The Annals of Statistics
1974, Vol, 2, No. 6, 1152-1174

MIXTURES OF DIRICHLET PROCESSES WITH APPLICATIONS TO BAYESIAN NONPARAMETRIC PROBLEMS[1]

BY CHARLES E. ANTONIAK

University of California, Berkeley

A random process called the Dirichlet process whose sample functions are almost surely probability measures has been proposed by Ferguson as an approach to analyzing nonparametric problems from a Bayesian viewpoint.

An important result obtained by Ferguson in this approach is that if observations are made on a random variable whose distribution is a random sample function of a Dirichlet process, then the conditional distribution of the random measure can be easily calculated, and is again a Dirichlet process.

This paper extends Ferguson's result to cases where the random measure is a mixing distribution for a parameter which determines the distribution from which observations are made. The conditional distribution of the random measure, given the observations, is no longer that of a simple Dirichlet process, but can be described as being a mixture of Dirichlet processes. This paper gives a formal definition for these mixtures and develops several theorems about their properties, the most important of which is a closure property for such mixtures. Formulas for computing the conditional distribution are derived and applications to problems in bio-assay, discrimination, regression, and mixing distributions are given.

1. Introduction. In certain statistical problems, the distribution F on the sample space (X, β) is only known to belong to some collection of distributions $\mathscr{F} = \{F_\alpha\}$. This collection $\mathscr{F}$ may be treated as a parameter space in a Bayesian analysis of these problems, but the Bayesian approach requires the placing of a prior distribution on $\mathscr{F}$. If $\mathscr{F}$ is a parametric family, the prior distribution can be given for the parameters. However, in many so-called nonparametric or distribution-free problems the set $\mathscr{F}$ may be too large to permit such treatment. For example, $\mathscr{F}$ may be the class of all distribution functions on the real line. In this context, discussing what he calls the "empirical Bayes" problem, Robbins [16] has said:

> A strictly Bayesian approach might be to start out with an *a priori* distribution of probability over the class $\mathscr{F}$ of all possible distributions $F \cdots$ this could possibly be converted into an *a posteriori* distribution after $X_1, X_2, \cdots, X_n$

Received December 1972; revised September 1973.

[1] This research was supported partly by the U.S. Navy Eletronics Lab, San Diego, California, and partly by the Office of Naval Research under Contract N00014-69-A-0200-1051 with the University of California.

AMS 1970 *subject classifications.* Primary 60K99; Secondary 60G35, 62C10, 62G99.

Key words and phrases. Dirichlet process, nonparametric, Bayes, empirical Bayes, mixing distribution, random measures, bio-assay, discrimination.

> have been observed. Whether anyone will espouse this view and recommend a procedure for carrying it out remains to be seen.

Some properties that would be desirable for such a procedure can be paraphrased from Raiffa and Schlaifer [14], page 44 for this situation as follows:

1. The class $\mathscr{D}$ of random prior distribution on $\mathscr{F}$ should be analytically tractable in three respects:
   (a) It should be reasonably easy to determine the posterior distribution on $\mathscr{F}$, given a "sample";
   (b) It should be possible to express conveniently the expectations of simple loss functions; and
   (c) The class $\mathscr{D}$ should be closed, in the sense that if the prior is a member of $\mathscr{D}$, then the posterior is a member of $\mathscr{D}$.
2. The class $\mathscr{D}$ should be "rich," so that there will exist a member of $\mathscr{D}$ capable of expressing any prior information or belief.
3. The class $\mathscr{D}$ should be parametrized in a manner which can be readily interpreted in relation to prior information and belief.

While these requirements are not mutually exclusive, they do seem, in the case we are considering, to be antagonistic in the sense that some may be obtained at the expense of others. For example, several authors, among them Dubins and Freedman [5], [6], Kraft and van Eeden (12), and Kraft [11], have described distribution function processes which satisfy the second requirement, but are deficient in the first and third.

More recently, Ferguson [9] has defined a process called the Dirichlet process which is particularly strong in satisfying the first and third requirements and is only slightly deficient with respect to the second requirement.

Although Ferguson was successful in using the Dirichlet process for Bayesian analyses of several nonparametric problems, there are some statistical models for which the closure property 1(c) does not hold. For example, the nature of sampling in mixing distribution problems and bio-assay problems can be such that the posterior distribution is not a simple Dirichlet process, but can be represented as a mixture of Dirichlet processes, i.e., as belonging in a sense to a linear manifold spanned by Dirichlet processes.

This paper reviews the basic properties of Ferguson's Dirichlet process in Section 2. Section 3 develops the additional structure necessary to define mixtures and derives some basic properties of mixtures, culminating in Theorem 3 which states, roughly, that mixtures of Dirichlet processes have the closure property 1(c).

The somewhat unusual behavior of samples from a Dirichlet process is examined in Section 4 and explicit formulas for the posterior distributions are derived. Section 5 illustrates the application of mixtures of Dirichlet processes

to estimation problems, including estimation of a mixing distribution and empirical Bayes estimation. The remaining Sections 6–8 give applications to a regression problem, a bio-assay problem, and a discrimination problem.

2. Ferguson's Dirichlet process. In a fundamental paper on a Bayesian approach to nonparametric problems, Ferguson [9] defines a random process, called the Dirichlet process, whose sample functions are almost surely probability measures, and he derives many important properties of this process. To save space, we will list below only the definition and those properties we need for this paper.

DEFINITION 1. Let Θ be a set, and $\mathscr{A}$ a σ-field of subsets of Θ. Let α be a finite, nonnull, nonnegative, finitely additive measure on $(\Theta, \mathscr{A})$. We say a random probability measure P on $(\Theta, \mathscr{A})$ is a Dirichlet process on $(\Theta, \mathscr{A})$ with parameter α, denoted $P \in \mathscr{D}(\alpha)$, if for every $k = 1, 2, \cdots$ and measurable partition $B_1, \cdots, B_k$ of Θ, the joint distribution of the random probabilities $(P(B_1), \cdots, P(B_k))$ is Dirichlet with parameters $(\alpha(B_1), \cdots, \alpha(B_k))$, denoted $(P(B_1), \cdots, P(B_k)) \in \mathscr{D}(\alpha(B_1), \cdots, \alpha(B_k))$. (When $\alpha(B_i) = 0$, $P(B_i) = 0$ with probability one.)

Ferguson shows that this definition satisfies the Kolmogorov criteria for the existence of a probability $\mathscr{P}$ on the space of all functions from $\mathscr{A}$ into $[0, 1]$ with the σ-field generated by the cylinder sets. Certain properties of the Dirichlet process obtained by Ferguson will be needed later.

(i) If $P \in \mathscr{D}(\alpha)$, and $A \in \mathscr{A}$, then $\mathscr{E}(P(A)) = \alpha(A)/\alpha(\Theta)$.

(ii) If $P \in \mathscr{D}(\alpha)$ and conditional given P, $\theta_1, \theta_2, \cdots, \theta_n$ are i.i.d. P, then $P \mid \theta_1, \theta_2, \cdots, \theta_n \in \mathscr{D}(\alpha + \sum_{i=1}^{n} \delta_{\theta_i})$ where δ_x denotes the measure giving mass one to the point x.

(iii) If $P \in \mathscr{D}(\alpha)$, then P is almost surely discrete. (See also Blackwell [2].)

We point out that property (ii) above satisfies desirable properties 1(a) and 1(c). The posterior distribution is easily computable and in the same class as the prior. Property (i) above fulfills desirable property 3 in two respects. Since $\mathscr{E}(P(A)) = \alpha(A)/\alpha(\Theta)$, one can choose the "shape" of α to reflect his prior guess at the shape of the distribution. Moreover, it follows from (i) and (ii) that $\mathscr{E}(P(\cdot) \mid \theta_1, \theta_2, \cdots, \theta_n)$ is $(n + \alpha(\Theta))^{-1}[\alpha(\Theta)\mathscr{E}(P(\cdot)) + nF_n(\cdot)]$, where F_n is the empirical df. Thus the magnitude of $\alpha(\Theta)$ represents, in a sense, the degree of faith in the prior guess, and appears in the formula as if it were "prior sample size." The statistician can choose the magnitude of $\alpha(\Theta)$ to represent the strength of his conviction, independent of his opinion about the "shape" of the distribution.

On the other hand, the almost sure discreteness of a Dirichlet selection given in (iii) would seem to be inconsistent with desirable property 2 for situations where one wants a prior on a class of continuous distributions. In many cases of interest, however, this discreteness is no more troublesome than the discreteness

of a sample cdf. The real concern of property 2 in this context is that distribution functions chosen by a Dirichlet process can be reasonably "close" to the kind of distribution functions one is interested in. In this regard, Ferguson has proved that the Dirichlet process *is* "rich" in the sense that there is positive probability that a sample function will approximate as closely as desired the measure given any specified collection of $\mathscr{A}$ measurable sets by any fixed distribution which is absolutely continuous with respect to the parameter measure α. The mixtures of Dirichlet processes which we define below can also be shown to be "rich" in this same sense, but we will make no use of the property in this paper.

3. Mixtures of Dirichlet processes. Because the basic Dirichlet process defined above does not encompass enough of the situations encountered in Bayesian analysis, we proceed now to the concept of a mixture of Dirichlet processes, which, roughly, is a Dirichlet process where the parameter α is itself random.

Before we can make this idea more precise we need a slight generalization of the usual definition of a transition probability, and later we will find it necessary to impose certain regularity conditions on the underlying spaces to assure the existence of some needed conditional distributions.

DEFINITION 2. Let $(\Theta, \mathscr{A})$ and $(U, \mathscr{B})$ be two measurable spaces. A *transition measure* on $U \times \mathscr{A}$ is a mapping α of $U \times \mathscr{A}$ into $[0, \infty)$ such that

(a) For every $u \in U$, $\alpha(u, \cdot)$ is a finite, nonnegative, nonnull measure on $(\Theta, \mathscr{A})$.

(b) For every $A \in \mathscr{A}$, $\alpha(\cdot, A)$ is measurable on $(U, \mathscr{B})$.

We note that this differs from the definition of a transition probability in that $\alpha(u, \Theta)$ need not be identically one. We make this change because we want $\alpha(u, \cdot)$ to be a parameter for a Dirichlet process.

