Convergence of Dirichlet Measures and
the Interpretation of Their Parameters

by

Jayaram Sethuraman and Ram C. Tiwari

The Florida State University
1. Introduction. The form of the Bayes estimate of the population mean with respect to a Dirichlet prior with parameter $\alpha$ has given rise to the interpretation that $\alpha(X)$ is the prior sample size. Furthermore, if $\alpha(X)$ is made to tend to zero, then the Bayes estimate mathematically converges to the classical estimator, namely the sample mean. This has further given rise to the general feeling that allowing $\alpha(X)$ to become small not only makes the 'prior sample size' small but also that it corresponds to no prior information. By investigating the limits of prior distributions as the parameter $\alpha$ tends to various values, we show that it is misleading to think of $\alpha(X)$ as the prior sample size and the smallness of $\alpha(X)$ as no prior information. In fact very small values of $\alpha(X)$ actually mean that we have very definite information concerning the unknown true distribution.

*Research supported by the U.S. Army Research Office under Grant No. DAAG 29-79-C-0158. The United States Government is authorized to reproduce and distribute reprints for governmental purposes.

†Now at Allahabad University, India.

AMS Classification Numbers: 60B10, 60C10.

KEY WORDS: Non-parametric Bayesian analysis, Dirichlet measures.
2. The Dirichlet measure. Let \((X, A)\) be separable metric space endowed with the corresponding Borel \(\sigma\)-field. Let \(P\) and \(M\) be the class of probability measures and finite measures (countably additive) on \((X, A)\). The natural \(\sigma\)-field, \(\sigma(P)\), on \(P\) is the smallest \(\sigma\)-field in \(P\) such that the function \(P \mapsto P(A)\) is measurable for each \(A\) in \(A\). There is also the notion of weak convergence in both \(P\) and \(M\), namely, \(\alpha \prec \gamma\) if and only if \(\int g d\alpha \leq \int g d\gamma\) for all bounded continuous functions on \(X\).

Under this convergence \(P\) becomes a separable complete metric space (Prohorov [4]) and the \(\sigma\)-field \(\sigma(P)\) above is the Borel \(\sigma\)-field in \(P\).

To each non-zero measure \(\alpha\) in \(M\), we denote by \(\overline{\alpha}\) the corresponding normalized measure, namely \(\overline{\alpha}(A) = \alpha(A)/\alpha(X), A \in A\).

In non-parametric Bayesian analysis, the 'true' probability measure \(\cal{P}\) takes values in \(P\), is random and has a prior distribution. To facilitate the use of standard probability theory we must view \(\cal{P}\) as a measurable map from some probability space \((\Omega, S, Q)\) into \((P, \sigma(P))\) and the induced measure \(QP^{-1}\) becomes the prior distribution. For any non-zero measure \(\alpha\) in \(M\), the Dirichlet prior measure \(D_\alpha\) with parameter \(\alpha\), is defined as follows (Ferguson [3]): For any finite measurable partition \((A_1, \ldots, A_k)\) of \(X\), the distribution of \((P(A_1), \ldots, P(A_k))\) under \(D_\alpha\) is the singular Dirichlet distribution \(D(\alpha(A_1), \ldots, \alpha(A_k))\) defined on the \(k\)-dimensional simplex as in Wilks [7] Section 7.7. Ferguson [3] used this definition and also an alternate definition (See Theorem 1 of Ferguson [3]), and derived many properties of Dirichlet priors and the corresponding Bayes estimates of population parameters. Blackwell [1] and Blackwell and MacQueen [2] have also given alternative definitions of the Dirichlet.
prior. We give below yet another definition of the Dirichlet prior which
is more general than the previous ones since we will not have to assume
that $X$ is separable metric. Let $\alpha$ be a non-zero measure in $\mathcal{M}$. Let
$(\Omega, S, Q)$ be a probability space rich enough to support two independent
sequences of i.i.d. random variables $Y_1, Y_2, \ldots$ and $\theta_1, \theta_2, \ldots$, where
$Y_1$ is $X$-valued and has distribution $\alpha$ and $\theta_1$ is real valued and has a
Beta distribution with parameters $1$ and $\alpha(X)$. Let $p_1 = \theta_1, p_2 = \theta_2(1-\theta_1),$
p_3 = $\theta_3(1-\theta_1)(1-\theta_2), \ldots$. For any $y$ in $X$ let $\delta_y$ stand for the degenerate
probability measure at $y$. Define the measurable map $P$ from $(\Omega, S)$ into
$(P, \sigma(P))$ as follows:

