G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh, eds., *Statistics and Probability: Essays in Honor of C.R. Rao*

A NOTE ON THE DIRICHLET PROCESS

D. BASU and R. C. TIWARI
The Florida State University, Tallahassee, FL, U.S.A.

Written mainly for its pedagogical interest, this expository note is concerned primarily with the question of existence of a Dirichlet process. A random probability measure on a measurable space $(\mathcal{X}, \mathcal{A})$ is a stochastic process $\{P(A): A \in \mathcal{A}\}$ – a collection of random variables indexed by the measurable sets in $\mathcal{X}$ – such that almost every realization of the process is a probability measure on $(\mathcal{X}, \mathcal{A})$. Given a finite measure α on $(\mathcal{X}, \mathcal{A})$, a Dirichlet process D^{α} is a random probability measure on $(\mathcal{X}, \mathcal{A})$ such that, for every partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$ into a finite number of measurable sets, the joint distribution of the random variables $(P(A_1), P(A_2), \ldots, P(A_k))$ is a singular Dirichlet distribution with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$.

Part one of this article deals with the familiar case where $\mathcal{X}$ is a finite set. Properties of the k-dimensional Dirichlet distribution are so exposited as to motivate Blackwell's (1973) constructive definition of the Dirichlet process. In part two, the case where $(\mathcal{X}, \mathcal{A})$ is a Borel space is discussed in some detail.

1. Introduction

This report is concerned primarily with the question of existence of a Dirichlet process on a Borel space $(\mathcal{X}, \mathcal{A})$. Part one of this report deals with the familiar case when $\mathcal{X}$ is a finite set. Properties of the k-dimensional Dirichlet distribution are so exposited as to motivate Blackwell's (1973) constructive definition of the Dirichlet process. In part two, the case where $(\mathcal{X}, \mathcal{A})$ is a Borel space is discussed in some detail.

Section 2 deals with Bayesian (parametric) inference and the family of natural conjugate priors. Section 3 is devoted to some characterizations of the Dirichlet distribution and elucidations of its useful properties. In Section 4, the results of Section 3 are extended and it is shown that there exists a Dirichlet process on $(\mathcal{X}, \mathcal{A})$, when $\mathcal{X}$ is a finite or a countably infinite set.

In Section 5, some preliminary material on the Dirichlet process is presented. Some useful results on a Borel space are given in Section 6. Blackwell's (1973) construction of a Dirichlet process on a Borel space $(\mathcal{X}, \mathcal{A})$ is discussed in Section 7. Section 8 is devoted to the study of some properties of a Dirichlet process.

PART I: THE DIRICHLET DISTRIBUTION

2. Bayesian inference and natural conjugate priors

Let X be an observable random variable (r.v.) with a statistical model that is characterized by a probability density function (p.d.f.) $f(\cdot|\theta)$, where $\theta \in \Theta$ is the

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_31,

unknown parameter of interest. Before any data are collected, a Bayesian represents his prior opinion about θ by a distribution on Θ, called a prior distribution. If he observes X n-times and denotes by $\boldsymbol{D}_1^n=(x_1, x_2, \ldots, x_n)$ the data so obtained, then his opinion about θ is represented by the distribution of θ given $\boldsymbol{D}_1^n$, called the posterior distribution.

Let q be the prior p.d.f. of θ. The posterior p.d.f. of θ given $\boldsymbol{D}_1^n$, namely $q(\cdot|\boldsymbol{D}_1^n)$, is expressed by the relation

$$q(\theta|\boldsymbol{D}_1^n)\propto q(\theta)L(\theta|\boldsymbol{D}_1^n). \tag{2.1}$$

Here, $L(\theta|\boldsymbol{D}_1^n)$ is the likelihood function of θ at point $\boldsymbol{D}_1^n$, and the proportionality symbol, $\propto$, is used to indicate that the posterior p.d.f. of θ given $\boldsymbol{D}_1^n$ is equal to the right side of (2.1) divided by the factor $\int_\Theta L(\theta|\boldsymbol{D}_1^n)q(\theta)\mathrm{d}\theta$ which does not involve θ.

This is how new knowledge, obtained through data, may be combined with prior knowledge. The Bayesian continually updates his knowledge as more observations are taken. Clearly,

$$\begin{aligned} q(\theta|\boldsymbol{D}_1^{n+m}) &\propto q(\theta)L(\theta|\boldsymbol{D}_1^{n+m}) \\ &\propto q(\theta)L(\theta|\boldsymbol{D}_1^n)L(\theta|\boldsymbol{D}_{n+1}^{n+m}) \\ &\propto q(\theta|\boldsymbol{D}_1^n)L(\theta|\boldsymbol{D}_{n+1}^{n+m}). \end{aligned}$$

Thus, the opinion $q(\theta|\boldsymbol{D}_1^{n+m})$ based on data $\boldsymbol{D}_1^{n+m}$ may be regarded as the posterior based on data $\boldsymbol{D}_{n+1}^{n+m}$ and prior $q(\theta|\boldsymbol{D}_1^n)$. This process of updating opinion may go through many stages.

It is clear from the relation (2.1) that the change of opinion about θ after the data are obtained is effected through the likelihood function. In the context of a chosen statistical model, a Bayesian will regard the likelihood function as the sole reservoir of all the relevant information about the parameter that is contained in the data. This is usually stated as: *The Likelihood Principle*. Two sets of data generating equivalent likelihood functions contain the same relevant information about the parameter. Two likelihood functions are said to be equivalent if one of them is a constant multiple of the other, where the constant may depend on the data. [See Basu (1975) for more on the likelihood principle.]

In many situations, it is convenient to access the prior within a family, C, of distributions. The class C should be large enough to accommodate various shades of opinion about the parameter. Further, if $q\in C$ is a prior p.d.f. of θ, then the posterior p.d.f. $q(\theta|\boldsymbol{D}_1^n)$ of θ given the data $\boldsymbol{D}_1^n$ ought to be in a simple computable form. If $q(\theta|\boldsymbol{D}_1^n)\in C$ for all $q\in C$ and data $\boldsymbol{D}_1^n$, then C is called a *conjugate family of priors*.

It frequently happens that a conjugate family of priors naturally coexists with a given statistical model for the observable r.v. X. Suppose the model is such that there exists an $n_0>0$ with the property that for all data $\boldsymbol{D}_1^n$ with $n\geqslant n_0$ the induced likelihood function $L(\theta|\boldsymbol{D}_1^n)$ is integrable (with respect to some integrating measure μ) over the parameter space Θ. Consider then the family C_0 of p.d.f.'s $q(\theta)$ of the

form:

$$q(\theta)=\frac{L(\theta|\boldsymbol{D}_1^n)}{\int_\theta L(\theta|\boldsymbol{D}_1^n)\,\mathrm{d}\mu(\theta)}.$$

If the prior p.d.f. $q(\theta)$ corresponds to a so-called prior data, $\boldsymbol{D}_1^n=(y_1, y_2,\ldots, y_n)$, then with the current data $\boldsymbol{D}_1^m=(x_1, x_2,\ldots, x_m)$. The posterior p.d.f. $q(\theta|\boldsymbol{D}_1^m)$ will correspond to the likelihood function $L(\theta|\boldsymbol{D}_1^{n+m})$, where $\boldsymbol{D}_1^{n+m}$ is the extended data $(y_1, y_2,\ldots, y_n, x_1, x_2,\ldots, x_m)$. Thus for each prior $q\in C_0$, the posterior $q(\cdot|\boldsymbol{D}_1^m)$ belongs to C_0 for all possible current data $\boldsymbol{D}_1^m$.

