The Annals of Statistics
1974, Vol. 2, No. 4, 615-629

SPECIAL INVITED PAPER

PRIOR DISTRIBUTIONS ON SPACES OF PROBABILITY MEASURES[1]

BY THOMAS S. FERGUSON

University of California, Los Angeles

Methods of generating prior distributions on spaces of probability measures for use in Bayesian nonparametric inference are reviewed with special emphasis on the Dirichlet processes, the tailfree processes, and processes neutral to the right. Some applications are given.

0. Introduction. Recently there has been active research in a narrow but important area, that of construction of prior distributions on spaces of probability measures for use in deriving Bayesian decision rules in nonparametric statistical problems. There are two desirable properties of such prior distributions: (1) the support of the prior with respect to some suitable topology on the space of probability measures should be large, and (2) the posterior distribution given a sample from the true probability measure should be manageable analytically.

One of the drawbacks of decision theory in general and of the Bayesian approach to it in particular is the difficulty of putting the cost of the computation into the model. This drawback is particularly severe in Bayesian nonparametric problems. There are no doubt examples in which "quick and easy" rules are preferable to "optimal" rules for a Bayesian simply because it costs less to perform the computations. On the other hand, Bayes rules are certainly desirable since generally they are admissible and have nice large sample properties [12]. Therefore, it behooves the statistician to suggest large classes of easily computable Bayes rules in the hope that users may find some rules to their liking.

It is the purpose of this paper to review the literature in this area and the somewhat limited success to date. Because the earlier papers of Freedman [19] and Fabius [14] were concerned mainly with other problems, their value in Bayesian nonparametric statistics has not been generally appreciated. It is hoped that this paper serves to recognize this work and to clarify the relationship with later work.

For ease of exposition, we restrict attention unless otherwise specified to prior distributions on the space of all probability measures on $(\mathbb{R}, \mathscr{B})$ where $\mathbb{R}$ is the

Received January 1973; revised October 1973.

[1] Special invited address at the annual meeting of the I.M.S., Hanover, August 28—September 1, 1972. This work was supported in part by NSF Grant No. GP-33431X.

AMS 1970 *subject classifications*. Primary 62C10; Secondary 60B05.

Key words and phrases. Bayesian nonparametric inference, prior distributions, Dirichlet process, tailfree processes, characterization of distributions, neutral to the right, adaptive sampling with recall, adaptive investment.

real line and $\mathscr{B}$ is the σ-algebra of Borel subsets of $\mathbb{R}$. Let

$$\mathscr{F} = \{P\colon P \text{ is a probability measure on } (\mathbb{R}, \mathscr{B})\}.$$

(In this paper, the word "measure" refers to a nonnegative σ-additive set function.) Let $\mathscr{A}$ denote some suitable σ-algebra of subsets of $\mathscr{F}$, for example the Borel sets with respect to the topology of weak convergence. We use $\mathscr{P}$ to denote a probability measure on $(\mathscr{F}, \mathscr{A})$, and $\mathscr{E}$ to denote the expectation with respect to $\mathscr{P}$. We let P denote the random probability measure chosen according to $\mathscr{P}$, so that $\int x\,dP(x)$ and $P(B)$ for $B \in \mathscr{B}$ are random variables. We let $X_1, X_2, \cdots, X_n$ denote a random sample chosen according to P. The information $X_1, X_2, \cdots, X_n$ is to be used to make inferences about the true value of P.

We may restate the problem. Find $\mathscr{P}$ so that (1) the support of $\mathscr{P}$ with respect to weak convergence, say, is $\mathscr{F}$, and (2) the posterior distribution of P given $X_1, \cdots, X_n$ is manageable analytically.

1. The Dirichlet process. The simplest of the prior distributions $\mathscr{P}$ described in this paper are the Dirichlet processes. There are a large number of Dirichlet processes on the real line, one for each finite non-null measure α on $(\mathbb{R}, \mathscr{B})$. We refer to α as the parameter of the process. There are various ways to describe this process, denoted by $\mathscr{D}(\alpha)$, of which two are presented below. The main source of the results is [17].

We use $\mathscr{G}(\alpha, \beta)$ for $\alpha > 0$ and $\beta > 0$ to represent the gamma distribution with density $\Gamma(\alpha)\beta^{-\alpha}e^{-x/\beta}x^{\alpha-1}I_{(0,\infty)}(x)$, where I denotes the indicator function. If $\alpha = 0$, $\mathscr{G}(\alpha, \beta)$ is defined to be degenerate at zero. We use $\mathscr{B}e(\alpha, \beta)$ for $\alpha > 0$ and $\beta > 0$ to denote the beta distribution with density $\Gamma(\alpha + \beta)(\Gamma(\alpha)\Gamma(\beta))^{-1}x^{\alpha-1}(1 - x)^{\beta-1}I_{(0,1)}(x)$. If $\alpha = 0$ and $\beta > 0$, $\mathscr{B}e(\alpha, \beta)$ is defined to be degenerate at zero, while if $\alpha > 0$ and $\beta = 0$, $\mathscr{B}e(\alpha, \beta)$ is defined to be degenerate at one.

For the purposes of this paper, it is convenient to define the m-dimensional Dirichlet distribution with parameter $(\alpha_1, \alpha_2, \cdots, \alpha_m)$, where $\alpha_i \geqq 0$ and $\sum_1^m \alpha_1 > 0$, as the distribution of $(Z_1/S, Z_2/S, \cdots, Z_m/S)$, where $Z_1, Z_2, \cdots, Z_m$ are independent random variables with $Z_i \in \mathscr{G}(\alpha_i, 1)$ $i = 1, \cdots, m$, and $S = \sum_1^m Z_i$. The one-dimensional marginal distributions of the Dirichlet are beta, for example, $Z_1/S \in \mathscr{B}e(\alpha_1, \sum_2^m \alpha_i)$. A convenient source of information on the Dirichlet distribution is Wilks [27].

DEFINITION 1. Let $\alpha(\cdot)$ be a finite non-null measure on $(\mathbb{R}, \mathscr{B})$, and let $P(\cdot)$ be a stochastic process indexed by elements of $\mathscr{B}$. We say P is a Dirichlet process with parameter α and write $P \in \mathscr{D}(\alpha)$, if for every finite measurable partition $\{B_1, \cdots, B_m\}$ of $\mathbb{R}$ (i.e. the B_i are measurable, disjoint, and $\bigcup_1^m B_i = \mathbb{R}$), the random vector $(P(B_1), \cdots, P(B_m))$ has a Dirichlet distribution with parameter $(\alpha(B_1), \cdots, \alpha(B_m))$.