$$P(A) = \sum_{j=1}^{\infty} p_j \delta_{Y_j}(A). \quad (1.1)$$

Then the induced distribution of $P$ is the Dirichlet measure $D_\alpha$
with parameter $\alpha$. The proof of this fact and that the standard properties
of Dirichlet measures can be deduced from this will be given elsewhere,
Sethuraman [S].

In the statistical problem of non-parametric Bayesian analysis we
have a random variable $P$ taking values in $P$ and whose distribution is $D_\alpha$.
We also have a sample $X_1, \ldots, X_n$, which are random variables taking values
in $X$. Given $P$, these are i.i.d. with common distribution $P$. It is re-
quired to estimate a function $\phi(P)$, and the Bayes estimator $\hat{\phi}$ with respect
to squared loss is given by

$$E(\phi(P)|X_1, \ldots, X_n).$$

In particular, if $\phi(P) = \phi_g(P)$ where
\( \phi_g(P) = \int g(x)P(dx) \quad (1.2) \)

where \( g \) is a real valued measurable function on \( X \) with \( \int g^2dx < \infty \), then the Bayes estimate is given by

\[
\hat{\phi}_g = \frac{\alpha(X) \int gd\alpha + n \int gdF_n}{\alpha(X) + n}, \quad (1.3)
\]

where \( F_n \) is the empirical d.f. of \( X_1, \ldots, X_n \) (Ferguson [3]). In this if we let \( \alpha(X) \to 0 \) we obtain the classical estimate \( \int gdF_n \). Also the denominator in this estimate is \( \alpha(X) + n \) which is \( \alpha(X) \) plus the sample size. These facts have given rise to the interpretation that \( \alpha(X) \) is the prior sample size and allowing \( \alpha(X) \) to tend to zero corresponds to no prior information.

In the next section we investigate what happens to Dirichlet measures when their parameters are allowed to converge to certain values. In section 4 we investigate what happens to Bayes estimates when the parameters of the corresponding Dirichlet priors are allowed to converge to the zero measure. From the results in these two sections it follows that small values of \( \alpha(X) \) actually correspond to certain definitive information about \( P \).

3. Convergence of Dirichlet measures. In this section we study the convergence of Dirichlet measures as their parameter is allowed to converge in appropriate ways. Since \( (P, \sigma(P)) \) is a separable complete metric space endowed with its Borel \( \sigma \)-field, we can talk about the usual weak convergence of probability measures on \( (P, \sigma(P)) \) and of Dirichlet measures, in particular.
THEOREM 3.1. Let \( \{\tau_r\} \) be a sequence of measures in \( M \) and let the sequence of normalized measures \( \{\alpha_r\} \) be tight. Then the sequence \( \{D_{\alpha_r}\} \) of Dirichlet measures is tight.

PROOF. Fix \( \epsilon > 0 \). There exists a sequence of compact sets \( K_d \) in \( X \) such that
\[
\sup \alpha_r(K_d^c) \leq 6\epsilon/d^3\pi^2, \tag{3.1}
\]
\( d = 1, 2, \ldots \). Let
\[
M_d = \{P: P(K_d^c) \leq 1/d\}, \tag{3.2}
\]
d = 1, 2, ... and let
\[
M = \bigcap_d M_d. \tag{3.3}
\]
Then clearly \( M \) is a compact subset of \( P \) in the weak topology. Now, by the Chebyshev inequality
\[
D_{\alpha_r}(M_d^c) = \|E_{\alpha_r}(P(K_d^c))\| = d \alpha_r(K_d^c) \leq 6\epsilon/\pi^2d^2 \tag{3.4}
\]
and
\[
D_{\alpha_r}(M^c) \leq \sum_d 6\epsilon/\pi^2d^2 = \epsilon, \text{ for all } \tau. \tag{3.5}
\]
This proves that \( \{D_{\alpha_r}\} \) is tight. \( \Box \)

