4

Dirichlet Processes

The Dirichlet process is of fundamental importance in Bayesian nonparametrics. It is a default prior on spaces of probability measures, and a building block for priors on other structures. In this chapter we define and construct the Dirichlet process by several methods, and derive many properties of interest, both simple and involved: expressions for moments, support, convergence, discreteness, self-similarity, thickness of tails, distribution of functionals, characterizations. We also study posterior conjugacy, joint and predictive distribution of observations and mixtures of Dirichlet processes.

4.1 Definition and Basic Properties

Throughout the chapter $\mathfrak{M}$ denotes the set of probability measures on a Polish space $(\mathfrak{X}, \mathscr{X})$. Unless stated otherwise, it is equipped with the weak topology and the corresponding Borel σ-field $\mathscr{M}$. The phrase *random measure on* $(\mathfrak{X}, \mathscr{X})$ means a probability measure on $(\mathfrak{M}, \mathscr{M})$.

Definition 4.1 (Dirichlet process) A random measure P on $(\mathfrak{X}, \mathscr{X})$ is said to possess a *Dirichlet process* distribution $\mathrm{DP}(\alpha)$ with *base measure* α, if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)). \tag{4.1}$$

In this definition α is a given finite positive Borel measure on $(\mathfrak{X}, \mathscr{X})$. We write $|\alpha| = \alpha(\mathfrak{X})$ its total mass (called *prior precision*), and $\bar{\alpha} = \alpha/|\alpha|$ the probability measure obtained by normalizing α (called *center measure*), and use the notation $P \sim \mathrm{DP}(\alpha)$ to say that P has a Dirichlet process distribution with base measure α; and also $P \sim \mathrm{DP}(MG)$ in the case that $\mathfrak{X} = \mathbb{R}^k$, $M = |\alpha|$ and $\bar{\alpha} = \alpha/M$ has cumulative distribution function G[1].

Existence of the Dirichlet process is not obvious, but a proof is deferred to Section 4.2. The finite-dimensional Dirichlet distribution in the right side of (4.1) is reviewed in Appendix G.

It is immediate from the definition that for any measurable function $\psi: \mathfrak{X} \to \mathfrak{Y}$ between standard Borel spaces, the random measure $P \circ \psi^{-1}$ on $\mathfrak{Y}$ obtained from a Dirichlet process distribution $P \sim \mathrm{DP}(\alpha)$ possesses the $\mathrm{DP}(\alpha \circ \psi^{-1})$-distribution.

[1] The definition could be extended to *finitely additive Dirichlet processes P on a subfield $\mathscr{X}_0 \subset \mathscr{X}$ by requiring that* (4.1) *hold for all partitions whose elements belong to $\mathscr{X}_0$. Then it suffices that α is finitely additive.*

By applying (4.1) to the partition $\{A, A^c\}$ for a Borel set A, the vector $(P(A), P(A^c))$ possesses the $\mathrm{Dir}(2; \alpha(A), \alpha(A^c))$-distribution; equivalently, the variable $P(A)$ is $\mathrm{Be}(\alpha(A), \alpha(A^c))$-distributed.

In particular $P(A) > 0$ almost surely for every measurable set A with $\alpha(A) > 0$. Because the exceptional null set depends on A, this does not imply that P and α are mutually absolutely continuous. In fact, in Theorem 4.14 they are shown to be almost surely singular unless α is atomic.

4.1.1 Expectations, Variances and Co-Variances

The center measure is the mean measure of the Dirichlet prior.

Proposition 4.2 (Moments) *If $P \sim \mathrm{DP}(\alpha)$, then, for any measurable sets A and B,*

$$\mathrm{E}(P(A)) = \bar{\alpha}(A), \tag{4.2}$$

$$\mathrm{var}\,(P(A)) = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1 + |\alpha|}, \tag{4.3}$$

$$\mathrm{cov}\,(P(A), P(B)) = \frac{\bar{\alpha}(A \cap B) - \bar{\alpha}(A)\bar{\alpha}(B)}{1 + |\alpha|}, \tag{4.4}$$

Proof The first two results are immediate from the properties of the beta distribution of $P(A)$. To prove the third relation, first assume that $A \cap B = \varnothing$. Then $\{A, B, A^c \cap B^c\}$ is a partition of $\mathfrak{X}$, whence $(P(A), P(B), P(A^c \cap B^c)) \sim \mathrm{Dir}(3; \alpha(A), \alpha(B), \alpha(A^c \cap B^c))$. By the expressions for the moments of finite Dirichlet distribution given by Corollary G.4,

$$\mathrm{cov}\,(P(A), P(B)) = -\frac{\alpha(A)\alpha(B)}{|\alpha|^2(1 + |\alpha|)},$$

which is the same as (4.4) in the special case $A \cap B = \varnothing$. For the proof of the general case we decompose $A = (A \cap B) \cup (A \cap B^c)$ and $B = (A \cap B) \cup (A^c \cap B)$, and write the left-hand side of (4.4) as

$$\mathrm{cov}\,(P(A \cap B) + P(A \cap B^c), P(A \cap B) + P(A^c \cap B)).$$

We write this as a sum of four terms and use both (4.3) and the special case of (4.4) twice to reduce this to the right side of (4.4). □

The first assertion of the proposition can also be rewritten as $\int P(A)\, d\mathrm{DP}_\alpha(P) = \bar{\alpha}(A)$. Interpreting P in this equation as the law of an observation X from P, we see

$$\text{if } P \sim \mathrm{DP}(\alpha) \text{ and } X|\, P \sim P, \text{ then } X \sim \bar{\alpha}. \tag{4.5}$$

Thus $\bar{\alpha}$ is the marginal law of "an observation from the Dirichlet process," as well as the prior mean. It is also called the *center measure* of the Dirichlet process.

By (4.4) the total mass $|\alpha|$ controls the variability of any $P(A)$ around its prior mean, and hence is appropriately called the *precision parameter*. For instance, if one expects the Nor(0, 1) model to hold, but is not absolutely confident about it, one may propose a Dirichlet process prior with center measure Nor(0, 1) and precision $|\alpha|$ reflecting the degree of confidence in the prior guess. A more precise assessment of $|\alpha|$ can be based on the posterior

distribution, which we obtain later on. However, we shall see later that the "precision parameter" also has another role and the rule "small $|\alpha|$ means less prior information" should be interpreted cautiously.

Proposition 4.2 implies that different choices of α generally lead to different Dirichlet processes: if $\alpha_1 \neq \alpha_2$, then $\mathrm{DP}(\alpha_1) \neq \mathrm{DP}(\alpha_2)$, unless $\bar{\alpha}_1 = \bar{\alpha}_2 = \delta_x$ for some x. This is obvious if $\bar{\alpha}_1 \neq \bar{\alpha}_2$, since the two random measures then have different means by (4.2). In case $\bar{\alpha}_1 = \bar{\alpha}_2$, but $|\alpha_1| \neq |\alpha_2|$, the variances of $P(A)$ under the two priors are different by (4.3), for every A with $0 < \bar{\alpha}_1(A) = \bar{\alpha}_2(A) < 1$; provided the base measures are nondegenerate, there is at least one such A.

Proposition 4.2 can be generalized to integrals of functions with respect to P.

Proposition 4.3 (Moments) *If $P \sim \mathrm{DP}(\alpha)$, then for any measurable functions ψ and ϕ for which the expression on the right-hand side is meaningful,*

$$\mathrm{E}(P\psi) = \int \psi \, d\bar{\alpha}, \tag{4.6}$$

$$\mathrm{var}(P\psi) = \frac{\int (\psi - \int \psi \, d\bar{\alpha})^2 \, d\bar{\alpha}}{1 + |\alpha|}, \tag{4.7}$$

$$\mathrm{cov}(P\psi, P\phi) = \frac{\int \psi\phi \, d\bar{\alpha} - \int \psi \, d\bar{\alpha} \int \phi \, d\bar{\alpha}}{1 + |\alpha|}. \tag{4.8}$$

Proof For indicator functions $\psi = \mathbb{1}_A$ and $\phi = \mathbb{1}_B$ the assertions reduce to those of Proposition 4.2. They extend by linearity (for the first) or bilinearity (for the second and third) to simple functions ψ and ϕ, and next by splitting in positive and negative parts and monotone convergence (for the first) or continuity in the $\mathbb{L}_2$-sense (for the second and third) to general measurable functions. □

Remark 4.4 In particular $\int |\psi| \, d\alpha < \infty$ implies that $P|\psi| < \infty$ a.s. $[\mathrm{DP}(\alpha)]$. We shall see in Sections 4.3.5 and 4.3.7 that the converse is false.

4.1.2 Self-Similarity

The Dirichlet process is tail-free in the sense of Definition 3.11, for any sequence of successive partitions of the sample space. In fact, it possesses much stronger conditional independence properties.

For a measure P and measurable set B, let $P|_B$ stand for the restriction measure $P|_B(A) = P(A \cap B)$, and P_B for the conditional measure $P_B(A) = P(A|\,B)$, for B with $P(B) > 0$.

Theorem 4.5 (Self-similarity) *If $P \sim \mathrm{DP}(\alpha)$, then $P_B \sim \mathrm{DP}(\alpha|_B)$, and the variable and processes $P(B)$, $(P_B(A)\colon A \in \mathscr{X})$ and $(P_{B^c}(A)\colon A \in \mathscr{X})$ are mutually independent, for any $B \in \mathscr{X}$ such that $\alpha(B) > 0$.*

Proof Because $P(B) \sim \mathrm{Be}(\alpha(B), \alpha(B^c))$, the condition that $\alpha(B) > 0$ implies that $P(B) > 0$ a.s., so that the conditional probabilities given B are well defined.

For given partitions $A_1, \ldots, A_r$ of B and $C_1, \ldots, C_s$ of B^c, the vector

$$X := (P(A_1), \ldots, P(A_r), P(C_1), \ldots, P(C_s))$$

possesses a Dirichlet distribution $\mathrm{Dir}(r+s; \alpha(A_1), \ldots, \alpha(A_r), \alpha(C_1), \ldots, \alpha(C_s))$. By Proposition G.3 the four variables or vectors

$$Z_1 := \sum_{i=1}^{r} X_i, \qquad Z_2 := \sum_{i=r+1}^{r+s} X_i, \qquad \Big(\frac{X_1}{Z_1}, \ldots, \frac{X_r}{Z_1}\Big), \qquad \Big(\frac{X_{r+1}}{Z_2}, \ldots, \frac{X_{r+s}}{Z_2}\Big)$$

are mutually independent, and the two vectors have Dirichlet distributions with the restrictions of the original parameters. These are precisely the variables $P(B)$, $P(B^c)$ and vectors with coordinates $P_B(A_i)$ and $P_{B^c}(C_i)$. □

Theorem 4.5 shows that the Dirichlet process "localized" by conditioning to a set B is again a Dirichlet process, with base measure the restriction of the original base measure. Furthermore, processes at disjoint localities are independent of each other, and also independent of the "macro level" variable $P(B)$. Within any given locality, mass is further distributed according to a Dirichlet process, independent of what happens to the "outside world." This property may be expressed by saying that locally a Dirichlet process is like itself; in other words it is *self similar*.

Theorem 4.5 also shows that the Dirichlet process is a special Pólya tree process: for any sequence of successively refined binary partitions $\{A_0, A_1\}, \{A_{00}, A_{01}, A_{10}, A_{11}\}, \ldots$, the splitting variables $V_{\varepsilon 0} = P(A_{\varepsilon 0} | A_\varepsilon)$ satisfy

$$V_{\varepsilon 0} \sim \mathrm{Be}(\alpha(A_{\varepsilon 0}), \alpha(A_{\varepsilon 1})), \tag{4.9}$$

and all variables $V_{\varepsilon 0}$, for $\varepsilon \in \mathcal{E}^*$, are mutually independent. A special property of the Dirichlet process is that it is a Pólya tree for *any* sequence of partitions.

Another special property is that the parameters of the beta-distributions are additive in the sense that $\alpha(A_{\varepsilon 0}) + \alpha(A_{\varepsilon 1}) = \alpha(A_\varepsilon)$, for every $\varepsilon \in \mathcal{E}^*$. Generally in a Pólya tree, the probabilities $P(A_\varepsilon)$ are products of independent beta variables (cf. (3.12)). In a Dirichlet process these probabilities are themselves beta-distributed. These properties are closely related (see Problem G.4). The additivity of the parameters is actually a distinguishing property of the Dirichlet process within the class of Pólya trees, as shown in Section 4.4.

4.1.3 Conjugacy

One of the most remarkable properties of the Dirichlet process prior is that the posterior distribution is again a Dirichlet process. Consider observations $X_1, X_2, \ldots, X_n$ sampled independently from a distribution P that was drawn from a Dirichlet prior distribution. By an abuse of language, which we shall follow, such observations are often termed a *sample from the Dirichlet process*.

Theorem 4.6 (Conjugacy) *The* $\mathrm{DP}(\alpha + \sum_{i=1}^{n} \delta_{X_i})$*-process is a version of the posterior distribution given an i.i.d. sample* $X_1, \ldots, X_n$ *from the* $\mathrm{DP}(\alpha)$*-process.*

Proof Because the Dirichlet process is tail-free by Theorem 4.5, and a given measurable partition $\{A_1, \ldots, A_k\}$ of $\mathfrak{X}$ can be viewed as part of a sequence of successive binary partitions, the posterior distribution of the vector $(P(A_1), \ldots, P(A_k))$ given $X_1, \ldots, X_n$ is the same as the posterior distribution of this vector given the vector $N = (N_1, \ldots, N_k)$ of cell counts, defined by $N_j = \#(1 \leq i \leq n : X_i \in A_j)$. Given P the vector N possesses a multinomial distribution with parameter $(P(A_1), \ldots, P(A_k))$, which has a $\mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k))$ prior distribution. The posterior distribution can be obtained using Bayes's rule applied to these finite-dimensional vectors, as in Proposition G.8. □

Theorem 4.6 can be remembered as the updating rule $\alpha \mapsto \alpha + \sum_{i=1}^n \delta_{X_i}$ for the base measure of the Dirichlet distribution. In terms of the parameterization $\alpha \leftrightarrow (M = |\alpha|, \bar{\alpha})$ of the base measure, this rule takes the form

$$M \mapsto M + n \quad \text{and} \quad \bar{\alpha} \mapsto \frac{M}{M+n}\bar{\alpha} + \frac{n}{M+n}\mathbb{P}_n, \tag{4.10}$$

where $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical distribution of $X_1, \ldots, X_n$. Combining the theorem with Proposition 4.2, we see that

$$\mathrm{E}(P(A) | X_1, \ldots, X_n) = \frac{|\alpha|}{|\alpha| + n}\bar{\alpha}(A) + \frac{n}{|\alpha| + n}\mathbb{P}_n(A). \tag{4.11}$$

Thus the posterior mean (the "Bayes estimator" of P) is a convex combination of the prior mean $\bar{\alpha}$ and the empirical distribution, with weights $M/(M+n)$ and $n/(M+n)$, respectively. For a given sample it is close to the prior mean if M is large, and close to the empirical distribution (which is based only on the data) if M is small relative to n. Thus M determines the extent to which the prior controls the posterior mean — a Dirichlet process prior with precision M contributes information equivalent to a sample of size M (although M is not restricted to integer values). This invites to view M as the *prior sample size*, or the "number of pre-experiment samples." In this interpretation the sum $M + n$ is the *posterior sample size*.

For a fixed prior (i.e. fixed M), the posterior mean (4.11) behaves asymptotically as $n \to \infty$ like the empirical distribution $\mathbb{P}_n$ to the order $O(n^{-1})$, a.s. Thus it possesses the same asymptotic properties regarding consistency (Glivenko-Cantelli theorem) and asymptotic normality (Donsker theorem) as the empirical distribution (see Appendix F). In particular, if $X_1, X_2, \ldots$ are sampled from a "true distribution" P_0, then the posterior mean will tend a.s. to P_0.

In addition the full posterior distribution will contract to its mean, whenever the posterior sample size tends to infinity. Indeed, by combining Theorem 4.6 and (4.3), we see, for $\tilde{\mathbb{P}}_n$ the posterior mean (4.11),

$$\mathrm{var}\,(P(A) | X_1, \ldots, X_n) = \frac{\tilde{\mathbb{P}}_n(A)\tilde{\mathbb{P}}_n(A^c)}{1 + M + n} \leq \frac{1}{4(1 + M + n)}. \tag{4.12}$$

In Sections 4.3.2 and 12.2 we shall strengthen this "setwise contraction" to contraction of the posterior distribution as a whole.

Consequently, if the data are sampled from a true distribution P_0, then the posterior distribution of P converges weakly to the measure degenerate at P_0. It is remarkable that this is true for *any* base measure α: the prior is not even required to "support the true value" P_0,

in the sense of assigning positive probabilities to (weak) neighborhoods of P_0. (It should be noted that "the" posterior distribution is uniquely defined only up to the marginal distribution of the observations. The remarkable consistency is true for the particular version given in Theorem 4.6, not for *every* version; cf. Example 6.5.)

4.1.4 Marginal and Conditional Distributions

The joint distribution of a sample $X_1, X_2, \ldots$ from a Dirichlet process (i.e. $P \sim \mathrm{DP}(\alpha)$ and $X_1, X_2, \ldots \mid P \overset{\text{iid}}{\sim} P$) has a complicated structure, but can be conveniently described by its sequence of *predictive distributions*: the laws of X_1, $X_2 \mid X_1$, $X_3 \mid X_1, X_2$, etc.

Because $\mathrm{P}(X_1 \in A) = \mathrm{EP}(X_1 \in A \mid P) = \mathrm{E}P(A) = \bar{\alpha}(A)$, the marginal distribution of X_1 is $\bar{\alpha}$.

Because $X_2 \mid (P, X_1) \sim P$ and $P \mid X_1 \sim \mathrm{DP}(\alpha + \delta_{X_1})$, we can apply the same reasoning again, but now conditionally given X_1, to see that $X_2 \mid X_1$ follows the normalization of $\alpha + \delta_{X_1}$. This is a mixture of $\bar{\alpha}$ and δ_{X_1} with weights $|\alpha|/(|\alpha|+1)$ and $1/(|\alpha|+1)$, respectively.

Repeating this argument, using that $P \mid X_1, \ldots, X_{i-1} \sim \mathrm{DP}(\alpha + \sum_{j=1}^{i-1} \delta_{X_j})$, we find that

$$X_i \mid X_1, \ldots, X_{i-1} \sim \begin{cases} \delta_{X_1}, & \text{with probability } 1/(|\alpha| + i - 1), \\ \vdots & \vdots \\ \delta_{X_{i-1}}, & \text{with probability } 1/(|\alpha| + i - 1), \\ \bar{\alpha}, & \text{with probability } |\alpha|/(|\alpha| + i - 1). \end{cases} \tag{4.13}$$

Being a mixture of a product of identical distributions, the joint distribution of $X_1, X_2, \ldots$ is exchangeable, so re-labeling does not affect the structure of (4.13).

The recipe (4.13) is called the *generalized Pólya urn scheme*, and can be viewed as a continuous analog of the familiar Pólya urn scheme. Consider balls which can carry a continuum $\mathfrak{X}$ of "colors." Initially the "number of balls" is $M = |\alpha|$, which may be any positive number, and the colors are distributed according to $\bar{\alpha}$. We draw a ball from the collection, observe its color X_1, and return it to the urn along with an additional ball of the same color. The total number of balls is now $M + 1$, and the colors are distributed according to $(M\bar{\alpha} + \delta_{X_1})/(M+1)$. We draw a ball from this updated urn, observe its color X_2, and return it to the urn along with an additional ball of the same color. The probability of picking up the ball that was added after the first draw is $1/(M+1)$, in which case $X_2 = X_1$; otherwise, with probability $M/(M+1)$, we make a fresh draw from the original urn. This process continues indefinitely, leading to the conditional distributions in (4.13).