The natural conjugate family C_0 of prior distributions takes on a simple form when, irrespective of the sample size n, there exists a sufficient statistic $T=(T_1, T_2,\ldots, T_k)$ of fixed and small dimension $k(k\geqslant 1)$. Then,

$$L(\theta|\boldsymbol{D}_1^n)\propto H_n(\theta, T_1, T_2,\ldots, T_k), \tag{2.2}$$

where $T_1, T_2,\ldots, T_k$ are functions of $\boldsymbol{D}_1^n$. In this case, the natural conjugate family C_0 of prior distributions is characterized by $k+1$ superparameters, namely, particular values of $T_1, T_2,\ldots, T_k$ and n.

For example, suppose each observation on an observable r.v. X belongs to one of the $k+1$ mutually exclusive and collectively exhaustive categories. Let $p_i(0<p_i<1)$ be the probability that an observation belongs to the ith category, $i=1,2,\ldots, k+1$, where $\sum_{i=1}^{k+1}p_i=1$. We may regard $(p_1, p_2,\ldots, p_k)$ as the model parameters. Suppose X is observed n times and let $\boldsymbol{D}_1^n$ be the data $(x_1, x_2,\ldots, x_n)$ collected. Furthermore, let n_i denote the number of x's that belong to the ith category, $i=1,2,\ldots, k+1$. Then each n_i is a non-negative integer and $\sum_{i=1}^{k+1}n_i=n$. Also, since $\sum_{i=1}^{k+1}n_i=n$, we may regard $T=(n_1, n_2,\ldots, n_k)$ as the k-dimensional sufficient statistic. Before the data are collected, $(n_1, n_2,\ldots, n_k)$ are r.v.'s having a multinomial distribution with parameters n and $(p_1, p_2,\ldots, p_k)$. The likelihood function $L(p_1, p_2,\ldots, p_k|\boldsymbol{D}_1^n)\propto \prod_{i=1}^k p_i^{n_i}(1-\sum_{i=1}^k p_i)^{n-\sum_{i=1}^k n_i}$, which is of the form (2.2) with $\theta=(p_1, p_2,\ldots, p_k)$ and $T=(n_1, n_2,\ldots, n_k)$.

The natural conjugate family of prior distributions for the parameters $(p_1, p_2,\ldots, p_k)$ is then the family $\mathcal{C}_0$ of distributions with p.d.f.'s of the form

$$q(p_1, p_2,\ldots, p_k)\propto \prod_{i=1}^k p_i^{a_i-1}\left(1-\sum_{i=1}^k p_i\right)^{a_{k+1}-1};\quad p_i>0,\quad i=1,2,\ldots, k,$$

$\sum_{i=1}^k p_i<1$, and each a_i is a positive integer.

Clearly, for any prior p.d.f. $q(p_1, p_2,\ldots, p_k)\in\mathcal{C}_0$ and any data $\boldsymbol{D}_1^n$ the posterior p.d.f.,

$$q(p_1, p_2,\ldots, p_k|\boldsymbol{D}_1^n)\propto \prod_{i=1}^k p_i^{a_i+n_i-1}\left(1-\sum_{i=1}^k p_i\right)^{a_{k+1}+(n-\sum_{i=1}^k n_i)-1}$$

and so $q(p_1, p_2,\ldots, p_k|\boldsymbol{D}_1^n)\in\mathcal{C}_0$.

The natural conjugate family $\mathcal{C}_0$ of prior distributions for the parameters $(p_1, p_2, \ldots, p_k)$ is a subfamily of the family of the Dirichlet distributions, defined in the next section.

3. The Dirichlet distribution

This section is devoted to the study of the family of the Dirichlet distributions, as the natural conjugate family for the parameters of a multinomial distribution, and its characterizations. The Dirichlet distribution is defined as follows.

Definition 3.1. Let $\alpha_i > 0, i = 1, 2, \ldots, k+1$. The r.v.'s $(Y_1, Y_2, \ldots, Y_k)$ are said to have a Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, denoted by $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, if the joint distribution of $(Y_1, Y_2, \ldots, Y_k)$ has the p.d.f. $f(y_1, y_2, \ldots, y_k) = \text{const}\, y_1^{\alpha_1 - 1} y_2^{\alpha_2 - 1}, \ldots, y_k^{\alpha_k - 1}(1 - y_1 - \cdots - y_k)^{\alpha_{k+1} - 1}$, over the k-dimensional simplex S_k defined by the inequalities $y_i > 0, i = 1, 2, \ldots, k, \sum_{i=1}^{k} y_i < 1$.

More generally, in the above definition one may assume $\alpha_i \geq 0$ for each i, and $\sum_{i=1}^{k+1} \alpha_i > 0$. However, if $\alpha_i = 0$ for some i, then the corresponding $Y_i = 0$ with probability one.

For $k = 1$, the Dirichlet distribution $D(\alpha_1, \alpha_2)$ for Y_1 is the familiar Beta distribution with parameters α_1 and α_2, Beta(α_1, α_2). The proof of the following basic proposition has already been outlined in the previous section.

Proposition 3.1. *Let $D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$ be the prior probability model for the parameters $(p_1, p_2, \ldots, p_k)$ in the statistical model of a $k+1$ valued r.v. X. Then, with n independent observations on X, giving rise to the sample frequencies $n_1, n_2, \ldots, n_{k+1}$ for the $k+1$ values of the r.v. X, the posterior distribution of $(p_1, p_2, \ldots, p_k)$ will be $D(\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_{k+1} + n_{k+1})$.*

The rest of this section is devoted to some characterizations of the Dirichlet distribution and elucidations of some of its more useful properties.

First of all, note that if we define Y_{k+1} as $1 - \sum_{i=1}^{k} Y_i$ then the joint distribution of $(Y_1, Y_2, \ldots, Y_{k+1})$ is singular with respect to the $k+1$ dimensional Lebesgue measure λ_{k+1} on R_{k+1}. The support of this singular distribution is the k-dimensional simplex E_{k+1} defined by the inequalities $y_i > 0, i = 1, 2, \ldots, k+1, \sum_{i=1}^{k+1} y_i = 1$. The joint p.d.f. (with respect to the k-dimensional Lebesgue measure on E_{k+1}) of the $k+1$ variables may be neatly represented as const $\prod_{i=1}^{k+1} y_i^{\alpha_i - 1}$.

The following result follows immediately.

Proposition 3.2. *If $i_1, i_2, \ldots, i_k$ is any sequence of distinct integers from the set $\mathcal{X} - \{1, 2, \ldots, k+1\}$ then $(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_k}) \sim D(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_{k+1}})$.*

A characterization of the Dirichlet distribution in terms of mutually independent Beta r.v.'s is given by

Proposition 3.3. *Let* $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$. *Let* $U_1 = Y_1$, *and* $U_i = Y_i/(1 - Y_1 - \ldots - Y_{i-1})$, $i = 2,3,\ldots,k$. *Then* $U_i \sim Beta(\alpha_i, \sum_{j=i+1}^{k+1} \alpha_j)$, $i = 1,2,\ldots,k$, *and* $U_1, U_2, \ldots, U_k$ *are mutually independent.*

Proof. The joint p.d.f. of $(Y_1, Y_2, \ldots, Y_k)$ is $f(y_1, y_2, \ldots, y_k) = \text{const} \prod_{i=1}^{k} y_i^{\alpha_i - 1}(1 - \sum_{i=1}^{k} y_i)^{\alpha_{k+1} - 1}; (y_1, y_2, \ldots, y_k) \in S_k$. Consider the one–one transformation of S_k onto the k-dimensional cube $(0,1)^k$ given by the relation $y_1 = u_1$, $y_i = u_i \prod_{j=1}^{i-1}(1 - u_j)$, $i = 2,3,\ldots,k$, the Jacobian of transformation being $\prod_{i=1}^{k-1}(1 - u_i)^{k-i}$. It follows then that the joint p.d.f. of $(U_1, U_2, \ldots, U_k)$ is $g(u_1, u_2, \ldots, u_k) = \text{const} \prod_{i=1}^{k} u_i^{\alpha_i - 1}(1 - u_i)^{\alpha_{i+1} + \alpha_{i+2} + \cdots + \alpha_{k+1} - 1}$.