THEOREM 3.2. Let \( \{\tau_r\} \) be a sequence of measures in \( M \) such that
\[
\sup_A |\alpha_r(A) - \alpha_0(A)| \rightarrow 0 \tag{3.6}
\]
where \( \alpha_0 \) is a non-zero measure in \( M \). Then \( D_{\alpha_r} \) converges to \( D_{\alpha_0} \) weakly.
PROOF. The proof of this result rests heavily on the constructive definition of the Dirichlet measure in (1.1) and the following result which is proved in Sethuraman [6].

Let \( \{ \beta_r \} \) be a sequence of probability measures on an arbitrary measurable space \((V, B)\) and let

\[
\sup_{B} |\beta_r(B) - \beta_0(B)| \to 0, \tag{3.7}
\]

where \( \beta_0 \) is a probability measure on \((V, B)\). Then there exists a sequence of \( V \)-valued random variables \( \{ Y_r \}_{r=0}^{\infty} \) with marginal distributions \( \{ \beta_r \}_{r=0}^{\infty} \) such that

\[
\text{Prob. } \{ Y_r \neq Y_0 \} \to 0 \text{ as } r \to \infty. \tag{3.8}
\]

From (1.1) and the above result, we can find independent sequences of i.i.d. random variables \( \{ Y_r^r \}, \{ \theta_j^r \}, r = 0, 1, 2, \ldots \) such that the distribution of \( Y_1^r \) is \( \alpha_r \), the distribution of \( \theta_j^r \) is Beta with parameters \( 1 \) and \( \alpha_r(X) \), \( r = 0, 1, \ldots \), and

\[
\text{Prob. } \{ Y_j^r \neq Y_j^0 \} \to 0 \tag{3.9}
\]

and

\[
\text{Prob. } \{ \theta_j^r \neq \theta_j^0 \} \to 0 \text{ as } r \to \infty, j = 1, 2, \ldots. \tag{3.10}
\]

Furthermore, if \( p_1^r = \theta_1^r \), \( p_j^r = \theta_j^r(1-\theta_j^{r-1}) \ldots (1-\theta_1^r) \) for \( j \geq 1 \), and

\[
P_r^r(A) = \sum_{j=1}^{\infty} p_j^r \delta_{Y_j^r}(A), \tag{3.11}
\]
then the distribution of $P^r$ is the Dirichlet measure $D_{\alpha^r}$, $r = 0, 1, \ldots$.

From (3.11) it can be easily shown that, for any integer $m$,

$$\sup_A |P^r(A) - p^o(A)| \leq \sum_{j=1}^m |p_j^r - p_j^o| + \sum_{j=1}^m I(Y_j^r \neq Y_j^o) + 2 \prod_{j=1}^m (1-\delta_j^o) - \prod_{j=1}^m (1-\delta_j^r).$$

(3.12)

From the construction above and (3.8), (3.9) and (3.12) and by first choosing $m$ appropriately and then allowing $r$ to tend to $\infty$ that

$$\sup_A |P^r(A) - p^o(A)| \to 0$$

in probability which is a stronger assertion than made in the theorem, namely that $D_{\alpha^r} \to D_{\alpha^o}$ weakly. \(\square\)

**THEOREM 3.3.** Let $(\alpha_r)$ be a sequence of measures in $M$ such that

$$\alpha_r(X) \to 0 \text{ and } \sup_A |\alpha_r(A) - \alpha_o(A)| \to 0 \text{ as } r \to \infty,$$

(3.13)

where $\alpha_o$ is a probability measure in $P$. Then the measures $D_{\alpha_r}$ converge to a random degenerate measure $\delta_{Y^o}$ where $Y^o$ has distribution $\alpha_o$.