Clearly the scheme (4.13) will produce ties in the values $X_1, \ldots, X_n$ with positive probability. The joint distribution of $(X_1, \ldots, X_n)$ is a mixture of distributions with components described by the collection of partitions $\{S_1, \ldots, S_k\}$ of the set of indices $\{1, 2, \ldots, n\}$. The coordinates X_i with index i belonging to the same partitioning set S_j are identical, with their common value distributed as $\bar{\alpha}$, independent of the other common values. For instance, if $n = 3$, then the partition $\{\{1, 2\}, \{3\}\}$ corresponds to a component that is supported on the "diagonal" $D_{12;3} = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 = x_2\}$, which can be described as the distribution of a vector (Y_1, Y_2, Y_3) with $Y_1 = Y_2 \sim \bar{\alpha}$ independent of $Y_3 \sim \bar{\alpha}$.

The induced random partition of $\{1, 2, \ldots, n\}$ is of interest by itself, and known as the *Chinese restaurant process*. It is discussed in the next section and in Chapter 14.

The following result gives an expression for the mean of certain product functions. Let $\mathscr{S}_n$ stand for the collection of all partitions of $\{1, \dots, n\}$, and write a partition as $\mathcal{S} = \{S_1, \dots, S_N\}$, for $N = \#\mathcal{S}$ the number of partitioning sets; and let $s^{[n]} = s(s+1)\cdots(s+n-1)$ stand for the *ascending factorial*.

Proposition 4.7 *If $X_1, \dots, X_n$ are a random sample from a Dirichlet process with base measure α, or equivalently arise from the generalized Pólya urn scheme* (4.13)*, then for any measurable functions $g_i \colon \mathscr{X} \to \mathbb{R}$ such that the right side is defined,*

$$\mathrm{E}\Big(\prod_{i=1}^{n} g_i(X_i)\Big) = \frac{1}{|\alpha|^{[n]}} \sum_{\mathcal{S}\in\mathscr{S}_n} \prod_{j=1}^{\#\mathcal{S}} \Big[(\#S_j - 1)! \int \prod_{l\in S_j} g_l(x)\, d\alpha(x)\Big]. \tag{4.14}$$

Proof By factorizing the joint distribution of $(X_1, \dots, X_n)$ as the product of its predictive distributions, we can write the left side as

$$\int \cdots \int \prod_{i=1}^{n} g_i(x_i) \prod_{j=1}^{n} \frac{d(\alpha + \delta_{x_{j-1}} + \cdots + \delta_{x_1})(x_j)}{|\alpha| + j - 1}.$$

We next expand the second product, thus obtaining a sum of $n!$ terms, corresponding to choosing α, or one of $\delta_{x_{j-1}}, \dots \delta_{x_1}$ in the jth term, for $j = 1, \dots, n$. Every choice of a Dirac measure sets two coordinates equal, while every choice of α "opens" a partitioning set S_j. Thus the $n!$ terms of the sum can be grouped according to the partitions of $\{1, 2, \dots, n\}$. That every partition appears as many times as claimed in the proposition is perhaps best proved by induction.

For $n = 1$ the proposition is true. By (4.13) the left side of the proposition for $n+1$ can be written as

$$\mathrm{E}\Big[\prod_{i=1}^{n} g_i(X_i) \int g_{n+1}(x)\, \frac{d(\alpha + \sum_{j=1}^{n} \delta_{X_j})(x)}{|\alpha| + n}\Big]$$
$$= \frac{1}{|\alpha| + n}\Big[\mathrm{E}\Big(\prod_{i=1}^{n} g_i(X_i)\Big) \int g_{n+1}\, d\alpha + \sum_{j=1}^{n} \mathrm{E}\Big(\prod_{i\neq j} g_i(X_i)\,(g_j g_{n+1})(X_j)\Big)\Big].$$

By the proposition for n we can replace both expectations by a sum over $\mathscr{S}_n$. The resulting expression can be rewritten as a sum over $\mathscr{S}_{n+1}$, where the first term delivers all partitions of $\{1, 2, \dots, n+1\}$ with the one-point set $\{n+1\}$ as one of the partitioning sets and the second term the remaining partitions, where $n+1$ is joined to the partitioning set of the partition of $\{1, \dots, n\}$ to which j belongs. □

4.1.5 Number of Distinct Values

In this section we investigate the number of distinct values in a random sample $X_1, \dots, X_n$ from the Dirichlet process, and other aspects of the induced partitioning of the set $\{1, \dots, n\}$. For simplicity we assume throughout that the base measure α is atomless, so that with probability one the ith value X_i in the Pólya scheme (4.13) is distinct from the previous $X_1, \dots, X_{i-1}$ precisely if it is drawn from $\bar{\alpha}$.

For $i \in \mathbb{N}$ define $D_i = 1$ if the ith observation X_i is a "new value," i.e. if $X_i \notin \{X_1, \ldots, X_{i-1}\}$, and set $D_i = 0$ otherwise. Then $K_n = \sum_{i=1}^n D_i$ is the number of distinct values among the first n observations. Let $\mathcal{L}(K_n)$ stands for its law (distribution).

Proposition 4.8 (Distinct values) *If the base measure α is atomless with total mass $|\alpha| = M$, then the variables $D_1, D_2, \ldots$ are independent Bernoulli variables with success probabilities $\mathrm{P}(D_i = 1) = M/(M + i - 1)$. Consequently, for fixed M, as $n \to \infty$,*

(i) $\mathrm{E}(K_n) \sim M \log n \sim \mathrm{var}(K_n)$,
(ii) $K_n / \log n \to M$, a.s.,
(iii) $(K_n - \mathrm{E}K_n)/\, sd(K_n) \rightsquigarrow \mathrm{Nor}(0, 1)$,
(iv) $d_{TV}(\mathcal{L}(K_n), \mathrm{Poi}(\mathrm{E}(K_n))) = O(1/\log n)$.

Proof The first assertion follows because given $X_1, \ldots, X_{i-1}$ the variable X_i is "new" if and only if it is drawn from $\bar{\alpha}$, which happens with probability $M/(M + i - 1)$. Then assertion (i) can be derived from the exact formulas

$$\mathrm{E}(K_n) = \sum_{i=1}^n \frac{M}{M + i - 1}, \qquad \mathrm{var}(K_n) = \sum_{i=1}^n \frac{M(i-1)}{(M + i - 1)^2}.$$

Furthermore, assertion (ii) follows from Kolmogorov's strong law of large numbers for independent variables, since

$$\sum_{i=1}^\infty \frac{\mathrm{var}(D_i)}{(\log i)^2} = \sum_{i=1}^\infty \frac{M(i-1)}{(M + i - 1)^2 (\log i)^2} < \infty.$$

Next (iii) is a consequence of the Lindeberg central limit theorem. Finally for (iv) we apply the Chen-Stein approximation (see Chen 1975, Arratia et al. 1990, Chen et al. 2011) to the distribution of a sum of independent Bernoulli variables to see that

$$d_{TV}(\mathcal{L}(K_n), \mathrm{Poi}(\mathrm{E}(K_n))) \le 2 \sum_{i=1}^n (\mathrm{E}(D_i))^2 \frac{1 - e^{-\mathrm{E}(K_n)}}{\mathrm{E}(K_n)}. \tag{4.15}$$

Here $\sum_{i=1}^n (\mathrm{E}(D_i))^2 = M^2 \sum_{i=1}^n (M + i - 1)^{-2}$ is bounded in n, whereas $\mathrm{E}(K_n)$ tends to infinity at order $\log n$, by (i). □

Thus the number of distinct values in a (large) sample from a distribution taken from a fixed Dirichlet prior is logarithmic in the sample size. Furthermore, the fluctuations of this number around its mean are of the order $\sqrt{\log n}$, and we might approximate its distribution by a Poisson.

The distribution of the number of distinct values depends on the base measure, but only through its prior strength $M = |\alpha|$. From the formula for the exact mean and variance, given in the preceding proof, we can derive that, for all M and n,

$$1 \vee M \log\Big(1 + \frac{n}{M}\Big) \le \mathrm{E}(K_n) \le 1 + M \log\Big(1 + \frac{n}{M}\Big),$$
$$\mathrm{var}(K_n) \le M \log\Big(1 + \frac{n}{M}\Big).$$

Thus small M allows few distinct observations. As $M \to 0$, the mean number tends to one and the variance to zero, and there can be only one distinct observation. (See also Theorem 4.16 in this context.) That the posterior distribution of P is a $(M, n)/(M+n)$-mixture of the base measure and the empirical distribution motivated us earlier to think of M as the *prior strength*. However, that all observations must be identical is a strong prior opinion, and hence this interpretation may be inappropriate as $M \to 0$. Especially in density estimation and clustering using Dirichlet mixtures, the limit as $M \to 0$ should not be seen as rendering a "noninformative prior."

The following proposition gives the exact distribution of K_n. The sum in the formula for $C_n(k)$ involves $\binom{n-1}{k-1}$ terms, and hence computation could be demanding for large n.

Proposition 4.9 (Distinct values) *If the base measure α is atomless with total mass M, then*

$$\mathrm{P}(K_n = k) = C_n(k)\, n!\, M^k \frac{\Gamma(M)}{\Gamma(M+n)}, \tag{4.16}$$

where, for $k = 1, 2, \ldots,$

$$C_n(k) = \frac{1}{n!} \sum_{S \subset \{1,\ldots,n-1\}:|S|=n-k} \prod_{j \in S} j. \tag{4.17}$$

Proof The event $K_n = k$ can happen in two ways: $K_{n-1} = k-1$ and the nth observation is new, or $K_{n-1} = k$ and the nth observation is a previous observation. This gives the following recursion relation for $p_n(k, M) := \mathrm{P}(K_n = k \,|\, M)$, for $1 \le k \le n$:

$$p_n(k, M) = \frac{M}{M+n-1} p_{n-1}(k-1, M) + \frac{n-1}{M+n-1} p_{n-1}(k, M). \tag{4.18}$$

For $C_n(k) := p_n(k, 1)$ this recurrence relation (with $M = 1$) implies

$$C_n(k) = \frac{1}{n} C_{n-1}(k-1) + \frac{n-1}{n} C_{n-1}(k). \tag{4.19}$$

We shall first verify (4.16) with $C_n(k) = p_n(k, 1)$, and next derive (4.17).

We prove (4.16) (for every k) by induction on n. For $n = 1$ this equation reduces to $p_1(1, M) = C_1(1)$ for $k = 1$, which is true by definition, and to $0 = 0$ for other k. Assuming the result for $n-1$, we can apply (4.18) to express $p_n(k, M)$ into $p_{n-1}(k-1, M)$ and $p_{n-1}(k, M)$, and next (4.16) for $n-1$ to express the latter two in $C_{n-1}(k-1)$ and $C_{n-1}(k)$. With some algebra and (4.19) this can next be reduced to the right side of (4.16) for n.

In terms of the generating function $A_n(s) = \sum_{k=1}^{\infty} C_n(k) s^k$ of the sequence $C_n(k)$ (which is actually a polynomial of finite degree), the recurrence (4.19) can be written $A_n(s) = n^{-1}(s A_{n-1}(s) + (n-1) A_{n-1}(s)) = ((s+n-1)/n)\, A_{n-1}(s)$. Repeatedly applying this and using the fact that $A_1(s) = s$, we see that

$$A_n(s) = A_1(s) \prod_{j=2}^{n} \frac{s+j-1}{j} = \frac{1}{n!} \prod_{j=0}^{n-1} (s+j).$$

Expanding the product and collecting the coefficients of s^k, we obtain (4.17). □

We say that $X_1, \ldots, X_n$ possesses *multiplicity class* $C(m_1, \ldots, m_n)$ if in the set $\{X_1, \ldots, X_n\}$, m_1 values appear exactly once, m_2 (different) values appear twice, and so on. Here $m_1, \ldots, m_n$ are nonnegative integers with $m_1 + 2m_2 + \cdots + nm_n = n$ (and many entries of $m_1, \ldots, m_n$ will be zero). Let $s^{[n]} = s(s+1)\cdots(s+n-1)$ stand for the ascending factorial. *Ewens's sampling formula* gives the probability that a sample from the Dirichlet process has a given multiplicity class.

Proposition 4.10 (Ewens's sampling formula) *A random sample $X_1, \ldots, X_n$ from a Dirichlet process with nonatomic base measure of strength $|\alpha| = M$ possesses multiplicity class $C(m_1, \ldots, m_n)$ with probability equal to*

$$\frac{n!}{M^{[n]}} \prod_{i=1}^{n} \frac{M^{m_i}}{i^{m_i} m_i!}.$$

Proof Consider the event that $C(m_1, \ldots, m_n)$ occurs with the m_1 values of multiplicity 1 appearing first (so $X_1, \ldots, X_{m_1}$ are distinct and occur once in the full sample), with the m_2 pairs of multiplicity 2 appearing next (so $X_{m_1+1} = X_{m_1+2} \neq \cdots \neq X_{m_1+2m_2-1} = X_{m_1+2m_2}$), followed by the m_3 triples of multiplicity 3 etc. From the Pólya scheme (4.13) it is straightforward to see that the probability of this event is given by

$$\frac{\prod_{i=1}^{n} ((i-1)!)^{m_i} M^{\sum_{i=1}^{n} m_i}}{M^{[n]}}.$$

Because the vector $(X_1, \ldots, X_n)$ is exchangeable, the probability that it possesses multiplicity class $C(m_1, \ldots, m_n)$ is the probability in the display multiplied by the number of configurations of distinct and equal values that give this multiplicity class. Here, "configuration" refers to the pattern of equal and distinct values, without taking the values into account.

We can represent the m_1 values of multiplicity 1 by the symbols $1^1, 1^2, \ldots, 1^{m_1}$, the m_2 pairs of multiplicity 2 by $2^1, 2^1, 2^2, 2^2, \ldots, 2^{m_2}, 2^{m_2}$, the m_3 triplets of multiplicity 3 by $3^1, 3^1, 3^1, 3^2, \ldots, 3^{m_3}, 3^{m_3}, 3^{m_3}$, etc. The configurations can then be identified with the permutations of these n symbols, where permutations that are identical after swapping superscripts within a multiplicity level are considered equivalent. There are $n!$ possible orderings if the n symbols are all considered different, but for every i the m_i groups of multiplicity i (given by different superscripts) can be rearranged in $m_i!$ orders, and every group of multiplicity i (of identical symbols, e.g. the group $3^1, 3^1, 3^1$) can be internally permuted in $i!$ ways. Consequently the total number of configurations is

$$\frac{n!}{\prod_{i=1}^{n} (m_i! (i!)^{m_i})}.$$

Multiplying the two preceding displays and using the relation $(i-1)!/i! = 1/i$, we obtain the result. □

A random sample $X_1, \ldots, X_n$ from the Dirichlet process also induces a nontrivial random partition of $\{1, 2, \ldots, n\}$, corresponding to the pattern of ties. (More precisely, we consider the equivalence classes under the relation $i \equiv j$ if and only if $X_i = X_j$.) In the following

proposition we obtain its distribution. In Chapter 14 we return to this topic and discuss the distribution of partitions induced by more general processes.

Proposition 4.11 (Dirichlet partition) *A random sample $X_1, \ldots, X_n$ from a Dirichlet process with atomless base measure of strength $|\alpha| = M$ induces a given partition of $\{1, 2, \ldots, n\}$ into k sets of sizes $n_1, \ldots, n_k$ with probability equal to*

$$\frac{M^k \Gamma(M) \prod_{j=1}^k \Gamma(n_j)}{\Gamma(M+n)}. \tag{4.20}$$

Proof By exchangeability the probability depends on the sizes of the partitioning sets only. The probability that the partitioning set of size n_1 consists of the first n_1 variables, the one of size n_2 of the next n_2 variables, etc. can be obtained by multiplying the appropriate conditional probabilities for the consecutive draws in the Pólya urn scheme in their natural order of occurrence. For $\bar{n}_j = \sum_{l=1}^j n_l$, it is given by

$$\frac{M}{M}\frac{1}{M+1}\cdots\frac{n_1-1}{M+n_1-1}\times\cdots\times\frac{M}{M+\bar{n}_{k-1}}\frac{1}{M+\bar{n}_{k-1}+1}\cdots\frac{n_k-1}{M+\bar{n}_{k-1}+n_k-1}.$$

This can be rewritten as in the proposition. □

4.2 Constructions

We describe several methods for constructing the Dirichlet process. The first construction, via a stochastic process, is sufficient to prove existence, but the other methods give additional characterizations.

4.2.1 Construction via a Stochastic Process

Definition 4.1 specifies the joint distribution of the vector $(P(A_1), \ldots, P(A_k))$, for any measurable partition $\{A_1, \ldots, A_k\}$ of the sample space. In particular, it specifies the distribution of $P(A)$, for every measurable set A, and hence the *mean measure* $A \mapsto \mathrm{E}(P(A))$. By Proposition 4.2, this is the center measure $\bar{\alpha}$, which is a valid Borel measure by assumption. Therefore Theorem 3.1 implies existence of the Dirichlet process $\mathrm{DP}(\alpha)$, provided the specification of distributions can be consistently extended to any vector of the type $(P(A_1), \ldots, P(A_k))$, for arbitrary measurable sets and not just partitions, in such a way that it gives a finitely-additive measure.

An arbitrary collection $A_1, \ldots, A_k$ of measurable sets defines a collection of 2^k atoms of the form $A_1^* \cap A_2^* \cap \cdots \cap A_k^*$, where A^* stands for A or A^c. These atoms $\{B_j : j = 1, \ldots, 2^k\}$ (some of which may be empty) form a partition of the sample space, and hence the joint distribution of $(P(B_j) : j = 1, \ldots, 2^k)$ is defined by Definition 4.1. Every A_i can be written as a union of atoms, and $P(A_i)$ can be defined accordingly as the sum of the corresponding $P(B_j)$s. This defines the distribution of the vector $(P(A_1), \ldots, P(A_k))$.

To prove the existence of a stochastic process $(P(A) : A \in \mathscr{X})$ that possesses these marginal distributions, it suffices to verify that this collection of marginal distributions is consistent in the sense of Kolmogorov's extension theorem. Consider the distribution of the

vector $(P(A_1), \ldots, P(A_{k-1}))$. This has been defined using the coarser partitioning in the 2^{k-1} sets of the form $A_1^* \cap A_2^* \cap \cdots \cap A_{k-1}^*$. Every set in this coarser partition is a union of two sets in the finer partition used previously to define the distribution of $(P(A_1), \ldots, P(A_k))$. Therefore, consistency pertains if the distributions specified by Definition 4.1 for two partitions, where one is finer than the other, are consistent.

Let $\{A_1, \ldots, A_k\}$ be a measurable partition and let $\{A_{i1}, A_{i2}\}$ be a further measurable partition of A_i, for $i = 1, \ldots, k$. Then Definition 4.1 specifies that

$$(P(A_{11}), P(A_{12}), P(A_{21}), \ldots, P(A_{k1}), P(A_{k2})) \\ \sim \mathrm{Dir}(2k; \alpha(A_{11}), \alpha(A_{12}), \alpha(A_{21}), \ldots, \alpha(A_{k1}), \alpha(A_{k2})).$$

In view of the group additivity of finite dimensional Dirichlet distributions given by Proposition G.3, this implies

$$\Big(\sum_{j=1}^{2} P(A_{1j}), \ldots, \sum_{j=1}^{2} P(A_{kj})\Big) \sim \mathrm{Dir}\Big(k; \sum_{j=1}^{2} \alpha(A_{1j}), \ldots, \sum_{j=1}^{2} \alpha(A_{kj})\Big).$$

Consistency follows as the right side is $\mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k))$, since α is a measure.

That $P(\varnothing) = 0$ and $P(\mathfrak{X}) = 1$ almost surely follow from the fact that $\{\varnothing, \mathfrak{X}\}$ is an eligible partition in Definition 4.1, whence $(P(\varnothing), P(\mathfrak{X})) \sim \mathrm{Dir}(2; 0, |\alpha|)$ by (4.1). That $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ almost surely for every disjoint pair of measurable sets A_1, A_2, follows similarly from consideration of the distributions of the vectors $(P(A_1), P(A_2), P(A_1^c \cap A_2^c))$ and $(P(A_1 \cup A_2), P(A_1^c \cap A_2^c))$, whose three and two components both add up to 1.

We have proved the existence of the Dirichlet process distribution $\mathrm{DP}(\alpha)$ for every Polish sample space and every base measure α.