DEFINITION 3. Let $(\Theta, \mathscr{A})$ be a measurable space, let $(U, \mathscr{B}, H)$ be a probability space, called the index space, and let α be a transition measure on $U \times \mathscr{A}$. We say P is a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$ with mixing distribution H on the index space $(U, \mathscr{B})$, and transition measure α, if for all $k = 1, \cdots$ and any measurable partition $A_1, A_2, \cdots, A_k$ of Θ we have

$$\mathscr{P}\{P(A_1) \leqq y_1, \cdots, P(A_k) \leqq y_k\} = \int_U D(y_1, \cdots, y_k \mid \alpha(u, A_1), \cdots, \alpha(u, A_k))\, dH(u),$$

where $D(\theta_1, \cdots, \theta_k \mid \alpha_1, \cdots, \alpha_k)$ denotes the distribution function of the Dirichlet distribution with parameters $(\alpha_1, \cdots, \alpha_k)$.

In concise symbols we use the heuristic notation:

$$(P(A_1), P(A_2), \cdots, P(A_k)) \in \int_U \mathscr{D}(\alpha(u, A_1), \cdots, \alpha(u, A_k))\, dH(u)$$

or simply $P \in \int_U \mathscr{D}(\alpha(u, \cdot))\, dH(u)$.

Roughly, we may consider the index u as a random variable with distribution H and conditional given u, P is a Dirichlet process with parameter $\alpha(u, \cdot)$. In fact $\mathscr{U}$ can be defined as the identity mapping random variable and we will use the notation $| u$ for "given $\mathscr{U} = u$." In alternative notation, $u \in H$, $P | u \in \mathscr{D}(\alpha_u)$, wher $\alpha_u = \alpha(u, \cdot)$.

An example of a mixture of Dirichlet processes derived from a simple Dirichlet process is given below.

EXAMPLE 1. Let P be a Dirichlet process on $(\Theta, \mathscr{A})$ with parameter α. Define $\alpha(u, A) = \alpha(A) + \delta_u(A)$, where $\delta_u(A) = 1$ if $u \in A$, 0 otherwise. Let H be a fixed probability measure on $(\Theta, \mathscr{A})$. Then the process P^* which choses u according to H, and P from a Dirichlet process with parameter $\alpha(u, A)$ is a mixture of Dirichlet processes as defined above. Moreover, for the mixture in this example, we note that if $(B_1, \cdots, B_k)$ is any measurable partition of Θ,

$$(1) \qquad (P(B_1), P(B_2), \cdots, P(B_k)) \in \sum_{i=1}^{k} H(B_i)\mathscr{D}(\alpha(B_1), \cdots, \alpha(B_i) + 1, \cdots, \alpha(B_k)) .$$

Relation (1), in fact, characterizes mixtures of this type, and we will encounter it frequently in this paper.

DEFINITION 4. Let P be a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$ with mixing distribution H on index space $(U, \mathscr{B})$ and transition measure α on $U \times \mathscr{A}$. We say that $\theta_1, \theta_2, \cdots, \theta_n$ is a sample of size n from P if for any $m = 1, 2, \cdots$ and measurable sets $A_1, A_2, \cdots, A_m, C_1, C_2, \cdots, C_n$ we have:

$$(2) \qquad \mathscr{P}\{\theta_1 \in C_1, \cdots, \theta_n \in C_n | u, P(A_1), \cdots, P(A_m), P(C_1), \cdots, P(C_n)\} = \prod_{i=1}^{n} P(C_i) \quad \text{a.s.}$$

This definition determines the joint distribution of $\theta_1, \cdots, \theta_n, P(A_1), \cdots, P(A_m)$ since the customary

$$\mathscr{P}\{\theta_1 \in C_1, \cdots, \theta_n \in C_n, P(A_1) \leqq y_1, \cdots, P(A_m) \leqq y_m\}$$

may be found by integrating (3) with respect to the known joint conditional distribution of $P(A_1), \cdots, P(A_m), P(C_1), \cdots, P(C_n)$ given u over the set $[0, y_1] \times \cdots \times [0, y_m] \times [0, 1] \times \cdots \times [0, 1]$ and then integrating the resulting function of u with respect to $H(u)$ over U.

An immediate consequence of this definition is the following useful result.

PROPOSITION 1. *If $P \in \int_U \mathscr{D}(\alpha(u, \cdot))\, dH(u)$ and θ is a sample of size one from P, then for any measurable set A,*

$$(3) \qquad \mathscr{P}(\theta \in A) = \int_U \frac{\alpha(u, A)}{\alpha(u, \Theta)}\, dH(u) .$$

PROOF. $\mathscr{P}(\theta \in A | u, P(A)) = P(A)$ a.s. hence

$$\mathscr{P}(\theta \in A | u) = \mathscr{E}\{\mathscr{P}(\theta \in A | u, P(A)) | u\} \quad \text{a.s.}$$

$$= \mathscr{E}\{P(A) | u\} = \frac{\alpha(u, A)}{\alpha(u, \Theta)} \quad \text{a.s.}$$

Finally

$$\mathscr{P}(\theta \in A) = \mathscr{E}\left[\frac{\alpha(u, A)}{\alpha(u, \Theta)}\right] = \int_U \frac{\alpha(u, A)}{\alpha(u, \Theta)}\, dH(u)\,. \qquad \square$$

THEOREM 1. *Let P be a Dirichlet process on $(\Theta, \mathscr{A})$, with parameter α. Let θ be a sample of size 1 from P, and $A \in \mathscr{A}$ be any measurable set such that $\alpha(A) > 0$. Then the conditional distribution of P, given $\theta \in A$, is a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$, with index space $(A, \mathscr{A} \cap A)$, and transition measure α on $A \times (\mathscr{A} \cap A)$, with distribution H_A on $(A, \mathscr{A} \cap A)$, where $H_A(\cdot) = \alpha(\cdot)/\alpha(A)$ on A and $\alpha(u, \cdot) = \alpha + \delta_u$ for $u \in A$.*

PROOF. Let $(B_1, B_2, \cdots, B_n)$ be any measurable partition of Θ. Since A is measurable we obtain a refined measurable partition by letting $B_i' = A \cap B_i$, and $B_i^0 = A^c \cap B_i$ so that $A = \bigcup_{i=1}^n B_i'$. If follows from property (ii) of the Dirichlet process listed in Section 2 that the conditional distribution of $P(B_1')$, $P(B_2')$, $\cdots$, $P(B_n')$, $P(B_1^0)$, $\cdots$, $P(B_n^0)$, given $\theta \in B_i'$ is Dirichlet with parameters $(\alpha(B_1'), \alpha(B_2'), \cdots, \alpha(B_i') + 1, \cdots, \alpha(B_n^0)$. Integrating this with respect to the conditional probability that $\theta \in B_i'$, given $\theta \in A$, we obtain the conditional distribution of $P(B_1')$, $P(B_2')$, $\cdots$, $P(B_n^0)$, given $\theta \in A$, is

$$\sum_{i=1}^n \frac{\alpha(B_i')}{\alpha(A)}\, D(\alpha(B_1'), \cdots, \alpha(B_i') + 1, \cdots, \alpha(B_n^0))\,.$$

which we recognize as relation (1) in Example 1 following Definition 3, with $H(\cdot) = \alpha(\cdot)/\alpha(A)$ on A. $\square$

COROLLARY 1.1. *Let P be a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$ with index space also $(\Theta, \mathscr{A})$, and transition measure $\alpha_u = \alpha + \delta_u$. If the distribution H on the index space $(\Theta, \mathscr{A})$ is given by $H(A) = \alpha(A)/\alpha(\Theta)$, then P is in fact a simple Dirichlet process on $(\Theta, \mathscr{A})$ with parameter α. In symbols*

$$\int_\Theta \mathscr{D}(\alpha + \delta_u)\, \frac{\alpha(du)}{\alpha(\Theta)} = \mathscr{D}(\alpha)\,.$$

We omit the proof and simply point out that $H(\cdot) = \alpha(\cdot)/\alpha(\Theta)$ is what we would obtain in Theorem 1 if we let $A = \Theta$. But being given $\theta \in \Theta$ is being given no information at all, so the posterior distribution is the same as the prior, a simple Dirichlet process. The usefulness of Corollary 1.1 is that it enables one to reduce some mixtures of Dirichlet processes to simple Dirichlet processes.

PROPOSITION 2. *Let P be a Dirichlet process on $(\Theta, \mathscr{A})$ with parameter α, and let $A \in \mathscr{A}$. Then given $P(A) = M$, the conditional distribution of $(1/M)P$ restricted to $(A, \mathscr{A} \cap A)$ is a Dirichlet process on $(A, \mathscr{A} \cap A)$ with parameter α restricted to A. That is, if $A_1, \cdots, A_k$ is any measurable partition of A, then the conditional distribution of $(P(A_1)/M, \cdots, P(A_k)/M)$, given $P(A) = M$, is a Dirichlet distribution with parameter $(\alpha(A_1), \cdots, \alpha(A_k))$.*

PROOF. From the definition of a Dirichlet distribution in terms of the gamma distribution as given in Ferguson [9] we know the distribution of

$P(A_1)/\sum_{i=1}^k P(A_i), \cdots, P(A_k)/\sum_{i=1}^k P(A_i)$ is Dirichlet $(\alpha(A_1), \cdots, \alpha(A_k))$. But $\sum_{i=1}^k P(A_i) = P(A) = M$ by hypothesis. □

Proposition 2 is an alternative way of stating that the Dirichlet process is "tailfree" as discussed by Doksum [4] and Fabius [7], who show further that the Dirichlet process is essentially the only process which is tailfree with respect to every tree of partitions. They also show that the Dirichlet process is the only one for which the posterior distribution of $P(A)$ given a sample $\theta_1, \theta_2, \cdots, \theta_n$, depends only on the *number* of observations falling in A, not *where* they fall in A.

We now combine the results of Theorem 1, Corollary 1.1, and Proposition 1 to get:

THEOREM 2. *Let P be a Dirichlet process on $(\Theta, \mathscr{A})$ with parameter α, and let θ be a sample from P. Let $A \in \mathscr{A}$. Then the conditonal distribution of P given $P(A)$ and $\theta \in A$ is the same as the conditional distribution of P given $P(A)$.*

Roughly speaking, if $P(A)$ is known, then the event $\theta \in A$ tells us nothing more about the process. This is consistent with the definition of sampling given in Definition 4, and since $P(\Theta) = 1$, a.s. it confirms the interpretation given for Corollary 1.1.

Before we can proceed to the most interesting theorems about mixtures of Dirichlet processes, we must stop to examine the measure theoretic structure we have created and insure the existence of certain conditional distributions by adding appropriate regularity conditions to the underlying spaces.

Starting with the measure space $(\Theta, \mathscr{A})$, index probability space $(U, \mathscr{B}, H)$, and transition measure α on $U \times \mathscr{A}$, we define a measure μ on the measurable product space $(\Theta \times U, \mathscr{A} \times \mathscr{B})$ as $\mu(A \times B) = \int_B \alpha(u, A)/\alpha(u, \Theta)\, dH(u)$.