In particular, for every $B \in \mathscr{B}$, $P(B) \in \mathscr{B}e(\alpha(B), \alpha(\mathbb{R}) - \alpha(B))$ and therefore $\mathscr{E}P(B) = \alpha(B)/\alpha(\mathbb{R})$.

For the second definition of the Dirichlet process on $(\mathbb{R}, \mathscr{B})$, it is convenient

to use the distribution function form of the measures. Let

$$\alpha(t) = \alpha((-\infty, t]) \qquad \text{and} \qquad F(t) = P((-\infty, t]) .$$

DEFINITION 2. We say $P \in \mathscr{D}(\alpha)$ (or $F \in \mathscr{D}(\alpha)$), if the process $F(t)$ may be written as Z_t/Z_∞ where Z_t is a process with independent increments, $Z_t \in \mathscr{G}(\alpha(t), 1)$ and $Z_\infty = \lim_{t\to\infty} Z_t \in \mathscr{G}(\alpha(\mathbb{R}), 1)$.

To contrast these definitions, note that although the existence of the process of Definition 1 requires a demonstration, the existence of the process of Definition 2 is immediate since the existence of the independent increment process with gamma distributions, sometimes called the gamma process, is well known. On the other hand, Definition 1 may be used to define the Dirichlet process on an arbitrary measurable space, whereas the extension of the gamma process to arbitrary spaces is not so well known. Nevertheless,

Fact 1. These two definition are equivalent.

An alternative definition, that views the Dirichlet process as a limit of Polya urn schemes, may be found in Blackwell and MacQueen [4].

Fact 2. If $F \in \mathscr{D}(\alpha)$, then with probability one F is discrete.

It is well known that the separable version of the gamma process on the real line increases only in jumps with probability one. Therefore F with probability one is a discrete distribution function. The gamma process, and hence the Dirichlet process on an arbitrary measurable space, also gives probability one to sums of point masses. This may be seen easily using a result of Ferguson and Klass [18]. Using the first definition, results of Blackwell [3] show that Dirichlet processes on arbitrary spaces concentrate on discrete distributions.

Fact 3. The support of $\mathscr{D}(\alpha)$ with respect to the topology of weak convergence is the set of all distributions whose support is contained in the support of α.

This is a version of desirable property 1 mentioned in the introduction. If the support of α is $\mathbb{R}$, then the support of $\mathscr{D}(\alpha)$ with respect to convergence in law is $\mathscr{F}$.

Fact 4. For any nonnegative measurable function g, $\int g(t)\, d\alpha(t) < \infty$ if, and only if, $\int g(t)\, dF(t) < \infty$ with probability one.

This exhibits a strong connection between the parameter α of the process and the random distribution function F. In particular, a kth moment of α exists if and only if a kth moment of F exists with probability one. The main result, that shows the Dirichlet process satisfies the second desirable property, is the following.

THEOREM 1. *If $F \in \mathscr{D}(\alpha)$ and if $X_1, \cdots, X_n$ is a sample from F, then the posterior distribution of F given $X_1, \cdots, X_n$ is $\mathscr{D}(\alpha + \sum_1^n \delta_{X_i})$, where δ_x is the measure giving mass one to x.*

In distribution function form, the posterior parameter of the process is

$$\alpha(t) + \sum_1^n I_{[X_i,\infty)}(t) .$$

EXAMPLES. Consider the problem of estimating an unknown distribution function F by a distribution function $\hat{F}$ with loss function $L(F, \hat{F}) = \int (F(t) - \hat{F}(t))^2 \, dW(t)$, where $W(t)$ is a given non-random weight function (finite measure). Suppose as a prior distribution, we take $F \in \mathscr{D}(\alpha)$. If we have no observations from F, then since $F(t) \in \mathscr{B}e(\alpha(t), \alpha(\mathbb{R}) - \alpha(t))$, the Bayes estimate is

$$\hat{F}(t) = \mathscr{E}F(t) = \alpha(t)/\alpha(\mathbb{R}) =_{\text{def}} F_0(t) . \tag{1.1}$$

We may consider $F_0(t)$ as our prior guess at $F(t)$. The Bayes estimate based on a sample $X_1, \cdots, X_n$ from F is therefore

$$\begin{aligned} \hat{F}_n(t) = \mathscr{E}(F(t) \mid X_1, \cdots, X_n) &= (\alpha(t) + \textstyle\sum_1^n I_{[X_i,\infty)}(t))/(\alpha(\mathbb{R}) + n) \\ &= p_n F_0(t) + (1 - p_n) F_n(t) \end{aligned} \tag{1.2}$$

where F_n is the sample distribution function,

$$F_n(t) = \frac{1}{n} \sum_1^n I_{[X_i,\infty)}(t) \tag{1.3}$$

and

$$p_n = \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + n) . \tag{1.4}$$

The Bayes estimate $\hat{F}_n$ is thus a mixture of the prior guess F_0 and the sample distribution function F_n. If $\alpha(\mathbb{R})$ is large compared to n, $\hat{F}_n$ gives most of its weight to F_0, while if $\alpha(\mathbb{R})$ is small compared to n, $\hat{F}_n$ gives most of its weight to F_n. Thus, one may consider $\alpha(\mathbb{R})$ as a measure of the strength of belief in the prior guess, measured in units of sample size. The parameter of a Dirichlet prior is specified by the function F_0, and the real parameter $\alpha(\mathbb{R})$.

Similarly, one may estimate the mean of an unknown distribution F with loss function $L(F, \hat{\mu}) = (\int t \, dF(t) - \hat{\mu})^2$. For the prior distribution $F \in \mathscr{D}(\alpha)$ with $\int t^2 \, d\alpha(t) < \infty$, the Bayes estimate based on a sample $X_1, \cdots, X_n$ from F is

$$\begin{aligned} \hat{\mu}_n &= \mathscr{E}(\int t \, dF(t) \mid X_1, \cdots, X_n) = \int t \, d\mathscr{E}(F(t) \mid X_1, \cdots, X_n) \\ &= \int t \, d\hat{F}_n(t) = p_n \mu_0 + (1 - p_n) \bar{X}_n \end{aligned}$$

where $\bar{X}_n$ is the sample mean and μ_0 is the prior guess at the mean

$$\mu_0 = \int t \, dF_0(t) .$$

It is interesting that the Bayes estimate of the mean depends on the parameter of the Dirichlet prior only through the values of μ_0 and $\alpha(\mathbb{R})$.