4.2.2 Construction through Distribution Function

When $\mathfrak{X} = \mathbb{R}$ we can also construct the Dirichlet process through its cumulative distribution function, following the scheme of Proposition 2.3.

For any finite set $-\infty < t_1 < t_2 < \cdots < t_k = \infty$ of rational numbers, let

$$(F(t_1), F(t_2) - F(t_1), \ldots, 1 - F(t_{k-1})) \sim \mathrm{Dir}(k+1, MG(t_1), \\ MG(t_2) - MG(t_1), \ldots, M - MG(t_{k-1})),$$

in agreement with (4.1). The resulting distributions of $(F(t_1), F(t_2), \ldots, F(t_k))$ can be checked to form a consistent system in the sense of (i) of Proposition 2.3, and (ii) of the proposition is satisfied, since the increments, which are Dirichlet distributed, are nonnegative.

For given $s_n \downarrow s$, the distribution of $F(s_n)$ is $\mathrm{Be}(MG(s_n), M - MG(s_n))$ and converges weakly to $\mathrm{Be}(MG(s), M - MG(s))$, which is the distribution of $F(s)$. Finally to verify (iv) of Proposition 2.3, we note that $0 \le \mathrm{E}(F(s)) = G(s) \to 0$ as $s \downarrow -\infty$, while $1 \ge \mathrm{E}(F(s)) = G(s) \to 1$ as $s \uparrow \infty$.

4.2.3 Construction through a Gamma Process

The *gamma process* (see Appendix J) is a random process $(S(u): u \geq 0)$ with nondecreasing, right continuous sample paths $u \mapsto S(u)$ and independent increments satisfying $S(u_2) - S(u_1) \sim \text{Ga}(u_2 - u_1, 1)$, for $u_1 < u_2$. Thus, for given $M > 0$ and cumulative distribution function G on $\mathfrak{X} = \mathbb{R}$, we can define a random distribution function by $F(t) = S(MG(t))/S(M)$.

The corresponding distribution is the DP(MG)-distribution. Indeed, for any partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$, we have $S(MG(t_i)) - S(MG(t_{i-1})) \overset{\text{ind}}{\sim} \text{Ga}(MG(t_i) - MG(t_{i-1}), 1)$, whence $(F(t_1) - F(t_0), \ldots, F(t_k) - F(t_{k-1})) \sim \text{Dir}(k; MG(t_1) - MG(t_0), \ldots, MG(t_k) - MG(t_{k-1}))$-distribution, by Proposition G.2. It follows that (4.1) holds for any partition in intervals. As intervals form a measure-determining class, the corresponding measure is the Dirichlet process (cf. Proposition A.5).

Rather than only the distribution function, we can also construct the full measure directly by viewing the gamma process as a *completely random measure*, as described in Appendix J. This is a random measure $(U(A): A \in \mathscr{X})$ such that $U(A_i) \overset{\text{ind}}{\sim} \text{Ga}(\alpha(A_i), 1)$, for every measurable partition $\{A_1, \ldots, A_k\}$. Then $P(A) = U(A)/U(\mathfrak{X})$ is a random probability measure that satisfies (4.1), in view of Proposition G.2.

The construction using a completely random measure works not only for the positive real line, but for every Polish sample space $\mathfrak{X}$.

4.2.4 Construction through Pólya Urn Scheme

A *Pólya sequence* with parameter α is a sequence of random variables $X_1, X_2, \ldots$ whose joint distribution satisfies

$$X_1 \sim \bar{\alpha}, \qquad X_{n+1} | X_1, \ldots, X_n \sim \frac{\alpha + \sum_{i=1}^n \delta_{X_i}}{|\alpha| + n}, \quad n \geq 1.$$

In Section 4.1.4 a random sample from a distribution generated from a Dirichlet process with base measure α is shown to be a Pólya sequence. Such a sequence can of course also be constructed without reference to the Dirichlet process prior. In this section we reverse the construction, and obtain the Dirichlet process from the Pólya urn sequence, invoking *de Finetti's theorem* (see Schervish 1995, Theorem 1.49).

It can be verified that a Pólya sequence $X_1, X_2, \ldots$ is exchangeable. Therefore, by de Finetti's theorem, there exists a random probability measure P such that $X_i | P \overset{\text{iid}}{\sim} P$, $i = 1, 2, \ldots$. We claim that (the law of) P is the DP(α)-process.

In fact, given existence of the Dirichlet process, we can refer to Section 4.1.4 to see that the Dirichlet process generates a Pólya sequence. Because the random measure given by de Finetti's theorem is unique, there is nothing more to prove.

A proof that does not appeal to an existence result needs further argumentation. It suffices to show that the joint distribution of $(P(A_1), \ldots, P(A_k))$ for any measurable partition $\{A_1, \ldots, A_k\}$, where P is the de Finetti measure, satisfies (4.1). These are the almost sure limits of the sequences of the empirical marginals $(\mathbb{P}_n(A_1), \ldots, \mathbb{P}_n(A_k))$. Define a sequence $Y_1, Y_2, \ldots$ of variables that register the partitioning sets containing $X_1, X_2, \ldots$, by $Y_i = j$ if $X_i \in A_j$. These variables take their values in the finite set $\{1, \ldots, k\}$ and can be seen to form

a Pólya sequence themselves, with parameter β given by $\beta\{j\} = \alpha(A_j)$. Because the existence of the finite-dimensional Dirichlet process $\mathrm{DP}(\beta)$ is clear, we can use the argument of the preceding paragraph to see that the de Finetti measure Q of the exchangeable sequence $Y_1, Y_2, \ldots$ satisfies $Q \sim \mathrm{DP}(\beta)$. Now $\mathbb{P}_n(A_j) = \mathbb{Q}_n\{j\} \to Q\{j\}$, almost surely, and hence $(P(A_1), \ldots, P(A_k)) = (Q\{1\}, \ldots, Q\{k\})$ possesses the correct Dirichlet distribution.

4.2.5 Stick-Breaking Representation

The *stick-breaking representation* of a Dirichlet process expresses it as a random discrete measure of the type discussed in Section 3.4.2, with stick-breaking weights, as in Section 3.3.2, based on the beta-distribution. The random support points are generated from the center measure.

The representation gives an easy method to simulate a Dirichlet process, at least approximately.

Even though the representation is explicit and constructive, the proof given below that it gives the Dirichlet process presumes existence of the Dirichlet process.

Theorem 4.12 (Sethuraman) *If $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ and $V_1, V_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ are independent random variables and $W_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$, then $\sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar{\alpha})$.*

Proof Because $\mathrm{E}(\prod_{l=1}^{j}(1 - V_l)) = (M/(M + 1))^j \to 0$, the stick-breaking weights W_j form a probability vector a.s. (c.f. Lemma 3.4), so that P is a probability measure a.s.

For $j \geq 1$ define $W_j' = V_{j+1} \prod_{l=2}^{j}(1 - V_l)$ and $\theta_j' = \theta_{j+1}$. Then $W_{j+1} = (1 - V_1)W_j'$ for every $j \geq 1$ and hence

$$P := W_1\delta_{\theta_1} + \sum_{j=2}^{\infty} W_j\delta_{\theta_j} = V_1\delta_{\theta_1} + (1 - V_1)\sum_{j=1}^{\infty} W_j'\delta_{\theta_j'}.$$

The random measure $P' := \sum_{j=1}^{\infty} W_j'\delta_{\theta_j'}$ has exactly the same structure as P, and hence possesses the same distribution. Furthermore, it is independent of (V_1, θ_1).

We conclude that P satisfies the distributional equation (4.21) given below, and the theorem follows from Lemma 4.13. □

The distributional equation for the Dirichlet process used in the preceding proof is of independent interest, and will be a valuable tool later on. For independent random variables $V \sim \mathrm{Be}(1, |\alpha|)$ and $\theta \sim \bar{\alpha}$, consider the equation

$$P =_d V\delta_\theta + (1 - V)P. \tag{4.21}$$

We say that a random measure P that is independent of (V, θ) is a solution to equation (4.21) if for every measurable partition $\{A_1, \ldots, A_k\}$ of the sample space the random vectors obtained by evaluating the random measures on its left and right sides are equal in distribution in $\mathbb{R}^k$[2].

[2] By Proposition A.5 this is equivalent to the random measures on either side being equal in distribution as random elements in $\mathfrak{M}$, justifying the notation $=_d$.

Lemma 4.13 *For given independent* $\theta \sim \bar{\alpha}$ *and* $V \sim \mathrm{Be}(1, |\alpha|)$*, the Dirichlet process* $\mathrm{DP}(\alpha)$ *is the unique solution of the distributional equation* (4.21)*.*

Proof That the Dirichlet process is a solution is immediate from (4.1) and Proposition G.11. Uniqueness is a consequence of Lemma L.2. □

4.3 Further Properties

4.3.1 Discreteness and Support

A realization from the Dirichlet process is discrete with probability one, also when the base measure is absolutely continuous. This is perhaps disappointing, especially if the intention is to model absolutely continuous probability measures. In particular, the Dirichlet process cannot be used as a prior for density estimation.

Theorem 4.14 (Discreteness) *Almost every realization from the* $\mathrm{DP}(\alpha)$ *is a discrete measure:* $\mathrm{DP}_\alpha(P \in \mathfrak{M}\colon P \text{ is discrete}) = 1$.

Proof The event in the theorem is a measurable subset of $\mathfrak{M}$, by Proposition A.7.

The stick-breaking representation exhibits the Dirichlet distribution explicitly as a random discrete measure. The representation through the gamma process also gives a direct proof of the theorem, as the gamma process increases only by jumps (see Appendix J).

We present a third proof based on the posterior updating formula for the Dirichlet process. A measure P is discrete if and only if $P\{x\colon P\{x\} > 0\} = 1$. If $P \sim \mathrm{DP}(\alpha)$ and $X \mid P \sim P$, then

$$\mathrm{P}(P\{X\} > 0) = \int P\{x\colon P\{x\} > 0\}\, d\mathrm{DP}_\alpha(P).$$

Hence $\mathrm{DP}(\alpha)$ gives probability one to the discrete measures if and only if $\mathrm{P}(P\{X\} > 0) = 1$. By conditioning in the other direction, we see that this is (certainly) the case if $\mathrm{P}(P\{X\} > 0 \mid X = x) = 1$ for every x. Because the posterior distribution of P given $X = x$ is $\mathrm{DP}(\alpha + \delta_x)$, given $X = x$ the variable $P\{X\}$ possesses a $\mathrm{Be}(\alpha\{x\}+1, \alpha(\{x\}^c))$-distribution, and hence is a.s. positive. □

Notwithstanding its discreteness, the support of the Dirichlet process relative to the weak topology is typically very large. It is the whole of $\mathfrak{M}$ as soon as the base measure is fully supported in $\mathfrak{X}$.

Theorem 4.15 (Support) *The weak support of* $\mathrm{DP}(\alpha)$ *is the set* $\{P \in \mathfrak{M}\colon \mathrm{supp}(P) \subset \mathrm{supp}(\alpha)\}$*. Further if* $\mathfrak{X}$ *is a Euclidean space and* P_0 *is atomless and belongs to the weak support of* $\mathrm{DP}(\alpha)$*, then it also belongs to the Kolmogorov-Smirnov support.*

Proof The set $\mathcal{H} := \{P\colon \mathrm{supp}(P) \subset \mathrm{supp}(\alpha)\}$ is weakly closed. Indeed, if $P_n \rightsquigarrow P$, then $P(F) \geq \limsup_{n\to\infty} P_n(F)$, for every closed set, by the Portmanteau theorem. For $F = \mathrm{supp}(\alpha)$ and $P_n \in \mathcal{H}$, we have $P_n(F) = 1$ and hence $P(F) = 1$. This shows that $\mathrm{supp}(P) \subset F$, and hence $P \in \mathcal{H}$.

We now first show that $\text{supp}(\text{DP}(\alpha)) \subset \mathcal{H}$. For $F = \text{supp}(\alpha)$ we have $\alpha(F^c) = 0$, and hence a random measure P with $P \sim \text{DP}(\alpha)$ satisfies $P(F^c) \sim \text{Be}(0, \alpha(F))$, whence $P(F^c) = 0$ a.s. This shows that $P \in \mathcal{H}$ a.s., so that $\text{DP}_\alpha(\mathcal{H}) = 1$, and hence $\text{supp}(\text{DP}(\alpha)) \subset \mathcal{H}$.

To prove the converse inclusion $\mathcal{H} \subset \text{supp}(\text{DP}(\alpha))$, we must show that $\text{DP}_\alpha(\mathcal{U}) > 0$ for every weakly open neighborhood $\mathcal{U}$ of any $P_0 \in \mathcal{H}$. Let $\mathscr{E}$ be the collection of all finite unions of sets from a countable base for the open sets in $\mathfrak{X}$. If $\mathscr{E}$ is enumerated arbitrarily as $\mathscr{E} = \{E_1, E_2, \ldots\}$, then $d(P, Q) = \sum_i 2^{-i} \big|P(E_i) - Q(E_i)\big|$ is a well-defined metric, which generates a topology that is stronger than the weak topology. Indeed, suppose that $P_n(E) \to P(E)$, for every $E \in \mathscr{E}$. For every open set G in $\mathfrak{X}$, there exists a sequence E^m in $\mathscr{E}$ with $E^m \uparrow G$, and hence $\liminf P_n(G) \geq \liminf P_n(E^m) = P(E^m)$, for every m. By the Portmanteau theorem, this implies that $P_n \rightsquigarrow P$. Because from all balls around a point, at most countably many can be *dis*continuity sets for a given measure P_0, we can also ensure that all sets in the collection $\mathcal{E}$ are P_0-continuity sets.

Thus it suffices to show that $\text{DP}_\alpha(P\colon d(P, P_0) < \delta) > 0$ for every $\delta > 0$, or equivalently that $\text{DP}_\alpha(P\colon \cap_{i=1}^k |P(E_i) - P_0(E_i)| < \delta) > 0$ for every finite collection $E_1, \ldots, E_k \in \mathscr{E}$ and $\delta > 0$. This certainly follows if the corresponding statement is true for the set of all intersections $F_1 \cap \cdots \cap F_k$, where $F_i \in \{E_i, E_i^c\}$, whence we may assume that $E_1, \ldots, E_k$ form a partition of $\mathfrak{X}$, so that $(P(E_1), \ldots, P(E_k)) \sim \text{Dir}(k; \alpha(E_1), \ldots, \alpha(E_k))$. If the vector $(\bar{\alpha}(E_1), \ldots, \bar{\alpha}(E_k))$ is in the interior of the simplex, then the corresponding Dirichlet distribution has full support, and hence it charges any neighborhood of $(P_0(E_1), \ldots, P_0(E_k))$. The cases that $\alpha(E_i) = 0$ or $\alpha(E_i^c) = 0$ for some i, when $P(E_i) = 0$ or $P(E_i) = 1$ a.s. under $\text{DP}(\alpha)$, must be considered separately.

If $\alpha(E) = 0$, then certainly $\alpha(E^o) = 0$, for E^o the interior of E, and hence $E^o \subset \text{supp}(\alpha)^c \subset \{P_0\}^c$, if $P_0 \in \mathcal{H}$. Thus $0 = P_0(E^o) = P_0(E)$, if E is a P_0 continuity set, so that $P(E) = P_0(E)$ almost surely under $\text{DP}(\alpha)$, and hence $\text{DP}_\alpha(|P(E) - P_0(E)| < \delta) = 1 > 0$. We may apply this with both E_i and E_i^c to remove those partitioning sets with α-measure 0. □

4.3.2 Convergence

The Dirichlet process depends continuously on its parameter relative to the weak topology. The following theorem also gives weak limits if the total mass of the base measure tends to 0 or ∞. In the latter case the "prior precision" increases indefinitely and the Dirichet process collapses to a degenerate measure at its mean measure.

Theorem 4.16 *Let α_m be a sequence of finite measures on $\mathfrak{X}$ such that $\bar{\alpha}_m \rightsquigarrow \bar{\alpha}$ for some Borel probability measure $\bar{\alpha}$ on $\mathfrak{X}$.*

(i) *If $|\alpha_m| \to 0$, then $\text{DP}(\alpha_m) \rightsquigarrow \delta_X$, for $X \sim \bar{\alpha}$.*
(ii) *If $|\alpha_m| \to M \in (0, \infty)$, then $\text{DP}(\alpha_m) \rightsquigarrow \text{DP}(\alpha)$, where $\alpha = M\bar{\alpha}$.*
(iii) *If $|\alpha_m| \to \infty$, then $\text{DP}(\alpha_m) \rightsquigarrow \delta_{\bar{\alpha}}$, the distribution degenerate at $\bar{\alpha}$.*

If, moreover, $\bar{\alpha}_m(A) \to \bar{\alpha}(A)$ for every Borel set A, then $\int \psi \, dP_m \rightsquigarrow \int \psi \, dP$ for every Borel measurable function $\psi\colon \mathfrak{X} \to \mathbb{R}$ that is uniformly integrable relative to α_m, where

$P_m \sim \mathrm{DP}(\alpha_m)$ and P is a random measure distributed according to the limit distribution given in each case (i)–(iii).

Proof By Theorem A.6 a sequence of random measures on $(\mathfrak{M}, \mathscr{M})$ is tight if and only if the corresponding sequence of mean measures on $(\mathfrak{X}, \mathscr{X})$ is tight. Since the sequence of mean measures $\bar{\alpha}_m$ of the given Dirichlet processes is weakly convergent by assumption, it is tight by Prohorov's theorem. Hence, by Prohorov's theorem in the other direction, it suffices to identify the weak limit point(s) of the sequence $\mathrm{DP}(\alpha_m)$ as the one given, in each case (i)–(iii).

If the random measure Q is distributed as a weak limit point, then by the continuous mapping theorem $\int \psi \, dQ$ is a weak limit of the sequence $\int \psi \, dP_m$, for $P_m \sim \mathrm{DP}(\alpha_m)$ and every continuous $\psi: \mathfrak{X} \to [0, 1]$. We shall show in each of the cases (i)–(iii) that $\int \psi \, dQ =_d \int \psi \, dP$, for P a random measure with the limit distribution claimed in (i)–(iii). By Proposition A.5 this implies that P and Q are equal in distribution, and the proof is complete.

A given continuous $\psi: \mathfrak{X} \to [0, 1]$ can be approximated from below by simple functions $-1 \le \psi_k \uparrow \psi$. Without loss of generality these can be chosen of the form $\sum_{i=1}^{k} a_i \mathbb{1}\{A_i\}$, for constants $a_i \ge -1$ and $\{A_1, \ldots, A_k\}$ a measurable partition of $\mathfrak{X}$ into α-continuity sets. (For instance $\sum_i \xi_{i-1,k} \mathbb{1}\{\xi_{i-1,k} < \psi \le \xi_{i,k}\}$, for $-1 < \xi_{1,k} < \xi_{2,k} < \cdots < \xi_{k.k} < 1$ a grid with meshwidth tending to zero and such that every set $\{x: \psi(x) = \xi_{i,k}\}$ is an α-continuity set. By the continuity of ψ the boundary of a set $\{x: a < \psi(x) \le b\}$ is contained in the set $\{x: \psi(x) = a\} \cup \{x: \psi(x) = b\}$. Because sets of the latter form are disjoint for different a or b, at most countably many b can fail to give a continuity set, and hence a grid $\{\xi_{i,k}\}$ exists.) Suppose that we can show that, as $m \to \infty$, for every partition in α-continuity sets,

$$(P_m(A_1), \ldots, P_m(A_k)) \rightsquigarrow (P(A_1), \ldots, P(A_k)). \tag{4.22}$$

Then $\int \psi_k \, dP_m \rightsquigarrow \int \psi_k \, dP$, by the continuous mapping theorem. Because $\int \psi \, dQ$ is the weak limit of $\int \psi \, dP_m \ge \int \psi_k \, dP_m$, it follows that $\int \psi \, dQ \ge_s \int \psi_k \, dP$ in the sense of stochastic ordering (e.g. by ordering of cumulative distribution functions), for every k. Letting $k \to \infty$, we conclude that $\int \psi \, dQ$ is stochastically larger than $\int \psi \, dP$. By applying the argument also to $1 - \psi$, we see that the two variables are equal in distribution. It remains to prove (4.22) in each of the three cases.