If P is mixture of Dirichlet processes on $(\Theta, \mathscr{A})$ with these parameters, and $\theta_1, \cdots, \theta_n$ is a sample of size n, we will need to be able to go from the known conditional distribution of P given $u, \theta_1, \theta_2, \cdots, \theta_n$ to the conditional distribution of u given $\theta_1, \cdots, \theta_n$. To assure the existence of this conditional distribution we will henceforth require that $(\Theta, \mathscr{A})$ and $(U, \mathscr{B})$ be standard Borel spaces, defined as follows:

DEFINITION 5. A standard Borel space is a measurable space $(\Theta, \mathscr{A})$, in which $\mathscr{A}$ is countably generated, and for which there exists a bi-measurable mapping between $(\Theta, \mathscr{A})$ and some complete separable metric space $(Y, \mathscr{C})$.

If $(\Theta, \mathscr{A})$ and $(U, \mathscr{B})$ are standard Borel spaces, then the product space $(\Theta \times U, \mathscr{A} \times \mathscr{B})$ is a standard Borel space, and the required conditional distributions specified above are known to exist (see, for example, Parthasarathy [13] Chapter 5).

With these definitions we can now state the last and most important theorem in this section which says, roughly, that if we sample from a *mixture* of Dirichlet processes, and the sample is distorted by random error, the posterior distribution

of the process is again a mixture of Dirichlet processes. We will see in the sections on applications that this situation occurs frequently.

THEOREM 3. *Let P be a mixture of Dirichlet processes on a standard Borel space $(\Theta, \mathscr{A})$ with standard Borel index space $(U, \mathscr{B})$, distribution H on $(U, \mathscr{B})$, and transition measure α on $U \times \mathscr{A}$. Let $(X, \mathscr{C})$ be a standard Borel sample space, and F a transition probability from $\Theta \times \mathscr{C}$ to $[0, 1]$. If θ is a sample from P, i.e., $\theta \mid P$, $u \in P$ and $X \mid P, \theta, u \in F(\theta, \cdot)$, then the distribution of P given $X = x$ is a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$, with index space $(\Theta \times U, \mathscr{A} \times \mathscr{B})$, transition measure $\alpha_u + \delta_\theta$ on $(\Theta \times U) \times \mathscr{A}$, and mixing distribution H_x on the index space $(\Theta \times U, \mathscr{A} \times \mathscr{B})$ where H_x is the conditional distribution of (θ, u) given $X = x$. In symbols, if*

$$u \in H\,, \qquad P \mid u \in \mathscr{D}(\alpha_u)\,, \qquad P \in \textstyle\int_U \mathscr{D}(\alpha_u)\, dH(u)\,,$$
$$\theta \mid P, u \in P \qquad \text{and} \qquad X \mid P, \theta, u \in F(\theta, \cdot)$$

then $(P \mid X = x) \in \int_{\Theta \times U} \mathscr{D}(\alpha_u + \delta_\theta)\, dH_x(\theta, u)$.

PROOF. The distribution of P given (θ, u, x) is $\mathscr{D}(\alpha_u + \delta_\theta)$, independent of x, the distribution of X given (θ, u) is $F(\theta, \cdot)$ independent of u; the distribution of θ given u is $\alpha(u, \cdot)/\alpha(u, \Theta)$; and the distribution of u is H. The last three define the joint distribution of (θ, u, X) as given in the theorem, and conditional distribution of P given X is obtained by integrating the known conditional distribution of P given (θ, u, X), namely $\mathscr{D}(\alpha_u + \delta_\theta)$, with respect to the conditional distribution of (θ, u) given X, yielding the formula given in the theorem, which we recognize as a mixture of Dirichlet processes. □

Lest the imposing notation obscure the basic principles involved, we illustrate Theorem 3 with an example where U and X are two-point spaces, and Θ is the unit interval.

EXAMPLE 2. Let X be a Bernoulli random variable with $P(X = 1) = \theta$, and let θ have as prior distribution an equal mixture of Beta distributions, $g(\theta) = \frac{1}{2}\mathscr{B}e(1, 2) + \frac{1}{2}\mathscr{B}e(2, 2)$. Then the posterior distribution of θ given $X = 1$ is a mixture of Beta distributions, $g(\theta \mid X = 1) = \frac{2}{5}\mathscr{B}e(2, 2) + \frac{3}{5}\mathscr{B}e(3, 2)$.

Notice that the posterior mixture gives more weight to $\mathscr{B}e(3, 2)$ since given $X = 1$, θ is more likely to have come from $\mathscr{B}e(2, 2)$ in the prior mixture. This is a specific example of a general property of mixtures of Dirichlet processes which will be stated formally as Corollary 3.2.

We now state two corollaries to Theorem 3 which treat cases that occur frequently in applications. No proof is given since they are simply special cases of Theorem 3.

COROLLARY 3.1. *Let P be a Dirichlet process on a standard Borel space $(\Theta, \mathscr{A})$, with parameter α and let θ be a sample from P. Let $(X, \mathscr{C})$ be a standard Borel sample space and F a transition probability from $\Theta \times \mathscr{C}$ to $[0, 1]$. If the conditional distribution of X given P and θ is $F(\theta, \cdot)$, then the conditional distribution of P given*

$X = x$ is a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$ with mixing distribution H on index space $(\Theta, \mathscr{A})$ and transition measure $\alpha(\theta, \cdot) = \alpha(\cdot) + \delta_\theta(\cdot)$, where the mixing distribution H on $(\Theta, \mathscr{A})$ considered as the index space is the conditional distribution of θ given $X = x$; in concise notation:

$$P \in \mathscr{D}(\alpha), \theta \in P, X \mid P, \theta \in F(\theta, \cdot) \Rightarrow P \mid X \in \int \mathscr{D}(\alpha + \delta_\theta)\, dH_x(\theta) .$$

Roughly, if the sample from a *simple* Dirichlet process is distorted by random error, then the posterior distribution of the process given the distorted sample is a *mixture* of Dirichlet processes.

COROLLARY 3.2. *Let P be a mixture of Dirichlet processes on a standard Borel space $(\Theta, \mathscr{A})$, with standard Borel index space $(U, \mathscr{B})$, distribution H on $(U, \mathscr{B})$, and transition measure α on $U \times \mathscr{A}$. If θ is a sample from P, then P given θ is a mixture of Dirichlet processes on $(\Theta, \mathscr{A})$, with transition measure $\alpha + \delta_\theta$, and distribution H_θ on $(U, \mathscr{B})$, where H_θ is the conditional distribution of u given θ. In symbols, if $P \in \int_U \mathscr{D}(\alpha_u)\, dH(u)$ and $\theta \in P$ then $(P \mid \theta) \in \int_U \mathscr{D}(\alpha_u + \delta_\theta)\, dH_\theta(u)$.*

The essential point of this corollary is that the observation θ affects each component of the mixture as one would expect it to, by adding δ_θ to α_u, and in addition, changes the relative weightings of components of the mixture to the conditional distribution of u given θ. An explicit expression for this conditional distribution is given in Lemma 1 following Proposition 4.

4. Properties of samples from Dirichlet processes. The preceding theorems were stated rather formally to reveal the underlying measure theoretic structure, but only for samples of size one, to avoid cumbersome notation. In what follows we will develop more useful formulas for samples of size n, with less regard for elaborate formalism. We will see that the joint distribution of multiple samples possesses some peculiar properties caused by a virtual "memory" of the process. Sometimes a sample of size two is sufficient to illustrate these peculiarities. For example, let P be a Dirichlet process on a standard Borel space $(\Theta, \mathscr{A})$ with parameter α, and assume that α is *nonatomic*. If θ_1 and θ_2 were a sample from $\alpha(\theta)/\alpha(\Theta)$ in the usual independent identically distributed sense, we would expect $\mathscr{P}\{\theta_1 = \theta_2\} = 0$. The following argument shows that in fact, for such a Dirichlet process, $\mathscr{P}\{\theta_1 = \theta_2\} = 1/(\alpha(\Theta) + 1)$. The reason for this is that although α may be nonatomic, the conditional distribution of P given θ, is a Dirichlet process with parameter $\alpha + \delta_\theta$, which is already atomic with an atom of measure 1 at θ. Hence, the probability that $\theta_2 = \theta_1$, given θ_1, is $1/(\alpha(\Theta) + 1)$ independent of θ_1. If we proceed to calculate $\mathscr{P}\{\theta_3 = \theta_2 = \theta_1 \mid \theta_2 = \theta_1\}$ we see that the conditional distribution of P given θ_1, θ_2, $\theta_1 = \theta_2$, is Dirichlet with parameter $\alpha + 2\delta_{\theta_1}$. Hence $\mathscr{P}\{\theta_3 = \theta_2 = \theta_1 \mid \theta_2 = \theta_1\} = 2/(\alpha(\Theta) + 2)$, and consequently the joint probability that $\theta_1 = \theta_2 = \theta_3$ is $2/[\alpha(\Theta) + 1)(\alpha(\Theta) + 2)]$, and by induction, $\mathscr{P}(\theta_1 = \theta_2 = \cdots = \theta_n) = 1^{(n-1)}/\{[\alpha(\Theta) + 1]^{(n-1)}\}$, where the definition of $a^{(0)} = 1$ and $a^{(n)} = a(a + 1) \cdots (a + n - 1)$, $n > 0$. This property of the Dirichlet process is very similar to what happens in Polya urn schemes

and in one sense characterizes Dirichlet processes (see Feller [8] and Blackwell and MacQueen [3]).

On the other hand, one can consider the probability that θ_n is a new value, distinct from any previous observations $\theta_1, \theta_2, \cdots, \theta_{n-1}$. By the same logic as before, one notes that $P(\theta_2 \neq \theta_1) = \alpha(\Theta)/(\alpha(\Theta) + 1)$. Similarly $P(\theta_3 \notin \{\theta_1, \theta_2\}) = \alpha(\Theta)/(\alpha(\Theta) + 2)$, regardless of whether $\theta_1 = \theta_2$ or not. If one defines W_i as a random variable which equals 1 if θ_i is a new, distinct value, and zero otherwise, then $P(W_i = 1) = \alpha(\Theta)/(\alpha(\Theta) + i - 1)$. If we further define $Z_n = \sum_{i=1}^{n} W_i$ then Z_n is simply the number of distinct values of θ which have occurred in the first n observations.

