ON THE CONSISTENCY OF BAYES' ESTIMATES

by

Persi Diaconis
Stanford University

and

David Freedman
University of California, Berkeley

TECHNICAL REPORT NO. 213
FEBRUARY 1984

PREPARED UNDER THE AUSPICES OF
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE CONSISTENCY OF BAYES' ESTIMATES

by

Persi Diaconis
Stanford University

and

David Freedman
University of California, Berkeley

TECHNICAL REPORT NO. 213
February 1984

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE CONSISTENCY OF BAYES' ESTIMATES

by

Persi Diaconis
Department of Statistics
Stanford University

and

David Freedman
Department of Statistics
University of California, Berkeley

Abstract

We discuss frequency properties of Bayes' rules, paying special attention to consistency. Some new and fairly natural counterexamples are given, involving nonparametric estimates of location. Even the Dirichlet prior can lead to inconsistent estimates if used too aggressively. Finally, we discuss reasons for Bayesians to be interested in frequency properties of Bayes' rules. As a part of the discussion we give a subjective equivalent to consistency, and compute the derivative of the map taking priors to posteriors.
1. CONSISTENCY OF BAYES' RULES

One of the basic problems in statistics can be put this way. Data is collected following a probability model with unknown parameters; the parameters are to be estimated from the data. Often, there is prior information about the parameters, for example, their probable sign or order of magnitude. Many statisticians express such information in the form of a prior probability over the unknown parameters. Estimates based on prior probabilities will be called Bayes' estimates in what follows.

This paper studies points of contact between frequentist and Bayesian statistics. We derive frequency properties of Bayes' estimates and suggest a Bayesian interpretation for some frequentist computations. Our main concern is consistency: as more and more data are collected, will the Bayes' estimates converge to the true value for the parameters?

If the underlying probability mechanism has only a finite number of possible outcomes (tossing a coin or die) and the prior probability does not exclude the true parameter values as impossible, it has long been known that Bayes' estimates are consistent. As will be discussed below, if the underlying mechanism allows an infinite number of possible outcomes (e.g., estimation of an unknown probability on the integers), Bayes' estimates can be inconsistent: as more and more data comes in, some Bayesian statisticians will become more and more convinced of the wrong answer. The class of tail-free and Dirichlet priors was introduced to insure consistency in such settings. We present examples showing that mechanical use of such priors still leads to inconsistent estimates.

In section 2 we review other points of contact between the mathematics of frequentist and Bayesian statistics. In section 3, we offer two Bayesian uses for frequentist computations. The first is a subjective equivalent to consistency involving intersubjective agreement. The second uses frequency computations
as a way of thinking about priors. As part of the discussion, we compute the
derivative of the map taking priors to posteriors.

To define things, consider a family of probabilities \( \{ Q_\theta : \theta \in \Theta \} \) on a
space \( X \). Write \( Q_\theta^\infty \) for the infinite product measure on \( X^\infty \), which makes
the coordinate random variables \( X_1, X_2, \ldots \), independent with common dis-
btribution \( Q_\theta \). We will assume throughout that \( X \) and \( \Theta \) are Borel subsets
of complete separable metric spaces. Let \( \mu \) be a prior probability on \( \Theta \).
Let \( P_\mu \) be the joint distribution of the parameter and the data:

\[
P_\mu(A \times B) = \int_A Q_\theta^\infty(B) \mu(d\theta)
\]

for Borel sets \( A \) and \( B \). The posterior is the \( P_\mu \)-distribution of the
parameter \( \theta \) given the data \( X_1, \ldots, X_n \); we denote this as

\[ \mu_n(d\theta | X_1 \ldots X_n) \].

The usual Bayes' estimate is just the mean of the posterior,

Here are a few typical examples:

- **coin-tossing.** \( X \) has two points, \( H \) and \( T \) for heads and tails, res-
  spectively. The parameter space \( \Theta \) can be taken as the unit interval: \( \theta \in \Theta \)
is the probability of a head. Then \( X^\infty \) is the space of sequences of heads
  and tails; \( Q_\theta^\infty \) makes the sequence independent; in any position, there is a
  chance \( \theta \) of getting a head and \( 1-\theta \) of getting a tail. We call this a "\( \theta \)-coin."
Informally, \( P_\mu \) may be described as follows: choose \( \theta \) at random from \( \mu \), then
toss a \( \theta \)-coin. Of course, the posterior \( \mu_n(d\theta | X_1, \ldots, X_n) \) has a density with re-
spect to the prior, proportional to the likelihood function:

\[
\theta^S(1-\theta)^{n-S}/[\int \theta^S(1-\theta)^{n-S} \mu(d\theta)],
\]

where \( S \) is the number of heads, \( n-S \) the number of tails. Thus, the Bayes' estimate is

\[
\int \theta^{S+1}(1-\theta)^{n-S} \mu(d\theta)/\int \theta^S(1-\theta)^{n-S} \mu(d\theta).
\]
• **rolling a die.** This is the same as coin tossing, except that
\( X = \{1, 2, 3, 4, 5, 6\} \), and \( \Theta \) is the "6-simplex," all sequences \( \theta_1, \ldots, \theta_6 \) of
length six whose terms are nonnegative and add to one:
\[
\theta_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{6} \theta_i = 1 .
\]

• **rolling an infinite die.** The same, except \( X = \{1, 2, 3, \ldots\} \) and \( \Theta \) is
an infinite-dimensional simplex. Despite the superficial similarity, all the
paradoxes of inconsistency already appear in this case. From our point of view,
this is the simplest natural example of a nonparametric problem: estimating an
unknown probability on the positive integers. On the other hand, we are still
in the dominated case, so the posterior can still be computed in the usual way:
\( \mu_n \) denotes this familiar version of the posterior.

• **the real line.** Now \( X = (-\infty, \infty) \), and \( \Theta \) is the set of all probabilities
on \( X \). This is another nonparametric problem: estimating an unknown proba-
bility on the line. There is an additional complication: \( \Theta \) is undominated.
The posterior still exists, but in general there is no explicit formula for it.
There will usually be many versions of the posterior, and this causes some tech-
nical difficulties.

We will use the weak-star topology: If \( \mu_n \) and \( \mu \) are probabilities
on \( \Theta \), then \( \mu_n \rightarrow \mu \) iff \( \int f d\mu_n \rightarrow \int f d\mu \) for all bounded continuous functions
\( f \) on \( \Theta \). We denote point mass at \( \theta \) by \( \delta_\theta \). Thus, \( \mu_n \rightarrow \delta_\theta \) iff \( \mu_n(U) \rightarrow 1 \)
for every open neighborhood \( U \) of \( \theta \). For example, in the coin tossing case, \( \Theta \)
is the unit interval; \( \mu_n \rightarrow \mu \) weak star if and only if \( \mu_n[0,x] \rightarrow \mu[0,x] \) for all
intervals such that \( \mu(x) = 0 \). Turn next to the infinite die. Then \( \mu_n \rightarrow \mu \) if
and only if
\[
\mu_n(\{\theta | \theta_1 \leq x_1 \text{ and } \theta_2 \leq x_2 \ldots \text{ and } \theta_k \leq x_k\}) \\
\rightarrow \mu(\{\theta | \theta_1 \leq x_1 \text{ and } \theta_2 \leq x_2 \ldots \text{ and } \theta_k \leq x_k\})
\]
for all $k$ and all $x$'s such that

$$\mu(\theta|\theta_1 = x_1 \text{ or } \theta_2 = x_2 \ldots \text{ or } \theta_k = x_k) = 0$$

Finally, take the line. Changing the imagery a little, $\mu_n \rightarrow \mu$ if and only if

$$\int_\Theta \left[ \prod_{i=1}^k \int f_i(x) \theta(dx) \right] \mu_n(d\theta)$$

$$+ \int_\Theta \left[ \prod_{i=1}^k \int f_i(x) \theta(dx) \right] \mu(d\theta)$$

for all $k$ and all bounded continuous functions $f_i$ on $X$.

The pair $(\theta, \mu_n)$ is consistent if for $Q_\theta^\infty$-almost all sequences, the posterior $\mu_n$ converges to point mass at $\theta$ in the weak-star topology. The weak-star topology has the fewest open sets of any natural topology, so it is fairly easy for posteriors to be consistent. Minor technical difficulties apart, if $(\theta, \mu_n)$ is consistent in our sense, the Bayes' estimate for $\theta$ -- in the sense of a posterior mean -- will be consistent too. We say $\mu_n$ is consistent if $(\theta, \mu_n)$ is consistent for all $\theta$. Notice that consistency depends to some extent on the version selected for $\mu_n$. Often, there is only one natural version, as in the dominated case. Sometimes, a natural version can be selected on the basis of continuity in the data: see Zabell (1979). Sometimes, however, there is no sensible way to resolve the ambiguity. When there is a natural version of $\mu_n$, as in the dominated case, we will say that $(\theta, \mu)$ is consistent, rather than $(\theta, \mu_n)$.