Proof of (ii). The vector on the left of (4.22) is $\mathrm{Dir}(k; \alpha_m(A_1), \ldots, \alpha_m(A_k))$-distributed. Because $\alpha_m(A_j) \to \alpha(A_j)$ for every j by assumption and construction, these distributions tend weakly to $\mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k))$ by Proposition G.6, which is the law of the right side of (4.22), if $P \sim \mathrm{DP}(\alpha)$.

Proof of (i) and (iii). We prove (4.22) by verifying the convergence of all (mixed) moments. By Corollary G.4, for any nonnegative integers $r_1, \ldots, r_k$ and $u^{[r]}$ the ascending factorial,

$$\mathrm{E}[P_m(A_1)^{r_1} \times \cdots \times P_m(A_k)^{r_k}] = \frac{\alpha_m(A_1)^{[r_1]} \times \cdots \times \alpha_m(A_k)^{[r_k]}}{|\alpha_m|^{[\sum_i r_i]}}.$$

If $\bar{\alpha}_m(A_j) \to \bar{\alpha}(A_j)$ and $|\alpha_m| \to \infty$ as in (iii), then this tends to $\bar{\alpha}(A_1)^{r_1} \cdots \bar{\alpha}(A_k)^{r_k}$, which is the corresponding mixed moment when P is identically $\bar{\alpha}$. If $|\alpha_m| \to 0$ as in (i), then there are two cases. If exactly one $r_j \ne 0$, then the moment tends to $\bar{\alpha}(A_j)$, which

is equal to $\mathrm{E}(P(A_j)^{r_j}) = \mathrm{E}[\delta_X(A_j)^{r_j}] = \bar{\alpha}(A_j)$, if $X \sim \bar{\alpha}$. On the other hand, if at least two of $r_1, \ldots, r_k$ are nonzero, then the mixed moment is zero under the limit measure δ_X, as $P(A_j) = \delta_X(A_j)$ is nonzero for exactly one j, and the mixed moments in the preceding display also tend to zero, as at least two factors in the numerator tend to zero, while only $|\alpha_m|$ does so in the denominator and is bigger; all other factors tend to positive integers.

For the proof of the final assertion consider first a bounded, measurable function ψ, and let L be a weak limit point of the sequence $\int \psi \, dP_m$. The function ψ can be approximated from below by a sequence of indicator functions $0 \le \psi_k \uparrow \psi$. Under the additional assumption of convergence of $P_m(A)$ for every Borel set, the convergence (4.22) is true for any measurable partition $\{A_1, \ldots, A_k\}$. Consequently $\int \psi_k \, dP_m \rightsquigarrow \int \psi_k \, dP$, for every k. Arguing as before, we find that $L \ge_s \int \psi \, dP$. Applying the argument also to $1 - \psi$, we see that $L =_d \int \psi \, dP$, and the proof for a bounded ψ is complete. A general measurable ψ is the sum of the bounded function $\psi \mathbb{1}\{|\psi| \le M\}$ and the function $\psi \mathbb{1}\{|\psi| > M\}$. Here $\mathrm{E} \int |\psi| \mathbb{1}\{|\psi| > M\} \, dP_m = \int |\psi| \mathbb{1}\{|\psi| > M\} \, d\alpha_m$ can be made arbitrarily small by choosing large M, uniformly in m. □

As a corollary, the above theorem implies limiting behavior of the posterior of a Dirichlet process in two different asymptotic scenarios – as the sample size goes to infinity and in the noninformative limit as the prior sample size goes to zero. Recall that the posterior distribution based on a random sample $X_1, \ldots, X_n$ from a distribution drawn from a $\mathrm{DP}(\alpha)$-prior is $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$.

Corollary 4.17 *Suppose $X_1, X_2, \ldots$ are an i.i.d. sample from P_0.*

(i) *If $n \to \infty$, then* $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i}) \rightsquigarrow \delta_{P_0}$, *a.s.,*
(ii) *If $|\alpha| \to 0$, then* $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i}) \rightsquigarrow \mathrm{DP}(\sum_{i=1}^n \delta_{X_i})$, *a.s.*

Proof The total mass of the base measure $\alpha + \sum_{i=1}^n \delta_{X_i}$ tends to infinity in case (i) and to n in case (ii). As the empirical distribution tends weakly to P_0 almost surely as $n \to \infty$, by Proposition F.3, the center measure tends to P_0 and $n^{-1} \sum_{i=1}^n \delta_{X_i}$, respectively. Thus the assertions follow from Theorem 4.16 (iii) and (ii). □

Assertion (i) means that the posterior is "consistent"; this is discussed more precisely in Chapter 6. Assertion (ii) gives the *noninformative limit* as the prior sample size goes to zero, so that the effect of the prior is eliminated from the posterior. The limiting posterior process is known as the *Bayesian bootstrap*. The result is one of the few concrete noninformative analyses in Bayesian nonparametrics. The Bayesian bootstrap generates random probability measures that are supported only at the observed points, and gives rise to a resampling scheme similar to the usual empirical (or Efron's) bootstrap. It is discussed further in Section 4.7.

4.3.3 Approximations

In this section we discuss two approximations to the Dirichlet process, both taking the form of a random discrete distribution, as discussed in Section 3.4.2, with finite support.

In the first, the weights are assigned by a finite Dirichlet distribution. Because this has only finitely many parameters, it is convenient for approximate inference on a computer.

Definition 4.18 (Dirichlet-multinomial) The *Dirichlet-multinomial process* of order N and parameters G and $(\alpha_1, \ldots, \alpha_N)$ is the random probability measure $\sum_{k=1}^N W_k \delta_{\theta_k}$, where $(W_1, \ldots, W_N) \sim \mathrm{Dir}(N; \alpha_1, \ldots, \alpha_N)$ and $\theta_1, \ldots, \theta_N \overset{\text{iid}}{\sim} G$ are independent.

The name "Dirichlet-multinomial process" stems from the fact that a random sample $X_1, \ldots, X_n$ generated from a distribution sampled from such a process can be represented as $X_i = \theta_{s_i}$, and $(R_1, \ldots, R_N)|W \sim \mathrm{MN}_N(n; W_1, \ldots, W_N)$, where $R_j = \#\{i : s_i = j\}$. This representation shows, in particular, that the sample $X_1, \ldots, X_n$ will tend to have ties even when n is much smaller than N.

The Dirichlet-multinomial assigns probability $\sum_{k:\theta_k \in A} W_k$ to a given set A. Because aggregation of Dirichlet weights to a smaller number of cells gives a Dirichlet distribution, by Proposition G.3, given the vector $\theta = (\theta_1, \ldots, \theta_N)$ the Dirichlet-multinomial process is a Dirichlet process with base measure $\sum_{k=1}^N \alpha_k \delta_{\theta_k}$.

In particular, this shows that $\mathrm{E}(\int \psi \, dP | \theta) = \int \psi \, d(\sum_k \bar{\alpha}_k \delta_{\theta_k}) = \sum_k \bar{\alpha}_j \psi(\theta_k)$, for any measurable function $\psi \in \mathbb{L}_1(G)$, for $\bar{\alpha}_k = \alpha_k / \sum_k \alpha_k$. In particular, the expectation $\mathrm{E}(\int \psi \, dP) = \int \psi \, dG$ does not depend on the parameters $\alpha_1, \ldots, \alpha_N$.

As it is finitely supported, a sample from the Dirichlet-multinomial is easily generated. If the parameters are chosen appropriately, this will be almost as good as generating from a Dirichlet process, as part (ii) of the following theorem shows.

Theorem 4.19 *Let P_N be a Dirichlet-multinomial process of order $N \to \infty$ with parameters $(\alpha_{1,N}, \ldots, \alpha_{N,N})$ and G, satisfying $\max_{1 \le k \le N} \alpha_{k,N}/\alpha_{.N} \to 0$, for $\alpha_{.N} = \sum_{k=1}^N \alpha_{k,N}$.*

(i) *If $\alpha_{.N} \to 0$, then $\int \psi \, dP_N \rightsquigarrow \psi(\theta)$, where $\theta \sim G$, for any bounded continuous ψ. In particular, this is true if $\alpha_{k,N} = \lambda_N$ and $N\lambda_N \to 0$.*

(ii) *If $\alpha_{.N} \to M$, then $\int \psi \, dP_N \rightsquigarrow \int \psi \, dP$, where $P \sim \mathrm{DP}(MG)$, for any $\psi \in \mathbb{L}_1(G)$.*

(iii) *If $\alpha_{.N} \to \infty$, then $\int \psi \, dP_N \to_p \int \psi \, dG$, for any $\psi \in \mathbb{L}_1(G)$. In particular, this is true if $\alpha_{k,N} = \lambda_N$ and $N\lambda_N \to \infty$.*

Furthermore, if $\theta_i \overset{iid}{\sim} G$ and the weights are given by $W_{k,N} = \Gamma_k / \sum_{j=1}^N \Gamma_j$, where $\Gamma_k \overset{ind}{\sim} \mathrm{Ga}(\lambda_k, 1)$ with $\sum_{k=1}^\infty \lambda_k^2/k^2 < \infty$ and $N^{-1} \sum_{k=1}^N \lambda_k \to \lambda > 0$ (hence including the case $\alpha_{k,N} = \lambda_k$), then the convergence in (iii) is also almost surely, for any $\psi \in \mathbb{L}_2(G)$.

Proof Given $\theta_1, \ldots, \theta_N$ the Dirichlet-multinomial process is a Dirichlet process with base measure $\sum_k \alpha_{k,N} \delta_{\theta_k}$. The integral of a function ψ relative to the center measure is equal to $\sum_k \bar{\alpha}_{k,N} \psi(\theta_k)$, for $\bar{\alpha}_{k,N} = \alpha_{k,N}/\alpha_{.N}$, and has expectation and variance

$$\mathrm{E}\Big(\sum_k \bar{\alpha}_{k,N} \psi(\theta_k) \Big) = \int \psi \, dG,$$

$$\mathrm{var}\Big(\sum_k \bar{\alpha}_{k,N} \psi(\theta_k) \Big) = \sum_k \bar{\alpha}_{k,N}^2 \, \mathrm{var}\, \psi(\theta_1) \le \max_k \bar{\alpha}_{k,N} \int \psi^2 \, dG.$$

In particular, $\int \psi \, d(\sum_k \bar{\alpha}_{k,N} \delta_{\theta_k})$ tends in probability to $\int \psi \, dG$, for every $\psi \in \mathbb{L}_2(G)$, in particular for every bounded, continuous ψ. It follows that conditionally on $\theta_1, \theta_2, \ldots$ the

sequence of center measures converges weakly to G, in probability. The total mass of the base measure is $\alpha_{\cdot N}$ and converges to 0, M and ∞ in the cases (i)–(iii). Therefore, Theorem 4.16 gives that the Dirichlet-multinomial $P_N = \sum_{k=1}^{N} W_{k,N}\delta_{\theta_k}$ tends in distribution in $\mathfrak{M}$ to δ_θ for $\theta \sim G$, DP(MG), and δ_G, in the three cases (i)–(iii), conditionally on $\theta_1, \theta_2, \ldots$ in probability, and then also unconditionally, as the three limits do not depend on the sequence $\theta_1, \theta_2, \ldots$ Moreover, the last convergence holds in probability since G is a fixed probability measure. Because the map $P \mapsto \int \psi\, dP$ from $\mathfrak{M}$ to $\mathbb{R}$ is continuous, for every bounded, continuous function $\psi\colon \mathfrak{X} \to \mathbb{R}$, the continuous mapping theorem shows that assertions (i)–(iii) of the theorem are true for every bounded continuous function ψ.

It suffices to improve the conclusion to $\psi \in \mathbb{L}_1(G)$ in cases (ii) and (iii), and to almost sure convergence in the special case of (iii) mentioned in the final assertion of the theorem. For the upgrade to integrable ψ we apply the final assertion of Theorem 4.16. Because $\mathrm{E}\int |\psi|\mathbb{1}\{|\psi| > M\}\, d(\sum_k \bar{\alpha}_{k,N}\delta_{\theta_k}) = \int |\psi|\mathbb{1}\{|\psi| > M\}\, dG$, any function $\psi \in \mathbb{L}_1(G)$ satisfies the uniform integrability assumption, in mean. Secondly, the argument given previously already showed that given $\theta_1, \theta_2, \ldots$ the center measures converge to G not only weakly, but also setwise.

To prove the final statement of the theorem we apply Kolmogorov's strong law. Because $N^{-1}\sum_k \mathrm{E}(\Gamma_k) \to \lambda$ and $\sum_k \mathrm{var}(\Gamma_k)/k^2 = \sum_k \lambda_k/k^2 < \infty$ by assumption, this gives first that $N^{-1}\sum_{k=1}^{N}\Gamma_k \to \lambda$ a.s. Second, $N^{-1}\sum_{k=1}^{N}\Gamma_k\psi(\theta_k) \to \lambda\int\psi\, dG$ a.s., since $\sum_k \mathrm{var}\,(\Gamma_k\psi(\theta_k))/k^2 \le \sum_k(\lambda_k + \lambda_k^2)\int\psi^2\, dG/k^2 < \infty$. □

The tempting "default choice" $\alpha_{1,N} = \cdots = \alpha_{N,N} = 1$ belongs to case (iii) of Theorem 4.19, and hence yields a sequence of Dirichlet-multinomial priors that converges to the prior for P degenerate at G. It is inappropriate for inference (unless G is equal to the data-generating measure). The corresponding Dirichlet-multinomial distribution with $N = n$ and weights at the observations is the same as the Bayesian bootstrap given in Corollary 4.17. This converges to the correct distribution as $n \to \infty$, but is a posterior distribution, with $\theta_1, \theta_2, \ldots$ observations, and not a prior.

A second approximation to the Dirichlet process is obtained by truncating the stick-breaking representation. To guarantee an error smaller than a predetermined level, the truncation point can be chosen random. For stick-breaking weights $W_i = \prod_{j=1}^{i-1}(1 - V_j)V_i$, for $V_1, V_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Be}(1, M)$, and independent $\theta_0, \theta_1, \ldots \overset{\text{iid}}{\sim} G$, consider for a given, small $\epsilon > 0$,

$$P_\epsilon = \sum_{i=1}^{N_\epsilon} W_i\delta_{\theta_i} + \Big(1 - \sum_{i=1}^{N_\epsilon} W_i\Big)\delta_{\theta_0}, \qquad N_\epsilon = \inf\Big\{N \ge 1\colon \sum_{i=1}^{N} W_i > 1 - \epsilon\Big\}. \tag{4.23}$$

We shall call the distribution of P_ϵ an *ϵ-Dirichlet process*, and denote it by DP(α, ϵ)[3].

Proposition 4.20 (ϵ-Dirichlet process) *For every $\epsilon > 0$ the total variation distance between the ϵ-Dirichlet process P_ϵ defined in* (4.23) *and the Dirichlet process $P = \sum_{i=1}^{\infty} W_i\delta_{\theta_i}$ satisfies $d_{TV}(P_\epsilon, P) \le \epsilon$ a.s. Consequently, the Lévy distance defined in* (A.1) *between the distributions* DP(α, ϵ) *and* DP(α) *of these random measures on* ($\mathfrak{M}, \mathscr{M}$),

[3] An alternative construction with the same properties is $P_\epsilon = \sum_{i\le N_\epsilon} W_i\delta_{\theta_i} / \sum_{i\le N_\epsilon} W_i$.

equipped with the total variation distance, satisfies $d_L(\mathrm{DP}(\alpha,\epsilon),\mathrm{DP}(\alpha)) \le \epsilon$. *Furthermore, as* $\epsilon \to 0$ *we have* $\int \psi\, dP_\epsilon \to \int \psi\, dP$ *a.s., for any measurable function* ψ *that is* P*-integrable a.s. Finally, the number* $N_\epsilon + 1$ *of support points of* P_ϵ *satisfies* $N_\epsilon - 1 \sim \mathrm{Poi}(M \log_- \epsilon)$.

Proof The first assertion is clear from the fact that $\big|P(A) - P_\epsilon(A)\big| \le \sum_{i>N_\epsilon} W_i < \epsilon$, for any measurable set $A \subset \mathfrak{X}$. We conclude that $P \in \mathcal{A}$ for an arbitrary $\mathcal{A} \in \mathscr{M}$ implies that $P_\epsilon \in \{Q\colon d_{TV}(Q,\mathcal{A}) < \epsilon\}$ and hence also $P_\epsilon \in \{Q\colon d_L(Q,\mathcal{A}) < \epsilon\}$, because the Lévy distance d_L (on $\mathfrak{M} = \mathfrak{M}(\mathfrak{X})$) satisfies $d_L \le d_{TV}$. Consequently $\mathrm{DP}_{\alpha,\epsilon}(\mathcal{A}) = \mathrm{P}(P_\epsilon \in \mathcal{A}) \le \mathrm{P}(P \in \mathcal{A}^\epsilon) = \mathrm{DP}_\alpha(\mathcal{A}^\epsilon)$. Similarly $\mathrm{DP}_\alpha(\mathcal{A}) \le \mathrm{DP}_{\alpha,\epsilon}(\mathcal{A}^\epsilon)$, and the assertion follows from the definition of the Lévy distance d_L (on $\mathfrak{M}(\mathfrak{M})$ this time).

The almost sure convergence of $\int \psi\, dP_\epsilon = \sum_{i=1}^{N_\epsilon} W_i \psi(\theta_i) + \bar{W}_\epsilon \psi(\theta_0)$, for $\bar{W}_\epsilon = 1 - \sum_{i=1}^{N_\epsilon} \le \epsilon$, to $\sum_{i=1}^{\infty} W_i \psi(\theta_i) = \int \psi\, dP$ is clear from the facts that $N_\epsilon \to \infty$ and $\bar{W}_\epsilon \to 0$.

Because $\sum_{i\le N} W_i = 1 - \prod_{i\le N}(1 - V_i)$, we have $N_\epsilon = \inf\{N\colon -\sum_{i\le N} \log(1 - V_i) > \log_- \epsilon\}$ by the definition of N_ϵ. Because $-\log(1 - V_i) \overset{\mathrm{iid}}{\sim} \mathrm{Exp}(M)$, the variable $N_\epsilon - 1$ can be identified with the number of events occurring in the time interval $[0, \log_- \epsilon]$ in a homogeneous Poisson process with rate M. This implies the final assertion of the proposition. □

The last assertion of Proposition 4.20 implies that on average $\mathrm{E}(N_\epsilon + 1) = 2 + M \log_- \epsilon$ terms are needed to attain a level of accuracy ϵ.

In the case that $\mathfrak{X} = \mathbb{R}$, we can also consider the Kolmogorov-Smirnov distance. Since this is bounded by the total variation distance, it follows that $d_{KS}(F_\epsilon, F) \le \epsilon$ almost surely, for F_ϵ and F the cumulative distribution functions of P_ϵ and P. This next implies that continuous functionals $\psi(F)$ can be approximated by the corresponding functional $\psi(F_\epsilon)$. Examples include the KS-distance itself, and quantiles of the distribution.

4.3.4 Mutual Singularity of Dirichlet Processes

The prior and posterior distributions given a random sample from a Dirichlet process with a atomless base measure are mutually singular. This follows from Theorem 3.22, since the Dirichlet process is a special case of a Pólya tree process (where we can choose a canonical binary partition with $\alpha_\varepsilon = \alpha(A_\varepsilon) = 2^{-|\varepsilon|}$, so that $\sum_{m=1}^{\infty} 2^m = \infty$). In this subsection, we show more generally, without referring to Pólya tree processes, that two Dirichlet processes with different base measures are typically mutually singular.

The mutual singularity of prior and posterior can arise, because the Dirichlet prior is not concentrated on a dominated set of measures (see Lemma 1.2). The following general theorem also implies the mutual singularity of the posterior distributions based on different Dirichlet priors. That (even slightly) different priors give quite different behavior is somewhat unusual.

A given base measure α can be decomposed as the sum $\alpha = \alpha_a + \alpha_c$ of a discrete measure α_a (i.e. $\alpha_a(D^c) = 0$ for a countable set D) and a measure α_c without atoms (i.e. $\alpha_c\{x\} = 0$ for every x). We refer to the two measures as the *atomic part* and *continuous part* of α, respectively.