That the converse of Proposition 3.3 is true, is established by reversing the above chain of arguments.

Remark 3.1. As a by-product of the above proposition we immediately have that the r.v. $Y_1 \sim \text{Beta}(\alpha_1, \alpha - \alpha_1)$ where $\alpha = \sum_{i=1}^{k+1} \alpha_i$. This fact together with Proposition 3.2 then gives:

Corollary 3.1. *If* $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, *then* $Y_i \sim Beta(\alpha_i, \alpha - \alpha_i)$, $i = 1,2,\ldots,k$.

This corollary may be generalized by using the converse of Proposition 3.3 to:

Corollary 3.2. *If* $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, *then* $(Y_1, Y_2, \ldots, Y_r) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_r, \alpha - \sum_{i=1}^{r} \alpha_i)$.

A more general extension is then given by Proposition 3.2 and the converse of Proposition 3.3 as

Corollary 3.3. *For any subset* $\{i_1, i_2, \ldots, i_r\}$ *of* $\mathfrak{X}$, $(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_r}) \sim D(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_r}, \alpha - \sum_{j=1}^{r} \alpha_{i_j})$.

The following proposition follows immediately from Proposition 3.3 and its converse.

Proposition 3.4. *Let* $(Y_1, Y_2, \cdots, Y_k) \sim D(\alpha_1, \alpha_2, \cdots, \alpha_{k+1})$. *Then, for any integer* r *such that* $2 \leqslant r \leqslant k$,

$$\left(\frac{Y_r}{1 - Y_1 - \cdots - Y_{r-1}}, \frac{Y_{r+1}}{1 - Y_1 - \cdots - Y_{r-1}}, \ldots, \frac{Y_k}{1 - Y_1 - \cdots - Y_{r-1}} \right)$$

is independent of $(Y_1, Y_2, \ldots, Y_{r-1})$. *Also,*

$$\left(\frac{Y_r}{1 - Y_1 - \cdots - Y_{r-1}}, \frac{Y_{r+1}}{1 - Y_1 - \cdots - Y_{r-1}}, \ldots, \frac{Y_k}{1 - Y_1 - \cdots - Y_{r-1}} \right) \sim D(\alpha_r, \alpha_{r+1}, \ldots, \alpha_{k+1}).$$

The Dirichlet distribution can also be characterized in terms of mutually independent Gamma r.v.'s. This is given by the following

Proposition 3.5. *Let $Z_1, Z_2, \ldots, Z_{k+1}$ be mutually independent Gamma r.v.'s with the common scale parameter $\beta>0$ and possibly different shape parameters $\alpha_i>0, i=1,2,\ldots,k+1$. Let $Z=\Sigma_i Z_i$ and $Y_i=Z_i/Z, i=1,2,\ldots,k$. Then $(Y_1, Y_2, \ldots, Y_k)\sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$. Also, $(Y_1, Y_2, \ldots, Y_k)$ is independent of Z.*

Proof. The joint p.d.f. of the Z_i's is

$$f(z_1, z_2, \ldots, z_{k+1}) \propto \exp\left\{-\beta \sum_i z_i\right\} \Pi_i z_i^{\alpha_i - 1}, \quad z_i>0, i=1,2,\ldots,k+1.$$

Consider the transformation $z=\Sigma_i z_i$, $y_i=z_i/z, i=1,2,\ldots,k$, the reverse transformation being $z_i=zy_i, i=1,2,\ldots,k$, and $z_{k+1}=z(1-\Sigma_{i=1}^k y_i)$. The Jacobian of transformation is z^k. It then follows that the joint p.d.f. of $(Z, Y_1, Y_2, \ldots, Y_k)$ is $g(z, y_1, y_2, \ldots, y_k) \propto \exp\{-\beta z\} z^{\alpha-1} \Pi_{i=1}^k y_i^{\alpha_i - 1}(1-\Sigma_{i=1}^k y_i)^{\alpha_{k+1}-1}$.

That $(Y_1, Y_2, \ldots, Y_k)$ is independent of Z in the above proposition can also be seen as follows. Regard the parameters $(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, where each $\alpha_i>0$, as known constants and $\beta>0$ as the unknown parameter. With $(Z_1, Z_2, \ldots, Z_{k+1})$ as the sample, $Z=\Sigma_i Z_i$ is a complete sufficient statistic. The vector valued statistic $((Z_1/Z),(Z_2/Z),\ldots,(Z_k/Z))$ is scale invariant. Since $\beta>0$ is a scale parameter, it follows that $((Z_1/Z),(Z_2/Z),\ldots,(Z_k/Z))$ is ancillary statistic. Hence it follows (Basu (1955)) that $((Z_1/Z),(Z_2/Z),\ldots,(Z_k/Z))=(Y_1, Y_2, \ldots, Y_k)$ is independent of Z.

The converse of Proposition 3.5 may be stated as follows. The proof is omitted.

Proposition 3.6. *If Z is a r.v. having a Gamma distribution with shape parameter $\alpha_i>0$ and scale parameter $\beta>0$, denoted by $Z\sim G(\alpha, \beta)$, and if Z is independent of $(Y_1, Y_2, \ldots, Y_k)$, where $(Y_1, Y_2, \ldots, Y_k)\sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$; then the r.v.'s $Z_i=ZY_i$, $i=1,2,\ldots,k$, and $Z_{k+1}=Z(1-\Sigma_{i=1}^k Y_i)$ are mutually independent with $Z_i\sim G(\alpha_i, \beta)$, $i=1,2,\ldots,k+1$.*

Remark 3.2. Proposition 3.4 can also be proved using the Basu theorem (Basu (1955)) and the Gamma characterization of the Dirichlet distribution.

For more on the Dirichlet distribution, see Wilks (1962).

4. Further properties of the Dirichlet distribution

Suppose A is any subset of the set $\mathcal{X}=\{1,2,\ldots,k+1\}$ and $(y_1, y_2, \ldots, y_{k+1})$ is any given point in the simplex E_{k+1}. Define $P(A)=\Sigma_{i\in A} y_i$. Then, P is a probability measure on $\mathcal{X}$, and P is identified by the point $(y_1, y_2, \ldots, y_{k+1})$ in E_{k+1}. Thus, E_{k+1} represents a class of probability measures on $\mathcal{X}$. If $(Y_1, Y_2, \ldots, Y_{k+1})$ is a

random point in E_{k+1} then $P(A)=\Sigma_{i\in A}Y_i$ is random probability measure of the set A, and P is a random probability measure on $\mathfrak{X}$. A random probability measure on $\mathfrak{X}$ can, therefore, be viewed as a probability measure on E_{k+1}.