In the coin tossing example, what does it mean for $(\theta, \mu_n)$ to be consistent? Just that a Bayesian with prior $\mu$, who happens to be tossing a $\theta$-coin, will eventually find this out: his posterior will concentrate in smaller and smaller intervals around $\theta$, as more and more data comes in. Likewise for the infinite die. More specifically, $(\theta, \mu)$ is consistent if and only if for any positive integer $k$ and any small positive $\varepsilon$, 

4
\[ \mu_n \{ N_{k\in} | X_1, \ldots, X_n \} \to 1 \text{ as } n \to \infty \text{ a.e. } Q_\theta \]

where

\[ N_{k\in} = \{ \phi : \phi \in \Theta \text{ and } |\phi_i - \theta_i| < \varepsilon \text{ for } i = 1, \ldots, k \} . \]

For coin tossing, \((\theta, \mu)\) is consistent iff \(\mu\) assigns positive mass to every open interval around \(\theta\). For the infinite die, the situation is much more complicated -- and that is what the present discussion is all about.

Doob (1948) proved a very general theorem on consistency: a reimplication is that: \((\theta, \mu_n)\) is consistent for \(\mu\)-almost all \(\theta\). Thus, a Bayesian with prior \(\mu\) can be sure that the posterior will converge. (This does not depend on the version, because null sets don't matter here. Doob's work has been extended by Breiman, LeCam, and Schwartz, 1964.) However, a frequentist using a Bayes' rule will want to know for which \(\theta\) the rule is consistent.

For smooth, finite dimensional families, \((\theta, \mu)\) is consistent if and only if \(\theta\) is in the support of \(\mu\). See Freedman (1963) or Schwartz (1965). (The support is the smallest closed set of probability 1.) But the assumption of finite dimensionality is really needed. For example, take \(X\) to be the positive integers. Take \(\Theta\) to be the set of all probabilities on \(X\). Take \(\Theta\) to be the geometric distribution with parameter 1/2. Freedman (1963) constructs a prior \(\mu\) with the following properties:

- Every open neighborhood of \(\theta\) has positive \(\mu\)-probability.
- For \(Q^\infty\)-almost all sequences, the posterior converges to a point mass at a geometric distribution with parameter 3/4.

This example is generic in a topological sense: for most priors \(\mu\) and most parameters \(\theta\), the pair \((\theta, \mu)\) is inconsistent. To make this precise,
we need the notion of "category." A set is of the first category if it is contained in a countable union of closed, nowhere dense sets. First-category sets are the topological equivalents of null sets. Put the weak-star topology on \( \pi(\Theta) \), the set of priors on \( \Theta \). Then the set of consistent pairs \((\Theta, \mu)\) is of the first category in \( \Theta \times \pi(\Theta) \). See Freedman (1965).

The existence of such counterexamples suggests the need for some careful investigation. Are there priors consistent for all parameters? (We will call such priors "consistent.") For countable \( X \), Freedman suggested using tail-free and Dirichlet priors, showing that these priors are consistent at all parameters. This work was extended by Fabius (1964), who showed how to construct consistent tail-free priors for any complete, separable metric \( X \). These ideas were further developed by Ferguson (1973, 1974) who gives an excellent survey of the literature. Other extensions include "neutral-to-the-right priors." See Doksum (1974).

One important feature of tail-free priors is that they allow a Bayesian to approximate any prior by one consistent at all parameter values. For definiteness, consider the problem of estimating an unknown probability \( \theta \) on the positive integers. Write \( \theta_i \) for the \( i \)th coordinate of \( \theta \). A Bayesian may have a reasonably clear opinion about \( \theta_i \) for some \( i \)'s. However, it seems unlikely that such an opinion can be carefully quantified for all \( i \). Now tail-free priors are arbitrary for any finite number of the \( \theta_i \), with the two restrictions that any open set have positive prior probability, and that some mass be left for the remaining \( \theta_i \)'s. The assignment of prior mass over the remaining \( \theta \)'s is done so the resulting prior is consistent. The construction will now be sketched.

We visualize the prior as randomly selecting a probability on the positive integers. Even more crudely, what the prior does is to randomly distribute a
total mass of 1 among the integers. Thus, $\theta_1$ is the randomly selected mass assigned to the integer 1, while $\theta_2$ is the mass assigned to 2, and so on.

For tail-free priors, the prior opinion on any finite number of $\theta_j$, say $1 \leq j \leq N$, can be specified in any way to reflect the prior opinion of a Bayesian statistician. Then, the prior specification is completed by "stick breaking," using a sequence of independent random variables $S_i$ with values in the open unit interval. These may have arbitrary distributions save for two restrictions: that $S_i$ falls into any open interval with positive probability, and that $\sum E(S_i) = \infty$. The "cuts" $S_i$ are used to distribute mass as follows: Mass $m = \sum_{j=1}^{N} \theta_j$ has been assigned and mass $1-m$ remains; now mass $(1-m)S_{N+1}$ is assigned to $N+1$ and mass

$$1 - m - (1-m)S_{N+1} = (1-m)(1-S_{N+1})$$

is left for the next move, which is carried out using $S_{N+2}$. And so on.

Dirichlet priors are tail-free, the $S_i$ having certain beta distributions. Usually, Dirichlet priors are parametrized in terms of a finite measure $\alpha$ on the observation space $X$, which is for the moment general. The Dirichlet prior $D(\alpha)$ with base measure $\alpha$ can be characterized as follows. Partition the observation space $X$ into a finite number of sets $A_1, A_2, \ldots, A_k$. Consider a probability $\theta$ on the observation space, selected at random from $D(\alpha)$. Then $\theta(A_1), \ldots, \theta(A_k)$ are random variables with respect to $D(\alpha)$. These have a Dirichlet distribution on the $k$-simplex, with parameter vector $\alpha(A_1), \ldots, \alpha(A_k)$. More concretely, these are distributed like

$$U_1/S, \ldots, U_k/S,$$

where $S = U_1 + \ldots + U_k$, the $U_i$ being independent gamma variables with shape parameter $\alpha(A_i)$. When the observation space is the integers, $\alpha$ is specified
by a countable collection of numbers \( \alpha_1, \alpha_2, \ldots \). Let \( ||\alpha|| = \sum \alpha_i \). By assumption, \( ||\alpha|| < \infty \). For a Dirichlet prior with parameter measure \( \alpha \) on the positive integers, the cut \( S_1 \) has a beta distribution with the two parameters \( \alpha_1 + \ldots + \alpha_i \) and \( ||\alpha|| - (\alpha_1 + \ldots + \alpha_i) \).

Dirichlet priors have been suggested for use in a wide variety of problems. Also, some writers have considered using mixtures of Dirichlet priors, see Antoniak (1974), or Dalal and Hall (1980). It is natural to ask whether mixtures of Dirichlet priors are consistent. The first step is easy; any mixture of a finite number of consistent priors is consistent. However, in Freedman and Diaconis (1983), we showed that a countable mixture of Dirichlet priors can be inconsistent. A starting point of our construction is an example showing that a mixture of a Dirichlet prior and a point mass at a certain long-tailed probability is inconsistent. Such priors are similar to ones suggested by Jeffreys (1967) for Bayesian hypothesis testing. On the positive side, we showed that if the mass of the parameter measure is bounded, then the mixture is consistent.

The inconsistent priors in Freedman (1963) were constructed with malice aforethought. Bayesian reaction seems to be, "Oh, nobody would ever use a prior like that." See for example the remarks in Box, Leonard, and Wu (1983, pg. xi). That a Jeffreys-style prior is inconsistent should therefore be of interest. Moreover, in the paper following this one, we give examples of priors suggested by practicing Bayesians, which turn out to be inconsistent. That paper is fairly technical, and the following heuristic discussion may be helpful.