Although $P(W_n = 1) = \alpha(\Theta)/(\alpha(\Theta) + n - 1)$ is monotone decreasing in n, nevertheless note that $E(Z_n) = \sum_{m=1}^{n} \alpha(\Theta)/(\alpha(\Theta)+m-1) = \alpha(\Theta) \sum_{m=1}^{n} 1/(\alpha(\Theta) + m - 1) \approx \alpha(\Theta)[\log((n + \alpha(\Theta))/\alpha(\Theta))]$. Hence $E(Z_n) \to \infty$, $n \to \infty$. In fact, Korwar and Hollander [10] show $Z_n \to_{\text{a.s.}} \infty$, as $n \to \infty$. Thus although new distinct values are increasingly rare, we are assured nonetheless of a steadily increasing number of distinct values. Moreover, since the distribution of the distinct values is simply $\alpha(\cdot)/\alpha(\Theta)$, this property can be used in the usual way to obtain information about the shape of $\alpha(\cdot)$ if it is unknown.

On the other hand, the rate at which new distinct values appear depends only on the magnitude of $\alpha(\Theta)$, and not the shape of $\alpha(\cdot)$, and this property should enable us to obtain information about the magnitude of $\alpha(\Theta)$ if it is unknown. We begin by obtaining an expression for $P(Z_n = k)$ as a function of $\alpha(\Theta)$. As an aid in this we define a sequence of polynomials $A_n(x)$ as

$$\begin{aligned} A_1(x) &= x \\ A_2(x) &= (x + 1)A_1(x) = x(x + 1) = x^{(2)} \\ &\vdots \\ A_n(x) &= (x + n - 1)A_{n-1}(x) = x(x + 1) \cdots (x + n - 1) = x^{(n)} . \end{aligned}$$

Hence $A_n(x)$ is a polynomial of degree n in x with integer coefficients, which we write

$$A_n(x) = {}_na_1 x + {}_na_2 x^2 + \cdots + {}_na_n x^n .$$

If we substitute $x = \alpha(\Theta)$ in $A_n(x)$, then considerations similar to those in the discussion of repeated values enable us to identify the kth term of this polynomial with the event of observing exactly k distinct values in a sample of size n, $P(Z_n = k) = {}_na_k \alpha(\Theta)^k / A_n(\alpha(\Theta))$.

The coefficients ${}_na_k$ of the polynomial are the absolute values of Stirling numbers of the first kind, tabulated in (Abramowitz and Stegun [1] page 833).

The significance of the foregoing is that if one knows he is sampling from some Dirichlet process with unknown parameter α, then he can obtain, independently, consistent estimates for $\alpha(\cdot)/\alpha(\Theta)$ and $\alpha(\Theta)$ by making use of the properties given above. If, however, there were some doubt that the process was in fact a Dirichlet process, then the only discriminating feature left is the

actual pattern of multiplicities observed. As a first step in such discrimination we show that we can, in fact, determine the fine structure of the probability of various patterns of multiplicities in a sample from a Dirichlet process. If we were interested, for example, in the probability that the first two observations were identical, the third was different from the first two, the fourth and fifth matched the third, and the sixth and seventh matched but were different from the previous values, then an argument similar to that given above yields $\mathscr{P}\{\theta_1 = \theta_2 \neq \theta_3 = \theta_4 = \theta_5 \neq \theta_6 = \theta_7 \neq \theta_1\} = \alpha(\Theta)^3 1^{(2)}/\alpha(\Theta)^{(7)}$.

Moreover, if one were only interested in characterizing the above event as one where, in a sample of size 7, only 3 distinct values of θ occurred, and that two were pairs and one a triplet, then the subscripts of the θ's are irrelevant, and we obtain the probability of the latter event by multiplying the previous expression by the appropriate combinatorial factor, in this case the familiar multinomial coefficient $\binom{7}{2,2,3}$, divided by 2! since the two pairs are indistinguishable. The result could be expressed

$$\mathscr{P}\{\theta_i = \theta_j \neq \theta_k = \theta_l \neq \theta_r = \theta_s = \theta_t \neq \theta_i\} = \frac{\binom{7}{2,2,3}}{2!} \frac{\alpha(\Theta)^3 1^{(2)}}{\alpha(\Theta)^{(7)}},$$

where $\{i, j, k, l, r, s, t\} = \{1, 2, 3, 4, 5, 6, 7\}$ in some order. The generalization of this example and the simplification of its expression are the goals of the following definition and proposition.

DEFINITION 6. Let $\theta_1, \theta_2, \cdots, \theta_n$ be a sample of size n from a Dirichlet process P. We will say that the sample belongs to the class $C(m_1, m_2, \cdots, m_n)$, and write $(\theta_1, \cdots, \theta_n) \in C(m_1, \cdots, m_n)$, if there are m_1 distinct values of θ that occur only once, m_2 that occur exactly twice, $\cdots$, m_n that occur exactly n times.

Two immediate consequences of this definition are that $n = \sum_{i=1}^n i m_i$, and the total number of distinct values of θ that occur is $Z_n = \sum_{i=1}^n m_i$. As an example of this notation we note that the sample in the preceding discussion belongs to the class $C(0, 2, 1, 0, 0, 0, 0)$.

PROPOSITION 3. *Let P be a Dirichlet process on a standard Borel space $(\Theta, \mathscr{A})$, with parameter α, and let α be nonatomic. Let $(\theta_1, \cdots, \theta_n)$ be a sample of size n from P. Then*

$$\mathscr{P}\{(\theta_1, \cdots, \theta_n) \in C(m_1, \cdots, m_n)\} = \frac{n!}{\prod_{i=1}^n i^{m_i}(m_i!)} \frac{\alpha(\Theta)^{\sum_{i=1}^n m_i}}{\alpha(\Theta)^{(n)}}. \tag{4}$$

PROOF. We calculate the probability of a particular sequence in the class $C(m_1, \cdots, m_n)$, observe that all such sequences are equally likely, and multiply by the number of these sequences.

Let $C_0(m_1, \cdots, m_n)$ be the event that $\theta_1, \cdots, \theta_{m_1}$, in that order, are unique in the sample and occur only once; that $\theta_{m_1+1}, \cdots, \theta_{m_1+2m_2}$ occur twice each, in the order $\theta_{m_1+1} = \theta_{m_1+2}$, $\theta_{m_1+3} = \theta_{m_1+4}$, etc. Then by the same argument given in the previous discussion $\mathscr{P}\{\theta_1, \cdots, \theta_n) \in C_0(m_1, \cdots, m_n)\} = [1^{(0)}\alpha(\Theta)]^{m_1}[1^{(1)}\alpha(\Theta)]^{m_2} \cdots [1^{(n-1)}\alpha(\Theta)]^{m_n}/\alpha(\Theta)^{(n)}$.

Next we must count the number of ways we can permute the indices of the θ's to obtain all essentially different sequences. But this number is well known as the number of elements in a conjugate class of the symmetric permutation group on n elements, and is equal to

$$\frac{1}{\prod_{i=1}^{n}(m_i!)}\binom{n}{\underbrace{1,\cdots,1}_{m_1},\underbrace{2,\cdots,2}_{m_2},\cdots,n}$$

where $\binom{n}{n_1,\cdots,n_m}$ denotes the well-known multinomial coefficient. For a derivation of this number see Scott [17]. Multiplying this combinatorial factor by the preceding probability and dividing numerator and denominator by $\prod_{i=1}^{n}[1^{(i-1)}]^{m_i}$, which is identical to $\prod_{i=1}^{n}[(i-1)!]^{m_i}$, the expression reduces to:

$$n!\,\alpha(\Theta)^{\sum_{i=1}^{n}m_i}/\alpha(\Theta)^{(n)}\prod_{i=1}^{n}i^{m_i}(m_i!)\,. \qquad \square$$

We see thus that the Dirichlet process induces a measure on the sample space Θ^n which gives positive mass to certain collections of sub-hyperplanes in Θ^n.

The relative magnitude of the mass concentrated on the sub-hyperplanes, compared to the mass distributed over the remainder of the sample space, is seen to be a function of the magnitude of $\alpha(\Theta)$. Consequently, for a given sample size, we would expect many more multiplicities if $\alpha(\Theta)$ is very small than if it is very large. Furthermore, if $\alpha(\Theta)$ is unknown, this property just described enables us to make an inference about the magnitude of $\alpha(\Theta)$ based on the number of multiplicities in a sample. We will see, in fact, that it is not the distribution of multiplicities, but the number of distinct values that occur in the sample, which is important in making inferences about $\alpha(\Theta)$. We begin by formalizing the results of the previous example.

PROPOSITION 4. *Let $P \in \int_U \mathscr{D}(\alpha_u)\,dH(u)$, where α_u is nonatomic for all $u \in U$, and let $(\theta_1, \cdots, \theta_n)$ be a sample of size n from P. Then the posterior distribution of u, given $(\theta_1, \cdots, \theta_n) \in C(m_1, \cdots, m_n) \in \mathscr{A}$ is determined by*

$$\mathscr{P}(u \in B \mid \underline{\theta} \in C(\underline{m})) = \frac{\int_B \frac{\alpha(u,\Theta)^{\sum_{i=1}^{n}m_i}}{\alpha(u,\Theta)^{(n)}}\,dH(u)}{\int_U \frac{\alpha(u,\Theta)^{\sum_{i=1}^{n}m_i}}{\alpha(u,\Theta)^{(n)}}\,dH(u)}\,.$$

PROOF. From Proposition 3, we have

$$\mathscr{P}(\underline{\theta} \in C(\underline{m}) \mid u) = \frac{n!}{\prod_{i=1}^{n}i^{m_i}(m_i!)}\,\frac{\alpha(u,\Theta)^{\sum_{i=1}^{n}m_i}}{\alpha(u,\Theta)^{(n)}}\,.$$

Integrating with respect to $H(u)$ over $u \in B$ gives the joint probability of $\underline{\theta} \in C(\underline{m})$ and $u \in B$. Since $\mathscr{P}(\underline{\theta} \in C(\underline{m})) > 0$ we obtain the usual conditional probability by application of Bayes formula, and note that the combinatorial coefficients cancel. $\square$

It is important to notice that a consequence of the combinatorial coefficients

cancelling is that the m_i occur in the above expression only as $\sum_{i=1}^{n} m_i$, that is, in a sum expressing the total number of distinct values, which we have previously denoted by Z_n. This confirms our earlier remark, that Z_n is sufficient for $\alpha(\Theta)$.

We can conclude from Proposition 4 that if α is nonatomic and $\alpha_u(\Theta) = M$, a constant independent of u, then the event $\underline{\theta} \in C(\underline{m})$ is independent of the event $u \in B$, and hence $\underline{\theta} \in C(\underline{m})$ provides no new information about u. This is a consequence of the assumption that α_u be nonatomic. If α_u is atomic the event $\underline{\theta} \in C(\underline{m})$ may still provide information about u even when $\alpha(u, \Theta)$ is independent of u. An analysis for a special case of atomic α is given later in Lemma 1, but some remarks of general validity are possible now.