Theorem 4.21 *If the continuous parts of α_1 and α_2 are unequal, or their atomic parts have different supports, then* $\mathrm{DP}(\alpha_1)$ *and* $\mathrm{DP}(\alpha_2)$ *are mutually singular.*

Proof Let $\alpha_1 = \alpha_{1,c} + \alpha_{1,a}$ and $\alpha_2 = \alpha_{2,c} + \alpha_{2,a}$ be the decompositions of the two base measures, S_1 and S_2 the supports of their atomic parts, and $\mathfrak{X}' = \mathfrak{X} \setminus (S_1 \cup S_2)$.

The Dirichlet measures $\mathrm{DP}(\alpha_1)$ and $\mathrm{DP}(\alpha_2)$ are probability measures on the space of measures $(\mathfrak{M}, \mathscr{M})$ on the sample space $(\mathfrak{X}, \mathscr{X})$. We shall show that there exist disjoint measurable subsets C_1, C_2 of $\mathfrak{X}^\infty$ such that $\mathrm{DP}_{\alpha_i}(P \in \mathfrak{M}\colon P^\infty(C_i) = 1) = 1$, for $i = 1, 2$. The theorem then follows as $\{P \in \mathfrak{M}\colon P^\infty(C_1) = 1\} \cap \{P \in \mathfrak{M}\colon P^\infty(C_2) = 1\} = \varnothing$.

For simplicity of notation consider P^∞ as the law of a random sequence $X_1, X_2, \ldots$ drawn from $P \sim \mathrm{DP}(\alpha_i)$. Let $Y_1, Y_2, \ldots$ be the distinct values in $X_1, X_2, \ldots$ that do not belong to $S_1 \cup S_2$, and let K_n be the number of these values among $X_1, \ldots, X_n$. (More precisely, define $D_i = 1$ if $X_i \notin \{X_1, \ldots, X_{i-1}\} \cup S_1 \cup S_2$; let $K_n = \sum_{i=1}^n D_i$; and put $Y_j = X_{\tau_j}$, for $\tau_1 = 1$ and $\tau_j = \min\{n\colon K_n = j\}$ for $j \in \mathbb{N}$.) Then $Y_1, Y_2, \ldots$ can be seen to form an i.i.d. sequence from $\bar{\alpha}_{i,c}$, the normalized continuous part of the base measure (cf. Problem 4.21). By the strong law of large numbers $n^{-1} \sum_{j=1}^n \mathbb{1}\{Y_j \in A\} \to \overline{\alpha}_{i,c}(A)$, for every measurable set A. Furthermore $K_n / \log n \to |\alpha_{i,c}|$ almost surely, by a slight extension of the result of Proposition 4.8(ii).

If $\alpha_{1,c} \neq \alpha_{2,c}$, then either $\bar{\alpha}_{1,c} \neq \bar{\alpha}_{2,c}$ or $|\alpha_{1,c}| \neq |\alpha_{2,c}|$, or both. In the first case there exists a measurable set A with $\bar{\alpha}_{1,c}(A) \neq \bar{\alpha}_{2,c}(A)$, and then the two events $\{n^{-1} \sum_{j=1}^n \mathbb{1}\{Y_j \in A\} \to \overline{\alpha}_{i,c}(A)\}$, for $i = 1, 2$, are disjoint, up to a null set. In the second case the two events $\{K_n / \log n \to |\alpha_{i,c}|\}$, for $i = 1, 2$, are disjoint up to a null set.

Finally if $S_1 \neq S_2$, then there exists x that is in exactly one S_i, say $x \in S_2 \setminus S_1$. Then the event $\cup_{j \geq 1}\{X_j = x\}$ has probability one under $P \sim \mathrm{DP}(\alpha_2)$ and probability zero if $P \sim \mathrm{DP}(\alpha_1)$. □

The conditions in the theorem cannot be relaxed further. If $\alpha_{1,c} = \alpha_{2,c}$ and $\mathrm{supp}(\alpha_{1,a}) = \mathrm{supp}(\alpha_{2,a})$, then $\mathrm{DP}(\alpha_1)$ and $\mathrm{DP}(\alpha_2)$ need not be mutually singular.

4.3.5 Tails of a Dirichlet Process

By Theorem 4.15 the support of a $\mathrm{DP}(MG)$-process on $\mathbb{R}$ is equal to the support of the center measure G. In this section we study the tails of the Dirichlet process, i.e. the behavior of $F(x)$ as $G(x) \downarrow 0$, or $1 - F(x)$ as $G(x) \uparrow 1$, for F the cumulative distribution function of the Dirichlet process. In view of the representation of $\mathrm{DP}(MG)$ given in Section 4.2.3, this is the same as that of a gamma process.

Theorem 4.22 *If F is the cumulative distribution function of the* $\mathrm{DP}(MG)$*-process, then*

$$\liminf_{G(x)\downarrow 0} F(x) \exp\Bigl(\frac{r \log |\log MG(x)|}{MG(x)}\Bigr) = \begin{cases} 0, & \text{if } r < 1, \\ \infty, & \text{if } r > 1, \end{cases} \quad \text{a.s.},$$

$$\limsup_{G(x)\downarrow 0} F(x) \exp\Bigl(\frac{1}{MG(x)|\log MG(x)|^r}\Bigr) = \begin{cases} 0, & \text{if } r > 1, \\ \infty, & \text{if } r \leq 1, \end{cases} \quad \text{a.s.}$$

The same results hold if F and G are replaced by $1 - F$ and $1 - G$.

Proof We can represent $F(x)$ as $\gamma(MG(x))/\gamma(M)$ for a standard gamma process γ (see Section 4.2.3 and Appendix J). The assertion on the lim inf is then immediate from Theorem J.19(ii).

According to Theorem J.19(iii), for a given convex, increasing function h, the variable $\limsup_{t\downarrow 0} \gamma(t)/h(t)$ is equal to 0 or ∞ if the integral $\int_0^1 \int_{h(t)}^\infty x^{-1}e^{-x}\,dx\,dt$ converges, or diverges, respectively. Combining the inequalities $e^{-1}\log(1/t) \le \int_t^1 x^{-1}e^{-x}\,dx \le \log(1/t)$, for $0 < t < 1$, and $0 \le \int_1^\infty x^{-1}e^{-x}\,dx \le e^{-1}$, we see

$$e^{-1}\log\frac{1}{h(t)} \le \int_{h(t)}^\infty x^{-1}e^{-x}\,dx \le e^{-1} + \log\frac{1}{h(t)}.$$

Thus the relevant integral converges or diverges if and only if $\int_0^1 \log_- h(t)\,dt$ converges or diverges. For $\log h(t) = t^{-1}|\log t|^{-r}$ this is the case if $r < 1$ or $r \ge 1$, respectively.

The result for the upper tail can be obtained by the same method, after noting that the stochastic process $s \mapsto \gamma(1) - \gamma(s)$ is equal in distribution to the process $s \mapsto \gamma(1-s)$. □

The theorem shows that the lim inf and lim sup of the tails have different orders. This is caused by the irregularity of the sample paths of the Dirichlet (or gamma) process, which increase by infinitely many (small) jumps on a countable dense set. This also causes that there is no deterministic function that gives the exact order of the lim inf, in the sense of existence of $\liminf_{G(x)\to 0} F(x)/h(x)$ for some deterministic function h as a number in $(0, \infty)$, rather than 0 or ∞ (cf. Theorem J.19(i)).

The message of the theorem can be summarized (imprecisely) in the bounds, for $r > 1$, a.s. eventually as $G(x)$ is small enough,

$$\exp\Big(-\frac{r\log|\log MG(x)|}{MG(x)}\Big) \le F(x) \le \exp\Big(-\frac{1}{MG(x)|\log MG(x)|^r}\Big). \tag{4.24}$$

Similar bounds are valid for the upper tail. These inequalities show that the tails of F are much thinner ("exponentially much") than the tails of the base measure G. This may supplement assertion (4.6), which shows that these tails are equal "on average." In particular, the variable $\int |\psi|\,dP$ may well be finite a.s. even if $\int |\psi|\,d\alpha = \infty$ (and the former variable is not integrable).

Example 4.23 (Normal distribution) The upper tail of the standard normal distribution G satisfies $1 - G(x) \sim x^{-1}(2\pi)^{-1/2}e^{-x^2/2}$. As $r > 1$ can be arbitrarily close to 1, we can absorb any constant in the power of x. Inequalities (4.24) imply, for fixed M as $x \to \infty$,

$$\exp\big[-2r(2\pi)^{1/2}M^{-1}e^{x^2/2}\log x\big] \le 1 - F(x) \le \exp\big[-e^{x^2/2}x^{-2r}\big].$$

Thus the tails are (much) thinner than the tails of the extreme value distribution.

Example 4.24 (Cauchy distribution) The upper tail of the standard Cauchy distribution G satisfies $1 - G(x) \sim \pi^{-1}x^{-1}$, for $x \to \infty$. Again we can absorb any constant in powers of $\log x$. Inequalities (4.24) imply, for fixed M as $x \to \infty$,

$$\exp\big[-r\pi M^{-1}x\log\log x\big] \le 1 - F(x) \le \exp\big[-x/|\log x|^r\big].$$

Thus the tails of F are almost exponential, even though G has no mean.

4.3.6 Distribution of Median

The *median* m_F of a DP(MG)-process on $\mathbb{R}$ is defined as any random variable m_F such that $F(m_F-) \le \frac{1}{2} \le F(m_F)$, for F the (random) cumulative distribution function of the DP(MG)-process. The following theorem gives an exact expression for its distribution, which is known as the *median-Dirichlet distribution* with parameters M and G. The theorem also shows that "any median of the random variable m_F is a median of the base measure G," a property that may be considered the analog of the equation $\mathrm{E}(\int \psi\, dF) = \int \psi\, dG$ for means.

Theorem 4.25 *Any median m_F of $F \sim \mathrm{DP}(MG)$ has cumulative distribution function H given by*

$$H(x) = \int_{1/2}^{1} \frac{\Gamma(M)}{\Gamma(MG(x))\Gamma(M(1-G(x)))} u^{MG(x)-1}(1-u)^{M(1-G(x))-1}du.$$

The cumulative distribution function H is continuous if G is so, and has the same (set of) medians as G.

Proof If $m_F \le x$, then $F(x) \ge F(m_F) \ge 1/2$, by the definition of m_F; furthermore, $m_F > x$ implies $F(x) \le F(m_F-) \le 1/2$. This shows that $\{F(x) > 1/2\} \subset \{m_F \le x\} \subset \{F(x) \ge 1/2\}$. As $F(x)$ has a Be($MG(x), M\bar{G}(x)$)-distribution, the probabilities of the events on left and right are equal and hence $\mathrm{P}(m_F \le x) = \mathrm{P}(F(x) \ge 1/2)$. This gives the formula for H by employing the formula for the beta distribution. The continuity of H is clear from the continuity of the beta distribution in its parameter.

The beta distributions $B_\theta := \mathrm{Be}(M\theta, M(1-\theta))$ form a one-dimensional exponential family in the parameter $0 < \theta < 1$, which possesses a strictly monotone likelihood ratio. Furthermore, $B_{1/2}$ is symmetric about $1/2$. Therefore, the preceding derivation gives $\mathrm{P}(m_F \le x) = B_{G(x)}([1/2, 1]) \ge 1/2$ if and only if $G(x) \ge 1/2$. By a similar derivation $\mathrm{P}(m_F > x) = \mathrm{P}(F(x) \le 1/2) = B_{G(x)}([0, 1/2])$. Replacing x by $x - \epsilon$ and taking the limit as $\epsilon \downarrow 0$, we find that $\mathrm{P}(m_F \ge x) = B_{G(x-)}([0, 1/2])$, whence $\mathrm{P}(m_F < x) \le 1/2$ if and only if $G(x-) \le 1/2$. Combining the preceding, we see that x is a median of m_F if and only if $G(x-) \le 1/2 \le G(x)$. □

4.3.7 Distribution of Mean

Consider the mean functional $\int \psi\, dP$, for a given measurable function $\psi: \mathfrak{X} \to \mathbb{R}$ and $P \sim \mathrm{DP}(\alpha)$.

Let $\mathrm{Re}\, z$ and $\mathrm{Im}\, z$ stand for the real and imaginary parts of a complex number z, and let $\log z = \log|z| + i \arg z$ denote the principal branch of the complex logarithm (with $\arg z \in [-\pi, \pi)$).

Theorem 4.26 *The mean $\int \psi\, dP$ for $P \sim \mathrm{DP}(\alpha)$ is defined as an a.s. finite random variable for any measurable function $\psi\colon \mathfrak{X} \to \mathbb{R}$ such that $\int \log(1+|\psi|)\, d\alpha < \infty$. Provided that $\beta := \alpha \circ \psi^{-1}$ is not degenerate its distribution is absolutely continuous with Lebesgue density h and cumulative distribution function H given by*

$$h(s) = \frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big[\exp\Big(-\int \log(1+it(s-x))\, d\beta(x)\Big)\int \frac{d\beta(x)}{1+it(s-x)}\Big]\, dt,$$
$$H(s) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{1}{t}\,\mathrm{Im}\Big[\exp\Big(-\int \log(1+it(s-x))\, d\beta(x)\Big)\Big]\, dt.$$

Proof Since $\int \psi\, dP = \int x\, dP\circ\psi^{-1}(x)$ and $P\circ\psi^{-1} \sim \mathrm{DP}(\alpha\circ\psi^{-1})$, it suffices to consider the special case where $\psi(x) = x$ on $\mathfrak{X} = \mathbb{R}$ and $P \sim \mathrm{DP}(\beta)$. We can then represent the cumulative distribution function of P as $\gamma(\beta(x))/\gamma(M)$, for β also denoting the cumulative distribution function of β and $M = |\alpha| = \beta(\infty)$.

The a.s. finiteness of $\int x\, dP(x)$ follows from Lemma 4.27.

Furthermore, the distribution function of the mean $\int x\, d\gamma(\beta(x))/\gamma(M)$ can be written $H(s) = \mathrm{P}(\int \psi\, d\gamma \le 0)$ for the function ψ defined by $\psi(x) = x - s$. Lemma 4.27 gives the characteristic function of the random variable $\int \psi\, d\gamma$, and an inversion formula (see Problem L.2) gives that

$$H(s) + H(s-) = 1 - \frac{2}{\pi}\lim_{\epsilon\downarrow 0, T\uparrow\infty}\int_\epsilon^T \frac{1}{t}\,\mathrm{Im}\Big[e^{-\int \log(1+it(s-x))\, d\beta(x)}\Big]\, dt. \tag{4.25}$$

To complete the proof we show that the integrand is integrable near 0 and ∞, so that the limit exists as an ordinary integral, and justify differentiation under the integral, which indeed gives the formula for the derivative h.

Because $\log z = \log|z| + i\arg z$, the integrand in the right side is given by

$$t^{-1}\exp\Big[-\int \log(1+t^2(s-x)^2)\, d\beta(x)/2\Big]\sin\Big(\int \tan^{-1}(t(s-x))\, d\beta(x)\Big).$$

Both the exponential and the sine functions are bounded by 1, and hence integrability is an issue only for t near 0 and ∞. Because $\int \log(1+t^2(s-x)^2)\, d\beta(x) \ge c_1 \log(1+c_2 t^2(s-c_3)^2)$, for some positive constants c_1, c_2 and constant c_3 (that depend on α), the expression is bounded above by

$$t^{-1}\exp\Big[-c_1\log(1+c_2t^2(s-c_3)^2)\Big] = \frac{1}{t(1+c_2t^2(s-c_3)^2)^{c_1}}.$$

The integral over $t > T$ converges and hence tends to zero as $T \uparrow \infty$. Because $|\tan^{-1}x| \lesssim |x|/(1+|x|)$ and $|\sin x| \le |x|$, the expression is bounded by a multiple of $\int |s-x|/(1+t|s-x|)\, d\alpha(x)$, whence

$$\int_0^\epsilon \int \frac{|s-x|}{(1+t|s-x|)}\, d\beta(x)\, dt = \int \log(1+\epsilon|s-x|)\, d\beta(x).$$

This tends to zero as $\epsilon \downarrow 0$. □

Lemma 4.27 *If γ is a gamma process with intensity measure α, then $\int |\psi|\, d\gamma < \infty$ a.s. for any measurable function $\psi\colon \mathbb{R} \to \mathbb{R}$ with $\int \log(1 + |\psi|)\, d\alpha < \infty$[4]. Furthermore, the characteristic function of this variable is given by*

$$\mathrm{E}\Big[\exp\Big(it \int \psi\, d\gamma\Big)\Big] = \exp\Big[-\int \log(1 - it\psi)\, d\alpha\Big]. \tag{4.26}$$

Proof First consider a function of the type $\psi = \sum_{i=1}^{k} a_i \mathbb{1}_{A_i}$, for constants $a_1, \ldots, a_k$ and disjoint measurable sets $A_1, \ldots, A_k$. As $\gamma(A_i) \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha(A_i), 1)$, the left side of (4.26) is then equal to

$$\prod_{i=1}^{k} \Big(\frac{1}{1 - ita_i}\Big)^{\alpha(A_i)} = \exp\Big[-\sum_{i=1}^{k} \alpha(A_i) \log(1 - ita_i)\Big].$$

This is identical to the right side of (4.26), and hence this equation is true for every simple function ψ of this type.

Given a general measurable function ψ satisfying the integrability condition of the lemma, let ψ_n be a sequence of simple, measurable functions with $0 \le \psi_n^+ \uparrow \psi^+$ and $0 \le \psi_n^- \uparrow \psi^-$. Because $|\log(1 - it)| \le \log(1 + |t|) + 2\pi$, we have $\int \log(1 - it\psi_n)\, d\alpha \to \int \log(1 - it\psi)\, d\alpha$, by the dominated convergence theorem, and hence also

$$\log \mathrm{E}\Big[\exp\Big(it \int \psi_n\, d\gamma\Big)\Big] = -\int \log(1 - it\psi_n)\, d\alpha \to -\int \log(1 - it\psi)\, d\alpha.$$

The limit on the right side is continuous at $t = 0$, and hence by Lévy's continuity theorem the sequence $\int \psi_n\, d\gamma$ converges weakly to a proper random variable whose characteristic function is given by the right side.

Because every realization of γ is a finite measure, we also have $\int \psi_n^+\, d\gamma \to \int \psi^+\, d\gamma$ a.s., by the monotone convergence theorem, and similarly for the negative parts. By the argument of the preceding paragraph applied to the positive and negative parts instead of the ψ_n, the two limit variables are proper, and hence we can take the difference to see that $\int \psi_n\, d\gamma \to \int \psi\, d\gamma$ a.s. Thus the weak limit found in the preceding paragraph is $\int \psi\, d\gamma$. We conclude that this variable possesses the right side of the preceding display as its log characteristic function. □

4.4 Characterizations

It was noted in Section 4.1.2 that the Dirichlet process is tail-free. It is also neutral in the sense defined below, and has the property that the posterior distribution of cell probabilities depends on the observations only through their cell counts. Each of these properties distinguishes the Dirichlet process from other random measures, except for the following three trivial types of random measures P:

(T1) $P = \rho$ a.s., for a deterministic probability measure ρ;
(T2) $P = \delta_X$, for a random variable $X \sim \rho$;

[4] This condition is also known to be necessary.

(T3) $P = W\delta_a + (1 - W)\delta_b$, for deterministic $a, b \in \mathfrak{X}$ and an arbitrary random variable W with values in $[0, 1]$.

The random measures (T1) and (T2) are limits of Dirichlet processes, as the precision parameter goes to infinity and zero, respectively. The third (T3) arises because random probability vectors of length two cannot be characterized by conditional independence properties.

The concept of a random measure that is tail-free relative to a given sequence of measurable (binary) partitions of the sample space (which are successive refinements) is introduced in Definition 3.11. In the following proposition we call a random measure *universally tail-free* if it is tail-free relative to *any* sequence of measurable binary partitions.