Similar results are obtained by these methods for the following problems: (1) Estimating moments, or a variance or covariance; (2) Estimating a median or other quantiles; (3) Estimating $P(X > Y)$ in a two-sample problem; (4) Estimating a quantile by a "tolerance" region; (5) Testing one-sided hypotheses concerning quantiles. Brunk and Pierce [5] have applied these methods to the estimation of a cumulative regression. G. J. Hall, Jr. [20] has discussed the

Dirichlet prior for use in an adaptive sequential search problem. Two other applications are presented in the last section of this paper.

These methods may be extended, though with less success, usually because of the relative difficulty of computing the Bayes decision rules, to such problems as bio-assay, regression or observation with error, and empirical Bayes. In these more complex problems, even if the prior is a Dirichlet process, the posterior often turns out to be a mixture of Dirichlet processes, wherein the parameter of the process α_u is indexed by a variable u taken to be random. This subject is treated in detail by Antoniak [1].

The bio-assay problem has also been treated by Ramsey [26] and by Kraft and van Eeden [23], the latter using a tailfree process (discussed in the next section) for a prior.

For certain other problems, particularly hypothesis testing problems, these methods turn out to be unsuitable. For example, consider the goodness-of-fit problem of testing the hypothesis that a distribution on [0, 1] is uniform. If for the alternative hypothesis we take $\mathscr{D}(\alpha)$ for a prior where α is uniform, $\alpha(t) = ct$ on [0, 1] for some $c > 0$, then the only non-trivial non-randomized Bayes rule is to reject the null hypothesis if and only if some two observations are exactly equal. This is really a test of continuity against discreteness.

2. Tailfree processes. The limitations of the Dirichlet process stem mainly from the fact that it chooses discrete distributions with probability one, so that we expect to see some observations repeated exactly. To avoid these limitations, we should try to find workable priors that choose continuous distributions with probability one. There are some among the tailfree processes of Freedman [19] and Fabius [14], which we now describe.

Let $\{\pi_m; m = 1, 2, \cdots\}$ be a tree of measurable partitions of $(\mathbb{R}, \mathscr{B})$; that is, let $\pi_1, \pi_2, \cdots$ be a sequence of measurable partitions such that π_{m+1} is a refinement of π_m for each $m = 1, 2, \cdots$, and $\bigcup_1^\infty \pi_m$ generates $\mathscr{B}$.

DEFINITION 3. The distribution of a random probability P on $(\mathbb{R}, \mathscr{B})$ is said to be tailfree with respect to $\{\pi_m\}$ if there exists a family of nonnegative random variables $\{V_{m,B}; m = 1, 2, \cdots, B \in \pi_m\}$ such that

(1) the families $\{V_{1,B}; B \in \pi_1\}, \{V_{2,B}; B \in \pi_2\}, \cdots$ are independent, and

(2) for every $m = 1, 2, \cdots$, if $B_j \in \pi_j$, $j = 1, \cdots, m$ is such that $B_1 \supset B_2 \supset \cdots \supset B_m$, then $P(B_m) = \prod_{j=1}^m V_{j,B_j}$.

Corresponding to Theorem 1 for Dirichlet processes we have the following for tailfree processes.

THEOREM 2. *If the distribution of P is tailfree with respect to $\{\pi_m\}$ and if $X_1, \cdots, X_n$ is a sample from P, then the posterior distribution of P given $X_1, \cdots, X_n$ is tailfree with respect to $\{\pi_m\}$.*

In addition, the posterior distributions of the V's of Definition 3 given $X_1, \cdots, X_n$

may easily be computed. A simple example illustrates the main ideas. We construct tailfree processes on the interval (0, 1] with respect to $\{\pi_m\}$ where π_m is the set of all dyadic intervals of length $1/2^m$, $\pi_m = \{((i-1)/2^m, i/2^m]; i = 1, \cdots, 2^m\}$.

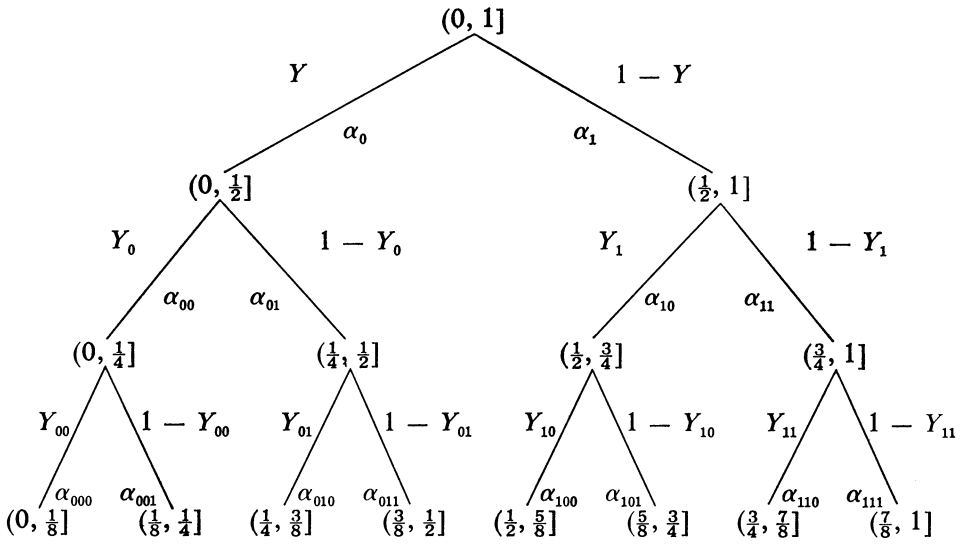

FIG. 1.