Consider estimating an unknown location parameter \( \theta \), with squared error as loss. The observations are modelled as

\[
X_i = \theta + \varepsilon_i \quad i = 1, 2, \ldots
\]

The \( \varepsilon_i \) are independent disturbance terms with a common distribution. If the
have known density $g$, and a prior $\pi$ is put on $\theta$, then the Bayes estimate is the mean of the posterior distribution:

$$
\hat{\theta} = \frac{\int_\Theta \pi g(x_i - \theta) \mu(d\theta)}{\int \pi g(x_i - \theta) \mu(d\theta)}.
$$

If $\mu(d\theta)$ is taken as Lebesgue measure, then this becomes the Pitman estimator. If the density $g$ is unknown, it can be estimated from the data. Often, $g$ is assumed to belong to some parametric family. See Fraser (1976), Box and Tiao (1973, Chapters 3,4), or Johns (1979).

A nonparametric approach is also natural. This involves putting a prior on $\theta$ and a prior on the law $F$ of $\varepsilon_i$. Dalal (1979a,b) has suggested using a Dirichlet prior for $F$. We will now show that for a Dirichlet prior with a Cauchy base measure $\alpha$, the Bayes' estimates are inconsistent.

To avoid identifiability problems, we will assume that the law $F$ of $\varepsilon$ is symmetric. To put a prior on symmetric $F$'s, we symmetrize the Dirichlet as follows: If $G$ is a distribution function for a random variable $X$, let $G^-$ be the distribution function of $-X$ and let $\bar{G} = \frac{1}{2}(G + G^-)$, so $\bar{G}$ is symmetric. Let $\bar{D}_\alpha$ be the law of $\bar{G}$, where $G$ has a Dirichlet distribution with base measure $\alpha$ on $\mathbb{R}$. This is a symmetrized Dirichlet. The construction was first suggested by Dalal (1979a,b). For further discussion see Hannum and Hollander (1983). We assume that the base measure $\alpha$ has density $\alpha'$, and write $||\alpha|| = \alpha(\mathbb{R})$. We make $\theta$ and $F$ independent; our prior for $\theta$ has density $f$ on $\mathbb{R}$.

The whole posterior distribution is computed in Theorem 4 of the paper which follows. For simplicity, we only discuss the Bayes' estimate here.

**Theorem A.** Suppose that $X_1, \ldots, X_n$ are all distinct. Let $\theta_{ij} = \frac{1}{2}(X_i + X_j)$, and suppose $\theta_{ij}$ are distinct. The Bayes' estimate $\theta$ is
\[ \hat{\theta}_n = \left[ \int \omega(\theta) \, d\theta + \sum_{i \leq j} \theta_{ij} \omega_{ij} \right] / \left[ \int \omega(\theta) \, d\theta + \sum_{i \leq j} \omega_{ij} \right] \]

with

\[
\omega(\theta) = ||\alpha|| f(\theta) \prod_{k=1}^{n} \alpha'(X_k - \theta) 
\]

\[
\omega_{ij} = \frac{1}{2} \left[ f(\theta_{ij}) / \alpha'(\delta_{ij}) \right] \prod_{k=1}^{n} \alpha'(X_k - \theta_{ij}) 
\]

\[
\delta_{ij} = \frac{1}{2} (X_i - X_j) 
\]

For the next result, take \( f \) to be standard normal and \( \alpha \) to be Cauchy. Neither choice is crucial: \( f \) can be any smooth density and \( \alpha \) can be any \( t \) distribution. We suppose that in fact the \( \epsilon_i \) have a density \( h \) which is symmetric about \( 0 \), with a strict maximum at \( 0 \); further, \( h \) is infinitely differentiable and vanishes off a compact interval.

**Theorem B.** For some \( h \), the Bayes' estimate is inconsistent. Indeed, for large \( n \), there is probability near \( \frac{1}{2} \) that \( \hat{\theta}_n \) is close to \( \gamma \) and probability near \( \frac{1}{2} \) that \( \hat{\theta}_n \) is close to \(-\gamma\), where \( \gamma \neq 0 \) depends on \( h \).

**Remarks.**

1) When sampling from a continuous density, \( X_i \) and \( \theta_{ij} \) will be distinct with probability 1.

2) The Bayes' estimate \( \hat{\theta} \) is a convex combination of two other estimates:

\[ \hat{\theta}_1 = \int \omega(\theta) \, d\theta / \int \omega(\theta) \, d\theta \]

and

\[ \hat{\theta}_2 = \sum_{i \leq j} \theta_{ij} \omega_{ij} / \sum_{i \leq j} \omega_{ij} \]
The first is a Bayes' estimate, as at (1.1), when sampling from known density \( \alpha' \). The second is a weighted average of the \( \hat{\theta}_{ij} \), which estimate \( \theta \) from the pair \( X_i \) and \( X_j \). If \( \alpha' \) is one of the standard unimodal densities such as normal or Cauchy, the weight \( \omega_{ij} \) is relatively large when \( \hat{\theta}_{ij} \) is in the center of the \( X_k \)'s. It turns out that \( \hat{\theta}_1 \) dominates \( \hat{\theta}_2 \) as \( n \to \infty \), for some \( h \); for others, \( \hat{\theta}_2 \) will dominate. Here, we focus on the first case.

3) The \( h \) constructed in Theorem B is trimodal, as in Figure 1 below. It has a unique maximum at 0, but the two other modes matter.

\[ \text{Figure 1. The Counterexample Density} \]

4) Any of the classical estimators, such as the mean or the median will be consistent in this situation, so the Bayes' estimates do worse than available frequentist procedures.

5) In the paper which follows, we prove that the Bayes' estimate is consistent for any strongly unimodal density \( h \). Further, we show that if \( \log \alpha' \) is convex, then the Bayes' estimate is consistent for any symmetric \( h \).

6) Doss (1983a,b,c) has carried out similar computations for neutral to the right and tail-free priors. He has also introduced some other methods of symmetrizing. Very roughly, his results parallel ours; it does not seem possible to find a consistent prior concentrated in a neighborhood of the normal location family.
We now sketch the argument for inconsistency. As explained in Remark 2, the Bayes' estimate is essentially

$$\hat{\theta}_1 = \frac{\int \tilde{\omega}(\theta) d\theta}{\int \omega(\theta) d\theta}$$

with $$\tilde{\omega}(\theta) = \prod_{i=1}^{n} \frac{1}{1 + (X_i - \theta)^2}$$, the $$X_i$$ being the data. We have

$$\tilde{\omega}(\theta) = \exp\left[\sum_{i=1}^{n} \log[1 + (X_i - \theta)^2]\right] \approx \exp\left[-n E[\log(1 + (X_i - \theta)^2)]\right].$$

In the approximation, the sum has been replaced by its mean under the true sampling distribution. Now consider

$$H(\theta) = E[\log[1 + (X-\theta)^2]].$$

If $$X$$ takes only the two values $$\pm 1$$, then $$H$$ has a local maximum at 0, a global minimum at $$\pm \gamma$$, and tends to infinity at $$\pm \infty$$. Of course, $$H$$ is symmetric. See Figure 2.

**Figure 2.** The function $$H$$.

Now we estimate the integrals in the numerator and denominator of $$\hat{\theta}_1$$ by Laplace's method. As a function, $$\omega(\theta)$$ is close to $$\exp[-nH(\theta)]$$, so only $$\theta$$'s near $$\pm \gamma$$ matter, and as $$n$$ tends to infinity, $$\hat{\theta}_1$$ oscillates between $$+\gamma$$ and $$-\gamma$$. The two-point distribution of $$X$$ can be smoothed out to the density $$h$$ shown in Figure 1.
This argument is like the one we used in Freedman and Diaconis (1982) to show inconsistency of M-estimators. The estimators proposed by Fraser (1976), Box and Tiao (1973) and Johns (1979) also seem to be inconsistent for some \( h \), for similar reasons.

It has been objected that the counterexample density \( h \) is not in the "support" of the Dirichlet, since the Dirichlet chooses a discrete measure with probability 1. A technical response is that the Dirichlet assigns positive mass to every open set of probabilities, so \( h \) is in the support — the smallest closed set of full prior mass. A broader response is that the Dirichlet assigns zero mass to any particular probability. After the fact, therefore, any troublesome probability can be differentiated in some qualitative way from a Borel set of full prior mass.

The foregoing discussion has all been asymptotic. What are the practical implications? Consider an estimation problem with a large amount of data — and a large number of parameters. A Bayes' rule will do well for most parameter values, "most" being defined relative to the prior measure itself. On the other hand, a set of parameters which is large as judged by one measure may be quite small as judged by another, so Bayes' rules may be quite unsatisfactory for a practical frequentist. Bayesians usually argue that the data will swamp the prior — but this may not happen in high-dimensional inference problems, or it may occur, but very slowly.