Whereas tail-freeness entails conditional independence across successive layers of ever finer partitions, "neutrality" refers to partitions at a given level. A random measure P is called *neutral* if for every finite measurable partition $\{A_1, \ldots, A_k\}$ the variable $P(A_1)$ is conditionally independent of the vector $(P(A_2)/(1 - P(A_1)), \ldots, P(A_k)/(1 - P(A_1)))$ given $P(A_1) < 1$[5]. As tail-freeness, neutrality has a constructive interpretation: after fixing the probability $P(A_1)$ of the first set, the *relative* sizes of the probabilities (or conditional probabilities) of the remaining sets of the partition are generated independently of $P(A_1)$. The neutrality principle can be applied a second time, to the coarsened partition $\{A_1 \cup A_2, A_3, \ldots, A_k\}$, to show that given the probability assigned to the first two sets the conditional probabilities of $A_3, \ldots, A_k$ are again generated independently, etc.

An alternative property is *complete neutrality*, which entails that the random variables $P(A_1)$, $P(A_2 | A_1^c)$, $P(A_3 | A_1^c \cup A_2^c)$, ... are independent (after properly providing for conditioning on events of probability zero). For a given partition in a given ordering, this appears to be slightly stronger, but when required for any ordering of the partitioning sets (or any partition) these concepts are equivalent (see Proposition G.7).

Theorem 4.28 *The following three assertions about a random measure P are equivalent:*

(i) *P is universally tail-free.*
(ii) *P is neutral.*
(iii) *P is a Dirichlet process, or one of the random measures (T1), (T2) or (T3).*

Proof That a Dirichlet process is universally tail-free is noted in Section 4.1.2; that it is also neutral follows most easily from the representation of the finite-dimensional Dirichlet distribution through gamma variables, given in Proposition G.2(i). That random measures of types (T1) and (T2) are universally tail-free and neutral follows by taking limits of Dirichlet processes. That a random measure (T3) is universally tail-free and neutral can be checked directly. Thus (iii) implies both (i) and (ii).

We now prove that (i) implies (ii). An arbitrary partition $\{A_1, \ldots, A_k\}$ can be built into a sequence of successively refined partitions $\{A_\varepsilon : \varepsilon \in \mathcal{E}^*\}$ as $B_0 = A_1$, $B_{10} = A_2$, $B_{110} = A_3$, etc. By (3.12) the probabilities of these sets satisfy $P(B_0) = V_0$, $P(B_{10}) = (1 - V_0)V_{10}$, $P(B_{110}) = (1 - V_0)(1 - V_{10})V_{110}$, etc. Under tail-freeness (i) the variables V_ε corresponding to ε of different lengths are independent. This implies that the variable $P(A_1)$ is neutral in the vector $(P(A_1), \ldots, P(A_k))$, whence P is neutral.

[5] See Section 13.4 for a related concept "neutral to the right."

It now suffices to prove that (ii) implies (iii). Under neutrality the variables $X_i = P(A_i)$ satisfy condition (i) of Proposition G.7, for any measurable partition $\{A_1, \ldots, A_k\}$. Consider two cases: there is, or there is not, a partition such that at least three variables $P(A_i)$ do not vanish with probability one. In the case that there is not, the support of almost every realization of P must consist of at most two given points, and then P is of type (T3). For the remainder of the proof assume that there is a partition $\{A_1, \ldots, A_k\}$ such that at least three variables $P(A_i)$ do not vanish with probability one. We may leave off the sets A_i such that $X_i = P(A_i)$ vanishes identically, and next apply Proposition G.7 (iv) to the vector of remaining variables. If this set consists of k variables we can conclude that this possesses a $\overline{\mathrm{Dir}}(k; M, \rho_1, \ldots, \rho_k)$-distribution, as defined preceding Proposition G.7. The vector $(\rho_1, \ldots, \rho_k)$ is the mean of this distribution; hence, ρ_i is equal to $\rho(A_i)$, for ρ the mean measure $\rho(A) = \mathrm{E}P(A)$ of P.

The parameter M is either positive and finite or takes one of the values 0 or ∞. This division in two cases is independent of the partition. Indeed, given another partition we can consider its common refinement with $\{A_1, \ldots, A_k\}$, whose probability vector must then also possess an extended Dirichlet distribution $\overline{\mathrm{Dir}}$. If this has M-parameter equal to 0 or ∞, then this distribution is discrete. This is possible if and only if the distribution of the aggregate vector $(P(A_1), \ldots, P(A_k))$ is also discrete.

The same argument combined with Proposition G.3 shows that the value of M in case $M \in (0, \infty)$ is the same for every partition. The random measure P is $\mathrm{DP}(M\rho)$ in this case.

If $M = \infty$, then the distribution of $(P(A_1), \ldots, P(A_k))$ is concentrated at the single point $(\rho(A_1), \ldots, \rho(A_k))$. If this is true for every partition, then P is of type (T1). If $M = 0$ for some partition, then the distribution of $(P(A_1), \ldots, P(A_k))$ is concentrated on the vertices of the k-dimensional unit simplex, and we can represent it as the distribution of $(\delta_X(A_1), \ldots, \delta_X(A_k))$, for a random variable X with $\mathrm{P}(X \in A_i) = \rho_i$, giving a random measure P of type (T2). □

For a tail-free process, the posterior distribution of a set depends on a random sample of observations only through the count of that set. By Theorem 3.14, this actually characterizes tail-freeness for a particular sequence of partitions. Combined with the fact that a Dirichlet process is tail-free for any sequence of partitions, we obtain the following characterization. The property may be considered a drawback of a Dirichlet process, since it indicates that the posterior distribution is insensitive to the location of the observations. In particular, the Dirichlet process provides no smoothing.

Corollary 4.29 *If for every n and measurable partition $\{A_1, \ldots, A_k\}$ of the sample space the posterior distribution of $(P(A_1), \ldots, P(A_k))$ given a random sample $X_1, \ldots, X_n$ from P depends only on $(N_1, \ldots, N_k)$, for $N_j = \sum_{i=1}^n 1\{X_i \in A_j\}$, then P is a Dirichlet process or one of the trivial processes (T1), (T2) and (T3).*

It was seen in Section 4.1.2 that a Dirichlet process is a Pólya tree. The following result shows that its distinguishing property within the class of Pólya tree processes is the additivity of the parameters of the beta-splitting variables across partitions.

Theorem 4.30 *A Pólya tree prior* $\mathrm{PT}(\mathcal{T}_m, \alpha_\varepsilon)$ *on a sequence of partitions* $\mathcal{T}_m$ *that generates the Borel sets is a Dirichlet process if and only if* $\alpha_\varepsilon = \alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}$ *for all* $\varepsilon \in \mathcal{E}^*$.

Proof That a Dirichlet process is a Pólya tree is already noted in Section 4.1.2, with the parameters given by $\alpha_\varepsilon = \alpha(A_\varepsilon)$ by (4.1.2), which are clearly additive.

Conversely, given a $\mathrm{PT}(\mathcal{T}_m, \alpha_\varepsilon)$-process P, define a set function α on the partitioning sets $\{A_\varepsilon : \varepsilon \in \mathcal{E}^*\}$ by $\alpha(A_\varepsilon) = \alpha_\varepsilon$. If the parameters α_ε are additive as in the theorem, then α is additive and extends to a finitely additive positive measure on the field generated by all partitions. By assumption the splitting variables $V_{\varepsilon 0} = P(A_{\varepsilon 0} \mid A_\varepsilon)$ of the Pólya tree process are independent and possess $\mathrm{Be}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$-distributions. In particular $P(A_0) = V_0 \sim \mathrm{Be}(\alpha_0, \alpha_1)$, whence $(V_0, 1 - V_0) \sim \mathrm{Dir}(2; \alpha_0, \alpha_1)$. By (3.12)

$$\begin{aligned}(P(A_{00}), P(A_{01}), P(A_{10}), P(A_{11})) \\ = (V_0 V_{00}, V_0(1 - V_{00}), (1 - V_0)V_{10}, (1 - V_0)(1 - V_{10})).\end{aligned}$$

This possesses a $\mathrm{Dir}(4; \alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11})$-distribution by the converse part of Proposition G.3, applied to the partition of $\{1, 2, 3, 4\}$ into the sets $\{1, 2\}$ and $\{3, 4\}$, since the required conditions (i)–(iii) are satisfied. In a similar manner $(P(A_\varepsilon) : \varepsilon \in \mathcal{E}^m)$ possesses a $\mathrm{Dir}(2^m; (\alpha_\varepsilon : \varepsilon \in \mathcal{E}^m))$-distribution, for any given m.

Thus the random measure P satisfies (4.1) for every of the partitions $\mathcal{T}_m$. Since this sequence of partitions is assumed to generate the Borel σ-field, this determines P as a Dirichlet process, by Proposition A.5. □

4.5 Mixtures of Dirichlet Processes

Application of the Dirichlet process prior requires a choice of a base measure α. It is often reasonable to choose the center measure $\bar{\alpha}$ from a specific family such as the normal family, but then the parameters of the family must still be specified. It is natural to give these a further prior. Similarly, one may put a prior on the precision parameter $|\alpha|$.

For a base measure α_ξ that depends on a parameter ξ, the Bayesian model then consists of the hierarchy

$$X_1, \ldots, X_n \mid P, \xi \overset{\text{iid}}{\sim} P, \qquad P \mid \xi \sim \mathrm{DP}(\alpha_\xi), \qquad \xi \sim \pi. \tag{4.27}$$

We denote the induced (marginal) prior on P by $\mathrm{MDP}(\alpha_\xi, \xi \sim \pi)$. Many properties of this *mixture of Dirichlet processes* (MDP) prior follow immediately from those of a Dirichlet process. For instance, any P following an MDP is almost surely discrete. However, unlike a Dirichlet process, an MDP is not tail-free.

Expressions for prior mean and variance are readily obtained from the corresponding ones for the Dirichlet process (see Proposition 4.2) and conditioning:

$$\begin{aligned}\mathrm{E}(P(A)) &= \int \bar{\alpha}_\xi(A)\, d\pi(\xi) =: \bar{\alpha}_\pi(A), \\ \mathrm{var}\,(P(A)) &= \int \frac{\bar{\alpha}_\xi(A)\bar{\alpha}_\xi(A^c)}{1 + |\alpha_\xi|}\, d\pi(\xi) + \int (\bar{\alpha}_\xi(A) - \bar{\alpha}_\pi(A))^2\, d\pi(\xi).\end{aligned}$$

The expressions for means and variances of functions $\int \psi\, dP$ generalize likewise.

Example 4.31 (Normal location) In a "robustified" location problem $X = \xi + \varepsilon$ we put priors both on the location parameter and the error distribution. The specification $\varepsilon | H \sim H$ and $H \sim \mathrm{DP}(\mathrm{Nor}(0, 1))$ is equivalent to specifying $X | P, \xi \sim P$ and $P | \xi \sim \mathrm{DP}(\mathrm{Nor}(\xi, 1))$. If $\xi \sim \mathrm{Nor}(0, 1)$, then after observing n i.i.d. observations $X_1, \ldots, X_n$,

$$\mathrm{E}(P(A) | X_1, \ldots, X_n) = \frac{1}{n+1} \int \Phi(A - \xi)\phi(\xi)\, d\xi + \frac{n}{n+1} \mathbb{P}_n(A),$$

where Φ is the cumulative distribution function of the standard normal distribution and ϕ is its density.

The posterior distribution of P in the scheme (4.27) turns out to be another MDP. Given ξ we can use the posterior updating rule for the ordinary Dirichlet process, and obtain that

$$P | \xi, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha_\xi + n\mathbb{P}_n). \tag{4.28}$$

To obtain the posterior distribution of P given $X_1, \ldots, X_n$, we need to mix this over ξ relative to its posterior distribution given $X_1, \ldots, X_n$. By Bayes's theorem the latter has density proportional to

$$\xi \mapsto \pi(\xi)\, p(X_1, \ldots, X_n | \xi).$$

Here the marginal density of $X_1, \ldots, X_n$ given ξ (the second factor) is described by the generalized Pólya urn scheme (4.13) with α_ξ instead of α. In general, this has a somewhat complicated structure due to the ties between the observations. However, for a posterior calculation we condition on the observed data $X_1, \ldots, X_n$, and know the partition that they generate. Given this information the density takes a simple form. In particular, if the observations are distinct (which happens with probability one if the observations actually follow a continuous distribution), then the Pólya urn scheme must have simply generated a random sample from the center measure $\bar{\alpha}_\xi$, in which case the preceding display becomes

$$\pi(\xi) \prod_{i=1}^{n} d\alpha_\xi(X_i) \prod_{i=1}^{n} \frac{1}{|\alpha_\xi| + i - 1},$$

for $d\alpha_\xi$ a density of α_ξ. This is somewhat awkward in a nonparametric model, and it will be seen later that it has negative implications towards consistency. Further calculations depend on the specific family and its parameterization.

The posterior mean and variance of $P(A)$ for a set A are those of an MDP process with parameters as in (4.28), and given by, for $\alpha_{\xi,n} = \alpha_\xi + n\mathbb{P}_n$ the base measure,

$$\begin{aligned}
\mathrm{E}(P(A) | X_1, \ldots, X_n) &= \int \bar{\alpha}_{\xi,n}(A)\, d\pi(\xi | X_1, \ldots, X_n) =: \tilde{P}_n(A), \qquad (4.29)\\
\mathrm{var}\,(P(A) | X_1, \ldots, X_n) &= \int \frac{\bar{\alpha}_{\xi,n}(A)\bar{\alpha}_{\xi,n}(A^c)}{1 + |\alpha_\xi| + n}\, d\pi(\xi | X_1, \ldots, X_n)\\
&\quad + \int (\bar{\alpha}_{\xi,n}(A) - \tilde{\mathbb{P}}_n(A))^2\, d\pi(\xi | X_1, \ldots, X_n).
\end{aligned}$$

If the precision $|\alpha_\xi|$ is bounded above uniformly in ξ, then $\bar{\alpha}_{\xi,n}$ and hence the posterior mean $\tilde{P}_n$ is equivalent to the empirical measure $\mathbb{P}_n$ up to order n^{-1} as $n \to \infty$, just as the posterior mean of the ordinary Dirichlet. Furthermore, the posterior variance is bounded

above by $5/(4(n+1)) \to 0$. Thus the posterior distribution of $P(A)$ contracts to a Dirac measure at the true probability of A whenever $|\alpha_\xi|$ is bounded above. (The latter condition can be relaxed substantially; see Problem 4.35.)

Typically the precision parameter M and center measure G in $\alpha = MG$ will be modeled as independent under the prior. The posterior calculation then factorizes in these two parameters. To see this, consider the following scheme to generate the parameters and observations:

(i) Generate M from its prior.
(ii) Given M generate a random partition $S = \{S_1, \ldots, S_{K_n}\}$ according to the distribution given in Proposition 4.11.
(iii) Generate G from its prior, independently of (M, S).
(iv) Given (S, G) generate a random sample of size K_n from G, independently of M, and set X_i with $i \in S_j$ equal to the jth value in this sample.

By the description of the Pólya urn scheme this indeed gives a sample $X_1, \ldots, X_n$ from the mixture of Dirichlet processes $\mathrm{MDP}(MG, M \sim \pi, G \sim \pi)$. We may now formally write the density of $(M, S, G, X_1, \ldots, X_n)$ in the form, with π abusively denoting prior densities for both M and G and p conditional densities of observed quantities,

$$\pi(M)\, p(S|\, M)\, \pi(G)\, p(X_1, \ldots, X_n|\, G, S).$$

Since this factorizes in terms involving M and G, these parameters are also independent under the posterior distribution, and the computation of their posterior distributions can be separated.

The term involving M depends on the data through K_n only (the latter variable is *sufficient* for M). Indeed, by Proposition 4.11 it is proportional to

$$M \mapsto \pi(M)\frac{M^{K_n}\Gamma(M)}{\Gamma(M+n)} \propto \pi(M) M^{K_n} \int_0^1 \eta^{M-1}(1-\eta)^{n-1}\, d\eta.$$

Rather than by (numerically) integrating this expression, the posterior density is typically computed by simulation. Suppose that $M \sim \mathrm{Ga}(a, b)$ a priori, and consider a fictitious random vector (M, η) with $0 \le \eta \le 1$ and joint (Lebesgue) density proportional to

$$\pi(M) M^{K_n} \eta^{M-1}(1-\eta)^{n-1} \propto M^{a+K_n-1} e^{-M(b-\log\eta)} \eta^{-1}(1-\eta)^{n-1}.$$

Then by the preceding display the marginal density of M is equal to its posterior density (given K_n, which is fixed for the calculation). Thus simulating from the distribution of (M, η) and dropping η simulates M from its posterior distribution. The conditional distributions are given by

$$M|\, \eta, K_n \sim \mathrm{Ga}(a + K_n, b - \log\eta), \qquad \eta|\, M, K_n \sim \mathrm{Be}(M, n). \tag{4.30}$$

We can use these in a *Gibbs sampling scheme*: given an arbitrary starting value η_0, we generate a sequence $M_1, \eta_1, M_2, \eta_2, M_3, \ldots$, by repeatedly generating M from its conditional distribution given (η, K_n) and η from its conditional distribution given (M, K_n), each time setting the conditioning variable (η or M) equal to its last value.

4.6 Modifications

In many applications the underlying probability measure is required to satisfy specific constraints, but still be nonparametric in nature. For instance, natural constraints on the error distribution in a location or regression problem are symmetry, zero mean or zero median, leading to the identifiability of the regression function. The first two subsections discuss methods of imposing constraints on a distribution obtained from a Dirichlet process. The third subsection discusses a modification of the Dirichlet process that can regulate correlation between probabilities of sets.

4.6.1 Invariant Dirichlet Process

Two methods to construct a nonparametric prior supported on the space of all probability measures symmetric about zero are to symmetrize a Dirichlet process $Q \sim \mathrm{DP}(\alpha)$ on $\mathbb{R}$ to $A \mapsto (Q(A) + Q(-A))/2$, where $-A = \{x : -x \in A\}$, and to "unfold" a Dirichlet process Q on $[0, \infty)$ to $A \mapsto (Q(A \cap [0, \infty)) + Q(-A \cap [0, \infty)))/2$. These methods can be generalized to constructing random probability measures that are invariant under a group of transformations of the sample space. Symmetry is the special case of the group $\{1, -1\}$ on $\mathbb{R}$ consisting of the identity and the reflection in zero.

Let $\mathfrak{G}$ be a compact metrizable group acting continuously on the Polish sample space $\mathfrak{X}$.[6] For a Borel measure Q and $g \in \mathfrak{G}$ the measure $A \mapsto \int \mathbb{1}\{g(x) \in A\}\, dQ(x)$ is well defined and measurable in g by Fubini's theorem; we denote it (as usual) by $Q \circ g^{-1}$. A measure P on $\mathfrak{X}$ is called *invariant* under the group $\mathfrak{G}$ if $P \circ g^{-1} = P$ for all $g \in \mathfrak{G}$. There are two basic methods of constructing invariant measures. Both utilize the (unique, both left and right) Haar probability measure on $\mathfrak{G}$, which we denote by μ.

The first method is to start with an arbitrary Borel measure Q on the sample space, and consider the measure

$$A \mapsto \int Q \circ g^{-1}(A)\, d\mu(g). \tag{4.31}$$

From the fact that $g \circ h \sim \mu$ for every $h \in \mathfrak{G}$ if g is distributed according to the Haar measure μ, it can be seen that this measure is invariant.

The second construction of an invariant random measure follows the idea of unfolding. The *orbit* of a point $x \in \mathfrak{X}$ is the set $\mathfrak{O}(x) := \{g(x) : g \in \mathfrak{G}\}$. The orbits are the equivalence classes under the relationship $x \equiv y$ if and only if $g(x) = y$ for some $g \in \mathfrak{G}$, and are invariant under the action of $\mathfrak{G}$. The set of all orbits is denoted by $\mathfrak{X}/\mathfrak{G}$ and is a measurable space relative to the *quotient σ-field*: the largest σ-field making the *quotient map* $x \mapsto \mathfrak{O}(x)$ measurable. If $\mathfrak{X}/\mathfrak{G}$ is a standard Borel space, then *Burgess's theorem* (cf. Theorem 5.6.1 of Srivastava 1998) allows us to choose a representative from each orbit to form a Borel measurable subset $\mathfrak{R}$ of $\mathfrak{X}$ that is isomorphic to $\mathfrak{X}/\mathfrak{G}$.[7] We may now start with a Borel probability measure Q on $\mathfrak{R}$ and "unfold" it to the full space as

[6] Every $g \in \mathfrak{G}$ is a map $g : \mathfrak{X} \mapsto \mathfrak{X}$, the group operation of $\mathfrak{G}$ is composition of maps $\circ$, and the *group action* $(g, x) \mapsto g(x)$ is continuous.