A simpler notation for the variables $V_{m,B}$ of Definition 3 is more appropriate for the tree of dyadic intervals. Let $.\varepsilon_1\varepsilon_2 \cdots \varepsilon_m$ denote the binary expansion of the dyadic rational $\sum_1^m \varepsilon_j 2^{-j}$, where each ε_j is zero or one. If $B \in \pi_m$ is of the form $(.\varepsilon_1 \cdots \varepsilon_m, .\varepsilon_1 \cdots \varepsilon_m + 2^{-m}]$, then for $\varepsilon_m = 0$ we use $Y_{\varepsilon_1\cdots\varepsilon_m}$ to denote $V_{m,B}$, while for $\varepsilon_m = 1$ we use $1 - Y_{\varepsilon_1\cdots\varepsilon_m}$ to denote $V_{m,B}$. This is possible since P is assumed to be a random probability. Then $P(B)$ is the product of all the variables associated with the path in the tree from (0, 1] to B, so that

$$P(B) = (\prod_{j=1;\varepsilon_j=0}^m Y_{\varepsilon_1\cdots\varepsilon_{j-1}})(\prod_{j=1;\varepsilon_j=1}^m (1 - Y_{\varepsilon_1\cdots\varepsilon_{j-1}})) . \tag{2.1}$$

For example $P((\frac{3}{8}, \frac{1}{2}]) = Y(1 - Y_0)(1 - Y_{01})$. The independence hypothesis of Definition 3 requires that the Y variables be independent between rows in Figure 1.

If we choose *all* the Y variables independently, with $Y_{\varepsilon_1\cdots\varepsilon_{m-1}} \in \mathscr{B}e(\alpha_{\varepsilon_1\cdots\varepsilon_{m-1}0}, \alpha_{\varepsilon_1\cdots\varepsilon_{m-1}1})$, then the posterior distributions of the Y variables given a sample $X_1, \cdots, X_n$ from P again has the same structure—independent with beta distributions. The posterior distribution of $Y_{\varepsilon_1\cdots\varepsilon_{m-1}}$ given $X_1, \cdots, X_n$ is $\mathscr{B}e(\alpha_{\varepsilon_1\cdots\varepsilon_{m-1}0} + M, \alpha_{\varepsilon_1\cdots\varepsilon_{m-1}1} + N)$, where M is the number of X_i's that fall in $(.\varepsilon_1 \cdots \varepsilon_{m-1}0, .\varepsilon_1 \cdots \varepsilon_{m-1}1]$ and N is the number of X_i's that fall in $(.\varepsilon_1 \cdots \varepsilon_{m-1}1, .\varepsilon_1 \cdots \varepsilon_{m-1}1 + 2^{-m}]$.

The Dirichlet process is tailfree with respect to every tree of partitions. The above scheme is a Dirichlet process if the parameters add in the following way. For every $\varepsilon_1 \cdots \varepsilon_m$,

$$\alpha_{\varepsilon_1\cdots\varepsilon_m} = \alpha_{\varepsilon_1\cdots\varepsilon_m0} + \alpha_{\varepsilon_1\cdots\varepsilon_m1} . \tag{2.2}$$

Thus, the tailfree processes are much more flexible than the Dirichlet. There are twice as many parameters at your disposal for each row in Figure 1. With this extra freedom, we can choose the parameters (α's) so that the random probability P is continuous singular or absolutely continuous with probability one. It is worthwhile to investigate three cases.

(a) $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 2^{-m}$. This yields a Dirichlet process and P is discrete with probability one. Conditions on the α's that lead to discrete P with probability one have been given by Blackwell [3].

(b) $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = 1$. This yields a random probability P of a type considered by Dubins and Freedman [13] and shown to be continuous singular with probability one.

(c) $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = m^2$. This yields a P absolutely continuous with probability one. Conditions on the α's that lead to absolutely continuous P with probability one may be obtained from the work of Kraft [22] and Metivier [24].

The drawback of (a) as a prior in the goodness of fit problem of Section 1 was due to the discreteness of P. Use of priors (b) and (c) partially overcomes this limitation, but unfortunately new drawbacks are introduced. A minor one is that $\mathscr{E}F(t)$ is now more difficult to compute. If t is not dyadic rational, $\mathscr{E}F(t)$ is an infinite series.

The main drawback is that the dyadic points of subdivision play a strong role in the posterior distributions. For each of (a), (b), and (c), the prior guess at F is, from symmetry, the uniform distribution, $\mathscr{E}F(t) = t$, $t \in (0, 1]$. The posterior expectation given a sample of size 1, $\mathscr{E}(F(t) \mid X = x)$ is still uniform in the dyadic intervals in which x does not lie, but it has corners at the dyadic rationals near x in cases (b) and (c). At x, $\mathscr{E}(F(t) \mid X = x)$ has a discontinuity in case (a), infinite slope in case (b), and bounded slope in case (c). We further note in case (c) that even though the density with respect to Lebesgue measure exists with probability one, the density has discontinuities at all the dyadic rationals with probability one.

It should be considered a liability that the points used to describe the process appear strongly in the almost sure properties or in the posterior expectations.

3. Characterization of the Dirichlet process. Are there tailfree processes other than the Dirichlet for which the points of subdivision chosen for the tree of partitions do not play an essential role in the behavior of the process? Except for three trivial types of processes, the answer is no. This and two other characterizations of the Dirichlet process given here are due to Doksum [11] and Fabius [15]. The three trivial types of processes seem to appear in all these characterizations. They are

T_1. P non-random ($F \equiv F_0$).

T_2. P degenerate at a random point ($F = I_{[X,\infty)}$ where X has distribution F_0).

T_3. P concentrated on two non-random points ($F = UI_{[a,\infty)} + (1 - U)I_{[b,\infty)}$ where U has an arbitrary distribution on [0, 1], and $a < b$).

Types T_1 and T_2 are limits of the Dirichlet process as $\alpha(\mathbb{R}) \to \infty$ and $\alpha(\mathbb{R}) \to 0$, respectively, with F_0 fixed.

Characterization 1. If P is tailfree with respect to every tree of partitions, then P is either a Dirichlet process or of types T_1, T_2 or T_3.