We note first that if α is atomic, then P assigns positive weight to certain specific points in the product space of observations Θ^n, and these points are contained in the collections of sub-hyperplanes described earlier. If $P_1 \in \mathscr{D}(\alpha_1)$ and $P_2 \in \mathscr{D}(\alpha_2)$, where $\alpha_1(\Theta) = \alpha_2(\Theta)$ but α_1 is nonatomic and α_2 has atoms, then there is a greater probability of duplicated values in a sample from P_2 than in a sample from P_1.

If $P \in \int_U \mathscr{D}(\alpha_u)\, dH(u)$, and some or all α_u are atomic, the posterior distribution of u given θ will depend on whether any of the observed samples θ_i match any of the atoms of any of the α_u, and if so, whether the matched atoms are common to more than one value of u, and if common to more than one α_u, whether the size of the atoms is the same, etc. To elucidate this idea consider the following example.

EXAMPLE 3. Let α_u be a geometric distribution on $u, u + 1, u + n, \cdots$ and $H(u)$ be uniform on $[0, 1]$. Then with only one sample θ from P, the distribution of u given θ is degenerate on $\theta(\text{mod } 1)$. Notice that here there is no dominating σ-finite meausre for the α_u, in fact, they are all mutually singular. Nevertheless, the fact that the joint distribution of (θ, u) is concentrated on a subspace of the space $\Theta \times U$ enables us to find rather easily the posterior distribution of u given θ.

We conclude this section with a partial extension of Corollary 3.2 to the case of a sample of size n.

LEMMA 1. *Let $P \in \int_U \mathscr{D}(\alpha_u)\, dH(u)$ as in Theorem 3, let $\underline{\theta} = (\theta_1, \cdots, \theta_n)$ be a sample of size n from P, and suppose there exists a σ-finite, σ-additive measure μ on $(\Theta, \mathscr{A})$ shch that for each $u \in U$,*

(i) *α_u is σ-additive and absolutely continuous with respect to μ, and*
(ii) *the measure μ has mass one at each atom of α_u. Then*

$$dH_{\underline{\theta}}(u) = \frac{\frac{1}{M_u^{(n)}} \prod_{i=1}^{r} \alpha_u'(\theta_i')(m_u(\theta_i') + 1)^{(n(\theta_i')-1)}\, dH(u)}{\int_U \frac{1}{M_u^{(n)}} \prod_{i=1}^{r} \alpha_u'(\theta_i')(m_u(\theta_i') + 1)^{(n(\theta_i')-1)}\, dH(u)} \tag{5}$$

where $\alpha_u'(\cdot)$ denotes the Radon–Nikodym derivative of $\alpha_u(\cdot)$ with respect to μ; $\theta_i'(\underline{\theta})$ is the ith distinct value of θ in $\underline{\theta}$; $n(\theta_i')$ is the number of times the value θ_i' occurs in $\underline{\theta}$; $M_u = \alpha_u(\Theta)$ and $m_u(\theta_i') = \alpha_u'(\theta_i')$ if θ_i' is an atom of α_u, zero otherwise.

PROOF. We obtain the joint distribution of $(u, \theta_1, \cdots, \theta_n)$ by calculating the likelihood of $\underline{\theta}$ and making the appropriate normalization. Referring to the proof of Proposition 3 we see that the likelihood of θ_{k+1}, given $u, \theta_1, \cdots, \theta_k$ is $\alpha_u'(\theta_{k+1})\, du/(M_u + k)$ for a value of θ_{k+1} which has not occurred previously in $\theta_1, \cdots, \theta_k$, and $[m_u(\theta_{k+1}) + j]\, d\mu/(M_u + k)$ for a value of θ_{k+1} which has occurred previously j times in $\theta_1, \cdots, \theta_k$. Hence the likelihood of $\underline{\theta} = (\theta_1, \cdots, \theta_n)$ given u is

$$\frac{1}{M_u^{(n)}} \prod_{i=1}^{r} \alpha_u'(\theta_i')(m_u(\theta_i') + 1)^{(n(\theta_i')-1)}\, d\mu^n .$$

We obtain $dH_{\underline{\theta}}(u)$ by multiplying the above by $dH(u)$ and dividing by the unconditional distribution of $\underline{\theta}$. □

We have thus established the validity of the term $dH_{\underline{\theta}}(u)$ in the following extension of Corollary 3.2.

COROLLARY 3.2′. *Let P and α_u be as in Lemma* 1 *and let $\underline{\theta} = (\theta_1, \cdots, \theta_n)$ be a sample of size n from P. Then*

$$P \mid \underline{\theta} \in \int_U \mathscr{D}(\alpha_u + \textstyle\sum_{i=1}^{n} \delta_{\theta_i})\, dH_{\underline{\theta}}(u) ,$$

where $H_{\underline{\theta}}$ is the conditional distribution of u given $\underline{\theta}$.

5. Estimation problems: parameters, mixing distributions, and empirical Bayes. In this section we present a general sampling model using mixtures of Dirichlet processes, and point out some cases where this Dirichlet model leads to estimates different from standard Bayesian analysis. Consider a mixture of Dirichlet processes where the index space, parameter space, and observation space are all the real line with the σ-field of Borel sets. Let $G(\theta)$ be a sample distribution function from a mixture with parameter α_u and mixing distribution $H(u)$. Let $\theta_1, \theta_2, \cdots, \theta_n$ be a sample of size n from $G(\theta)$ and let $X_{i1}, \cdots, X_{im_i}$ be a sample of size m_i from $F_{\theta_i}(x)$. We consider the following problems:

(a) *Estimating the index of the parameter.* If we wish to estimate u with squared error loss, then the Bayes estimate is simply $\hat{u} = E(U \mid \theta_1, \theta, \cdots, \theta_n)$ if the θ_i are observed directly, and $\hat{u} = E(u \mid X_{11}, \cdots, X_{nm_n})$ if we only observe the X_{ij}. The theory is given in Theorem 3 and the required conditional distribution of u can be obtained using Lemma 1. The positive probability of duplications among the θ_i causes some complications that will be illustrated in an example at the end of this section.

(b) *Estimating the (mixing) distribution function.* Suppose we wish an estimate $\hat{G}$ for G with squared error loss, weighted according to some finite measure W on $(-\infty, \infty)$, i.e., $L(G, \hat{G}) = \int_{-\infty}^{\infty} (G - \hat{G})^2\, dW$. Then $\hat{G} = E(G \mid \theta_1, \theta_2, \cdots, \theta_n)$

is the Bayes estimate when the θ_i are observed, and $\hat{G} = E(G \mid X_{11}, \cdots, X_{nm_n})$ when only the X_{ij} are observed, the latter being the case when $G(\theta)$ is an unknown mixing distribution.

(c) *The empirical Bayes problem.* When $G(\theta)$ is an unknown mixing distribution, we may wish to estimate θ_i given $X_{11}, \cdots, X_{im_i}$ with squared error loss. The Bayes estimate is $\hat{\theta}_i = E(\theta_i \mid X_{11}, \cdots, X_{im_i})$. We illustrate the problems discussed above with an example where the prior guess is that the distribution is approximately normal. For conciseness of notation we let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution function or measure as the context requires.

EXAMPLE 4. Let $G(\theta)$ be a sample distribution function from a mixture of Dirichlet processes on $(-\infty, \infty)$ with transition measure $\alpha_u = M\mathcal{N}(u, \sigma^2)$, mixing distribution $H = \mathcal{N}(0, \rho^2)$, and sampling distribution $F_\theta = \mathcal{N}(\theta, \tau^2)$. Before treating problems (a), (b), and (c) in turn, we list some of the conditional distributions we will need in the analysis, for a sample of size 2, which will be sufficient to illustrate the interesting features of this model. To save space we omit the derivations, and to simplify many of the expressions we define the following constants: $\alpha = (\rho^2 + \sigma^2 + \tau^2)^{-1}$, $\alpha' = (\sigma^2 + \tau^2)^{-1}$, $\beta = (2\rho^2 + \sigma^2 + \tau^2)^{-1}$, $\beta' = (2\rho^2 + 2\sigma^2 + \tau^2)^{-1}$.

$$G \mid \theta_1, \theta_2 \in \textstyle\int_{-\infty}^{\infty} \mathcal{D}(M\mathcal{N}(u, \sigma^2) + \delta_{\theta_1} + \delta_{\theta_2})\, dH(u \mid \theta_1, \theta_2) \tag{6}$$

where $H(u \mid \theta_1, \theta_2) = \mathcal{N}(\mu_1, \sigma_1^2)$, $\mu_1 = 2\bar{\theta}\rho^2/(2\rho^2 + \sigma^2)$ and $\sigma_1^2 = \rho^2\sigma^2/(2\rho^2 + \sigma^2)$ when $\theta_1 \neq \theta_2$; $\mu_1 = \bar{\theta}\rho^2/(\rho^2 + \sigma^2)$, $\sigma_1^2 = \rho^2\sigma^2/(\rho^2 + \sigma^2)$ when $\theta_1 = \theta_2$.

$$G \mid X_1 \in \textstyle\int_{-\infty}^{\infty} \mathcal{D}(M\mathcal{N}(u, \sigma^2) + \delta_\theta)\, dH_{X_1}(\theta, u) \tag{7}$$

where $H_{X_1}(\theta, u)$ is bivariate Normal with $\mu_1 = X_1\alpha(\rho^2 + \sigma^2)$, $\mu_2 = X_1\alpha\rho^2$, $\sigma_1^2 = \alpha\tau^2(\rho^2 + \sigma^2)$, $\sigma_2^2 = \alpha\rho^2(\sigma^2 + \tau^2)$, $\sigma_{21} = \alpha\tau^2\rho^2$.

$$G \mid X_1, X_2 \in \textstyle\int_{-\infty}^{\infty} \mathcal{D}(M\mathcal{N}(u, \sigma^2) + \delta_{\theta_1} + \delta_{\theta_2})\, dH_{X_1, X_2}(\theta_1, \theta_2, u) \tag{8}$$

where $H_{X_1, X_2}(\theta_1, \theta_2, u) = p_d\mathcal{N}(\underline{\mu}, \Sigma) + p_s\mathcal{N}(\underline{\mu}^*, \Sigma^*)$, a mixture of trivariate Normal distributions. The mixing coefficient $p_s = P(\theta_1 = \theta_2 \mid X_1, X_2) = (1 + \tau M(\alpha'\beta/\beta')^{\frac{1}{2}} \exp\{\sigma^2((X_1 - X_2)^2\alpha'/4\tau^2 - \bar{X}^2\beta'\beta)\})^{-1}$ and $p_d = 1 - p_s$. The means, variances, and covariances of the distribution $\mathcal{N}(\underline{\mu}, \Sigma)$ are $\mu_1 = \alpha'(X_1\sigma^2 + \bar{X}2\beta\rho^2\tau^2)$, $\mu_2 = \alpha'(X_2\sigma^2 + \bar{X}2\beta\rho^2\tau^2)$, $\mu_3 = \bar{X}2\beta\rho^2$, $\sigma_1^2 = \sigma_2^2 = \alpha'\beta\tau^2(\sigma^4 + 2\sigma^2\rho^2 + \tau^2\rho^2 + \sigma^2\tau^2)$, $\sigma_{21} = \alpha'\beta\tau^4\rho^2$, $\sigma_{31} = \sigma_{32} = \alpha'\beta\tau^2\rho^2(\sigma^2 + \tau^2)$, $\sigma_3^2 = \alpha'\beta\rho^2(\sigma^2 + \tau^2)^2$.

