{E}[\log(1 - V_l)] = -\infty$ is necessary and sufficient to obtain a probability measure.]

3.2 Suppose we want to construct a prior on the space of discrete distributions with decreasing probability mass function. Construct a random element ξ on $\mathbb{S}_\infty$ and

put $p_k = \sum_{j=k}^{\infty} \xi_j / j$. Show that every prior must be the law of a random vector $(p_1, p_2, \ldots,)$ of this form.

3.3 Construct a prior on each of the following spaces: (a) space of finite nonnegative measures on $\mathbb{N}$, (b) ℓ_1, (c) ℓ_2.

3.4 (Ghosh and Ramamoorthi 2003) Suppose that for every measurable partition $\{B_1, \ldots, B_k\}$ of a Polish space $(\mathfrak{X}, \mathscr{X})$, a probability $\Pi_{B_1,\ldots,B_k}$ has been specified with the property that if $\{A_1, \ldots, A_l\}$ is a coarser partition, i.e. $A_j = \cup_{i:B_i \subset A_j} B_i$ for all j, then the distribution of

$$\Big(\sum_{i:B_i \subset A_1} P(B_i), \ldots, \sum_{i:B_i \subset A_l} P(B_i) \Big),$$

where $(P(B_1), \ldots, P(B_k)) \sim \Pi_{B_1,\ldots,B_k}$, is $\Pi_{A_1,\ldots,A_l}$. If further $\Pi_{B_n} \rightsquigarrow \delta_0$ for any $B_n \downarrow \varnothing$ and $\Pi_{\mathfrak{X}} = \delta_1$, then there exists a unique probability measure Π on $\mathfrak{M}(\mathfrak{X})$ such that the induced distribution of $(Q(B_1), \ldots, Q(B_k))$, when $Q \sim \Pi$, is equal to $\Pi_{B_1,\ldots,B_k}$ for every measurable partition $\{B_1, \ldots, B_k\}$ of $\mathfrak{X}$.

3.5 Let P and Q be random probability measures on a Polish space $(\mathfrak{X}, \mathscr{X})$. Let $\mathscr{C}$ be a generator for the Borel σ-field on $\mathfrak{X}$. Suppose that for every finite collection $C_1, \ldots, C_n \in \mathscr{C}$,

$$(P(C_1), \ldots, P(C_n)) =_d (Q(C_1), \ldots, Q(C_n)).$$

Show that this identity then holds for every finite collection of Borel sets $C_1, \ldots, C_n$.

3.6 (Mauldin et al. 1992) Derive the posterior updating formula for a Pólya tree process prior through the urn scheme representation. Derive the marginal distribution, and the conditions for its continuity from the urn scheme representation.

3.7 (Mauldin et al. 1992) Generalize the Pólya urn scheme to k types of balls, and hence construct a Pólya tree process where the mass of any set at any level is distributed to k children of the set following a k-dimensional Dirichlet distribution. The construction is useful in defining Pólya tree processes on $\mathbb{R}^d$ with $k = 2^d$.

3.8 (Dubins and Freedman 1963) Consider the prior of Subsection 3.4.3. The mean cumulative distribution function $\tilde{\mu}$ can be characterized as the unique fixed point of the map $T_\mu : \mathfrak{M}[0, 1] \to \mathfrak{M}[0, 1]$ where

$$\begin{aligned}(T_\mu F)(x) = &\int_0^1 \int_x^1 \beta F(x/\alpha)\, \mu(d\alpha, d\beta) \\ &+ \int_0^1 \int_0^x \big[\beta + (1-\beta) F((x-\alpha)/(1-\alpha))\big]\, \mu(d\alpha, d\beta).\end{aligned}$$

Show that if $\mu = \delta_r \times \nu$, where ν has mean w, then $\tilde{\mu}$ satisfies the equation

$$F(x) = \begin{cases} wF(x/r), & 0 \le x \le r, \\ w + (1-w)F((x-r)/(1-r)), & r \le x \le 1. \end{cases}$$

In particular, when μ is the uniform distribution on the vertical line $x = 1/2$, as in case (a) of Subsection 3.4.3, then $\tilde{\mu}$ is the uniform distribution on $[0, 1]$. Show that $\tilde{\mu}(x) = (2/\pi) \sin^{-1} \sqrt{x}$ for cases (b) and (c).

3.9 Consider a canonical Pólya tree process $P \sim \mathrm{PT}^*(\lambda, cr_m)$ and observations $X_1, \ldots, X_n | P \overset{\text{iid}}{\sim} P$. Show that P under the posterior converges weakly to δ_λ as $c \to$

∞ and converges weakly to the random measure $\sum_{i=1}^{n} p_i \delta_{X_i}$, where $(p_1, \ldots, p_n) \sim \text{Dir}(n; 1, \ldots, 1)$, as $c \to 0$.

3.10 Derive conditions for the existence of higher-order derivatives of the posterior expected density everywhere for a location mixture of Pólya tree process.

3.11 Derive conditions for the existence of higher-order derivatives of the posterior expected density everywhere except at 0 for scale mixtures of Pólya tree process.