**PROOF.** As before we can construct independent sequences of i.i.d. random variables $\{Y_j^r\}$ and $\{\theta_j^r\}$, and an independent random variable $Y^o$, such that $Y_1^r$ has distribution $\alpha_r$, $Y^o$ has distribution $\alpha_o$, the distribution of $\theta_1^r$ is Beta with parameters 1 and $\alpha_r(X)$, $r = 1, 2, \ldots$, and

$$\text{Prob. } (Y_1^r \neq Y^o) \to 0 \text{ as } r \to \infty.$$

(3.14)

Furthermore, if $p_1^r = \theta_1^r$, $p_j^r = \theta_j^r(1-\theta_{j-1}^r) \ldots (1-\theta_1^r)$, for $j \geq 1$, and
\[ P^r(A) = \sum_{j=1}^{\infty} P^r \delta_{Y_j}(A), \quad (3.15) \]

then the distribution of \( P^r \) is the Dirichlet measure with parameter \( \alpha_r \), \( r = 1, 2, \ldots \).

From (3.15), it is easily seen that

\[ \sup_{A} |P^r(A) - \delta_{Y^0}(A)| \leq I(Y_1^r \neq Y^0) + 2(1-p_1^r). \quad (3.16) \]

From (3.14) and the fact that \( \alpha_r(X) \to 0 \), it follows that

\[ \sup_{A} |P^r(A) - \delta_{Y^0}(A)| \to 0 \] in probability which again is stronger than the assertion of the theorem. \( \square \)

From Theorem 3.2 it is clear that allowing \( \alpha_r(X) \) to tend to zero does not correspond to no information on \( P \). In fact if \( \alpha_r(X) \to 0 \) and the normalized measure \( \overline{\alpha}_r \) converges in the strong sense of (3.13) to a probability measure \( \overline{\alpha}_o \), then the information about \( P \) is that it is a probability measure concentrated at a particular point in \( X \) which is chosen at random according to \( \overline{\alpha}_o \). This is definitely very strong information about \( P \) and most probably not of the type any statistician would be willing to make.

4. Convergence of Bayes estimates. In this section we are mainly interested in the limits of Bayes estimates of various function \( \phi(P) \) as \( \alpha(X) \to 0 \). We will therefore make the following assumption throughout this section:

\[ \alpha_r(X) \to 0 \text{ and } \sup_{A} |\overline{\alpha}_r(A) - \overline{\alpha}_o(A)| \to 0, \quad (4.1) \]

where \( \overline{\alpha}_o \) is a probability measure in \( P \). We will also be mainly concerned
with a special class of functions \( \phi(P) \) as defined below. Let \( g \) be a permutation invariant measurable function from \( X^k \) into \( \mathbb{R}^l \) such that

\[
\int |g(x_1, \ldots, x_1, x_2, \ldots, x_m, \ldots, x_m)| \, d\alpha(x_1) \ldots d\alpha(x_m) < \infty \tag{4.2}
\]

for all possible combinations of arguments \((x_1, \ldots, x_1, x_2, \ldots, x_2, \ldots, x_m, \ldots, x_m)\) from all distinct \((m = k)\) to all identical \((m = 1)\). When the function \( g \) vanishes whenever any two coordinates are equal, condition (4.2) reduces to the simple condition

\[
\int |g(x_1, \ldots, x_k)| \, d\alpha(x_1) \ldots d\alpha(x_k) < \infty. \tag{4.3}
\]

Define the parametric function

\[
\phi_g(P) = \int g(x_1, \ldots, x_m) \, dP(x_1) \ldots dP(x_m) \tag{4.4}
\]

for all those \( P \)'s for which it exists. Let \( P \) have \( D_\alpha \) as the prior distribution and let \((X_1, \ldots, X_n)\) be a sample from \( P \). Under further assumptions concerning the second moment of \( g \) under \( D^k \), the Bayes estimate (with respect to squared error loss) of \( \phi_g(P) \) based on the sample is

\[
\hat{\phi}^n_{g,\alpha} = E_{D_\alpha} (\phi_g(P)|X_1, \ldots, X_n), \tag{4.5}
\]

and based on no sample is

\[
\hat{\phi}^0_{g,\alpha} = E_{D_\alpha} (\phi_g(P)). \tag{4.6}
\]