The component $\mathcal{N}(\underline{\mu}^*, \Sigma^*)$ is a singular trivariate Normal where all the mass of the distribution is concentrated on the $\theta_1 = \theta_2$ plane. In this case, $\mu_1^* = \mu_2^* = 2\bar{X}(\rho^2 + \sigma^2)\beta'$, $\mu_3^* = 2\bar{X}\rho^2\beta'$, $\sigma_1^{*2} = \sigma_2^{*2} = \sigma_{21}^* = \tau^2(\rho^2 + \sigma^2)\beta'$, $\sigma_3^{*2} = \rho^2(2\sigma^2 + \tau^2)\beta'$, $\sigma_{13}^* = \sigma_{23}^* = \rho^2\tau^2\beta'$.

We can now write down directly the solutions to the problems posed above. For (a) we get $\hat{u}_\theta = \bar{\theta}2\rho^2/(\sigma^2 + 2\rho^2)$ when $\theta_1 \neq \theta_2$, and $u_\theta = \bar{\theta}\rho^2/(\sigma^2 + \rho^2)$ when $\theta_1 = \theta_2$. Thus the estimate $\hat{u}_\theta$ is discontinuous at $\theta_1 = \theta_2$. The heuristic explanation for this is that when $\theta_1 = \theta_2$, it is almost surely due to the atom at θ_1, and gives no information about u, hence the estimate for u reduces to that given only

θ_1. If we are given X_1 and X_2, but do not get to see θ_1 and θ_2, we do not know whether $\theta_1 = \theta_2$, so we must consider two estimates, $\hat{u} = \bar{X}2\rho^2/(2\rho^2 + \sigma^2 + \tau^2)$, which is appropriate when $\theta_1 \neq \theta_2$, and $\hat{u}^* = \bar{X}2\rho^2/(2\rho^2 + 2\sigma^2 + \tau^2)$ when $\theta_1 = \theta_2$. Since we do not observe whether $\theta_1 = \theta_2$ or not, we must weight these two estimates according to the posterior probability, given X_1 and X_2, that $\theta_1 = \theta_2$. This gives the estimate $\hat{u}' = p_d\hat{u} + p_s\hat{u}^*$.

The solution to problem (b), estimating $G(\theta)$, requires that we compute $E(G(\theta) \mid \theta_1, \theta_2)$ and $E(G(\theta) \mid X_1, X_2)$. From Proposition 1 we get

$$E(G(\theta) \mid \theta_1, \theta_2) = \frac{M}{M+2} \int_{-\infty}^{\infty} \Phi\left(\frac{\theta - u}{\sigma}\right) dH(u \mid \theta_1, \theta_2) + \frac{2}{M+2} F_2(\theta) .$$

For $\theta_1 \neq \theta_2$, we get

$$\hat{G}(\theta) = \frac{M}{M+2} \mathcal{N}\left(\frac{\bar{\theta}2\rho^2}{2\rho^2 + \sigma^2}, \frac{\sigma^4 + 3\rho^2\sigma^2}{2\rho^2 + \sigma^2}\right) + \frac{\delta_{\theta_1} + \delta_{\theta_2}}{M+2} . \tag{9}$$

If $\theta_1 = \theta_2$, then for the reason given in (a) above, we get

$$\hat{G}(\theta) = \frac{M}{M+2} \mathcal{N}\left(\frac{\bar{\theta}\rho^2}{\rho^2 + \sigma^2}, \frac{\sigma^4 + 2\sigma^2\rho^2}{\rho^2 + \sigma^2}\right) + \frac{2}{M+2} \delta_{\theta_1} .$$

If we observe only the X_1, X_2, then

$$\begin{aligned} &E(G(\theta) \mid X_1, X_2) \\ &\quad = p_d\left[\frac{M}{M+2} \mathcal{N}\left(\frac{\bar{X}2\rho^2}{2\rho^2 + \sigma^2 + \tau^2}, \frac{3\rho^2\sigma^2 + \sigma^4 + \sigma^2\tau^2 + \rho^2\tau^2}{2\rho^2 + \sigma^2 + \tau^2}\right)\right. \\ &\qquad + \frac{1}{M+2} \mathcal{N}\left(\frac{X_1\sigma^2 + \bar{X}2\beta\rho^2\tau^2}{\sigma^2 + \tau^2}, \frac{\tau^2(\sigma^4 + 2\sigma^2\rho^2 + \sigma^2\tau^2 + \tau^2\rho^2)}{(2\rho^2 + \sigma^2 + \tau^2)(\sigma^2 + \tau^2)}\right) \\ &\qquad \left. + \frac{1}{M+2} \mathcal{N}\left(\frac{X_2\sigma^2 + \bar{X}2\beta\rho^2\tau^2}{\sigma^2 + \tau^2}, \frac{\tau^2(\sigma^4 + 2\sigma^2\rho^2 + \sigma^2\tau^2 + \tau^2\rho^2)}{(2\rho^2 + \sigma^2 + \tau^2)(\sigma^2 + \tau^2)}\right)\right] \\ &\qquad + p_s\left[\frac{M}{M+2} \mathcal{N}\left(\frac{\bar{X}2\rho^2}{2\rho^2 + 2\sigma^2 + \tau^2}, \frac{4\rho^2\sigma^2 + 2\sigma^4 + \sigma^2\tau^2 + \rho^2\tau^2}{2\rho^2 + 2\sigma^2 + \tau^2}\right)\right. \\ &\qquad \left. + \frac{2}{M+2} \mathcal{N}\left(\frac{\bar{X}2(\rho^2 + \sigma^2)}{2\rho^2 + 2\sigma^2 + \tau^2}, \frac{\tau^2(\rho^2 + \sigma^2)}{2\rho^2 + 2\sigma^2 + \tau^2}\right)\right] . \end{aligned} \tag{10}$$

Finally, we write down the solution to (c), the empirical Bayes problem. We have immediately $\hat{\theta}_1 = E(\theta \mid X_1) = X_1\rho^2/(\rho^2 + \sigma^2 + \tau^2)$ and $\hat{\theta}_2 = E(\theta_2 \mid X_1, X_2) = p_d(X_2\sigma^2 + \bar{X}2\beta\rho^2\tau^2/(\sigma^2 + \tau^2) + p_s 2\bar{X}(\rho^2 + \sigma^2)/(2\rho^2 + 2\sigma^2 + \tau^2)$. Moreover, if the formulation of the problem allows "hindsight," then given X_1 and X_2 we can obtain an expression for a revised estimate of θ_1 by replacing the subscript 2 by 1 in the expression for $\hat{\theta}_2$. The effect of the possibility that $\theta_1 = \theta_2$ is to shift $\hat{\theta}_1 \mid X_1, X_2$ away from $\hat{\theta}_1 \mid X_1$ and toward $\hat{\theta}_2$.

Certain features of the preceding example illustrate an important difference between simple Dirichlet processes and mixtures of Dirichlet processes. Consider $G \in \mathcal{D}(M\mathcal{N}(u, \sigma^2))$ as above and assume for the moment that u is known, so that we have a simple Dirichlet process. Let $\bar{G}_0 = E(G)$ and $\bar{G}_2 = E(G \mid \theta_1, \theta_2)$,

and let A denote the open interval (θ_1, θ_2), (assume without loss of generality that $\theta_1 < \theta_2$). Note that $\bar{G}_2(A) = \bar{G}_0(A)M/(M+2)$ and in general, for any set which does not include any of the observations, the expected value of the probability assigned that set under the posterior distribution is smaller by a factor $M/(M+2)$ than under the prior, even though the set may be "close" to the observations, as (θ_1, θ_2) is.

However, if we let u be an index with mixing distribution $\mathcal{N}(0, \rho^2)$ as originally in the example, then although θ_1 and θ_2 still do not contribute anything directly to $\bar{G}_2(A)$, they do cause the *mean* of the posterior distribution of u given θ_1 and θ_2 to be shifted away from zero toward A, and the *variance* to be decreased to $\rho^2\sigma^2/(2\rho^2 + \sigma^2)$. These changes may more than compensate for the factor $M/(M+2)$, especially if ρ is large compared to σ, $(\rho \gg \sigma)$, as seen from (9).

Next consider the case where we only observe X_1 and X_2. Let $B = (X_1, X_2)$, (assume $X_1 < X_2$), and $\bar{G}_2^* = E(G \mid X_1, X_2)$. Referring to (10), we can see several terms which contribute to $\bar{G}_2^*(B)$. There is a change in the posterior distribution of u similar to that described for θ_1 and θ_2 above. There is a component $p_d/(M+2)$ which would have been concentrated on θ_1 if θ_1 had been observed, but which is now spread out around the estimated value of θ_1, and some of this mass falls in the interval B; similarly for θ_2. Finally there is a component $2p_s/(M+2)$ spread out around the midpoint of the estimates for θ_1 and θ_2. If $\rho \gg \sigma \gg \tau$, then $\hat{\theta}_i$ is near X_i, $i = 1, 2$ for the p_d components, and $\hat{\theta}$ is near $\bar{X}$ for the p_s component. Thus $\bar{G}_2^*(B)$ may be much larger than $\bar{G}_0(B)$, even though B does not contain X_1 or X_2.

6. A regression problem. The problems considered in this section are similar to those in Section 5, in that the goal is a Bayes estimate of an unknown distribution function G on $[0, 1]$, with loss function $L(G, \hat{G}) = \int_{-\infty}^{\infty} (G(t) - \hat{G}(t))^2\, dW(t)$, where $W(t)$ is some finite measure. Again we assume G to be chosen by a Dirichlet process, but this time the sampling technique is more like that used in regression problems. If $G(t)$ represents the cdf of G, then various values of t, say $0 \leqq t_1 < \cdots < t_k \leqq 1$ are chosen and the unknown value of $G(t_i)$ becomes a parameter in a distribution $F(x \mid G(t_i))$. Samples from $F(x \mid G(t_i))$ are used to make inferences about the value of $G(t_i)$.