[7] Since $\mathfrak{G}$ is compact in our case, the full power of Burgess's theorem is not necessary. The same result can be derived from the simpler Effros theorem (cf. Theorem 5.4.2 of Srivastava 1998).

$$A \mapsto \int Q(g^{-1}(A) \cap \mathfrak{R})\, d\mu(g). \tag{4.32}$$

The formula has the same structure as (4.31), and hence the measure is invariant.

The second construction is the special case of the first where the measure Q is concentrated on the subset $\mathfrak{R} \subset \mathfrak{X}$. If the action of $\mathfrak{G}$ is *free* ($g(x) \neq x$ for any $x \in \mathfrak{X}$ and any $g \in \mathfrak{G}$ not equal to the identity), then $\mathfrak{X}$ can be identified with the product space $\mathfrak{R} \times \mathfrak{G}$ through the map $g(r) \leftrightarrow (r, g)$, and any invariant probability measure P on $\mathfrak{X}$ can be represented as $Q \times \mu$, giving $P(A) = Q \times \mu((r, g)\colon g(r) \in A) = \int Q(g^{-1}(A))\, d\mu(g)$.[8] Thus every invariant measure can be constructed using either method. This may be extended to the situation that a single point $x \in \mathfrak{X}$ is invariant under *every* $g \in \mathfrak{G}$, as is the case for the symmetry group on $\mathbb{R}$, or the group of rotations.

The preceding constructions apply equally well to *random* measures. Thus we may start with a Dirichlet process $P \sim \mathrm{DP}(\alpha)$ on the full space $\mathfrak{X}$ and make it invariant using (4.31), or take a Dirichlet process Q on $\mathfrak{R} = \mathfrak{X}/\mathfrak{G}$ and apply (4.32). Both constructions lead to an "invariant Dirichlet process," as given in the following definition. Say that a set $A \subset \mathfrak{X}$ is *invariant* if $g(A) = A$ for every $g \in \mathfrak{G}$, and call a random measure *invariant* if all its realizations are invariant.

Definition 4.32 (Invariant Dirichlet process) An invariant random measure P is said to follow an *invariant Dirichlet process* $\mathrm{IDP}(\alpha, \mathfrak{G})$ with (invariant) base measure α, if for every measurable partition of $\mathfrak{X}$ into $\mathfrak{G}$-invariant sets $A_1, \ldots, A_k$,

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)).$$

The base measure α in this definition is a finite positive Borel measure on $\mathfrak{X}$, which without loss of generality is taken to be invariant. Since by definition an invariant set A satisfies $g^{-1}(A) = A$ for any $g \in \mathfrak{G}$, it is obvious that the invariant version of Q in (4.31) coincides with Q on invariant sets. Hence if $Q \sim \mathrm{DP}(\alpha)$ for an invariant base measure, then the invariant version (4.31) is an $\mathrm{IDP}(\alpha, \mathfrak{G})$-process. Alternatively we may use the second construction (4.32), starting from a $\mathrm{DP}(\tilde{\alpha})$-process Q on the quotient space $\mathfrak{R}$, with arbitrary base measure $\tilde{\alpha}$ on $\mathfrak{R}$. For an invariant set A the integral (4.32) is simply $Q(A \cap \mathfrak{R})$ and hence the marginal distributions of the process $P(A)$ in this construction are $\mathrm{Dir}(k; \tilde{\alpha}(A_1 \cap \mathfrak{R}), \ldots, \tilde{\alpha}(A_k \cap \mathfrak{R}))$. By unfolding $\tilde{\alpha}$ these can also be written as in Definition 4.32 for an invariant measure α on $\mathfrak{X}$.[9]

It should be noted that an invariant Dirichlet process is itself not a Dirichlet process (as the name might suggest), but is in fact a mixture of Dirichlet processes. A Dirichlet process $P \sim \mathrm{DP}(\alpha)$ for an invariant base measure α possesses some distributional invariance properties, but is not invariant. For instance, in the symmetry case a Dirichlet process will split the mass to the negative and positive half lines by a $\mathrm{Be}(1/2, 1/2)$-variable, whereas the invariant Dirichlet process (4.31) makes a deterministic $(1/2, 1/2)$ split.

[8] Every $\mathfrak{G}$-invariant set A can be represented as $\tilde{A} \times \mathfrak{G}$, where $\tilde{A} = \{x \in \mathfrak{R}\colon \mathfrak{O}(x) \cap A \neq \varnothing\}$.

[9] Both constructions can be extended to any distribution on $\mathfrak{M}$, including the Pólya tree process or the Dirichlet mixture process. The correspondence between the two constructions is then not as neat as in the Dirichlet case, where the same invariant random measure is obtained by making the base measures α and $\tilde{\alpha}$ correspond.

Many of the properties of an invariant Dirichlet process are similar to those of a Dirichlet process. Suppose $P \sim \mathrm{IDP}(\alpha, \mathfrak{G})$ for an invariant measure α.

(i) For any invariant α-integrable or nonnegative measurable function ψ we have $\mathrm{E}(\int \psi \, dP) = \int \psi \, d\bar{\alpha}$. In particular, the marginal distribution of $X | P \sim P$ is $\bar{\alpha}$.
(ii) Any $\mathfrak{G}$-invariant probability measure P_0 with $\mathrm{supp}(P_0) \subset \mathrm{supp}(\alpha)$ belongs to the weak support of $\mathrm{IDP}(\alpha, \mathfrak{G})$.
(iii) If $X_1, \ldots, X_n | P \overset{\mathrm{iid}}{\sim} P$ and $P \sim \mathrm{IDP}(\alpha, \mathfrak{G})$, then

$$P | X_1, \ldots, X_n \sim \mathrm{IDP}\Big(\alpha + \sum_{i=1}^{n} \int \delta_{g(X_i)} \, d\mu(g), \mathfrak{G}\Big). \tag{4.33}$$

Unlike the Dirichlet process, the invariant Dirichlet process is not tail-free: the partitions in a tail-free representation get mixed up by the action of the group $\mathfrak{G}$.

Property (i) follows easily from the representation $P = \int Q \circ g^{-1} \, d\mu(g)$ for $Q \sim \mathrm{DP}(\alpha)$, while property (ii) can be proved along the lines of Theorem 4.15. To prove (iii), observe that by the first representation of the IDP, the model may be described hierarchically as a mixture of Dirichlet processes:

$$X_i | Q, g_1, \ldots, g_n \overset{\mathrm{ind}}{\sim} Q \circ g_i, \qquad Q \sim \mathrm{DP}(\alpha), \qquad g_i \overset{\mathrm{iid}}{\sim} \mu, \quad i = 1, \ldots, n.$$

It follows that $(g_1(X_1), \ldots, g_n(X_n)) | (Q, g_1, \ldots, g_n) \overset{\mathrm{iid}}{\sim} Q$ and hence $Q | (X_1, \ldots, X_n, g_1, \ldots, g_n) \sim \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{g_i(X_i)})$ by Theorem 4.6 applied conditionally given $g_1, \ldots, g_n$. The law P of the observations is the deterministic transformation $P = \int Q \circ g^{-1} \, d\mu(g)$ of Q and hence its posterior is obtained by substituting the posterior of Q in this integral. This posterior is certainly invariant, as it has the form (4.31). It remains to show that its marginal distribution on an invariant partition is a Dirichlet with base measure as given. Because $P(A) = Q(A)$ for every invariant set A, we may concentrate on the posterior law of $(Q(A_1), \ldots, Q(A_k))$ for an invariant partition. By the preceding, it is a mixture of $\mathrm{Dir}((\alpha + \sum_{i=1}^n \delta_{g_i(X_i)})(A_1), \ldots, (\alpha + \sum_{i=1}^n \delta_{g_i(X_i)})(A_k))$-distributions relative to the posterior distribution of $(g_1, \ldots, g_n)$. Since $\delta_{g(x)}(A) = \delta_x(A)$ for invariant set A and every g, the components of this mixture are actually identical and the mixture is a Dirichlet distribution with parameters $((\alpha + \sum_{i=1}^n \delta_{X_i})(A_1), \ldots, (\alpha + \sum_{i=1}^n \delta_{X_i})(A_1))$. For a representation through an invariant base measure we replace each term $\delta_{X_i}(A_j)$ by its invariant version $\int \delta_{g(X_i)}(A_j) \, d\mu(g)$, and we arrive at the form (4.33).

The probability measure $\int \delta_{g(x)} \, d\mu(g)$, which appears in the base measure of the posterior distribution in (4.33), is restricted to the orbit $\mathfrak{O}(x)$ of x. If the action of $\mathfrak{G}$ on $\mathfrak{X}$ is free, then it can be regarded as the "uniform distribution on this orbit".

Example 4.33 (Symmetry) Distributions on $\mathbb{R}$ that are symmetric about 0 are invariant under the reflection group $\mathfrak{G} = \{1, -1\}$ consisting of the map $x \mapsto g \cdot x$. The $\mathrm{IDP}(\alpha, \{-1, 1\})$-process is the *symmetrized Dirichlet process* considered at the beginning of the section. The posterior mean of the cumulative distribution function F based on a random sample of size n is given by, for $\alpha(x) = \alpha(-\infty, x]$,

$$\mathrm{E}(F(x)|\, X_1, \ldots, X_n) = \frac{\alpha(x) + \frac{1}{2}\sum_{i=1}^n (\mathbb{1}(X_i \le x) + \mathbb{1}(X_i \ge -x))}{|\alpha| + n}. \tag{4.34}$$

This is the natural Bayes estimator for F.

Example 4.34 (Rotational invariance) A rotationally invariant random measure on $\mathbb{R}^k$ may be constructed as an invariant Dirichlet process under the action of the group of orthogonal $k \times k$ matrices. The action of the orthogonal group is free if the origin is removed and hence the posterior base measure is the sum of the base measure and the uniform distributions on the orbits of $X_1, \ldots, X_n$. This is one of the natural examples of a group of infinite cardinality.

Example 4.35 (Exchangeability) The exchangeable measures on $\mathbb{R}^k$ are invariant measures under the action of the permutation group Σ_k of order k on the order of the coordinates of the vectors. Therefore an invariant Dirichlet process under the action of the permutation group is a suitable prior on the exchangeable measures. The permutation group action is free. The posterior expectation of the cumulative distribution function given a random sample of observations $X_1 = (X_{1,1}, \ldots, X_{1,k}), \ldots, X_n = (X_{n,1}, \ldots, X_{n,k})$ from an IDP(α, Σ_k)-prior, is given by

$$\mathrm{E}(F(y_1, \ldots, y_k)|\, X_1, \ldots, X_n) = \frac{\alpha(y_1, \ldots, y_k) + \frac{1}{k!}\sum_{i=1}^n \sum_{\sigma\in\Sigma_k} \mathbb{1}\{X_{i,\sigma(1)} \le y_1, \ldots, X_{i,\sigma(k)} \le y_k\}}{|\alpha| + n}.$$

Example 4.36 (Orthant symmetry) The group consisting of the coordinatewise multiplications $(x_1, \ldots, x_k) \mapsto (g_1 x_1, \ldots, g_k x_k)$ of vectors in $\mathbb{R}^k$ by elements $g \in \{-1, 1\}^k$ generates measures that are invariant under relocation of orthants. It generalizes the notion of symmetry on the line. The posterior expectation of the cumulative distribution function of an IDP(α, $\{-1, 1\}^k$)-random measure, given observations $X_1, \ldots, X_n$ as in the preceding example, is given by

$$\mathrm{E}(F(y_1, \ldots, y_k)|\, X_1, \ldots, X_n) = \frac{\alpha(y_1, \ldots, y_k) + \frac{1}{k!}\sum_{i=1}^n \sum_{g\in\{-1,1\}^k} \mathbb{1}\{g_1 X_{i,1} \le y_1, \ldots, g_k X_{i,k} \le y_k\}}{|\alpha| + n}.$$

4.6.2 Constrained Dirichlet Process

Given a finite measurable partition $\{A_1, \ldots, A_k\}$ of the sample space (called *control sets*), the conditional distribution of a DP(α)-process P given $(P(A_1), \ldots, P(A_k)) = w$ for a given vector $w = (w_1, \ldots, w_k)$ in the k-dimensional unit simplex is called a *constrained Dirichlet process*.

Such a process can actually be constructed as a finite mixture $P = \sum_{j=1}^k w_j P_j$ of independent DP($w_j \alpha|_{A_j}$)-processes P_j (with orthogonal supports). This follows, since by the self-similarity of Dirichlet processes (see Theorem 4.5),

$$(P(A_1), \ldots, P(A_k)) \perp\!\!\!\perp P_{A_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp P_{A_k},$$

where each P_A is a $\mathrm{DP}(\alpha|_A)$-process. The posterior distribution can be obtained as for general mixtures of Dirichlet processes (see Section 4.5). In particular the posterior mean follows (4.29).

The particular case of a binary partition $A_1 = (-\infty, 0)$, $A_2 = [0, \infty)$ of $\mathbb{R}$ with $w_1 = w_2 = 1/2$ yields a random distribution with median 0. Constraints on multiple quantiles can be placed in a similar manner.

Restriction on moments are possible as well. For instance, a random measure P with mean 0 can be generated as $P(A) = Q(A - \int x\, dQ(x))$ for $Q \sim \mathrm{DP}(\alpha)$. However, such a scheme does not lead to a simple representation in terms of an MDP, because the change of location is random and dependent on Q.

4.6.3 Penalized Dirichlet Process

That the Dirichlet process selects discrete distributions with probability one (see Theorem 4.14) may be construed as an absence of smoothing. This lack of smoothing leads to anomalous relationships between the random probabilities of collections of sets.

Given a measurable set A and an observation $X = x$ with $x \notin A$, we have under the $\mathrm{DP}(\alpha)$ prior on the distribution of X, since $\delta_x(A) = 0$,

$$\mathrm{E}(P(A)|\, X = x) = \frac{\alpha(A)}{|\alpha| + 1} < \frac{\alpha(A)}{|\alpha|} = \mathrm{E}(P(A)). \tag{4.35}$$

By itself it is natural that the expectation of $P(A)$ decreases, since $x \notin A$ makes A "less likely." However, the posterior expectation is smaller than the prior expectation no matter the proximity of x to A: the Dirichlet process does not respect closeness. This is counterintuitive if one is used to continuity, as one tends to believe that an observation in a locality enhances the probability of the locality.

Another anomalous property is the negative correlation between probabilities of disjoint sets, as is evident from (4.4), even when the sets are adjacent to each other. This blanket attachment of negative correlation again indicates a complete lack of smoothing. This may for instance be considered undesirable for random histograms. The histograms obtained from a continuous density should have similar heights across neighboring cells, and a random histogram should comply with this criterion by assigning positive correlation to probability ordinates of neighboring cells. The Dirichlet process is too flexible for this purpose. Although binning takes care of smoothing within a cell, no smoothing across bins is offered, which will result in a less efficient density estimator.

In order to rectify this problem we may introduce positive correlation among the probabilities of neighboring cells, by penalizing too much variation between neighboring cells.

Consider density estimation on a bounded interval by random histograms. Because the binning at any stage gives rise to only finitely many cell probabilities, it is enough to modify the finite-dimensional Dirichlet distribution. Instead of the finite-dimensional Dirichlet density, consider the density proportional to $(p_1, \ldots, p_k) \mapsto p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} e^{-\lambda \Delta(p)}$, where $\Delta(p)$ is a penalty term for roughness. The choice $\lambda = 0$ returns the ordinary Dirichlet distribution. Meaningful choices of the penalty $\Delta(p)$ are $\sum_{j=1}^{k-1} (p_{j+1} - p_j)^2$, which helps control variation of successive cell probabilities, $\sum_{j=2}^{k-1} (p_{j+1} - 2p_j + p_{j-1})^2$, which helps control

the second-order differences of cell probabilities, and $\sum_{j=1}^{k-1}(\log p_{j+1} - \log p_j)^2$, which helps control the ratios of successive cell probabilities. Interestingly, the resulting posterior distribution based on a random sample $X_1, \ldots, X_n | p \overset{\text{iid}}{\sim} p$ has the same form as the prior, with α_i updated to $\alpha_i + \sum_{j=1}^n \mathbb{1}\{X_j = i\}$. The posterior mean, if desired, could be obtained by numerical integration or by the Metropolis-Hastings algorithm.

4.7 Bayesian Bootstrap

The weak limit $\text{DP}(\sum_{i=1}^n \delta_{X_i})$ of the posterior distribution $\text{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$ in the non-informative limit as $|\alpha| \to 0$ is called the *Bayesian bootstrap* (BB) distribution. Its center measure is the empirical measure $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ and a random probability generated from it is necessarily supported on the observation points. In this sense it compares to Efron's bootstrap, justifying the name. In fact, both the Bayesian and Efron's bootstrap can be represented as $\sum_{i=1}^n W_i \delta_{X_i}$, where $(W_1, \ldots, W_n) \sim \text{Dir}(n; 1, \ldots, 1)$ for the Bayesian bootstrap (the uniform distribution on the unit simplex) and $n(W_1, \ldots, W_n) \sim \text{MN}_n(n; n^{-1}, \ldots, n^{-1})$ for Efron's bootstrap. Thus the Bayesian bootstrap assigns continuous weights (and in particular, puts positive weight to every observation on every realization), and leads to smoother empirical distributions of resampled variables than Efron's bootstrap, which typically assigns zero values to some observations.

When n is moderate or large, the empirical part in the Dirichlet posterior $\text{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$ will dominate, and hence the posterior distribution of a linear or non-linear function will be close to that under the Bayesian bootstrap. The biggest advantage of the Bayesian bootstrap relative to the true posterior is that samples are easily generated from the Bayesian bootstrap, allowing the approximation of the posterior distribution of any quantity of interest by simple Monte Carlo methods. The representation $W_i = Y_i / \sum_{j=1}^n Y_j$, where $Y_i \overset{\text{iid}}{\sim} \text{Exp}(1)$, for the weights of the Bayesian bootstrap (cf. Proposition G.2) is convenient here. The distributions of resampled variables may be used to compute both estimates and credible intervals.

By Theorem 4.16, the Bayesian bootstrap based on a random sample $X_1, \ldots, X_n$ from a distribution P_0 converges weakly to the degenerate measure at P_0, as $n \to \infty$. This suggests that the Bayesian bootstrap-based estimators and credible intervals have frequentist validity like Efron's bootstrap. This has indeed been shown for the more general class of *exchangeable bootstrap* processes (see van der Vaart and Wellner 1996, Chapter 3.6). We discuss a large sample property of the Bayesian bootstrap in Chapter 12.

The density of a mean functional $\mu = \int \psi \, dP$ under the Bayesian bootstrap may also be obtained analytically from Theorem 4.26. With $Y_i = \psi(X_i)$, it leads to the following formula

$$p(\mu | X_1, \ldots, X_n) = (n-1) \sum_{i=1}^n \frac{(Y_i - \mu)_+^{n-2}}{\prod_{j \neq i}(Y_i - Y_j)},$$

provided that $Y_i \neq Y_j$ a.s. for all $i \neq j$. The density is a spline function of order $(n-2)$ with knots at $Y_1, \ldots, Y_n$, and has $(n-3)$ continuous derivatives.