On the other hand, even the most dedicated subjectivist will usually not insist on quantifying all the precise details of a prior opinion in a high-dimensional situation. Our positive results indicate that small changes in the details of a prior can lead to Bayes' rules with much better operating characteristics.

Quantifying these ideas is hard, and that is why we give asymptotic results. We hope to study the finite sample problem elsewhere, and in particular we hope to quantify the extent to which small changes in the prior can make
small changes in the posterior, but big improvements in rates of convergence. The derivative of the posterior with respect to the prior is a relevant concept, and will be discussed in section 3.

2. OTHER CONNECTIONS BETWEEN BAYESIANS AND FREQUENTISTS

Frequentists often discuss Bayes' rules for the following reasons. First, the complete class theorem, as in Wald (1950), LeCam (1955), Stein (1955), Sacks (1963), or Brown (1981), implies that all admissible procedures are approximately Bayes'. Similarly, all minimax procedures are approximately Bayes'. So not much is lost by confining attention to Bayes' procedures. Thus, Bayes' procedures are convenient, tractable, and close to optimal. For instance, in some problems confidence sets are difficult to obtain. Welch and Peers (1963), Hinkley (1980), and Stein (1982) suggest using regions of high posterior mass, the prior being chosen so that Bayesian and frequentist coverage probabilities agree to several terms in an asymptotic expansion. Similarly, Bayesian techniques are used as a way of eliminating nuisance parameters. Berk (1970) gives examples of this in sequential analysis.

Second, Bayesian techniques can be helpful in proving frequentist theorems. For example, consider estimating the mean of a univariate normal with known scale and squared error loss. Any admissible estimator must be an analytic function of the observations. The only available proof uses complete class theorems of Sacks (1963) and Stein (1958) to represent the estimator as a formal Bayes' rule. It is easy to show that formal Bayes' rules are suitably smooth. Another example of this sort is in Matthes and Truax (1967).

Naturally, frequentists have been interested in frequentist properties of Bayes' procedures. Asymptotic results dating back to Laplace (1847) show that the posterior distribution can be approximated by the sampling distribution of
the maximum likelihood estimator. Modern versions of these theorems can be found in Bernstein (1934), Von Mises (1964), Johnson (1967, 1970), LeCam (1982), or Ghosh et al (1982). Bayesians use these results to show that standard frequentist procedures are nearly Bayes'. See Lindley (1965).

There are many other points of contact between the frequentist and Bayesian schools. Often, there is some incompatibility between Bayes' rules and frequentist desiderata. For example, Blackwell and Girshick (1954) followed by Blackwell and Bickel (1967) and Noorbaloochi and Meedan (1983) have shown that there are essentially no unbiased Bayes' procedures. Other points of difference are emphasized in survey articles by Pratt (1965), Cornfield (1969), Lindley (1972), and Neyman (1977).

Sufficiency gives a point of technical contact. Kolmogorov (1942) introduced the notion of Bayesian sufficiency -- a statistic is Bayes' sufficient if for any prior the posterior only depends on the data through the sufficient statistic. For smooth, finite-dimensional problems, Bayesian sufficiency is the same as the frequentist concept. Recently, Blackwell and Ramamoorthi (1982) give an infinite-dimensional example where the two notions disagree.

Points of agreement arise in Bayesian discussions of robustness as in Berger (1982), or Kadane and Chuang (1978). Here frequentist properties of Bayes' procedures are used as a means of protection from naive specification in the tail of the prior. A similar compromise was suggested by Hodges and Lehmann (1952).

3. BAYESIAN INTERPRETATIONS OF CONSISTENCY

It is useful to separate Bayesians into two groups: we will call them "classical" and "subjectivist." Classical Bayesians, like Laplace or Bayes himself, believe there is a true but unknown parameter which is to be estimated from data. This parameter is part of an objective probability model for the data. Prior opinion about the parameter is expressed as a probability distribution. Subjective Bayesians like de Finetti and Savage, reject such ideas; for them,
probabilities represent degrees of belief, and there are no objective probability models. (Freedman used to be a classical Bayesian, while Diaconis was a subjectivist.)

Consistency properties of Bayes' rules are clearly of interest to classical Bayesians; as data comes in, the posterior should converge to point mass at the true parameter. We will now argue that frequency properties of Bayes' rules are also of interest to subjectivists. The first reason has to do with "intersubjective agreement." In some circumstances, Bayesians learn from experience, so opinions based on very different priors will merge as data accumulates; the data swamps the prior.

A general result of this type was proved by Blackwell and Dubins (1962). To state their result, assume that $P$ and $Q$ are probabilities governing a process $(X_1, X_2, \ldots)$. Let $P_n$ and $Q_n$ be regular conditional probabilities for the future $(X_{n+1}, X_{n+2}, \ldots)$ given the past $(X_1, X_2, \ldots, X_n)$. We think of $P$ and $Q$ as the priors of two Bayesians, and assume $P \equiv Q$: the Bayesians agree on what is possible or impossible. Blackwell and Dubins show that almost surely, $P_n$ and $Q_n$ merge in variation distance.

Reverting to the context of section 1, assume that $\Theta \rightarrow \Theta_0$ is a homeomorphism and consider two Bayesians with priors $\mu$ and $\nu$ respectively. Now $\mu \equiv \nu$ if and only if $P_\mu \equiv P_\nu$, and $\mu \perp \nu$ if and only if $P_\mu \perp P_\nu$. If $\mu \perp \nu$ (as is usually the case with Dirichlet priors), then the Blackwell-Dubins result does not apply. However, even if merging in variation distance does not happen, "weak star merging" is still a possibility. We say two sequences $\{\alpha_n\}, \{\beta_n\}$ of probabilities merge weak star if and only if $\alpha_n$ and $\beta_n$ become indistinguishable from the point of view of integrating bounded continuous functions: more precisely, for every weak star continuous function $R$ defined on pairs of probabilities with $R(\alpha, \alpha) = 0$, $R(\alpha_n, \beta_n) \rightarrow 0$. As an example, $R$ might be a metric for weak star convergence; or $R(\alpha, \beta)$ might be $\int f d\alpha - \int f d\beta$.

We show that the version $\eta_n$ is consistent if and only if $P_n$ and $P_\nu$ merge weak-star for any $\nu$, in the following sense. Recall that $P_\mu$ is the joint distribution of $\Theta$ and $X_1, X_2, \ldots$. Let $P_{\mu n}$ be the law of $X_{n+1}, X_{n+2}, \ldots$,
given \( X_1, \ldots, X_n \), determined by the formula

\[ P_{\mu n} = \int Q_\theta^{\infty} \mu_n(d\theta). \]

A Bayesian with prior \( \mu \) holds the opinion \( P_{\mu n} \) about the future \( X_{n+1}, X_{n+2}, \ldots \), after seeing the data \( X_1, \ldots, X_n \).

**Theorem C.** Suppose \( \Theta \rightarrow P_\Theta \) is continuous, 1-1, and continuously invertible. Suppose the version \( \mu_n \) is consistent. Let \( \nu \) be any other prior. As \( n \rightarrow \infty \), \( P_{\mu n} \) and \( P_{\nu n} \) merge weak-star along \( P_\nu \)-almost all sequences. Conversely, if \( P_{\mu n} \) and \( P_{\nu n} \) merge for all \( \nu \), then \( \mu_n \) is consistent.

Informally, the prior \( \mu \) is consistent if and only if any other subjectivist with prior \( \nu \) is sure that \( P_\mu \) and \( P_\nu \) will merge on the future, as more and more data comes in. Theorem C is proved in Appendix A. Similar issues were considered by Lockett (1971).

This completes our discussion of "intersubjective agreement." We turn to our second reason for thinking that frequency properties should interest subjectivists: as a way of specifying priors. The idea is simple: after specifying a prior distribution, generate imaginary data sequences, compute the posterior, and consider whether the posterior would be an adequate representation of the updated prior. This is quite close to the checks for coherence proposed by de Finetti and Savage; see Savage (1971) for a review. The use of fictitious samples was suggested by the "device of imaginary results;" Good (1950, p. 35) uses this device as a method of roughly quantifying a prior in difficult situations. We call it the "what if" method: what if the data came out that way?

Take the prior suggested by Dalal (1979), for use in symmetric location problems, as discussed in section 1. If data were generated from the density \( h \) of Theorem B, the posterior would oscillate and never converge to the real \( \theta \). The "what if" method strongly suggests modifying the prior. We give further examples and discussion in Diaconis and Freedman (1983).
The "what if" method suggests various interesting mathematical problems. For example, to what extent do the Bayes' estimates determine the prior? Diaconis and Ylvisaker (1979) show that the posterior mean of the natural parameter in an exponential family determines the prior, under suitable regularity conditions. Diaconis and Ylvisaker (1983) give counterexamples in location problems.