From Definition 1 it follows that the Dirichlet process is the only random probability measure for which the distribution of $(P(B_1), \cdots, P(B_m))$ is Dirichlet for every measurable partition $\{B_1, \cdots, B_m\}$. Therefore, characterizations of the (finite-dimensional) Dirichlet distributions lead to characterizations of the Dirichlet process. The characterization of the Dirichlet distribution given by Darroch and Ratcliff [8] may be so used. If the Darroch–Ratcliff characterization of the Dirichlet distribution is strengthened as in Fabius [16], the resulting characterization of the Dirichlet process contains Characterization 1.

A related characterization of the Dirichlet distribution based on a different independence condition, called neutrality, leads to a second characterization of the Dirichlet process. The concept of neutrality is due to Connor and Mosimann [7]. The independence condition used below is essentially in the form given by Fabius [15], and weakened slightly in Fabius [16].

Characterization 2. If P is neutral with respect to every finite measurable partition (that is, if for every measurable partition $\{B_1, \cdots, B_m\}$, $P(B_1)$ and the vectors $(P(B_2)/(1 - P(B_1)), \cdots, P(B_m)/(1 - P(B_1)))$ are conditionally independent given $P(B_1) \neq 1$), then P is either a Dirichlet process or of types T_1, T_2 or T_3.

Although the above characterizations may be considered as attractive properties of the Dirichlet process, the next characterization may be considered as a drawback.

Characterization 3. If for every measurable set B, the posterior distribution of $P(B)$ given a sample $X_1, \cdots, X_n$ from P, depends on $X_1, \cdots, X_n$ only through the number of observations that fall in B, then P is either a Dirichlet process or of types T_1, T_2 or T_3.

One would like to have a prior distribution for P with the property that if X is a sample from P and $X = x$, then the posterior guess at P gives more weight to values close to x than the prior guess at P does. For the Dirichlet process prior, the posterior guess at P gives more weight to the point x itself, but it treats all other points equally. In particular, the posterior guess at P actually gives less weight to points near x but not equal to x. From this point of view, the tailfree process prior that chooses absolutely continuous distributions with probability one would seem to be more appropriate.

4. Processes neutral to the right. A large class of prior distributions on $\mathscr{F}$ that are not dependent on the partition points used to describe the process has recently been discovered by Doksum [11].

DEFINITION 4. A random distribution function $F(t)$ on the real line is said to

be neutral to the right if for every m and $t_1 < t_2 < \cdots < t_m$, there exist independent random variables $V_1, V_2, \cdots, V_m$, such that $(1 - F(t_1), 1 - F(t_2), \cdots, 1 - F(t_m))$ has the same distribution as $(V_1, V_1V_2, \cdots, \prod_1^m V_i)$.

Essentially, this says that F is neutral to the right if $1 - F(t_1), (1 - F(t_2))/(1 - F(t_1)), \cdots, (1 - F(t_m))/(1 - F(t_{m-1}))$ are independent when $t_1 < t_2 < \cdots < t_m$. Because of the possibility of the denominators being zero with positive probability, we prefer to use Definition 4 instead. Note that if F is neutral to the right then $Y_t = -\log(1 - F(t))$ has independent increments. Doksum gives a simple characterization of processes neutral to the right.

Let Y_t be a process with independent increments non-decreasing a.s., right continuous a.s., $\lim_{t\to-\infty} Y_t = 0$ a.s. and $\lim_{t\to+\infty} Y_t \doteq +\infty$ a.s. We allow the possibility that $Y_t = \infty$ with positive probability for finite t. Then $F(t) = 1 - e^{-Y_t}$ is a random distribution function neutral to the right.

For such a process Y_t there exist at most countably many fixed points of discontinuity $t_1, t_2, \cdots$. The corresponding jumps $S_1, S_2, \cdots$, are independent non-negative (possible infinite-valued) random variables with corresponding densities $f_{t_1}, f_{t_2}, \cdots$ (with respect to some convenient measure). The difference

$$Z_t = Y_t - \textstyle\sum_j S_j I_{[t_j,\infty)}(t) \tag{4.1}$$

is a non-decreasing process with independent increments without fixed points of discontinuity and therefore has Lévy formula

$$\log Ee^{uZ_t} = ub(t) + \textstyle\int_0^\infty (e^{uz} - 1)\, dN_t(z) \tag{4.2}$$

where b is a non-decreasing continuous function with $\lim_{t\to-\infty} b(t) = 0$, and where N_t is a continuous Lévy measure, that is, for every $B \in \mathscr{B}$, $N_t(B)$ is non-decreasing and continuous in t, and for each t, N_t is a measure on the Borel sets of $(0, \infty)$. We define Z_t to be $+\infty$ unless

$$\textstyle\int_0^\infty z/(1 + z)\, dN_t(z) < \infty\,. \tag{4.3}$$

The process neutral to the right is specified by the four quantities $\{t_1, t_2, \cdots\}$, $\{f_{t_1}, f_{t_2}, \cdots\}$, b, and N_t.

The function b corresponds to the non-random part of the process Y_t. If $b \equiv 0$, then Y_t, and hence $F(t)$, increase only in jumps a.s., so that F is discrete with probability one.

Corresponding to Theorems 1 and 2, the main result for processes neutral to the right is

THEOREM 3. *If F is neutral to the right, and if $X_1, \cdots, X_n$ is a sample from F, then the posterior distribution of F given $X_1, \cdots, X_n$ is neutral to the right.*

The paper of Doksum also contains a description of the posterior distributions of F in terms of the prior distribution of F. We give below an alternative description of Doksum's result in terms of the distribution of the process Y_t. It is sufficient to give this description for $n = 1$ since the description for arbitrary sample size would follow by repeated application.

Let the prior distribution of Y_t be specified by $\{t_1, t_2, \cdots\}$, $\{f_{t_1}, f_{t_2}, \cdots\}$, b, N_t, and let X be a sample from $F(t) = 1 - e^{-Y_t}$. The posterior distribution of Y_t given $X = x$ is best treated in two cases.

Case 1. x is one of the prior fixed points of discontinuity, say $x = t_k$. The posterior distribution of Y_t given $X = x = t_k$ is specified by

(1) the same set of fixed points of discontinuity and the same deterministic component function, b,

(2) the posterior Lévy measure

$$\begin{aligned} dN_t(z \mid x) &= e^{-z}\, dN_t(z) && \text{for } t < x \\ &= e^{-z}\, dN_x(z) + d[N_t(z) - N_x(z)] && \text{for } t > x, \end{aligned}$$

(3) for $i \neq k$, the posterior distribution of the jump at t_i,

$$\begin{aligned} f_{t_i}(s \mid x) &= ce^{-s} f_{t_i}(s) && \text{for } t_i < x \\ &= f_{t_i}(s) && \text{for } t_i > x, \end{aligned}$$

(4) the posterior distribution of the jump at $x = t_k$

$$f_{t_k}(s \mid t_k) = c(1 - e^{-s}) f_{t_k}(s)$$

(where c represents a normalizing constant).