From now on, we consider the particular case where the r.v.'s $(Y_1, Y_2, \ldots, Y_{k+1})$ have a singular Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$. To simplify matters, we introduce $k+1$ mutually independent Gamma r.v.'s $Z_1, Z_2, \ldots, Z_{k+1}$ with $Z_i \sim G(\alpha_i, \beta)$, $i=1,2,\ldots,k+1$, and regard $Y_i=Z_i/Z$, where $Z=Z(\mathfrak{X})=\Sigma_i Z_i$. For any subset A of $\mathfrak{X}$ we write $Z(A)=\Sigma_{i\in A}Z_i$, and $\alpha(A)=\Sigma_{i\in A}\alpha_i$. Then $P(A)=\Sigma_{i\in A}Y_i=(Z(A))/(Z(\mathfrak{X}))$.

For any subsets A and B of $\mathfrak{X}$, let $P(A|B)$ be the (random) conditional probability of A given B defined as

$$P(A|B)=\begin{cases}P(AB)/P(B), & \text{if } P(B)>0\\ 0, & \text{if } P(B)=0.\end{cases}$$

Note that for any collection $(A_1, A_2, \ldots, A_n)$ of disjoint subsets of $\mathfrak{X}$, the r.v.'s $Z(A_1), Z(A_2), \ldots, Z(A_n)$ are mutually independent and $Z(A_i)\sim G(\alpha(A_i), \beta)$, $i=1,2,\ldots,n$.

The following is a general property of the Dirichlet distribution.

Proposition 4.1. *Let $\mathfrak{X}$ be partitioned into non-empty subsets $A_1, A_2, \ldots, A_{m+1}$, $1\leqslant m\leqslant k$. Then, $(P(A_1), P(A_2), \ldots, P(A_m))\sim D(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_{m+1}))$. Also, $(P(A_1), P(A_2), \ldots, P(A_m))$ is independent of $Z(\mathfrak{X})$.*

This follows immediately from Proposition 3.5.

For $m=1$, the above proposition can be stated as: For any two disjoint subsets A_1 and A_2 of $\mathfrak{X}$, $(Z(A_1))/(Z(A_1\cup A_2))$ is independent of $Z(A_1\cup A_2)$. Also, $(Z(A_1))/(Z(A_1\cup A_2))\sim \text{Beta}(\alpha(A_1), \alpha(A_2))$, and $Z(A_1\cup A_2)\sim G(\alpha(A_1\cup A_2), \beta)$. As a direct consequence of Proposition 4.1, we have the following:

Corollary 4.1. *The marginal distribution of the sum of any r, $1\leqslant r\leqslant k$, r.v.'s $Y_{i_1}, Y_{i_2}, \ldots, Y_{i_r}$ is Beta$(\alpha_{i_1}+\alpha_{i_2}+\cdots+\alpha_{i_r}, \alpha-\Sigma_{j=1}^{r}\alpha_{i_j})$, where $i_1, i_2, \ldots, i_r$ is any sequence of r distinct integers from $\mathfrak{X}=\{1,2,\ldots,k+1\}$.*

For any subset B of $\mathfrak{X}$ we denote by $\bar{B}$ the complement of B. The next result is a preliminary to Proposition 4.2.

Lemma 4.1. *Let B_1 and B_2 be any two subsets of $\mathfrak{X}$. Then, the r.v.'s $P(B_1)$, $P(B_2|B_1)$, $P(B_2|\bar{B}_1)$ are mutually independent.*

Proof. It suffices to show that the r.v.'s $(Z(B_1))/(Z(\mathfrak{X}))$, $(Z(B_1B_2))/(Z(B_1))$, $(Z(\bar{B}_1B_2))/(Z(\bar{B}_1))$ are mutually independent. Since the r.v.'s $Z(B_1B_2)$, $Z(B_1\bar{B}_2)$, $Z(\bar{B}_1B_2)$ and $Z(\bar{B}_1\bar{B}_2)$ are mutually independent, the pairs $(Z(B_1B_2), Z(B_1\bar{B}_2))$ and $(Z(\bar{B}_1B_2), Z(\bar{B}_1\bar{B}_2))$ of r.v.'s are independent both "within and between". Applying Proposition 4.1 to each of the two pairs we have then that the pairs $((Z(B_1B_2))/(Z(B_1)), Z(B_1))$ and $((Z(\bar{B}_1B_2))/(Z(\bar{B}_1)), Z(\bar{B}_1))$ of r.v.'s are

independent both "within and between". Thus, the r.v.'s $(Z(B_1B_2))/(Z(B_1))$, $(Z(\bar{B}_1B_2))/(Z(\bar{B}_1))$, $Z(B_1)$ and $Z(\bar{B}_1)$ are mutually independent. Applying Proposition 4.1 to the last two r.v.'s we finally conclude that the r.v.'s $(Z(B_1B_2))/(Z(B_1))$,$(Z(\bar{B}_1B_2))/(Z(\bar{B}_1))$, $(Z(B_1))/(Z(\mathcal{X}))$, and $Z(\mathcal{X})$ are mutually independent.

For any subset B of $\mathcal{X}$, define B^t as B when $t=1$ and as $\bar{B}$ when $t=0$, $i=1,2,\ldots,n$. We now state Proposition 4.2.

Proposition 4.2. *For any collection $B_1, B_2, \ldots, B_n$ of subsets of $\mathcal{X}$, the (2^n-1) r.v.'s $P(B_1), \{P(B_2|B_1^{t_1})\}, \ldots, \{P(B_n|B_1^{t_1}B_2^{t_2}\ldots B_{n-1}^{t_{n-1}})\}$ are mutually independent with $P(B_1) \sim Beta(\alpha(B_1), \alpha(\bar{B}_1))$ and $P(B_{r+1}|B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}) \sim Beta(\alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}B_{r+1}), \alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}\bar{B}_{r+1}))$, $r=1,2,\ldots,n-1$.*

Proof. For $n=2$ the proof is established in Lemma 4.1. The rest follows by induction.

Proposition 4.3. *If for any collection $B_1, B_2, \ldots, B_n$ of subsets of $\mathcal{X}$ the (2^n-1) r.v.'s $P(B_1), \{P(B_2|B_1^{t_1})\}, \ldots, \{P(B_n|B_1^{t_1}B_2^{t_2}\ldots B_{n-1}^{t_{n-1}})\}$ are mutually independent with $P(B_1) \sim Beta(\alpha(B_1), \alpha(\bar{B}_1))$ and $P(B_{r+1}|B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}) \sim Beta(\alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}B_{r+1}), \alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}\bar{B}_{r+1}))$, $r=1,2,\ldots,n-1$; then the joint distribution of 2^n r.v.'s $\{P(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$ is singular Dirichlet with parameters $\{\alpha(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$.*