For a fixed observation \( x \), Bayes' theorem gives the relationship between the posterior and the prior. This defines a map from measures to measures. In the next theorem we calculate the derivative of this map and the norm of the derivative. This helps to identify data sets \( x \) where small changes in the prior cause large changes in the posterior. These may be the most informative \( x \)'s for the "what if" method. (Of course, they may also be unlikely \( x \)'s.)

In the theorem, probabilities are considered as a subset of all signed measures, with distance defined by the variation norm: the distance between \( \mu \) and \( \nu \) is the total mass of the signed measure \( \mu - \nu \). Equivalently,

\[
\|\mu - \nu\| = \int \left| \frac{d\mu}{d\sigma} - \frac{d\nu}{d\sigma} \right| d\sigma
\]

where \( \sigma \) is any dominating \( \sigma \)-finite measure, e.g., \( \sigma = \mu + \nu \). With this norm, the measures form a Banach space.

Suppose \( \{ Q_\theta : \theta \in \Theta \} \) is a "dominated family;" there exists a \( \sigma \)-finite measure \( \lambda \) on \( X \) with all the \( Q_\theta \)'s absolutely continuous with respect to \( \lambda \). Let \( f(x|\theta) \) be the density of \( Q_\theta \) with respect to \( \lambda \); we write \( dQ_\theta = f(\cdot|\theta) d\lambda \).

We will assume that \( f \) is jointly measurable, and \( \sup_\Theta f(x|\theta) < \infty \) for every \( x \).

For a probability \( \mu \) on \( \Theta \), define

\[
T(\mu) = N(\mu)/D(\mu)
\]

where

\[
N(\mu)(d\theta) = f(x|\theta) \mu(d\theta) \quad \text{and} \quad D(\mu) = \int_\Theta f(x|\theta) \mu(d\theta) .
\]

Thus, \( N(\mu) \) is a measure and \( D(\mu) \) is a number; \( T(\mu) \) is the posterior distribution.
The dependence on $x$ is suppressed for the moment. The map $\mu \to T(\mu)$ takes priors into posteriors. It has a derivative $\dot{T}_\mu$ at $\mu$. This $\dot{T}_\mu$ is a linear map on signed measures such that

$$T(\mu + \delta) = T(\mu) + \dot{T}_\mu(\delta) + o(\|\delta\|) \quad \text{as} \quad \|\delta\| \to 0,$$

where $\delta$ is a signed measure with signed mass 0. The norm of $\dot{T}_\mu$ will be used as a measure of the effect of a small change in $\mu$. This is defined by

$$\|\dot{T}_\mu\| = \sup_{\|\delta\|=1} \|\dot{T}_\mu(\delta)\|.$$

Let $\sup_0 f(x|\theta) = \sup_{\emptyset} \{ f(x|\theta) : \mu(\emptyset) = 0 \}$, the sup over an empty set being 0 by convention. Under the given conditions:

**Theorem E.** a) $\dot{T}_\mu(\delta) = [N(\delta)/D(\mu)] - [N(\mu)D(\delta)/D(\mu)^2]$. 

b) $\|\dot{T}_\mu\| \leq \sup_0 f(x|\theta)/D(\mu)$.

c) $\|\dot{T}_\mu\| \geq \sup_0 f(x|\theta)/D(\mu)$.

Theorem E is proved in Appendix B. In (a), the quantity $N(\mu)$ is a signed measure, normalized by $D(\mu)$. Likewise, $N(\mu)$ is a signed measure, normalized by the factor $D(\delta)/D(\mu)^2$. Thus, $\dot{T}_\mu(\delta)$ is a signed measure. As is easily seen, this measure has signed mass 0. Conclusion (b) has a simple interpretation in terms of likelihood ratios: The x's where the posterior is most sensitive to small changes in the prior are the x's which have high ratio of objectivist likelihood to subjectivist likelihood. These x's are the ones where the "what if" method will be most informative: at such x, small changes in the prior can make big changes in the posterior.

As an example, consider a normal location problem with known variance $\sigma^2$. Without loss, suppose there is only one observation. Let $\mu$ be a normal prior for the location parameter $\theta$. Suppose that $\mu$ has mean $\mu_0$ and variance $\sigma_0^2$. 

19
Then $D(\mu)$ at $x$ turns out to be the normal density at $x$ with mean $\mu_0$ and variance $\sigma_0^2 + \sigma^2$. The norm computed in (b) is

$$\frac{\sigma^2 + \sigma_0^2}{\sigma^2} \exp \frac{1}{2} \{(x-\mu_0)^2/(\sigma^2 + \sigma_0^2)\}.$$ 

This is large when $x$ is far from $\mu_0$. We have carried out similar computations for the other standard one-dimensional exponential families with conjugate priors. The results are similar: values of $x$ far from the mean of the prior lead to large norms.

On the other hand, for examples of the type considered in Freedman and Diaconis (1983), the posterior concentrates at some distance from the maximum likelihood estimate, so $\int f d\mu$ will be many orders of magnitude smaller than $\max f$ for most data sets, relative to the time sampling distribution. Such data sets will have high leverage, and the "what if" method might prompt revision of an inconsistent prior.
APPENDIX A. Merging of posteriors and the weak-star topology.

We suppose \( \Theta \) is a Borel subset of a complete separable metric space -- a Borel set for short. Next, for each \( \theta \in \Theta \), we have a probability \( Q_\theta \) on the Borel subsets of another Borel set \( X \). Thus, \( \Theta \) is the parameter space and \( X \) the observation space. The map \( \theta \mapsto Q_\theta \) is assumed to be 1-1 and Borel. For a prior probability \( \mu \) on \( \Theta \), we write \( P_\mu \) for the probability on \( \Theta \times X^\infty \) defined by

\[
P_\mu(A \times B) = \int_A Q_\theta(B) \, \mu(d\theta)
\]

where \( A \) is Borel in \( \Theta \) and \( B \) is Borel in \( X^\infty \). As is easily verified, if \( H \) is a Borel subset of \( \Theta \times X^\infty \), then

\[(\text{A1}) \quad P_\mu(H) = \int_\Theta (\delta_\theta \times Q_\theta)(H) \, \mu(d\theta) .
\]

As usual, \( \delta_\theta \) is point mass at \( \theta \).

Fix a version \( \mu_n = \mu_n(d\phi|x_1, \ldots, x_n) \) of the \( P_\mu \)-law of \( \theta \) given \( X_1 = x_1, \ldots, X_n = x_n \). Here, as elsewhere, \( \theta \) denotes the coordinate map

\[
(\theta, x_1, x_2, \ldots) \mapsto \theta .
\]

The posterior \( \mu_n \) may be considered as a function on \( \Theta \times X^\infty \), or just \( X^\infty \), or even \( X^N \), according to convenience.

Consistency of \( (\theta, \mu_n) \) depends not only on \( \mu \), but also on the choice of the version \( \mu_n \), which is well-defined only a.e. An artificial example may clarify the point.

**Example A1.** Let \( \Theta = [0,1] \), \( X = [0,1] \), \( Q_\theta = \delta_\theta \), and let \( \mu \) be Lebesgue measure on \( \Theta \). Let

\[
\mu_n(d\phi|x_1, \ldots, x_n) = \delta_{x_1}
\]
unless \( x_1 = 0 \) in which case

\[
\mu_n \{d\phi | x_1, \ldots, x_n \} = \delta_{\frac{1}{2}}.
\]

This \( \mu_n \) is a (stupid) version of the law of \( \theta \) given \( X_1, \ldots, X_n \) because

\[
P_{\mu} \{X_1 = 0, \ldots, X_n = 0\} = 0.
\]

For this version, \( \mu_n \) is consistent except at \( \theta = 0 \) when \( \mu_n \rightarrow \delta_{\frac{1}{2}} \).

Bayes' estimates of an unknown probability on the line provide a less artificial example where \( \{Q_\theta\} \) is not dominated by a \( \sigma \)-finite measure. Then, the posterior can be changed on a set of measure zero, so consistency is determined for \( \mu \) almost all \( \theta \), but not for all \( \theta \).

The next topic is the weak-star topology for probability measures on \( \Theta \).