If G is a sample function from a Dirichlet process with parameter α, let $Y_1 = G(t_1)$, $Y_2 = G(t_2) - G(t_1)$, $\cdots$, $Y_{k+1} = 1 - G(t_k)$, and $\beta_1 = \alpha(t_1)$, $\beta_2 = \alpha(t_2) - \alpha(t_1)$, $\cdots$, $\beta_{k+1} = \alpha(1) - \alpha(t_k)$. Then the joint distribution of $Y_1, \cdots, Y_{k+1}$ is Dirichlet with parameters $\beta_1, \cdots, \beta_{k+1}$. Hence the observations for different values of i will not be stochastically independent in general. We illustrate the effect of this dependence with an example where $F(x \mid G(t))$ has a simple density function, and where $k = 2$.

EXAMPLE 5. Let P be a Dirichlet process on $([0, 1], \mathcal{B})$, $\mathcal{B}$ the Borel sets, with strictly monotone parameter α and let $F(x \mid G(t))$ have density

$$f(x \mid G(t)) = 2[xG(t) + (1 - x)(1 - G(t))] \qquad 0 < x < 1\,,$$

and zero elsewhere. For given t_1 and t_2 define Y_1 and Y_2 as above and let X_1 and X_2 given Y_1 and Y_2 be independent samples from $F(x \mid G(t_1))$ and $F(x \mid G(t_2))$ respectively. The joint density becomes

$$f(x_1, x_2, y_1, y_2) = 4[x_1 y_1 + (1 - x_1)(1 - y_1)][x_2(y_1 + y_2) + (1 - x_2)(1 - y_1 - y_2)] \times y_1^{\beta_1 - 1} y_2^{\beta_2 - 1}(1 - y_2 - y_1)^{\beta_3 - 1} \frac{\Gamma(M)}{\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_3)}.$$

A straightforward, but tedious, algebraic manipulation, using Bayes formula, yields

$$\begin{aligned}(Y_1, Y_2 \mid X_1 = x_1, X_2 = x_2) \\ \in c^{-1}(x_1, x_2)\{&x_1 x_2 \beta_1(\beta_1 + 1)\mathscr{D}(\beta_1 + 2, \beta_2, \beta_3) \\ &+ (1 - x_1)x_2 \beta_2(\beta_2 + 1)\mathscr{D}(\beta_1, \beta_2 + 2, \beta_3) \\ &+ (1 - x_1)(1 - x_2)\beta_3(\beta_3 + 1)\mathscr{D}(\beta_1, \beta_2, \beta_3 + 2) \\ &+ x_2 \beta_1 \beta_2 \mathscr{D}(\beta_1 + 1, \beta_2 + 1, \beta_3) \\ &+ (x_1 + x_2 - 2x_1 x_2)\beta_1 \beta_3 \mathscr{D}(\beta_1 + 1, \beta_2, \beta_3 + 1) \\ &+ (1 - x_1)\beta_2 \beta_3 \mathscr{D}(\beta_1, \beta_2 + 1, \beta_3 + 1)\}\end{aligned}$$

where $c(x_1, x_2) = (x_1 x_2 \beta_1(\beta_1 + 1) + \cdots + (1 - x_1)\beta_2 \beta_3)$.

The algebraic manipulation referred to above makes frequent use of property (ii) of the Dirichlet distribution given in Section 2. The technique is not limited to linear densities, but the example shows that even with linear densities and small k, the posterior distribution is somewhat unwieldy for hand computations.

Nevertheless, it is possible to transform any density which can be expressed as a finite polynomial in $G(t)$ into a polynomial in the Y_i's and absorb it into the Dirichlet density function. This leaves open the possibility of approximating density functions of interest by polynomials, and performing the required computations on a computer.

7. A bio-assay problem. The next problem we treat is a type of bio-assay problem. We wish to estimate the dose response curve of some animals to a certain drug. We assume, for simplicity, that an animal's response to a given dose of the drug is either positive, or negative (no response), and that each animal has a threshold that the dose given him must exceed to produce a positive response. However, this threshold varies from one animal to the next, so we treat it as a random variable with unknown distribution G. Kraft and van Eeden [12] have treated this problem using a process of the Dubins and Freedman type [5] to choose the distribution G. Our approach will be to let G be a sample function from a Dirichlet process. Essentially the same model has been considered by Ramsey [15].

Let $(\Theta, \mathscr{A}) = ([0, 1], \mathscr{B})$ be the unit interval with Borel sets and assume, then, that G is chosen by a Dirichlet process with parameter α, and we select some dosage level t and administer this dosage to n animals. $G(t)$ is the expected proportion of the animals whose response threshold is less than or equal to t,

that is, we "expect" $nG(t)$ animals to give a positive response. Of course, $G(t)$ is unknown, since G was chosen at random, but since G was chosen by a Dirichlet process, the prior distribution of $G(t)$ is $\mathscr{B}e(\alpha(t), M - \alpha(t))$. And, as is well known, the posterior distribution of $G(t)$ given k positive responses in n trials is $\mathscr{B}e(\alpha(t) + k, M + n - k - \alpha(t))$ where $\mathscr{B}e(\alpha_1, \alpha_2)$ is the Beta distribution.

This analysis for one dosage level was deceptively simple, since things become much more complicated with the use of more than one level, as may be seen in the case when two levels, t_1 and t_2 are used, with n_1 and n_2 trials respectively. Assume that $t_1 < t_2$, and $[\alpha(t_2) - \alpha(t_1)] > 0$. Let K_i be the number of successes among the n_i trials at t_i. Then the joint density of K_1, K_2, $G(t_1)$, $G(t_2)$ is most easily expressed in terms of Y_i's, where $Y_1 = G(t_1)$, $Y_2 = G(t_2) - G(t_1)$ and $Y_3 = 1 - G(t_2)$; and β_i's, where $\beta_1 = \alpha(t_1)$, $\beta_2 = \alpha(t_2)$, $\beta_3 = \alpha(1) - \alpha(t_2)$. The joint density of (K_1, K_2, Y_1, Y_2) then becomes

$$\binom{n_1}{k_1}\binom{n_2}{k_2} y_1^{k_1}(1 - y_1)^{n_1-k_1}(y_1 + y_2)^{k_2}(y_3)^{n_2-k_2} \frac{\Gamma(M)}{\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_3)} y_1^{\beta_1-1} y_2^{\beta_2-1} y_3^{\beta_3-1}$$

on $S = \{y_1, y_2 \mid y_1 \geqq 0, y_2 \geqq 0, y_1 + y_2 \leqq 1\}$ and where $y_3 = 1 - y_1 - y_2$.

We can transform this into an expression which is recognizable as a mixture of Dirichlet distributions by making the substitutions $(1 - y_1) = y_2 + y_3$, and expanding $(y_2 + y_3)^{n_1-k_1}$ and $(y_1 + y_2)^{k_2}$ using the Binomial formula. This leads finally to an expression for the conditional distribution of Y_1, Y_2 given $K_1 = k_1$, $K_2 = k_2$ as $\sum_{i=0}^{k_2} \sum_{j=0}^{n_1-k_1} a_{ij} \mathscr{D}(\beta_1 + k_1 + i, \beta_2 + n_1 - k_1 + k_2 - i - j, \beta_3 + n_2 - k_2 + j)$ where $a_{ij} = b_{ij} / \sum_{i=0}^{k_2} \sum_{j=0}^{n_1-k_1} b_{ij}$ and

$$b_{ij} = \binom{n_1 - k_1}{j}\binom{k_2}{i} \frac{\Gamma(\beta_1 + k_1 + i)\Gamma(\beta_2 + n_1 - k_1 + k_2 - i - j)\Gamma(\beta_3 + n_2 - k_2 + j)}{\Gamma(\beta_1)\Gamma(\beta_2)\Gamma(\beta_3)}.$$

Before proceeding to our goal of a Bayes estimate of G we examine the expressions above in more detail to determine the source of the increased complexity. It is helpful to consider the problem from a slightly different viewpoint. Saying that K_1 of the n_1 trials at t_1 were successes is statistically equivalent to saying that a sample $X_1, \cdots, X_{n_1}$ of size n_1 from G was taken, but all that was recorded was that K_1 of the X_i's were less than or equal to t_1. In particular, of the $n_1 - K_1$ values of the X_i which were greater than t_1, it is known how many fell in the interval $(t_1, t_2]$ and how many in $(t_2, 1]$. Hence we let J denote the number of the X_i's falling in the interval $(t_2, 1]$. Since the true value of J is unknown, we must enumerate and weight appropriately all possible values of J from 0 to $n_1 - K_1$. Similarly, if $Y_1 Y_2, \cdots, Y_{n_2}$ is a sample of size n_2 from G, and K_2 values of Y_i are less than or equal to t_2, we let I denote the number of the Y_i's that fell in the interval $[0, t_1]$. Now if it were known that $I = i$ and $J = j$, the joint distribution of $G(t_1)$, $G(t_2) - G(t_1)$, and $1 - G(t_2)$, given $K_1 = k_1$, $J = j$, $K_2 = k_2$, $I = i$ would be Dirichlet with parameters $(\beta_1 + k_1 + i, \beta_2 + n_1 - k_1 + k_2 - i - j, \beta_3 + j + n_2 - k_2)$. Since I and J are unknown, the proper expression is a mixture over all possible values of I and J as given by the double summation above.

By making a similar analysis for the case where there are three thresholds, one can show the corresponding summation runs over 6 indices, and in general, for m thresholds there would be $m(m-1)$ indices to be summed over. Generally such tasks are best left to the computer.

To resume our treatment of this problem we compute our estimate $\hat{G}$ of G. Recall that for squared error loss, the Bayes estimate is the mean of the posterior distribution of G. Moreover, since $\hat{G}$ is linear in the intervals $[0, t_1)$, $[t_1, t_2)$, $[t_2, 1]$, we need only calculate $\hat{G}(t_1)$ and $\hat{G}(t_2)$.

$$\hat{G}(t_1) = E(G(t_1)) = \sum_{i=0}^{k_2} \sum_{j=0}^{n_1-k_1} a_{ij} \frac{\beta_1 + k_1 + i}{M + n_1 + n_2}$$

$$\hat{G}(t_2) = E(G(t_2)) = \sum_{i=0}^{k_2} \sum_{j=0}^{n_1-k_1} a_{ij} \frac{\beta_1 + \beta_2 + n_1 + k_2 - j}{M + n_1 + n_2}$$

$$\begin{aligned} G(t) &= \hat{G}(t_1)t/t_1 && 0 \leqq t \leqq t_1 \\ &= \hat{G}(t_1) + [(t - t_1)/(t_2 - t_1)]\{\hat{G}(t_2) - \hat{G}(t_1)\}, && t_1 \leqq t \leqq t_2 \\ &= \hat{G}(t_2) + [(t - t_2)/(1 - t_2)]\{1 - \hat{G}(t_2)\}, && t_2 \leqq t \leqq 1. \end{aligned}$$

We recognize $\hat{G}(t_1)$ as a weighted average of the various estimates of $G(t_1)$ we would make if we knew I and J. It can be shown that for large M this estimate approaches the intuitively reasonable value $(\alpha_1 + k_1 + (\alpha_1/\alpha_2)k_2)/(M + n_1 + n_2)$. For small values of M, however, the situation is quite different, and contrary to intuition. Recall that M is, in a sense, a measure of our confidence in the parameter α. Very small values of M can be interpreted as meaning we have practically no prior information about G, and our estimate of the distribution will be heavily weighted in favor of the empirical distribution function. A less obvious consequence of a small M is the implication that the process chooses distributions with most of their mass concentrated at a few points.