Proof. Let $\{Y_{t_1t_2\ldots t_n}\}$ be a collection of 2^n r.v.'s having a singular Dirichlet distribution with parameters $\{\alpha(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$. Let $\mathfrak{y}=\{(t_1t_2\ldots t_n)\}$ be the set consisting of 2^n points. Define $C_i=\{(t_1t_2\ldots t_n): t_i=1\}$, $i=1,2,\ldots,n$, and for any subset C of $\mathfrak{y}$ write

$$Q(C)=\sum_{(t_1t_2,\ldots,t_n)\in C} Y_{t_1t_2\ldots t_n}.$$

Then Q is a random probability measure on $\mathfrak{y}$, and the joint distribution of 2^n r.v.'s $\{Q(C_1^{t_1}, C_2^{t_2}, \ldots, C_n^{t_n})\}$ is singular Dirichlet with parameters $\{\alpha(B_1^{t_1}, B_2^{t_2}, \ldots, B_n^{t_n})\}$. Furthermore, it follows from Proposition 4.2 that the (2^n-1) r.v.'s $Q(C_1), \{Q(C_2,|C_1^{t_1})\}, \ldots, \{Q(C_n|C_1^{t_1}C_2^{t_2}\ldots C_{n-1}^{t_{n-1}}\}$ are mutually independent with $Q(C_1) \sim \text{Beta}(\alpha(B_1), \alpha(\bar{B}_1))$ and $Q(C_{r+1}|C_1^{t_1}C_2^{t_2}\ldots C_r^{t_r}) \sim \text{Beta}(\alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}B_{r+1}), \alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}\bar{B}_{r+1}))$, $r=1,2,\ldots,n-1$. Thus, the joint distribution of (2^n-1) r.v.'s $\{P(B_1), P(B_2|B_1), P(B_2|\bar{B}_1), \ldots\}$ is the same as the joint distribution of (2^n-1) r.v.'s $\{Q(C_1), Q(C_2|C_1), Q(C_2|\bar{C}_1), \ldots\}$. It then follows that the joint distribution of 2^n r.v.'s $\{P(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$ is the same as the joint distribution of 2^n r.v.'s $\{Q(C_1^{t_1}C_2^{t_2}\ldots C_n^{t_n})\}$.

Remark 4.1. If the collection $\{(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$ is such that every single point subset of $\mathcal{X}=\{1,2,\ldots,k+1\}$ appears in the collection (in other words, $B_1, B_2, \ldots, B_n$ is a separating sequence), then the random probability measure P on $\mathcal{X}$ is a Dirichlet process (see Definition 5.1).

Up to this stage, only finite dimensional Dirichlet distributions were considered. A more general Dirichlet distribution may be defined as follows.

Let $\{\alpha_n\}$ be a sequence of numbers satisfying $\alpha_i > 0$ for each i and $\Sigma\alpha_i < \infty$. A sequence $\{Y_n\}$ of r.v's such that $0 < Y_i < 1$ for each i and $\Sigma Y_i = 1$ is said to have a Dirichlet distribution with parameters $\{\alpha_n\}$ if for each $k, (Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_k, \Sigma_{i=k+1}^{\infty}\alpha_i)$.

In the above definition one may assume $\alpha_i \geq 0$ for each i and $0 < \Sigma\alpha_i < \infty$. However, if $\alpha_i = 0$ for some i, then the corresponding $Y_i = 0$ with probability one.

The Dirichlet process on a countable infinite set $\mathcal{X} = \{1, 2, \ldots\}$ may now be defined as follows. Let $\{\alpha_n\}$ be a convergent sequence of non negative numbers. Consider the separating sequence $\{B_n\}$ of sets in $\mathcal{X}$ where $B_n = \{n\}$. Consider a sequence of mutually independent Beta r.v.'s $\{P(B_1), P(B_2|B_1), P(B_2|\bar{B}_1), \ldots\}$, where $P(B_1) \sim \text{Beta}(\alpha(B_1), \alpha(\bar{B}_1))$, $P(B_2|B_1) \sim \text{Beta}(\alpha(B_1B_2), \alpha(B_1\bar{B}_2))$, $P(B_2|\bar{B}_1) \sim \text{Beta}(\alpha(\bar{B}_1B_2), \alpha(\bar{B}_1\bar{B}_2))$, and so on. Then from Remark 4.1 it defines a Dirichlet distribution on $\{1, 2, \ldots, n\}$ for every n in a consistent manner.

In Part II we demonstrate how this constructive approach to the Dirichlet process also works in the case where $\mathcal{X}$ is a Borel space.

PART II: THE DIRICHLET PROCESS

5. Dirichlet process preliminaries

In the Bayesian analysis of non-parametric problems there is a sequence $\{X_n\}$ of independent identically distributed (i.i.d.) random variables with a common unknown distribution P, that is, given $P = \mathrm{P}$ the X_n's are i.i.d. P. Here P is regarded as the parameter and belongs to $\mathcal{P}$, the class of all probability measures on a given space $(\mathcal{X}, \mathcal{A})$. A prior for P is a probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$, where $\sigma(\mathcal{P})$ is the smallest σ-field of subsets of $\mathcal{P}$ such that the map $\mathrm{P} \to \mathrm{P}(A)$ from $\mathcal{P}$ into $[0, 1]$ is $\sigma(\mathcal{P})$-measurable $\forall A \in \mathcal{A}$. This prior may be viewed as a stochastic process $\{P(A): A \in \mathcal{A}\}$ whose sample functions are probability measures on $(\mathcal{X}, \mathcal{A})$. As in the parametric case, a class of processes satisfying the following properties is desired:

(I) It is wide enough to accommodate various shades of opinion about P.

(II) If a prior is selected from this class, then the posterior distribution given a sample of observations from P is manageable analytically, and it belongs to the class, i.e. the class is closed under "the Bayesian operation".

The class of Dirichlet process introduced by Ferguson (1973) is especially convenient since it satisfies the properties (I) and (II).

Let us look back at the definition of a random probability measure as given in the abstract of the paper. A random probability measure P on an arbitrary measurable

space $(\mathcal{X}, \mathcal{A})$ may be viewed as a measurable map from a probability space $(\Omega, \mathcal{F}, \mu)$ to the space $(\mathcal{P}, \sigma(\mathcal{P}))$. It may also be regarded as a transition function from $(\Omega, \mathcal{F})$ into $(\mathcal{X}, \mathcal{A})$. In otherwords, $P(\cdot, \cdot)$ is a measurable map from $\Omega \times \mathcal{A}$ into $[0,1]$ such that (i) for every ω in Ω, $P(\omega, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{A})$, and (ii) for every set A in $\mathcal{A}$, $P(\cdot, A)$ is a measurable function on $(\Omega, \mathcal{F})$, i.e. $P(\cdot, A)$ is a random variable with values in $[0,1]$. The distribution of P, namely μP^{-1}, is the prior probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$. Therefore, this paper can be thought of as dealing with a class of random probability measures, with a class of stochastic processes, or with a class of prior probabilities. The Dirichlet process is defined as follows.

Definition 5.1. Let α be a finite measure on $(\mathcal{X}, \mathcal{A})$. A Dirichlet process D^{α} is a random probability measure on $(\mathcal{X}, \mathcal{A})$ such that, for every partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$ into a finite number of measurable sets, the joint distribution of random variables $(P(A_1), P(A_2), \ldots, P(A_k))$ is singular Dirichlet with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$.

Ferguson (1973) shows through the Kolmogorov extension theorem that there exists a probability measure on $([0,1]^{\mathcal{A}}, \sigma([0,1]^{\mathcal{A}}))$ yielding the above finite dimensional Dirichlet distributions. Here $[0,1]^{\mathcal{A}}$ is the product space having for each of its factors the closed unit interval $[0,1]$, there being as many factors as elements of $\mathcal{A}$. Also, $\sigma([0,1]^{\mathcal{A}})$ is the product σ-field for $[0,1]^{\mathcal{A}}$, the σ-field generated by the measurable cylinders having a finite base. Viewing $[0,1]^{\mathcal{A}}$ as a class of set functions defined on $\mathcal{A}$ with values in $[0,1]$, each set in $\sigma([0,1]^{\mathcal{A}})$ may be defined by restrictions on a countable collection $\{p(A_n); n = 1, 2, \ldots\}$, where $\{A_n\}$ is a given countable subset of $\mathcal{A}$ and p denotes an element of $[0,1]^{\mathcal{A}}$. Observe that with $\mathcal{A}$ uncountable the single point sets in $[0,1]^{\mathcal{A}}$ are not in $\sigma([0,1]^{\mathcal{A}})$. Also, the class $\mathcal{P}$ of all probability measures on $(\mathcal{X}, \mathcal{A})$ does not belong to $\sigma([0,1]^{\mathcal{A}})$; it is not determined by a countable number of restrictions when $\mathcal{A}$ is uncountable. Thus a statement like "a Dirichlet process gives probability one to the class $\mathcal{P}$" is not meaningful.