Recall that \( \Theta \) is a separable metric space by assumption, but \( \Theta \) need not be complete, or compact, or locally compact: for example, \( \Theta \) might be the irrational numbers in the unit interval. Let \( C \) be the set of bounded, continuous functions on \( \Theta \). If \( \alpha_n \) and \( \alpha \) are probabilities on \( \Theta \), then \( \alpha_n \rightarrow \alpha \) weak star if and only if \( \int fd\alpha_n \rightarrow \int fd\alpha \) for all \( f \in C \).

A technical problem is that \( C \) will not usually be separable. However, it is enough to consider only the uniformly continuous \( f \in C \): to be specific, let \( \Theta_0 \) be a countable dense subset of \( \Theta \). For each \( \theta_0 \in \Theta_0 \) and nonnegative rationals \( r < s \) and arbitrary rationals \( u, v \), define a function \( f = f_{r,s,u,v} \) on \( \Theta \) as follows:

\[
f(\theta) = u \text{ for } \theta \text{ with } \rho(\theta_0, \theta) \leq r
\]

\[
eq v \text{ for } \theta \text{ with } \rho(\theta_0, \theta) > s
\]

while

\[
f(\theta) = u \frac{s - \rho(\theta_0, \theta)}{s - r} + v \frac{\rho(\theta_0, \theta) - r}{s - r}
\]

22
for \( \theta \) with \( r \leq \rho(\theta_0, \theta) \leq s \). Here, \( \rho \) is the given metric on \( \Theta \). Clearly, \( f \) is bounded and uniformly continuous. Let \( C_0 = \{f\} \), a countable collection. Let \( C_\vee = \{f_1 \vee \ldots \vee f_k : \text{for } f_i \in C_0 \text{ and } k = 1,2,\ldots\} \). Let \( C_\wedge = \{f_1 \wedge \ldots \wedge f_k : \text{for } f_i \in C_0 \text{ and } k = 1,2,\ldots\} \). Here \( \vee \) and \( \wedge \) are pointwise max and min, respectively. Now \( C_\vee \) and \( C_\wedge \) are countable subsets of \( C \), and pointwise dense, as follows.

**Lemma A1.** If \( f \in C \), there is a sequence \( f_k \in C_\vee \) with \( f_k \uparrow f \) pointwise; and another sequence \( g_k \in C_\wedge \) with \( g_k \downarrow C \) pointwise.

**Corollary A1.** If \( \int f d\alpha_n \to \int f d\alpha \) for all \( f \in C_\vee \cup C_\wedge \), then \( \alpha_n \to \alpha \) weak star.

We can now prove Doob's theorem in our context.

**Corollary A2.** \( \mu_n \to \delta_\theta \) weak star, \( P_\mu \)-almost surely.

**Proof.** Fix an \( f \in C_\vee \cup C_\wedge \). By Corollary A1, it is enough to prove that \( \int f d\mu_n \to f(\theta) \) a.e. \( P_\mu \). But \( \int f d\mu_n \) is a version of \( E[f(\theta) | X_1 \ldots X_n] \), so \( \int f d\mu_n \to E[f(\theta) | X_1, X_2, \ldots] \) a.e., by martingales. The last step is to prove identifiability: that \( \theta \) can be computed measurably from all the \( X \)'s, a.e. \( P_\mu \).

The identifiability argument will only be sketched: it is at this point that "1-1 ness" matters. Since \( \theta \to Q_\theta \) is 1-1 and Borel, so is the inverse: see Kurtowski (1958, Section 35.5). So we only need to compute \( Q_\theta \) from \( X_1, X_2, \ldots \). Using Corollary A1 again, we only need to compute \( \int g dQ_\theta \) for \( g \in C_\wedge \cup C_\vee \). But

\[
(A2) \quad \int g dQ_\theta = \lim_{n \to \infty} \frac{1}{n}[g(X_1) + \ldots + g(X_n)]
\]
a.e., by the law of large numbers.

In a bit more detail, let \( H_\theta \) be the subset of \( \Theta \times X^\infty \) where (A2) holds. Then \( H_\theta \) is Borel, and \( \delta_\theta \times Q_\theta(H_\theta) = 1 \) for all \( \theta \), by the law of large numbers. So \( P_\mu(H_\theta) = 1 \) by (A1). \( \square \)
**Lemma A2.** A Borel set is homeomorphic to a dense Borel subset of a compact metric set.

**Proof.** This standard fact can be proved by appealing to Urnsohn's lemma (Kurtowski, 1958, page 119) and then (Kurtowski, 1958, p. 397).

**Corollary A3.** \( \theta \to Q_\theta^\infty \) is weak-star continuous.

**Proof.** Let \( f \) be bounded continuous on \( X^\infty \), and

\[
g(\theta) = \int_{X^\infty} f dQ_\theta^\infty.
\]

We have to show that \( g \) is continuous.

**Case 1.** \( X \) is compact. Then \( f \) can be approximated by finite linear combinations of functions \( \prod_{i=1}^k f_i(x_i) \), with \( f_i \) bounded continuous on \( X \).

**The general case.** Embed \( X \) as a dense Borel subset of the compact metric space \( (\tilde{X}, \tilde{\rho}) \) and metrize \( X \) by \( \rho : \) see Lemma A2. The bounded, \( \rho \)-uniformly continuous function on \( X \) are those which extend to continuous functions on \( \tilde{X} \). Now use Corollary A1 to reduce the general case to Case 1, the point being that a function \( g \) which is both upper and lower semicontinuous is continuous.

The next topic is "weak-star merging". Let \( \alpha_n \) and \( \beta_n \) be probabilities on \( \Theta \); the sequences \( \{\alpha_n\} \) and \( \{\beta_n\} \) merge weak-star if and only if \( R(\alpha_n, \beta_n) \to 0 \) for all continuous functions \( R \) which satisfy \( R(\alpha, \alpha) = 0 \). Some possible choices for \( R \):

\[
R(\alpha, \beta) = \int f d\alpha - \int f d\beta \quad \text{for a fixed bounded continuous } f,
\]

\[
R(\alpha, \beta) = \lambda(\alpha, \beta) \quad \text{with } \lambda \text{ a metric for the weak star topology.}
\]

24
This is the "predictive" distribution of \( \mathcal{X}_{n+1}, \mathcal{X}_{n+2}, \ldots \), given \( \mathcal{X}_1, \ldots, \mathcal{X}_n \). Of course, \( P_{\mu_n} = P_{\mu_n}(\cdot|\mathcal{X}_1, \ldots, \mathcal{X}_n) \) is a probability on \( \mathcal{X}^\infty \), depending on the first \( n \) data points.

We will prove somewhat more than Theorem C. To state the result, let \( G \subset \Theta \times \mathcal{X}^\infty \) be the set of \( (\theta, \mathcal{X}_1, \mathcal{X}_2, \ldots) \) such that \( \mu_n(d\theta|\mathcal{X}_1, \ldots, \mathcal{X}_n) \to \delta_{\theta} \) weak star. Notice that \( G \) depends on \( \mu_n \), and \( \mu_n \) is consistent if and only if \( (\delta_{\theta} \times Q_{\theta})(G) = 1 \) for all \( \theta \).

**Theorem A1.** Let \( \theta \to Q_{\theta} \) be 1-1 and Borel. Fix a prior \( \mu \) on \( \Theta \) and a version \( \mu_n \) of the posterior. The following three conditions are equivalent.

(i) \( \mu_n \) is consistent.

(ii) \( P_{\mu}(G) = 1 \) for all probabilities \( \nu \) on \( \Theta \).

(iii) \( \mu_n(d\theta|\mathcal{X}_1, \ldots, \mathcal{X}_n) \) and \( \nu_n(d\theta|\mathcal{X}_1, \ldots, \mathcal{X}_n) \) merge weak star as \( n \to \infty \), with \( P_{\nu} \) probability one, for all probabilities \( \nu \) on \( \Theta \).

Suppose that \( \theta \to Q_{\theta} \) is continuous and has a continuous inverse. Then (i) is also equivalent to

(iv) \( P_{\mu_n} \) and \( P_{\nu_n} \) merge weak star as \( n \to \infty \), with \( P_{\nu} \) probability one, for all probabilities \( \nu \) on \( \Theta \).

**Proof.** Fix \( \nu \). We will prove that (i) \( \Rightarrow \) (ii) \( \Rightarrow \) (iii). The first implication is trivial: from (A1),

\[
P_{\nu}(G) = \int (\delta_{\theta} \times Q_{\theta})(G) \nu(d\theta)
\]

and \( (\delta_{\theta} \times Q_{\theta})(G) = 1 \) for all \( \theta \) from the definition of consistency. For the next implication, suppose (ii) for \( \nu \). We will argue that a.e. \( P_{\nu} \),

\[
\mu_n \to \delta_{\theta},
\]

\[
\nu_n \to \delta_{\theta}.
\]
The notation may be a bit confusing: \( \theta \) is being used for the coordinate function \((\theta, x_1, x_2, \ldots) \rightarrow \theta \) and its value \( \theta \). It would follow from (A5-6) that \( R(u_n, v_n) \rightarrow 0 \).