Case 2. x is not one of the prior points of discontinuity. The posterior distribution of Y_t given $X = x$ is the same as in Case 1 (1), (2), (3) except that x may now be a fixed point of discontinuity, and (4) is replaced by

(4′) define the measure μ_B on $(\mathbb{R}, \mathscr{B})$ for each fixed Borel subset B of $[0, \infty)$ to satisfy

$$\mu_B((-\infty, t]) = \int_B (1 - e^{-z})\, dN_t(z) + b(t) I_B(0)$$

where as usual I_B denotes the indicator function. Note that $\mu_B \ll \mu_{[0,\infty)}$. The posterior distribution of the jump S at x is given by $P(S \in B \mid x) = \varphi_B(x)$, where φ_B is the Radon–Nikodym derivative of μ_B with respect to $\mu_{[0,\infty)}$.

One may state conditions (1), (2) and (3) more simply as the condition that the distribution of the increments $Y_t - Y_{t-\varepsilon}$ for $t < x$ and $\varepsilon > 0$ are changed by multiplying the density by e^{-z} and renormalizing, while the distribution of the increments $Y_{t+\varepsilon} - Y_t$ for $t > x$ and $\varepsilon > 0$ remain unchanged.

It is best to take an example of the use of (4′). Suppose there are no fixed points of discontinuity, and suppose $dN_t(z) = \gamma(t) N(z)$ where $N(z)$ is a fixed measure such that $\int_0^\infty z/(1 + z)\, dN(z) < \infty$, and where $\gamma(t)$ is non-decreasing and continuous with $\gamma(t) \to 0$ as $t \to -\infty$ and $\gamma(t) \to \infty$ as $t \to \infty$. Suppose further that $b(t)$ and $\gamma(t)$ are absolutely continuous. Then

$$\mu_B((-\infty, t]) = \gamma(t) \int_B (1 - e^{-z})\, dN(z) + b(t) I_B(0), \tag{4.4}$$

so that

$$\varphi_B(t) = (1 - q(t)) \frac{\int_B (1 - e^{-z})\, dN(z)}{\int_0^\infty (1 - e^{-z})\, dN(z)} + q(t) I_B(0) \tag{4.5}$$

where

$$q(t) = b'(t)/(\gamma'(t) \int_0^\infty (1 - e^{-z})\, dN(z) + b'(t)) \tag{4.6}$$

represents the probability that the jump at the new fixed point of discontinuity is zero. If $b \equiv 0$, then $q \equiv 0$ and the distribution of the jump is independent of where it occurs. Note that the distribution of the new jump is not necessarily infinitely divisible.

It is interesting to view the Dirichlet process as a process neutral to the right. If $F \in \mathscr{D}(\alpha)$, then F is neutral to the right, and if α is continuous then $Y_t = -\log(1 - F(t))$ has no fixed points of discontinuity. This implies that if $X \in \mathscr{B}e(\alpha, \beta)$ then $Y = -\log(1 - X)$ is infinitely divisible. The density of Y is

$$f_Y(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} e^{-\beta y}(1 - e^{-y})^{\alpha - 1} I_{(0,\infty)}(y) \tag{4.7}$$

and the moment generating function of Y is

$$M_Y(u) = Ee^{uY} = \frac{\Gamma(\alpha + \beta)\Gamma(\beta - u)}{\Gamma(\beta)\Gamma(\alpha + \beta - u)} \qquad \text{for } u < \beta, \tag{4.8}$$

an unlikely looking function to be the moment generating function of an infinitely divisible distribution. The follow lemma gives the Lévy representation of this moment generating function.

LEMMA 1. *If* $Y = -\log(1 - X)$ *where* $X \in \mathscr{B}e(\alpha, \beta)$, *then*

$$\log M_Y(u) = \int_0^\infty (e^{uz} - 1) \frac{e^{-\beta z}(1 - e^{-\alpha z})}{(1 - e^{-z})z}\, dz .$$

PROOF. Using the formula $\Gamma(x) = x^{-1}\Gamma(x + 1)$, one may write (4.8) as

$$M_Y(u) = \left(\prod_{k=0}^{n-1} \frac{(\beta + k)(\alpha + \beta - u + k)}{(\alpha + \beta + k)(\beta - u + k)}\right) \cdot \frac{\Gamma(\alpha + \beta + n)\Gamma(\beta - u + n)}{\Gamma(\beta + n)\Gamma(\alpha + \beta - u + n)} . \tag{4.9}$$

Stirling's formula, $\Gamma(x) \sim (2\pi x)^{\frac{1}{2}}(x/e)^x$ for large x, implies that the term involving the gamma function on the right side of (4.9) tends to 1 as $n \to \infty$. Hence,

$$\log M_Y(u) = \sum_{k=0}^\infty \log \frac{(\beta + k)(\alpha + \beta - u + k)}{(\alpha + \beta + k)(\beta - u + k)} \qquad \text{for } u < \beta .$$

Consequently, the Lévy representation

$$\log \frac{\lambda}{\lambda - u} = \int_0^\infty (e^{uz} - 1) \frac{e^{-\lambda z}}{z}\, dz \qquad \text{for } u < \lambda$$

for the moment generating function of the negative exponential distribution with parameter λ implies

$$\log M_Y(u) = \sum_{k=0}^\infty \int_0^\infty (e^{uz} - 1) \frac{e^{-\beta z}(1 - e^{-\alpha z})}{z} e^{-kz}\, dz \qquad \text{for } u < \beta$$

from which the lemma follows immediately.