Berk and Savage (1979) discuss other technical problems relating to measurability in addition to some fundamental difficulties with Ferguson's definition of a Dirichlet process. However, as proved by Blackwell (1973), none of these difficulties arise when $(X, \mathcal{A})$ is a Borel space. A full discussion on Blackwell's construction is given in Section 7. Some basic results on a Borel space is given in the next section.

6. Some useful results on a Borel space

Let $(\mathcal{X}, \mathcal{A})$ be a Borel space — a complete separable metric space, $\mathcal{X}$, with $\mathcal{A}$ being the σ-field generated by the open subsets of $\mathcal{X}$. Since $\mathcal{X}$ is a separable metric space, $\mathcal{A}$ is countably generated. We may, therefore, assume that there exists a countable field $\mathcal{B} = \{B_1, B_2, \ldots\}$ such that its Borel extension is $\mathcal{A}$. The family $\mathcal{B}$ forms a separating sequence, i.e. for any distinct points x_1 and x_2 in $\mathcal{X}$ there exists a set $B_n \in \mathcal{B}$ which contains either x_1 or x_2 but not both.

Consider all sequences $t = (t_1, t_2, \ldots)$ such that each $t_i = 0$ or 1. Let $T = \{t\}$ be the class of all such sequences and $\mathcal{T}$ be the σ-field for T — the σ-field generated by the

cylinders having a finite base, the so called Kolmogorov's sets. Then $(T, \mathcal{T})$ is a Borel space.

Consider the map $\xi: \mathcal{X} \to T$ defined by $\xi(x)=(\xi_1(x), \xi_2(x), \ldots)$, where ξ_i is the indicator of B_i, $i=1,2, \ldots$. Notice that ξ is a measurable map since each coordinate is measurable. Also, ξ is one–one since any two x's that agree on all ξ_i's are the same. Then we have the following

Lemma 6.1. *$\xi(\mathcal{X})$ is a Borel subset of T.*

The proof of the above lemma is a direct consequence of *The Kuratowski Theorem* (Parthasarathy (1967), Theorem 3.9). If ρ is a one–one measurable map from a Borel subset E_1 of a complete separable metric space into another complete separable metric space with $\rho(E_1)=E_2$, then E_2 is a Borel set. Also, the map ρ from E_1 onto E_2 is one–one and bimeasurable.

Let $[0,1]^\infty$ be the set of all sequences $(w_1, w_2, \ldots)$ with $0 \leqslant w_n \leqslant 1$ for each n. Note that $([0,1]^\infty, \sigma([0,1]^\infty))$ is a Borel space. Consider the map $\eta: \mathcal{P} \to [0,1]^\infty$ defined as $\eta(\mathrm{P})=\{\mathrm{P}(B_1), \mathrm{P}(B_2), \ldots\}$. The map is one–one since $\mathcal{B}$ is a field. And it is measurable since each of its coordinates is a measurable map of $\mathcal{P}$ into $[0,1]$. Let $\mathcal{S}$ be the range of the map η. From Kuratowski's theorem it then follows:

Lemma 6.2. *$\mathcal{S}$ is a Borel subset of $[0,1]^\infty$, and the map η from $\mathcal{P}$ onto $\mathcal{S}$ is one–one and bimeasurable.*

For the remainder of this section we need only to assume $\mathcal{A}$ is countably generated and contains the single point sets. Let $\mathcal{B}$ be a countable field that generates $\mathcal{A}$.

The following result is a preliminary to Proposition 6.1.

Lemma 6.3. *Let* P *be a probability measure on $(\mathcal{X}, \mathcal{A})$. Then, for every $x \in \mathcal{X}$ we have*

$$\inf_{n:\, x \in B_n} \mathrm{P}(B_n)=\mathrm{P}(\{x\}).$$

Proof. Let $C_1, C_2, \ldots$ be an enumeration of sets in $\mathcal{B}$ that contain x. Then, $\bigcap_{n=1}^{\infty} C_n=\{x\}$. Defining $D_n=C_1 C_2 \ldots C_n$, we have $\mathrm{P}(D_n) \downarrow \mathrm{P}(\{x\})$.

Consider the set of all pairs (P, x) such that $\mathrm{P} \in \mathcal{P}$ and $x \in \mathcal{X}$, that is, the product space $\mathcal{P} \times \mathcal{X}$. Equip this with product σ-field $\sigma(\mathcal{P}) \times \mathcal{A}$. Let $E=\{(\mathrm{P}, x): \mathrm{P}(\{x\})>0\}$. The following result is useful.

Proposition 6.1. $E \in \sigma(\mathcal{P}) \times \mathcal{A}$.

Proof. It suffices to show that the map $(\mathrm{P}, x) \to \mathrm{P}(\{x\})$ from $\mathcal{P} \times \mathcal{X}$ into $[0,1]$ is $\sigma(\mathcal{P}) \times \mathcal{A}$-measurable. Consider the map $H: R_2 \to R_1$ defined as

$$H(a, b)=\begin{cases} a, & \text{if } b \neq 0, \\ 1, & \text{if } b=0. \end{cases}$$

Now, the map $P \to P(B_n)$ is $\sigma(\mathcal{P})$-measurable $\forall n$, and the map $x \to \xi_n(x)$ is $\mathcal{A}$-measurable $\forall n$. Also, observe that H is a measurable map from R_2 into R_1. Therefore, the map $(P, x) \to H(P(B_n), \xi_n(x))$ is $\sigma(\mathcal{P}) \times \mathcal{A}$-measurable $\forall n$, and so is $\inf_n H(P(B_n), \xi_n(x))$. Note that $\inf_n H(P(B_n), \xi_n(x)) = \inf_{n: x \in B_n} P(B_n) = P(\{x\})$, where the last equality follows from Lemma 6.3. Thus the map $(P, x) \to P(\{x\})$ is $\sigma(\mathcal{P}) \times \mathcal{A}$-measurable.

For $P \in \mathcal{P}$, let $E_P = \{x : P(\{x\}) > 0\}$ be the P-section of E. E_P is the discrete mass points of P. Also, if P is a discrete probability measure, then E_P is the support of P. We have the following:

Proposition 6.2. *The map $\psi : \mathcal{P} \to [0, 1]$ defined as $\psi(P) = P(E_P)$, the discrete mass of P, is $\sigma(\mathcal{P})$-measurable.*

Note that the maps $P \to P_d$, the discrete part of P, and $P_d \to P_d(\mathcal{X})$ are measurable. Thus the map $P \to P_d(\mathcal{X}) = P(E_P)$ is $\sigma(\mathcal{P})$-measurable.

Corollary 6.1. *The class $\mathcal{P}_0$ of all discrete probability measures on $(\mathcal{X}, \mathcal{A})$ is $\sigma(\mathcal{P})$-measurable.*

Observe that $\mathcal{P}_0 = \{P \in \mathcal{P} : \psi(\mathcal{P}) = 1\} \in \sigma(\mathcal{P})$. For further details refer to Dubins and Freedman (1964).