Now (A5) holds on \( \mathcal{G} \), which has \( P_\mathcal{G} \)-probability 1, by assumption. And (A6) holds a.e. \( P_\mathcal{V} \) by Doob's theorem, Corollary A2 above. Thus (i) \( \Rightarrow \) (ii) \( \Rightarrow \) (iii). Clearly, (iii) \( \Rightarrow \) (i): take \( \nu \) to be point mass and \( R \) as in (A3).

Now assume that \( \theta \rightarrow Q_\theta \) is 1-1 and continuous. We will argue (i) \( \Rightarrow \) (iv).

Indeed, suppose (i). We will show that a.e. \( P_\mathcal{V} \),

\[(A7) \quad P_{\mu n} \rightarrow Q_\theta^\infty,\]

\[(A8) \quad P_{\nu n} \rightarrow Q_\theta^\infty.\]

For (A7), let \( f \) be bounded and continuous on \( X^\infty \). By definition,

\[\int_{X^\infty} f dP_{\mu n} = \int_{\Theta} \int_{X^\infty} f dQ_\theta^\infty \mu_n(\theta).\]

Now \( g: \theta \rightarrow \int f dQ_\theta^\infty \) is a bounded, continuous function on \( \Theta \), in view of Corollary A3. In view of (i), \( \int g d\mu_n \rightarrow g(\theta) \) a.e. \( Q_\theta^\infty \) for all \( \theta \). Thus, \( \int g d\mu_n \rightarrow g(\theta) \) a.e. \( P_\mathcal{V} \). This proves (A7) via Corollary A1, letting \( f \) run through the analogs of \( C_\mathcal{V} \) and \( C_\lambda \) on \( X^\infty \). For (A8), by Doob's theorem, \( \int f dP_\nu \rightarrow g(\theta) \) a.e. \( P_\mathcal{V} \), and the rest of the argument is the same. Thus, (i) \( \Rightarrow \) (iv), in the presence of continuity.

Finally, assume (iv), and let \( \nu = \delta_\theta \), so \( P_\nu = \delta_\theta \times Q_\theta^\infty \) and \( P_{\nu n} = Q_\theta^\infty \) a.e. \( Q_\theta^\infty \). Also, \( P_{\mu n} = \int Q_\theta^\infty \mu_n(\theta) = Q_\theta^\infty \) a.e. \( Q_\theta^\infty \) by (iv). Thus \( U(\mu_n) \rightarrow U(\delta_\theta) \) a.e. \( Q_\theta^\infty \) and \( \mu_n \rightarrow \delta_\theta \) a.e. \( Q_\theta^\infty \), by Proposition A1 below, where \( U \) is defined. \( \square \)

Remarks. (1) Consider a Bayesian with prior \( \mu \) who chooses \( \mu_n \) as the posterior. Consider a second Bayesian with the generic prior \( \nu \). Now \( \mathcal{G} \) is the set where the first Bayesian gets the parameters \( \theta \) right. Condition (ii) in the theorem is
that any second Bayesian will be sure that the first Bayesian gets $\theta$ right -- whatever $\theta$ may be. Condition (iii) is that any second Bayesian is sure that his posterior will merge weak star with that of the first Bayesian. (This is well defined, since $P_\nu$-null sets don't matter here: any version $\nu_n$ can be used.) Condition (iv) is that any second Bayesian is sure that his conditional opinion of the future given the past will merge weak star with that of the first Bayesian.

(2) Our definition of merging is consistent with the finest uniformity compatible with the weak star topology: see Kelley (1955, Chapter 6). If two sequences merge in our sense, it follows that they merge in any metric for the weak star topology. The implications (iii) $\Rightarrow$ (i) or (iv) $\Rightarrow$ (i) are valid with much weaker notions of merging. Essentially, only one $R$ is needed, provided $R$ vanishes on the diagonal and is positive off the diagonal: a metric will do very nicely.

(3) The continuity conditions are needed to conclude (iv) from (i). To see this, take $\Theta = [0,1]$ and $X = \{0,1\}$ and let

$$f(\theta) = \begin{cases} 
\frac{1}{2} & \text{for } \theta \in [0,\frac{1}{2}) \\
\theta & \text{for } \theta = \frac{1}{2} \\
\frac{1}{2} + \frac{1}{2} \theta & \text{for } \theta \in (\frac{1}{2},1] 
\end{cases}$$

Let $Q_\theta$ be $f(\theta)$-coin tossing. Let $\Theta$ be the uniform distribution on $X$. Straightforward analysis shows that $\mu$ is consistent, but the predictive distribution $P_{\mu n}$ does not merge with $P_{\nu n}$ when $\nu = \delta_{\frac{1}{2}}$. Lockett (1971) gives a similar example involving geometric variables.

(4) The continuity conditions are needed to conclude (i) from (iv). To see this, take $\Theta = [0,2\pi)$ and $X$ as the unit circle. Let $F(\theta) = e^{i\theta}$ map $\Theta$ onto $X$. This $F$ is 1-1 and continuous, but does not have a continuous inverse at $\xi = (1,0) \in X$. Define $Q_\theta = \delta_{F(\theta)}$. Let $\mu$ be uniform on $\Theta$. Define a posterior, maliciously, as
\[ \mu_n(\cdot | x_1 \ldots x_n) = \delta_{F^{-1}(x_1)} \text{ if } x_1 \neq \xi \]
\[ = \delta_{2^{n-1}/n} \text{ if } x_1 = \xi \]

So \((0, \mu_n)\) is not consistent. But \(P_{\mu_n} + Q_0^\infty\) a.e. \(Q_0^\infty\), for all \(\theta\), even \(\theta = 0\), and so merges with \(P_{\nu_n}\) a.e. \(P_{\nu}\), for any \(\nu\).

The following is required to complete the proof of Theorem A1. If \(S\) is a Borel set, we write \(\pi(S)\) for the set of probabilities on \(S\), endowed with the weak-star topology. The set \(\pi(S)\) is Borel too. Let \(S_0\) be a Borel subset of \(S\). Let \(P_0 = \{\mu | \mu \in \pi(S) \text{ and } \mu(S_0) = 1\}\). Then \(P_0\) is a Borel subset of \(\pi(S)\). Let \(M\) map \(P_0\) onto \(\pi(S_0)\) as follows: if \(\alpha \in P_0\), then \(M_\alpha\) is the restriction of \(\alpha\) to the Borel subsets of \(S_0\). Thus, \(M\) maps probabilities on \(S\) to probabilities on \(S_0\). We write \(M = M(S, S_0)\) to show the dependence on the spaces \(S\) and \(S_0\).

**Lemma A3.** \(M\) is a homeomorphism of \(P_0\) onto \((S_0)\).

**Proof.** Use Corollary A1.

Using Lemma A3 we can view \(X\) as a dense Borel subset of the compact metric space \((\tilde{X}, \rho)\); metrize \(X\) by \(\rho\). Then \(\pi(X)\) can be viewed as a Borel subset of the compact metric set \(\pi(\tilde{X})\). For \(\mu\) a probability on \(\Theta\), let

\[ U(\mu) = \int_\Theta Q_0^\infty \mu(d\theta) . \]

So \(U(\mu)\) is a probability on \(X^\infty\).

**Proposition.** \(U\) is a homeomorphism of \(\pi(\Theta)\) into \(\pi(X^\infty)\).

**Proof.** Let \(\bar{\Theta} = \pi(\tilde{X})\). Let \(\bar{U}\) map \(\pi(\bar{\Theta})\) into \(\pi(\tilde{X}^\infty)\) as follows:

\[ \bar{U}(\mu) = \int_{\bar{\Theta}} Q_0^\infty \mu(d\theta) . \]

Then \(\bar{U}\) is continuous, and 1-1 by de Finetti's theorem (Diaconis and Freedman
Since $\pi(\mathcal{G})$ is compact, $\mathcal{U}^{-1}$ is continuous. Next, suppose $X$ is a Borel subset of $\tilde{X}$, $\Theta$ is a Borel subset of $\pi(X)$, and $Q_\Theta = \Theta$. In this setting, $U(\mu) = \int_\Theta \tilde{\Theta}^\infty \mu(d\Theta)$. Now apply Lemma A2 twice, with $M_1 = M(\tilde{\Theta}, \Theta)$ and $M_2 = M(\tilde{X}^\infty, X^\infty)$. Then $U = M_2 \circ \tilde{U} \circ M_1^{-1}$ is a homeomorphism on $\pi(\Theta)$. To see that the composition makes sense, let $\mu \in \pi(\Theta)$, and $\nu = M_1^{-1} \mu$; then $\nu \in \pi(\tilde{\Theta})$ and $\nu(\Theta) = 1$; so $\tilde{U}(\nu)(X^\infty) = 1$ because $\Theta(X) = 1$ for $\Theta \in \Theta$.