The multidimensional mean functional $\mu = (\int \psi_1 \, dP, \ldots, \int \psi_s \, dP)^\top$, can similarly be shown to be an $(n-s-2)$-times continuously differentiable if $Y_i := (\psi_j(X_i); j =$

$1, \ldots, s)^\top$, $i = 1, \ldots, n$, do not lie on any lower-dimensional hyperplane. The resulting density function is called the s-variate B-spline with knots at $Y_1, \ldots, Y_n$. Some recursive formula permit writing an s-variate B-spline in terms of $(s-1)$-variate ones; see Micchelli (1980). Nevertheless, the resulting density is very complicated, so it is easier to adopt a simulation based approach. Choudhuri (1998) discussed a simulation technique exploiting the log-concavity of the density of multivariate mean functional of the Bayesian bootstrap, and also showed that the sampling based approach with N resamples estimates a credible set within $O(N^{-1/(s+2)} \log N)$ in the metric defined by the Lebesgue measure of symmetric difference of sets.

The Bayesian bootstrap may be smoothed by convolving it with a kernel giving a substitute for posterior of density function which is computable without any MCMC technique. However, the bandwidth parameter must be supplied depending on the sample size.

The weights $W_i = Y_i / \sum_{j=1}^n Y_j$ in the Bayesian bootstrap may be generalized to Y_is other than i.i.d. standard exponential, leading to what is called the *Bayesian bootstrap clone*. The particular choice Ga(4, 1) for the distribution of the Y_i leads to $(W_1, \ldots, W_n) \sim \text{Dir}(n; 4, \ldots, 4)$. This is sometimes used to achieve higher-order agreement between credibility and confidence of a Bayesian bootstrap interval. The Bayesian bootstrap clone in turn is a special case of the *exchangeable bootstrap*, which only assumes that the weights are exchangeable.

4.8 Historical Notes

Ferguson (1973) introduced the Dirichlet process, which put Bayesian nonparametrics on firm ground by exhibiting practical feasibility of posterior computations. The importance of the Dirichlet process in Bayesian nonparametrics is comparable to that of the normal distribution in probability and general statistics. Ferguson (1973) also obtained prior moments, showed posterior conjugacy, and discussed the naive construction as well as the construction through a gamma process, based on an earlier idea of Ferguson and Klass (1972). Blackwell and MacQueen (1973) noticed the connection between the Pólya urn scheme and the Dirichlet process and presented the construction of a Dirichlet process through the Pólya urn scheme. The stick-breaking representation first appeared in McCloskey (1965), and was rediscovered in the statistical context by Sethuraman (1994). It has turned out to be very helpful in dealing with complicated functionals of a Dirichlet process. The ϵ-Dirichlet process approximation was studied by Muliere and Tardella (1998) and Gelfand and Kottas (2002). The discreteness property of the Dirichlet process was inferred by Ferguson (1973) from the pure jump nature of the gamma process, and by Blackwell and MacQueen (1973) as a consequence of the Pólya urn scheme. The proof presented here is possibly due to Savage. The weak support of the Dirichlet process was characterized by Ferguson (1973). Weak convergence of a sequence of Dirichlet processes were discussed in Sethuraman and Tiwari (1982). Ganesh and O'Connell (2000) observed that the posterior distribution of a Dirichlet process satisfies a large deviation principle under the weak topology (see Problem 4.14). Finer moderate deviation properties of a random probability measure "exchangeable with respect to a sequence of partitions," centered at the mean and scaled by an appropriately growing sequence, were studied by Eichelsbacher and Ganesh (2002). Such random processes include the Dirichlet process as well as Pólya tree processes. Convergence of the

multinomial-Dirichlet process to the general Dirichlet process was shown by Ishwaran and Zarepour (2002), and is especially useful for computation using BUGS; see Gilks et al. (1994). The marginal distribution of a Dirichlet sample and the distribution of the number of distinct observations were discussed by Antoniak (1974). Identifiability of the Dirichlet base measure based on infinitely many random samples and mutual singularity under non-atomicity were shown by Korwar and Hollander (1973). Tails of a Dirichlet process were studied by Doss and Sellke (1982). Formulas for the distribution of the mean functional of a Dirichlet process, including numerical approximations to the analytic formulas, were obtained by Cifarelli and Regazzini (1990), Regazzini et al. (2002) using the method to study ratios developed in Gurland (1948), and generalized to normalized independent increment processes by Regazzini et al. (2003). Several characterizations of the Dirichlet process were stated by Ferguson (1974) without proof. Antoniak (1974) studied mixtures of Dirichlet processes. Dalal (1979) studied symmetrized Dirichlet processes. The Dirichlet process conditioned to have median zero has been used in many contexts, such as in Doss (1985a,b) in modeling of error distributions in a location problem. The penalization idea to the Dirichlet distribution and its use in histogram smoothing was discussed by Hjort (1996). The Bayesian bootstrap was introduced by Rubin (1981) and later studied by many authors. Gasparini (1996) and Choudhuri (1998) studied the multidimensional mean functional in detail. Large sample properties of the Bayesian bootstrap were studied in Lo (1987) and Præstgaard and Wellner (1993); also see van der Vaart and Wellner (1996), Chapter 3.7.

Problems

4.1 Show that if $P \sim \mathrm{DP}(\alpha)$ on $\mathfrak{X}$ and $\psi: \mathfrak{X} \to \mathfrak{Y}$ is measurable, then $P \circ \psi^{-1} \sim \mathrm{DP}(\alpha \circ \psi^{-1})$ on $\mathfrak{Y}$.

4.2 Find an expression for $\mathrm{E}[\prod_{i=1}^{k} \int \psi_i \, dP]$ for $k = 3, 4, \ldots$

4.3 Let $P \sim \mathrm{DP}(\alpha)$ on a product space $\mathfrak{X} \times \mathfrak{Y}$, and let P_1 and $P_{2|1}$, and $\bar{\alpha}_1$ and $\bar{\alpha}_{2|1}$, the marginal distributions on the first coordinate and the conditional distributions of the second given the first coordinate of P and $\bar{\alpha}$.

(a) Show that $P_1 \sim \mathrm{DP}(|\alpha|\bar{\alpha}_1)$. [Hint: Apply Problem 4.1 with the projection function or the stick-breaking representation of a Dirichlet process.]

(b) Show that almost surely $P_{2|1}$ is given by $\mathrm{DP}(\alpha(\{X\} \times \mathfrak{Y})\bar{\alpha}_{2|1})$ if $\alpha(\{X\} \times \mathfrak{Y}) > 0$, and is given by δ_Y for $Y \sim \bar{\alpha}_{2|1}$ otherwise, and is distributed independently of P_1. [Hint: For the first case, use the fact that P conditioned on $\{X\} \times \mathfrak{Y}$ is a Dirichlet process in view of the self-similarity property. Use the stick-breaking representation in the second case and observe that there can be at most one support point in this case.]

Note that if $P_1(\{x\}) = 0$, then $P_{2|1}(\cdot \mid x)$ can be defined arbitrarily. Furthermore, for $\alpha_1(\{x\}) = 0$, the measure δ_Y with $Y \sim \bar{\alpha}_{2|1}(\cdot|X)$ can be viewed as a limiting Dirichlet process with precision parameter $\alpha_1(\{x\}) = 0$ and center measure $\bar{\alpha}_{2|1}(\cdot|X)$. Hence the two different forms of the conditional distribution can be unified with a broader notation.

4.4 (Two sample testing, Ferguson 1973) Let $X \sim F$ and $Y \sim G$ independently, and suppose that we want to estimate $\mathrm{P}(X \leq Y)$ with m independent X observations and

n independent Y observations. Put independent Dirichlet priors on F and G. Find an expression for the Bayes estimate of $\mathrm{P}(X \le Y)$ and show that it reduces to the classical Mann-Whitney U-statistic under a noninformative limit.

4.5 (Tolerance interval, Ferguson 1973) Let F be an unknown c.d.f. on $\mathbb{R}$ and $0 < q < 1$ be given. Consider the problem of estimating the qth quantile $t_q(F)$ of F with the loss function $L(a, t_q(F)) = pF(a) + q\mathbb{1}\{a < t_q(F)\}$, where $0 < p < 1$ is also given, indicating relatively high penalty for underestimating $t_q(F)$. Let $F \sim \mathrm{DP}(MG)$.

Show that the Bayes risk of the no-data estimation problem is given by the rth quantile of G, where r is the unique minimizer of

$$u \mapsto pu + q \int_0^q \frac{\Gamma(M)}{\Gamma(Mu)\Gamma(M(1-u))} z^{Mu-1}(1-z)^{M(1-u)-1} dz,$$

to be denoted by $r(p, q, M)$.

Now if a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ is available, obtain the Bayes estimator of $t_q(F)$.

4.6 Prove Theorem 4.6 by directly checking that

$$\mathrm{E}(\mathrm{DP}_{\alpha+\delta_X}(C)\mathbb{1}\{X \in A\}) = \mathrm{P}\{P \in C, X \in A\},$$

where $C \in \mathscr{M}$, $A \in \mathscr{X}$, X has marginal distribution $\bar{\alpha}$ on the left-hand side, and (X, P) follows the Dirichlet model $X|\, P \sim P$ and $P \sim \mathrm{DP}(\alpha)$. [Hint: Consider C a finite dimensional projection set and then show that the two sides lead to the same sets of moments.]

4.7 Theorem 4.5 can be generalized to more than two localities. If $P \sim \mathrm{DP}(\alpha)$ and $A_1, \ldots, A_k \in \mathscr{X}$ are disjoint measurable sets such that $\alpha(A_i) > 0$, then $\{P(A_i): i = 1, \ldots, k\} \perp\!\!\!\perp P_{A_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp P_{A_k}$ and $P_{A_i} \sim \mathrm{DP}(\alpha_{|A_i})$, for $i = 1, \ldots, k$.

4.8 Using the distributional equation (4.21), construct a Markov chain on $\mathfrak{M}$ which has stationary distribution $\mathrm{DP}(\alpha)$.

4.9 Prove Proposition 4.3 using (4.21).

4.10 Prove the posterior updating formula using the stick-breaking representation.

4.11 Let $E_1, E_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Exp}(1)$. Define $\Gamma_0 = 0$, $\Gamma_k = \sum_{j=1}^k E_j$, $k \in \mathbb{N}$. Let $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$. Show that $P := \sum_{k=1}^\infty (e^{-\Gamma_{k-1}/M} - e^{-\Gamma_k/M})\delta_{\theta_k}$ is a random probability measure having distribution $\mathrm{DP}(MG)$.

4.12 (Muliere and Tardella 1998) Obtain sharp probability inequalities for the number of terms present in an ϵ-Dirichlet process defined in Subsection 4.3.3.

4.13 Consider a stick-breaking process identical to the Dirichlet process except that the stick-breaking distribution is generalized to $\mathrm{Be}(a, b)$. Find expressions for prior expectation and variance.

4.14 (Ganesh and O'Connell 2000) Large deviation principle: If $X_i \overset{\text{iid}}{\sim} P_0$, then for all Borel subsets $\mathcal{B}$ of $\mathfrak{M}$,

$$-\inf_{P \in \mathcal{B}^o} K(P_0; P) \le \liminf_{n\to\infty} \frac{1}{n} \log \mathrm{DP}_{\alpha+\sum_{i=1}^n \delta_{X_i}}(\mathcal{B})$$
$$\le \limsup_{n\to\infty} \frac{1}{n} \log \mathrm{DP}_{\alpha+\sum_{i=1}^n \delta_{X_i}}(\mathcal{B}) \le -\inf_{P \in \bar{\mathcal{B}}} K(P_0; P), \quad a.s.[P_0^\infty].$$

4.15 Write down the density of the marginal distribution of a sample of size $n = 3$ from a Dirichlet process explicitly. Develop a systematic notation for the description of the density of the joint distribution for a general n.

4.16 Obtain (4.14) using mathematical induction.

4.17 (Bell number) Compute the number of terms in Proposition 4.7; (this is known as a *Bell number* or *Bell's exponential number*).

4.18 Derive an asymptotic formula for $\mathrm{var}(K_n)$ when n and M both can vary (cf. Proposition 4.8(i)).

4.19 Show Part (iii) of Proposition 4.8 by computing moment-generating functions.

4.20 For $M = 1, 10$, and $n = 30, 100$, numerically evaluate the exact distribution of K_n, the number of distinct observations in a sample of size n from a Dirichlet process using Proposition 4.9.

Evaluate the same expression through a simulation scheme involving independent Bernoulli variables.

Compare these with the normal and the Poisson approximations described in Subsection 4.1.5.

4.21 Show that the *distinct* values in a random sequence $X_1, X_2, \ldots$ sampled from a distribution generated from a $\mathrm{DP}(\alpha)$-prior with atomless base measure form an i.i.d. sequence from $\bar{\alpha}$. (Thus define $\tau_1 = 1$ and $\tau_j = \min\{n: \sum_{i=1}^{n} D_i = j\}$ for $D_i = 1$ if $X_i \notin \{X_1, \ldots, X_{i-1}\}$, and put $Y_j = X_{\tau_j}$.)

4.22 While sampling from a Dirichlet process with precision parameter M, show that the maximum likelihood estimator for M based on data $K_n = k$ is unique, consistent and asymptotically equivalent to $K_n / \log n$.

4.23 For the center measure $G = \mathrm{Unif}(0, 1)$, show that the rth raw moment of m_F is given by $\int_0^1 B(\frac{1}{2}; Mx^{1/r}, (1 - x^{1/r}))\, dx$, where $B(\cdot; a, b)$ stands for the incomplete beta function.

Numerically evaluate and plot the mean, variance, skewness and kurtosis of this distribution as functions of M.

For $M = 2, 20$, numerically evaluate and plot the density function of m_F.

4.24 Obtain the distribution of the pth quantile of a random $F \sim \mathrm{DP}(MG)$.

4.25 In Theorem 4.26, show that if G is symmetric, so will be H about the same point.

4.26 (Diaconis and Kemperman 1996) If $P \sim \mathrm{DP}(G)$, where G is $\mathrm{Unif}(0, 1)$, then show that the density of $\int x\, dP$ is given by $e\pi^{-1} \sin(\pi y) y^{-y} (1 - y)^{y-1} \mathbb{1}\{0 < y < 1\}$.

4.27 (Regazzini et al. 2002) If $P \sim \mathrm{DP}(G)$, where G is $\mathrm{Exp}(\lambda)$, then show that the density of $\int x\, dP$ is given by $\pi^{-1} y^{-1} \sin(\pi - \pi e^{-\lambda y}) \exp[e^{-\lambda y} \mathrm{PV}(\int_{-\infty}^{\lambda y} t^{-1} e^t\, dt)]$, where PV stands for the principal value of the integral.

4.28 (Yamato 1984, Cifarelli and Regazzini 1990, Lijoi and Regazzini 2004) Let $P \sim \mathrm{DP}(G)$, where G has Cauchy density given by $\sigma\pi^{-1}[1 + \sigma^2(x - \mu)^2]^{-1}$. Using the identity $\int \log|x-y| dG(x) = \frac{1}{2} \log[\{1+\sigma^2(y-\mu)^2\}/\sigma^2]$, show that the distribution of the Dirichlet mean $\int x\, dP$ is again G. Further, the property characterizes the Cauchy distribution.

4.29 Let $\alpha = \theta_0\delta_0 + \theta_1\delta_1$. If $P \sim \mathrm{DP}(\alpha)$, then $\int x\, dP = P(\{1\}) \sim \mathrm{Be}(\theta_1, \theta_0)$. Derive the same result from the general formula for the distribution of a Dirichlet mean.

4.30 Let $P \sim \text{DP}(MG)$ and $L = \int x\, dP(x)$. In view of (4.21), it follows that $L =_d V\theta + (1-V)L$, where $\theta \sim G$ and $V \sim \text{Be}(1, M)$. Use this relation to simulate a Markov chain whose stationary distribution is the distribution of L. For G chosen as normal, Cauchy, exponential or uniform, and $M = 1, 10$, obtain the distribution of L by the MCMC method, and compare with their theoretical distribution given by Theorem 4.26.

4.31 (Lo 1991) Let $X_1, X_2, \ldots | P \overset{\text{iid}}{\sim} P$, where P has an arbitrary prior distribution. Suppose that there exists a nonstochastic sequence $a_n > 0$ and nonstochastic probability measures P_n such that $\text{E}(P | X_1, \ldots, X_n) = (1-a_n)P_n + a_n n^{-1} \sum_{i=1}^n \delta_{X_i}$ for all n. Show that $a_n = n/(a_1^{-1} + n - 1)$, $P_n = \text{E}(P)$ and $P \sim \text{DP}(a_1 \text{E}(P))$.

4.32 For a location mixture of a Dirichlet process, obtain expressions for covariances of probabilities of sets.

4.33 Using conditioning arguments, show that

$$\text{MDP}\Big(M+1, \frac{MG+\delta_\theta}{M+1}, \theta \sim G\Big) \equiv \text{DP}(MG) \equiv \text{MDP}(M, G, \theta \sim \delta_0).$$

Thus unlike the DP, the parameters of MDP are not necessarily identifiable.

4.34 Using the monotonicity of distribution functions, show that if M_θ is bounded in a mixture of Dirichlet process model, then the posterior distribution of P is consistent at any true distribution P_0 in the Kolmogorov-Smirnov sense.

4.35 Consider samples generated from a atomless true distribution P_0. Let the modeled P be given a mixture of Dirichlet process distribution $\text{MDP}(MG_\xi, (M, \xi) \sim \pi_M \times \pi_\xi)$. Show that, with probability one, the posterior distribution of M is always equal to its prior π_M, and the relative weight $M/(M+n)$ of the first term in the posterior base measure $\bar{\alpha}_{\xi,n}$ in (4.29) always converges to zero. Conclude that the posterior mean is a consistent estimator.

Under the same condition, show that the posterior variance converges to zero. Thus the posterior distribution of any set, as well as the posterior distribution of P in the Kolmogorov-Smirnov sense, is consistent. [Hint: For the last part use a uniformization argument as in the proof of Glivenko-Cantelli theorem.]

4.36 Obtain the posterior distribution for a constrained Dirichlet process prior.

4.37 Describe a stick-breaking representation of an invariant Dirichlet process.

4.38 Let $X_1, \ldots, X_n | P \overset{\text{iid}}{\sim} P$ and $P \sim \text{DP}(\sum_{i=1}^N \delta_{\theta_i})$, the Bayesian bootstrap distribution based on the sample $(\theta_1, \ldots, \theta_N)$. Assume that $\theta_1, \ldots, \theta_n$ are all distinct, and let K_n be the number of distinct elements in $\{X_1, \ldots, X_n\}$. Show that K_n follows the *Bose-Einstein distribution*:

$$\text{P}(K_n = k) = \frac{\binom{N}{k}\binom{n-1}{k-1}}{\binom{N+n-1}{N-1}}, \quad k = 1, \ldots, \min(n, N).$$

4.39 (Ishwaran and James 2001) Consider a truncated stick-breaking process prior Π_N defined by $P_N = \sum_{k=1}^N V_k \delta_{\theta_k}$, where $V_k = [\prod_{j=1}^{k-1}(1-Y_j)]Y_k$, $Y_1, \ldots, Y_{N-1} \overset{\text{iid}}{\sim} \text{Be}(1, M)$, $Y_N = 1$ and $\theta_1, \ldots, \theta_N \overset{\text{iid}}{\sim} G$. Let $\psi(\cdot | \theta)$ be a parametric family of probability densities, $X_i | Z_i \sim \psi(\cdot | Z_i)$, where $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P_N$ and $P_N \sim \Pi_N$. Let $m_N^{(n)}$ stand for the marginal distribution of $(X_1, \ldots, X_n)$. Let $m_\infty^{(n)}$ be the marginal

distribution of $(X_1, \ldots, X_n)$ when $X_i \mid Z_i \sim \psi(\cdot \mid Z_i)$, $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P$ and $P \sim \Pi := \text{DP}(MG)$. Show that $d_{TV}(m_N^{(n)}, m_\infty^{(n)}) \approx 4ne^{-(N-1)/M}$.

4.40 Let $U_1, \ldots, U_{n-1} \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. Define the order statistics $U_{0:n-1} = 0$, $U_{n:n-1} = 1$ and $U_{j:n-1}$ the jth smallest of $\{U_1, \ldots, U_{n-1}\}$. If $W_j = U_{j:n-1} - U_{j-1:n-1}$, $j = 1, \ldots, n$, show that $(W_1, \ldots, W_n) \sim \text{Dir}(n; 1, \ldots, 1)$, the weights of the Bayesian bootstrap distribution.