7. Existence of a Dirichlet process

We proceed to prove the existence of a Dirichlet process D^α on a Borel space $(\mathcal{X}, \mathcal{A})$ corresponding to any finite measure α on $\mathcal{A}$. Choose and fix a countable field $\mathcal{B} = \{B_1, B_2, \dots\}$ of sets in $\mathcal{X}$ such that $\mathcal{B}$ is a generator of the σ-field $\mathcal{A}$. The map $x \to \xi(x) = \{\xi_1(x), \xi_2(x), \dots\}$, where ξ_n is the indicator of B_n, is then a one–one bimeasurable map of $(\mathcal{X}, \mathcal{A})$ into $(T, \mathcal{T})$. A probability measure Q on $(T, \mathcal{T})$ defines a probability measure $P = Q\xi$ on $(\mathcal{X}, \mathcal{A})$ provided $Q[\xi(\mathcal{X})] = 1$.

To simplify our notations we denote a typical point $(t_1, t_2, \dots, t_n)$ of the product space $T_n = \{0, 1\}^n$ by s_n. By $s_n 0$ we denote the point in T_{n+1} that is obtained by augmenting s_n by 0, that is, $s_n 0 = (t_1, t_2, \dots, t_n, 0)$, and similarly for $s_n 1$. Finally, we denote by $[s_n]$ the cylinder set of all points in T whose first n coordinates form the vector s_n. For example, [0] is the set of all $t \in T$ such that $t_1 = 0$.

It is easily seen that a probability measure Q on $(T, \mathcal{T})$ is uniquely defined by a sequence of blocks ω of numbers in the closed unit interval [0, 1]:

$$\omega = \{u, (u_0, u_1), (u_{00}, u_{01}, u_{10}, u_{11}), \dots\}, \tag{7.1}$$

where $u = Q([1])$, $u_0 = Q([0, 1]|[0])$, $u_1 = Q([1, 1]|[1])$, $u_{00} = Q([0, 0, 1]|[0, 0])$ and so on, a typical term of the $(n+1)$th block of the sequence of blocks being

$$u_{s_n} = Q([s_n 1]|[s_n]), \quad s_n \in T_n.$$

Let Ω denote the space of all sequence of blocks ω with its coordinates lying in [0, 1].

The probability measure on $(T, \mathcal{T})$ that coexists with each $\omega \in \Omega$ is denoted by Q_ω. If Ω is equipped with the product σ-field $\sigma(\Omega)$, then the map $\omega \to Q_\omega$ defines a transition function from $(\Omega, \sigma(\Omega))$ to $(T, \mathcal{T})$. If $(\Omega, \sigma(\Omega))$ is equipped with a probability measure μ, then we have a random probability measure on $(T, \mathcal{T})$ which we denote by Q_μ. How do we choose μ so that $Q_\mu \xi$ is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α?

For an arbitrary but fixed n, consider the partition $\{B^{s_n}: s_n \in T_n\}$ of $\mathcal{X}$, where by B^{s_n} we denote the set $B_1^{t_1} B_2^{t_2} \dots B_n^{t_n}$. If P_μ is D^α on $(\mathcal{X}, \mathcal{A})$, then the joint distribution of the 2^n r.v.'s $\{P_\mu(B^{s_n}): s_n \in T_n\}$ is singular Dirichlet with parameters $\{\alpha(B^{s_n}): s_n \in T_n\}$. Invoking Proposition 4.2 we then have

$$P_\mu(B_1), P_\mu(B_2|\bar{B}_1), P_\mu(B_2|B_1), P_\mu(B_3|\bar{B}_1\bar{B}_2), \dots \tag{7.2}$$

are mutually independent random variables with

$$P_\mu(B_1) \sim \text{Beta}\big(\alpha(B_1), \alpha(\bar{B}_1)\big), \quad \text{and}$$

$$P_\mu(B_{m+1}|B^{s_m}) \sim \text{Beta}\big(\alpha(B^{s_m 1}), \alpha(B^{s_m 0})\big), \quad s_m \in T_m, \quad m = 1, 2, \dots n. \tag{7.3}$$

Observe that the map $\xi: \mathcal{X} \to T$ transforms B_1 to [1], B^{s_m} to $[s_m]$, $P_\mu(B_1)$ to $Q_\mu([1])$, and so on. It is, therefore, clear that if under μ the coordinates of ω are mutually independent and are distributed as

$$u \sim \text{Beta}\big(\alpha(B_1), \alpha(\bar{B}_1)\big), \quad \text{and}$$

$$u_{s_n} \sim \text{Beta}\big(\alpha(B^{s_n 1}), \alpha(B^{s_n 0})\big), \quad n = 1, 2, \dots, \tag{7.4}$$

then (7.2) and (7.3) hold true for all n.

Theorem 7.1 (Blackwell (1973)). *If under μ coordinates of ω are mutually independent and* (7.4) *holds, then P_μ is D^α on $(\mathcal{X}, \mathcal{A})$.*

Proof. Since (7.2) and (7.3) hold, it follows (Remark 4.1) that $P_\mu(B_n) \sim \text{Beta}(\alpha(B_n), \alpha(\bar{B}_n))$ for all n. Since $\mathcal{B}$ is a field, $\mathcal{X} = B_n$ some n. Therefore, $P_\mu(\mathcal{X}) = 1$ a.s. $[\mu]$. This proves that P_μ is a random probability measure on $(\mathcal{X}, \mathcal{A})$.

To prove that $P_\mu(A) \sim \text{Beta}(\alpha(A), \alpha(\bar{A}))$ for each $A \in \mathcal{A}$ we proceed as follows: The map $\mathrm{P} \to (\mathrm{P}(B_1), \mathrm{P}(B_2), \dots)$ from $\mathcal{P}$ to $\mathcal{S}$ is one–one and bimeasurable (Lemma 6.2). For any $A \in \mathcal{A}$, the map $\mathrm{P} \to \mathrm{P}(A)$ from $\mathcal{P}$ to $[0, 1]$ is measurable. Hence there exists a measurable map $h_A: \mathcal{S} \to [0, 1]$ such that $h_A(\mathrm{P}(B_1), \mathrm{P}(B_2), \dots) = \mathrm{P}(A)$ for all $\mathrm{P} \in \mathcal{P}$. For each n, the joint distribution of $P_\mu(B_1), P_\mu(B_2), \dots, P_\mu(B_n)$ is well defined in terms μ. And for different n these joint distributions are mutually consistent. The Kolmogorov extension theorem, therefore, guarantees that the joint distribution of the whole sequence $P_\mu(B_1), P_\mu(B_2), \dots$, is well defined. If we denote this joint distribution by Π_μ, then $P_\mu(A) \sim \Pi_\mu h_A^{-1}$.

Consider now the hypothetical situation where we might have started with $\mathcal{B}^* = \{A, B_1, B_2, \dots\}$ as generator of $\mathcal{A}$. Proceeding as before, we would then have defined a random probability measure P_μ^* on $(\mathcal{X}, \mathcal{A})$. Under μ, the random probability measures P_μ and P_μ^* on $(\mathcal{X}, \mathcal{A})$ are the same, and therefore the joint

distribution of $(P_\mu(A), P_\mu(B_1), \ldots)$ is the same as the joint distribution of $(P_\mu^*(A), P_\mu(B_1), \ldots)$. For the random probability measure P_μ^* it is clear that $P_\mu^*(A) \sim$ Beta$(\alpha(A), \alpha(\bar{A}))$ and $(P_\mu^*(B_1), P_\mu^*(B_2), \ldots) \sim \Pi_\mu$. Therefore, $\Pi_\mu h_A^{-1}$ is Beta$(\alpha(A), \alpha(\bar{A}))$. This proves that $P_\mu(A) \sim$ Beta$(\alpha(A), \alpha(\bar{A}))$ for all $A \in \mathcal{A}$.