EXAMPLE 6. Let $(\Theta, \mathscr{A}) = ([0, 1], \mathscr{B})$, $\alpha(t) = Mt$, $n_1 = n_2 = 100$, $k_1 = 1$, $k_2 = 99$, $t_1 = \frac{1}{3}$, $t_2 = \frac{2}{3}$. In words, of 100 samples where thresholds were compared with $t_1 = \frac{1}{3}$, one was less than or equal to $\frac{1}{3}$, 99 were greater than $\frac{1}{3}$. Of 100 samples compared with $t_2 = \frac{2}{3}$, 99 were less than or equal to $\frac{2}{3}$, and 1 was greater than $\frac{2}{3}$. Recalling that because M is small we expect the process to have chosen a distribution with most of its mass on a few points, we could readily explain the sample result by deciding there is a very large jump in G somewhere on $(\frac{1}{3}, \frac{2}{3}]$, and significantly smaller jumps on $[0, \frac{1}{3}]$ and $(\frac{2}{3}, 1]$, a total of three jumps. When we evaluate the weights a_{ij}, however, we find that the product $\Gamma(M/3 + 1 + i)\Gamma(M/3 + 198 - i - j)\Gamma(M/3 + 1 + j)$ is largest, as $M \to 0$, for $i = 99, j = 99$, since then the center factor becomes $\Gamma(M/3)$ and $\Gamma(M/3) \to \infty$ as $M \to 0$. But when we check the interpretation of these values for I and J, we discover they correspond to estimating $F(\frac{1}{3}) = F(\frac{2}{3}) = \frac{1}{2}$. As $M \to 0$, the dominant posterior distribution is the one that gives a jump $\simeq \frac{1}{2}$ to $[0, \frac{1}{3})$, and a similar jump to $(\frac{2}{3}, 1]$, and practically no weight to $[\frac{1}{3}, \frac{2}{3})$! Roughly speaking, the posterior Dirichlet process gives most of its weight to those distributions

which can "explain" the sample with the *fewest* number of jumps, in this case two jumps instead of the more reasonable sounding three.

We conclude our treatment of the bio-assay problem with a brief discussion of the design problem: Given the model of the bio-assay problem described above, and a budget, say, of n observations total, what are the "best" thresholds t_i to test at, and what are the optimal numbers n_i to test at each threshold? While this problem for general α and W is intractable, we nevertheless feel that for the special case α linear on [0, 1] and W uniform on [0, 1], the following conjecture is true.

Conjecture. Let Θ be [0, 1] and $\mathscr{A}$ the Borel sets, and let the bio-assay response curve $G(\theta)$ be a sample function from a Dirichlet process on $(\Theta, \mathscr{A})$ with parameter $\alpha(t) = Mt$. If we have a total $n = \sum_{i=1}^{k} n_i$ of observations n_i to be compared with thresholds t_i, and we wish to estimate G by $\hat{G}$ to minimize $\int_0^1 (G - \hat{G})^2 \, dW$, $W = t$ on [0, 1], then

(i) If k is fixed beforehand, the Bayes design is to set $t_i = i/(k + 1)$ and make the n_i as nearly equal as possible, i.e., $|n_i - n_j| \leqq 1$ for all i, j.

(ii) If k is not fixed, the Bayes design is to let $k = n$ and take one observation each at the thresholds $t_i = i/(n + 1)$.

Ramsey [15] reaches this same conclusion independently for a similar model. A more intriguing question is what the optimal sequential design strategy is for fixed sample size, and for variable sample size, with the goal of achieving confidence bands for $G(t)$. The author has not had much success in seeking answers to these questions.

8. A discrimination problem. Statistical discrimination problems may be described as follows. We are given samples $X_{i1}, \cdots, X_{ik_i}$ from unknown distributions p_i, $i = 1, \cdots, k$ and a sample $Y_1, \cdots, Y_n$ known only to come from one of the p_i's. The problem is to decide which one.

We might model such a problem by assuming the p_i's are independent samples from a Dirichlet process with parameter α, $\alpha(\Theta) = M$. A little preliminary analysis then reveals that the solution is very sensitive to the assumptions we make about α. If we assume α to be continuous, then the posterior distributions of the p_i will have parameters α_i with atoms of various sizes at each distinct value of the X_{ij}. Moreover, with probability one, the samples from each p_i will be disjoint from the samples from all the other p's, but with probability $1 - (k_i/M + k_i)^n$, at least one sample point Y_i will match some sample point X_{ij} when the Y's come from distribution p_i, in which case the choice of i is clear. If, however, none of the Y_i matches any of the values of the X_{ij}, then the choice of the p_i to ascribe the Y_i to becomes biased in favor of the p_i from which we have the smallest sample of X's, i.e., smallest k_i. If, in this case, the k_i are all equal, the sample of Y's has not given us any information, and a random decision is made.

For these reasons we will assume that α is purely discrete, and for more generality we assume $p_1 \in \mathscr{D}(\alpha_1), \cdots, p_k \in \mathscr{D}(\alpha_k)$, where α_1 through α_k are all discrete with the same support.

In the notation we have been using, let Θ be the nonnegative integers and $\mathscr{A}$ the σ-algebra generated by the singleton sets. Let α_i be the finite nonnegative σ-additive measure defined on $\mathscr{A}$ through its values on the singleton sets, $\underline{\alpha}_i = (\alpha_{i0}, \alpha_{i1}, \cdots, \alpha_{in}, \cdots)$ where $\underline{\alpha}_i(\{j\}) = \alpha_{ij}$, and $|\underline{\alpha}_i| = \sum_{j=0}^{\infty} \alpha_{ij}$. Let $P_i \in \mathscr{D}(\underline{\alpha}_i)$ and $X_{i1}, \cdots, X_{ik_i}$ be a sample from P_i. Let $Y_1, \cdots, Y_n$ be a sample from some P_j, where the prior probability that $Y_1, \cdots, Y_n \in P_j$ is π_j, $j = 1, \cdots, k$. Let $L(i, j)$ denote the loss associated with deciding the Y's come from P_i when they come from P_j. We seek a nonrandomized decision rule which minimizes our expected loss given the observations.

First we notice that the distribution of P_i given $X_{i1}, \cdots, X_{in_i}$ is just $\mathscr{D}(\underline{\alpha}_i')$, where

$$\begin{aligned}\underline{\alpha}_i' &= (\alpha_{i0} + m_{i0}, \alpha_{i1} + m_{i1}, \cdots, \alpha_{in} + m_{in}, \cdots) \\ &= (\alpha_{i0}', \alpha_{i1}', \cdots, \alpha_{in}', \cdots)\end{aligned}$$

and m_{ij} is the number of X_i's equal to j, $j = 0, 1, \cdots$. Hence the problem reduces to deciding which of k different Dirichlet processes with known parameters the samples $Y_1, \cdots, Y_n$ come from. The discrimination problem has been reduced to a classification problem. To treat this we calculate the Bayes risk r_i for each i, where

$$r_i = \sum_{j=1}^{k} L(i, j)\pi(j \mid Y_1, \cdots, Y_n), \qquad \pi(i \mid Y_1, \cdots, Y_n) = \pi_i P(Y \mid i)/P(\underline{Y})$$
$$P(Y \mid i) = \prod_{j=0}^{\infty} \alpha_{ij}'^{(k_j)}/\alpha_j'^{(n)}, \qquad P(\underline{Y}) = \sum_{i=1}^{k} \pi_i P(\underline{Y} \mid i),$$

where k_j = number of Y's equal to j. The Bayes decision rule chooses s where $r_s = \min r_i$. If, for example $L(i, j) = 0$, $i = j$; 1, $i \neq j$, and $\pi_i = 1/k$ for all i, then the Bayes decision rule chooses that s for which $P(Y \mid s) = \max_i P(Y \mid i)$.

9. Acknowledgments. The author is immeasurably indebted to Professor Thomas Ferguson for introducing him to the concept of Dirichlet processes and providing countless helpful suggestions during the course of the research, and to David Blackwell and Jim MacQueen for helpful discussions. A referee and an associate editor for the *Annals of Statistics* made many helpful suggestions for revisions of the manuscript. This paper is an extension of part of the author's dissertation submitted in partial fulfillment of the requirements for the Ph. D. degree at U.C.L.A., written under the supervision of Professor Ferguson.

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards.

[2] BLACKWELL, DAVID (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1** 356–358.

[3] BLACKWELL, D. and MACQUEEN, J. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **1** 353–355.

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

[5] DUBINS, LESTER E. and FREEDMAN, DAVID A. (1966). Random distribution functions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **3** 183–214. Univ. of California Press.

[6] DUBINS, LESTER E. and FREEDMAN, DAVID A. (1963). Random distribution functions. *Bull. Amer. Math. Soc.* **69** 548–551.

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

[8] FELLER, WILLIAM (1968). *An Introduction to Probability Theory and its Applications*, **1**. Wiley, New York.

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

[10] KORWAR, R. and HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Prob.* **1** 705–711.

[11] KRAFT, CHARLES H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Probability* **1** 385–388.

[12] KRAFT, CHARLES H. and VAN EEDEN, CONSTANCE (1964). Bayesian bio-assay. *Ann. Math. Statist.* **35** 886–890.

[13] PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press, New York.

[14] RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* MIT Press, Cambridge.

[15] RAMSEY, F. L. (1972). A Bayesian approach to Bio-assay. *Biometrics* **28** 841–858.

[16] ROBBINS, HERBERT (1963). An empirical Bayes approach to testing statistical hypotheses. *Rev. Internat. Statist. Inst.*

[17] SCOTT, W. R. (1964). *Group Theory.* Prentice-Hall, Englewood Cliffs, New Jersey.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720