Returning to the Dirichlet process as a process neutral to the right, let $F \in \mathscr{D}(\alpha)$ and $Y_t = -\log(1 - F(t))$. Then $F(t) \in \mathscr{B}e(\alpha(t), \alpha(\mathbb{R}) - \alpha(t))$ so that the Lévy measure for Y_t is

$$dN_t(z) = \frac{e^{-(\alpha(\mathbb{R})-\alpha(t))z}(1 - e^{-\alpha(t)z})}{(1 - e^{-z})z}\,dz$$
$$= e^{-\alpha(\mathbb{R})z}(e^{\alpha(t)z} - 1)(1 - e^{-z})^{-1}z^{-1}\,dz$$

provided $0 < \alpha(t) < \alpha(\mathbb{R})$. Therefore,

$$\mu_B((-\infty, t]) = \int_B e^{-\alpha(\mathbb{R})z}(e^{\alpha(t)z} - 1)z^{-1}\,dz$$

so that

$$\varphi_B(t) = \alpha(\mathbb{R}) \int_B e^{-\alpha(\mathbb{R})z}\,dz\,.$$

Thus, the distribution of the jump in Y_t at $t = x$, the new fixed point of discontinuity, is independent of where it occurs and is negative exponential with density $\alpha(\mathbb{R})e^{-\alpha(\mathbb{R})s}I_{(0,\infty)}(s)$.

If in further samples from F, k more observations fall at x, r more observations fall above x, and the rest of the observations fall below x, the posterior distribution of the jump in Y_t at x has density $ce^{-(\alpha(\mathbb{R})+r)s}(1 - e^{-s})^k I_{(0,\infty)}(s)$, a density of the form (4.7).

It is useful to note that when F is neutral to the right, $\mathscr{E}F(t)$ can be computed directly from the moment generating functions of the variables involved in Y_t.

$$\mathscr{E}F(t) = 1 - \mathscr{E}\exp\{-Y_t\} = 1 - \mathscr{E}\exp\{-Z_t - \textstyle\sum_{t_j \le t} S_j\}$$
$$= 1 - M_{Z_t}(-1)\textstyle\prod_{t_j \le t} M_{S_j}(-1)\,.$$

Furthermore, the moment generating functions for the posterior distributions of the increments $Y_t - Y_{t-\varepsilon}$ for $t < x$ and $\varepsilon > 0$ and $Y_{t+\varepsilon} - Y_t$ for $t > x$ and $\varepsilon > 0$ may easily be found from the corresponding moment generating functions of the prior. However, to compute the distribution of the jumps at new fixed points of discontinuity it is useful to have knowledge of the Lévy form of the moment generating function of Z_t.

5. Applications. 1. *Adaptive sampling with recall.* Let $F \in \mathscr{D}(\alpha)$ where $\int t^2\,d\alpha(t) < \infty$, and let $X_1, X_2, \cdots$ be independent identically distributed observations from F. Consider the problem of finding a stopping rule N to maximize

$$\mathscr{E}(\max_{0 \le j \le N} X_j - Nc)$$

where $c > 0$ and where $X_0 = 0$.

The interpretation of this problem is as follows. You have an object to sell. Bids for the object come in one at a time chosen independently from some distribution F you do not know exactly. A Dirichlet process with parameter α expresses your prior knowledge of F. It costs you an amount c to wait from one bid to the next. You may stop viewing bids at any time and either sell the object for the maximum of the bids you have seen so far, or throw the object away and receive zero. Looking at a bid is costly, but it always gives information about the true value of F and it may be a large bid. When do you stop looking?

The condition $\int t^2\, d\alpha(t) < \infty$ implies that $\mathscr{E}X_i^2 < \infty$, and Theorem 1, page 352 of De Groot [9] implies that there is an optimal rule. Since the problem is monotone (see Chow, Robbins and Siegmund [6]), and one can show that the expected return can be approximated by the expected returns of truncated problems, this rule is the one-stage look-ahead rule. It is optimal to stop at the first n for which your present return is greater than or equal to your conditional expected return if you continue one more stage and stop. That is, stop at the first n for which

$$\max_{0\leqq j\leqq n} X_j - nc \geqq \mathscr{E}(\max_{0\leqq j\leqq n+1} X_j - (n+1)c \mid X_1, \cdots, X_n)\,.$$

Let M_n denote $\max_{0\leqq j\leqq n} X_j$, and rewrite this condition as: Stop at the first n for which

$$\begin{aligned} c &\geqq \mathscr{E}(\max(0, X_{n+1} - M_n) \mid X_1, \cdots, X_n) \\ &= \int \max(0, x - M_n)\, d\hat{F}_n(x) \\ &= p_n \int \max(0, x - M_n)\, dF_0(x) + (1 - p_n) \int \max(0, x - M_n)\, dF_n(x) \\ &= p_n \int \max(0, x - M_n)\, dF_0(x) \end{aligned}$$

where $\hat{F}_n$, F_0, F_n and p_n are as in Section 1. This application was pointed out to me by J. B. MacQueen. A similar application may be made to the exponential version of this model in which $Y_n = \beta^n \max_{i\leqq n} X_i$. The optimal rule is the one-stage look-ahead rule. It stops at the first n for which $(1-\beta)M_n \geqq \beta p_n \int \max(0, x - M_n)\, dF_0(x)$.

2. *An adaptive investment model.* Consider an investor with initial resources X_0 who makes investments during discrete time periods in m different investment opportunities. Denote by $\mathbf{b}_i$ the m-vector whose jth component b_{ij} is the amount invested during time period i in investment opportunity j. If X_{i-1} denotes the investor's fortune at the beginning of the ith period, we require of $\mathbf{b}_i$ that

$$\sum_{j=1}^m b_{ij} \leqq X_{i-1} \quad \text{and} \quad b_{ij} \geqq 0 \tag{5.1}$$

for all i. Denote by $\mathbf{Y}_i$ the random m-vector of returns, whose jth component $Y_{ij} \geqq 0$ is the return per unit invested during time period i in investment opportunity j. Given $\mathbf{Y}_i$ we may compute X_i from X_{i-1} and $\mathbf{b}_i$ by the formula

$$X_i = (X_{i-1} - \sum_{j=1}^m b_{ij}) + \sum_{j=1}^m b_{ij} Y_{ij}\,. \tag{5.2}$$

It is the objective of the investor to maximize the expected value of the utility of his fortune at the end of n time periods. For logarithmic utility, adaptive problems can be handled. Therefore we desire to maximize

$$\mathscr{E} \log X_n\,. \tag{5.3}$$

We assume that $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ are independent identically distributed from some unknown distribution function F. As a prior distribution for F, we take the Dirichlet process whose parameter α is a finite non-null measure on m-dimensional Euclidean space, $F \in \mathscr{D}(\alpha)$.