The above argument goes through, word for word, for an arbitrary measurable partiation $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$ leading to the conclusion that the r.v.'s $(P_\mu(A_1), P_\mu(A_2), \ldots, P_\mu(A_k))$ have a singular Dirichlet distribution with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$. This proves that P_μ is D^α on $(\mathcal{X}, \mathcal{A})$.

8. Support of the Dirichlet process

The existence theorem of the previous section may be restated as:

Theorem 8.1. *If $(\mathcal{X}, \mathcal{A})$ is a Borel space, then, for each finite measure α on $(\mathcal{X}, \mathcal{A})$, there exists a probability measure D^α on $(\mathcal{P}, \sigma(\mathcal{P}))$ such that with $P \sim D^\alpha$, the r.v.'s $(P(A_1), P(A_2), \ldots, P(A_k))$ have singular Dirichlet distribution with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$ for any measurable partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$.*

Let $\mathcal{P}_0$ be the family of discrete probability measure on $(\mathcal{X}, \mathcal{A})$. That $\mathcal{P}_0$ belongs to $\sigma(\mathcal{P})$ has been noted in Corollary 6.1.

Theorem 8.2. *If $\mathrm{P} \sim D^\alpha$, then almost every realization of P is a discrete probability measure on $(\mathcal{X}, \mathcal{A})$, that is,*

$$D^\alpha(\mathcal{P}_0) = 1.$$

Historical Note: Ferguson (1973) gave a rather involved argument to prove this result. Blackwell (1973) and Blackwell and MacQueen (1973) gave alternative arguments for the same result. The proof given here is a streamlined version of an ingenious argument given by Berk and Savage (1979).

Consider the pair (P, X) of random entities such that (i) $P \sim D^\alpha$ and (ii) $X|P \sim P$, that is, conditional on $P = \mathrm{P}$, the probability distribution of X on $(\mathcal{X}, \mathcal{A})$ is P. Let Δ^α denote the joint distribution of (P, X) on the product space $(\mathcal{P} \times \mathcal{X}, \sigma(\mathcal{P}) \times \mathcal{A})$. The marginal distribution of X is then easily verified to be the normalized measure $\tilde{\alpha} = \alpha/\alpha(\mathcal{X})$. It is well known (see Ferguson (1973)) that $P|X = x \sim D^{\alpha + \delta_x}$, where δ_x denotes the degenerate probability measure with its whole mass concentrated at x.

We have noted earlier (Proposition 6.1) that $E = \{(\mathrm{P}, x) : \mathrm{P}(\{x\}) > 0\}$ belongs to $\sigma(\mathcal{P}) \times \mathcal{A}$. The following proposition is a preliminary to the proof of Theorem 8.2.

Proposition 8.1. $\Delta^\alpha(E) = 1$.

Proof. Writing E^x for the x-section of E, we have

$$\Delta^\alpha(E) = \int D^{\alpha + \delta_x}(E^x) \, \mathrm{d}\tilde{\alpha}(x).$$

Now, E^x is the set of all $\mathrm{P} \in \mathcal{P}$ such that $\mathrm{P}(\{x\}) > 0$. Under the distribution $D^{\alpha + \delta_x}$, the random variable $P(\{x\})$ is positive with probability one— this is because the $\alpha + \delta_x$ measure of the set $\{x\}$ is positive. Therefore, $D^{\alpha + \delta_x}\{\mathrm{P} : \mathrm{P}\{x\} > 0\} = 1$ for all x. In other words, $\Delta^\alpha(E) = 1$.

Proof of Theorem 8.2. Consider now the P-section $E_{\mathrm{P}}=\{x:\mathrm{P}(\{x\})>0\}$ of the set E. Since X, given $P=\mathrm{P}$, is distributed as P we have

$$\Delta^{\alpha}(E)=\int\psi(\mathrm{P})\,\mathrm{d}D^{\alpha}(\mathrm{P}),$$

where $\psi(\mathrm{P})=\mathrm{P}(E_{\mathrm{P}})$ is the discrete mass of P. Since $\Delta^{\alpha}(E)=1$, we at once have $\psi(P)=1$ a.s. $[\mathrm{D}^{\alpha}]$. But $\{\mathrm{P}:\psi(\mathrm{P})=1\}=\mathcal{P}_0$.

Let $\mathcal{V}=\{V\}$ be the collection of all open sets in $\mathcal{X}$. Since $\mathcal{X}$ is a separable metric space, there exists a countable subcollection $\{V_1,V_2,\dots\}$ of open sets such that every V contains some V_n. Let $\mathcal{P}'$ be the collection of all $\mathrm{P}\in\mathcal{P}$ such that $\mathrm{P}(V)>0$ for all $V\in\mathcal{V}$. Similarly, let $\mathcal{P}_n=\{\mathrm{P}:\mathrm{P}(V_n)>0\}$. It is then clear that $\mathcal{P}'=\bigcap_{n=1}^{\infty}\mathcal{P}_n$.

Theorem 8.3. *If $\alpha(V)>0$ for all $V\in\mathcal{V}$, then $D^{\alpha}(\mathcal{P}')=1$.*

Proof. Since $P(V_n)\sim\mathrm{Beta}(\alpha(V_n),\alpha(\overline{V}_n))$ and $\alpha(V_n)>0$ it follows that $P(V_n)>0$ a.s. $[D^{\alpha}]$, that is, $D^{\alpha}(\mathcal{P}_n)=1$. Therefore, $D^{\alpha}(\mathcal{P}')=1$.

The set $\mathcal{P}_0\cap\mathcal{P}'$ is the collection of all discrete probability measure P on $(\mathcal{X},\mathcal{A})$ such that the mass points of P are everywhere dense in $\mathcal{X}$. Putting Theorems (8.2) and (8.3) together we finally have:

Theorem 8.4. *If $P\sim D^{\alpha}$ and the α-measure of every open subset of $\mathcal{X}$ is positive, then for almost every realization* P *of P it is true that* P *is discrete with its mass points everywhere dense in $\mathcal{X}$.*

Further properties of the Dirichlet process will be discussed in a forthcoming note.

Acknowledgement

The authors are greatly indebted to Professor David Blackwell for the benefit of some prolonged discussions and consultations during the Fall and Winter Quarters of 1978–1979 when he was visiting the Florida State University. The authors are also indebted to Professor J. Sethuraman for his useful discussions.

References

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā A*, **15**, 337–380.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā A* **37**, 1–71.

Berk, R.H. and Savage, I.R. (1979). Dirichlet process produce discrete measures: An elementary proof. *Contributions to Statistics. Jaroslav Hájek Memorial Volume*, pp. 25–31. Academia, North-Holland, Prague.

Blackwell, D. (1973). Discreteness of Ferguson Selections. *Ann. Statist.* **1**, 356–358.

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

Dubins, L. and Freedman, D. (1964). Measurable sets of measures. *Pacific J. Math.* **14**, 1211–1222.

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

Parthasarathy, K.R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.

Wilks, S.S. (1962). *Mathematical Statistics*. Wiley, New York.