Since the conditional distribution of \( P \) given \((X_1, \ldots, X_n)\) is \( D_{\alpha+nF_n} \), where \( F_n \) is the empirical distribution function of \((X_1, \ldots, X_n)\), we have

\[
\hat{\phi}^n_{g,\alpha} = \hat{\phi}^0_{g,\alpha+nF_n}. \tag{4.7}
\]

Suppose that we substitute \( \alpha = \alpha_r \) where \( \{\alpha_r\} \) satisfies (4.1). From the results of section 3 we know that
and

\[ D_{\alpha_R + nF_n} \rightarrow D_{nF_n} \]  

as \( r \to \infty \). The main result of this section pertains to the convergence of

the Bayes estimates \( \hat{\phi}^{\alpha}_{g, \alpha_R} \) and \( \hat{\phi}^{\alpha}_{g, \alpha_R + nF_n} \).

**THEOREM 4.1.** Let condition (4.1) hold. Let \( g \) be a continuous function

from \( x^k \) into \( \mathbb{R}^1 \). Let \( g(x_1, \ldots, x_1, x_2, \ldots, x_2, \ldots, x_m, \ldots, x_m) \) be

uniformly integrable with respect to \( \tilde{\alpha}_m \), for all combinations of arguments

\( (x_1, \ldots, x_1, x_2, \ldots, x_2, \ldots, x_m, \ldots, x_m) \) from all distinct to all

identical. Then

\[ \hat{\phi}^{\alpha}_{g, \alpha_R} = \int g(x, \ldots, x) d\tilde{\alpha}_0(x) \]  

and

\[ \hat{\phi}^{\alpha}_{g, \alpha_R + nF_n} + \hat{\phi}^{\alpha}_{g, nF_n} = E_{D_{nF_n}}(g(Z_1, \ldots, Z_k)) \]  

where \( (Z_1, \ldots, Z_k) \) is a sample from \( P \) where \( P \) has the distribution \( D_{nF_n} \).

**PROOF.** The easiest way to prove this result is to use the repre-

sentation (1.1) for the random probability measure \( P \) with a Dirichlet

distribution. The uniform integrability conditions on \( g \) with respect to

\( \alpha_R \) immediately show that \( \hat{\phi}_g(P^R) \) is uniformly integrable with respect to \( D_{\alpha_R} \)

since it is the convex combination of uniformly integrable functions as
given below:
\[ \phi_g(p^r) = \sum_{(j_1, \ldots, j_k)} p_{j_1} \cdots p_{j_k} g(y_{j_1}, \ldots, y_{j_k}), \]

where \( y_{j_1}, \ldots \) are i.i.d. with common distribution \( \overline{a}_r \). This fact and (4.8) and (4.9) establish the results (4.10) and (4.11) of the theorem. \( \square \)

The results of this theorem generalize those of Ferguson [3] Section 5b and Se and Yamato [8], [9]. Also when \( g(x_1, \ldots, x_k) \) is such that it vanishes whenever two coordinates are equal, it is easy to see that

\[ \hat{\phi}^{(k)}_{g,nF_n} = \frac{n^{(k)}}{n^{k-1}} U_{g,n} \]

where \( U_{g,n} \) is the usual \( U \) statistic based on \( g \) and the sample \( (X_1, \ldots, X_n) \). This result is also contained in Yamato [8], [9].

5. Acknowledgements. I wish to thank Professor Roger Berger and Wai Chan for their valuable assistance in the preparation of this paper.
REFERENCES


The form of the Bayes estimate of the population mean with respect to a Dirichlet prior with parameter \( \alpha \) has given rise to the interpretation that \( \alpha(X) \) is the prior sample size. Furthermore, if \( \alpha(X) \) is made to tend to zero, then the Bayes estimate mathematically converges to the classical estimator, namely the sample mean. This has further given rise to the general feeling that allowing \( \alpha(X) \) to become small not only makes the 'prior sample size' small but also that it corresponds to no prior information. By investigating the limits of prior distributions as the parameter \( \alpha \) tends to various values, we show that it is misleading to think of \( \alpha(X) \) as the prior sample size and the smallness of \( \alpha(X) \) as no prior information. In fact very small values of \( \alpha(X) \) actually mean that we have very definite information concerning the unknown true distribution.