In addition to (5.1) there are further constraints on the choice of the $\mathbf{b}_i$ that reflect the requirement that the rule be non-anticipatory; that is, $\mathbf{b}_i$ may be a function only of $X_0, \mathbf{b}_1, \mathbf{Y}_1, \mathbf{b}_2, \mathbf{Y}_2, \cdots, \mathbf{b}_{i-1}, \mathbf{Y}_{i-1}$, and not of $\mathbf{Y}_i, \cdots, \mathbf{Y}_n$. The problem then is to find a sequence $\mathbf{b}_1, \cdots, \mathbf{b}_n$ subject to these constraints to maximize $\mathscr{E} \log X_n$.

This problem when F is known and multinomial ($\mathbf{Y}_j$ takes only the values $(1, 0, \cdots, 0), (0, 1, \cdots, 0), \cdots, (0, 0, \cdots, 1)$) was introduced by Kelly [21]. That Kelly's model could be extended to the adaptive case was noticed by Bellman and Kalaba [2]. Further results along these lines were obtained by Murphy [25].

The optimal rule for these problems is the rule that, at each stage, maximizes the expected log of the resources one stage ahead. The proof as in Kelly is by backward induction. At the beginning of the nth stage one is to choose $\mathbf{b}_n$ to maximize $\mathscr{E}(\log X_n \mid \mathbf{Y}_1, \cdots, \mathbf{Y}_{n-1})$. Defining $\mathbf{c}_n$ by $\mathbf{b}_n = \mathbf{c}_n X_{n-1}$, we may compute

$$\begin{aligned} &\mathscr{E}(\log X_n \mid \mathbf{Y}_1, \cdots, \mathbf{Y}_{n-1}) \\ &\quad = \log X_{n-1} + \mathscr{E}(\log (1 - \textstyle\sum_{j=1}^m c_{nj} + \sum_{j=1}^m c_{nj} Y_{nj}) \mid \mathbf{Y}_1, \cdots, \mathbf{Y}_{n-1}) \\ &\quad = \log X_{n-1} + \int \log (1 - \textstyle\sum_{j=1}^m c_{nj}(y_j - 1))\, d\hat{F}_{n-1}(\mathbf{y}) \end{aligned} \tag{5.4}$$

where $\hat{F}_{n-1}$ is $\mathscr{E}(F \mid \mathbf{Y}_1, \cdots, \mathbf{Y}_{n-1})$ as in Section 1. The optimal rule chooses $\mathbf{c}_n$ to minimize this integral subject to the constraints $\sum_1^m c_{nj} \leqq 1$ and $c_{nj} \geqq 0$. Therefore, at the beginning of stage $n - 1$, the investor wants to choose $\mathbf{b}_{n-1}$ to minimize the expectation of (5.4) given $\mathbf{Y}_1, \cdots, \mathbf{Y}_{n-2}$. Since the last term of (5.4) does not depend on $\mathbf{b}_{n-1}$, this is equivalent to minimizing $\mathscr{E}(\log X_{n-1} \mid \mathbf{Y}_1, \cdots, \mathbf{Y}_{n-2})$. This procedure obviously continues down to the first stage. The optimal rule is therefore: Choose $\mathbf{c}_i$, $i = 1, \cdots, n$ to maximize

$$\int \log (1 + \textstyle\sum_{j=1}^m c_{ij}(y_j - 1))\, d\hat{F}_{i-1}(\mathbf{y})$$

subject to the constraints $\sum_{j=1}^m c_{ij} \leqq 1$ and $c_{ij} \geqq 0$, and invest $\mathbf{b}_i = \mathbf{c}_i X_{i-1}$. This is a convex programming problem for each $i = 1, \cdots, n$.

REFERENCES

[1] ANTONIAK, C. (1969). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. Ph. D. dissertation, Univ. of California, Los Angeles.

[2] BELLMAN, R. and KALABA, R. (1958). On communication processes involving learning and random duration. *IRE Nat. Convention Record* Part IV 16-21.

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

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

[5] BRUNK, H. D. and PIERCE, D. A. (1972). Note on Bayesian approaches to estimation of a cumulative regression. Technical Report, Oklahoma State Univ.

[6] CHOW, Y. S., ROBBINS, H. and SIEGMUND, D. (1971). *Great Expectations: The Theory of Optimal Stopping.* Houghton-Mifflin, Boston.

[7] CONNOR, R. J. and MOSIMANN, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **64** 194-206.

[8] DARROCH, J. N. and RATCLIFF, D. (1971). A characterization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **66** 641-643.

[9] DEGROOT M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

[10] DOKSUM, K. A. (1972). Decision theory for some nonparametric models. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **1** 331–343.

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

[12] DOOB, J. L. (1949). Application of the theory of martingales. *Colloq. Internat. du CNRS* 23–27.

[13] DUBINS, L. E. and FREEDMAN, D. A. (1966). Random distribution functions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **2** 183–214.

[14] FABIUS, J. (1964). Asymptotic behavior of Bayes estimates. *Ann. Math. Statist.* **35** 846–856.

[15] FABIUS, J. (1972). Neutrality and Dirichlet distributions. *Trans. Sixth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes.* To appear.

[16] FABIUS, J. (1973). Two characterizations of the Dirichlet distribution. *Ann. Statist.* **1** 583–587.

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

[18] FERGUSON, T. S. and KLASS, M. J. (1972). A representation of independent processes without Gaussian components. *Ann. Math. Statist.* **43** 1634–1643.

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

[20] HALL, G. J., Jr. (1973). Sequential search with random overlook probabilities. Ph. D. thesis in mathematics, Univ. of California, Los Angeles.

[21] KELLY, J. L. Jr. (1956). A new interpretation of information rate. *Bell System Tech. J.* **35** 917–926.

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

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

[24] METIVIER, M. (1971). Sur la construction de mesures aléatoires presque sûrement absolument continues par rapport à une mesure donnée. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **20** 332–344.

[25] MURPHY, R. E. Jr. (1965). *Adaptive Processes in Economic Systems.* Academic Press, New York.

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

[27] WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA 90024