The general case is almost immediate: $\Theta \rightarrow Q_\Theta$ being continuous and 1-1 on $\Theta$, the image $\tilde{\Theta}$ under this mapping of $\Theta$ is a Borel subset of $\pi(X)$. Apply the previous argument with $\tilde{\Theta}$ in place of $\Theta$. Let $M_0$ map $\pi(\Theta)$ onto $\pi(\tilde{\Theta})$ by the recipe $(M_0 \mu)(A) = \mu(\Theta : Q_\Theta \in A)$ for Borel $A \subset \tilde{\Theta}$. This $M_0$ is a homeomorphism because $\Theta \rightarrow Q_\Theta$ is. Now $U = M_2 \circ \tilde{U} \circ M_1^{-1} \circ M_0$ is a homeomorphism too. □
APPENDIX B. The derivative of the posterior with respect to the prior.

Proceeding heuristically for a moment, the derivative of the ratio
\[ T(\mu) = N(\mu)/D(\mu) \] is
\[ \dot{T}_\mu = [D(\mu)\dot{N}_\mu - N(\mu)\dot{D}_\mu]/D(\mu)^2. \]

Now \( N(\mu) \) is linear: \( N(\mu+\delta) = N(\mu) + N(\delta) \). So \( \dot{N}_\mu = N(\ast) \). Likewise, \( D(\mu) \) is a linear functional, so \( \dot{D}_\mu = D(\ast) \). For example, in our sense, the derivative of the linear function \( \phi: x \rightarrow 3x \) is just \( \phi \), because \( \phi(x+h) = \phi(x) + \phi(h) \).

The upshot is \( \dot{T}_\mu = R_\mu \), where
\[ R_\mu(\ast) = \frac{N(\ast)}{D(\mu)} - \frac{N(\mu)}{D(\mu)^2}D(\ast). \]

This is part a) of Theorem E. For a rigorous proof, we must show that for any \( \delta \) with signed mass \( 0 \),
\[ T(\mu+\delta) = T(\mu) + R_\mu(\delta) + o(||\delta||). \]

The difference \( T(\mu+\delta) - T(\mu) - R_\mu(\delta) \) is easily seen to equal
\[ -\left( \frac{N(\delta)}{D(\mu)} - \frac{D(\delta)}{D(\mu)} \right) \left( \frac{\frac{D(\delta)}{D(\mu)}}{\frac{D(\mu)}{D(\mu+\delta)}} \right). \]

The norm of this is smaller than
\[ (B1) \quad \left| \frac{D(\delta)}{D(\mu)D(\mu+\delta)} \right| \left( ||N(\delta)|| + ||N(\mu)|| \cdot \frac{D(\delta)}{D(\mu)} \right). \]

Let \( C = \sup_{\theta} f(x|\theta) < \infty \) by assumption. Then \( |D(\delta)| \leq C||\delta|| \) and \( ||N(\delta)|| \leq C||\delta|| \). Further, \( D(\mu+\delta) \) tends to \( D(\mu) \) as \( \delta \) tends to \( 0 \). It follows that the bound (B1) is smaller than \( C(\mu) ||\delta||^2 \) for \( ||\delta|| \) small. This completes the proof of part (a).

To prove part (b) of Theorem E, fix \( x \) and write \( f \) for \( f(x|\theta) \). Let \( \tilde{f} = f/D(\mu) \) so \( \int \tilde{f}du = 1 \). Then
\[ d\hat{T}_\mu(\delta) = \hat{F}d\delta - (\int \hat{F}d\delta)\hat{F}d\mu. \]

Choose \( \sigma \) to dominate both \( |\delta| \) and \( \mu \), e.g., \( \sigma = |\delta| + \mu \). Then \( d\delta = \delta d\sigma \) and \( d\mu = \mu d\sigma \) and

\[ d\hat{T}_\mu(\delta) = [\delta - (\int \hat{F}d\delta)\mu]\hat{F}d\sigma. \]

We must show

\[ ||\hat{T}_\mu|| \leq ||\delta|| \cdot \sup_\delta \hat{F}. \]

Indeed,

\[ \int \hat{F}d\sigma = \int \hat{F}d\delta \quad \text{and} \quad \int \mu \hat{F}d\sigma = \int \hat{F}d\mu = 1. \]

Thus, \( \hat{T}_\mu(\delta) \) has signed mass 0, namely, \( \hat{T}_\mu(\delta)[\Theta] = \int \hat{F}d\sigma - \int \hat{F}d\sigma = 0 \). It follows that

\[ ||\hat{T}_\mu(\delta)|| = 2\int [\delta - (\int \hat{F}d\delta)\mu]^+ \hat{F}d\sigma = \int [\delta - (\int \hat{F}d\delta)\mu]d\sigma. \]

Assume without loss of generality that \( \int \hat{F}d\delta > 0 \). For \( a > 0 \), clearly \( (x-a)^+ \leq x^+ \). So

\[ ||\hat{T}_\mu(\delta)|| \leq 2\int \delta^+ \hat{F}d\sigma \leq 2(\int \delta^+ d\sigma)(\sup F). \]

But

\[ \int \delta^+ d\sigma = \frac{1}{2}||\delta|| \cdot d\sigma = \frac{1}{2}||\delta|| \]

because \( \int \delta d\sigma = \int d\delta = 0 \). This completes the proof of (B2), and hence part b) of Theorem E.

To prove part c) of Theorem E, we must show that for every \( \varepsilon > 0 \) there is signed measure \( \delta \) with signed mass 0 and total mass 1, satisfying

\[ ||\hat{T}_\mu(\delta)|| \geq (1-\varepsilon) \sup_0 \hat{F}. \]
Choose $\theta_0$ with $\mu(\theta_0) = 0$ and $f(\theta_0) > (1-\varepsilon)\sup_0 f$. Let $\delta = \frac{1}{2}(\delta_{\theta_0} - \mu)$, where $\delta_{\theta_0}$ is point mass at $\theta_0$. Let $\sigma = \mu + \delta$. Then the right-most expression in (B3) can be evaluated, and is $f(\theta_0)$, so

$$||\hat{T}_\mu(\delta)|| = f(\theta_0) \geq (1-\varepsilon)\sup_0 f.$$ □

Remarks.

(1) Ordinarily, $\sup_0 f = \sup f$, so the theorem determines $||\hat{T}_\mu||$. If e.g. $\mu(\theta_0) > 0$ and $f(\theta_0) > \sup\{f(\theta) : \theta \neq \theta_0\}$, then $\sup f > \sup_0 f$. In this case, $||\hat{T}_\mu||$ is hard to compute. However, it can be shown that $||\hat{T}_\mu|| = \sup||\hat{T}_\mu(\delta)||$ where $\delta = \frac{1}{2}\delta_{\theta_1} - \frac{1}{2}\delta_{\theta_2}$ and $\theta_1 \neq \theta_2$ vary over $\Theta$.

(2) If $\delta$ is required to be absolutely continuous with respect to $\mu$, a similar argument shows that $||\hat{T}_\mu(\delta)|| \leq \mu$-ess. $\sup f$, the inequality being sharp if e.g. $\mu$ is continuous.

(3) We have chosen to differentiate in the set of signed measures. Our strong derivative is called the "Frechet derivative." Another standard way of perturbing $\mu$ is to consider the mixture $(1-\varepsilon)\mu + \varepsilon \nu$ for some probability $\nu$ as $\varepsilon$ tends to zero: the "Gateaux derivative." The mixture can be written as $\mu + \varepsilon(\nu - \mu)$ and the notions of derivative coincide -- for bounded likelihood functions.

(4) Similar computations can be carried out for Bayes' rules. If $\Theta$ is a real parameter, the mapping $M$:

$$\mu \rightarrow \frac{\int f(x|\theta) \mu(d\theta)}{\int f(x|\theta) \mu(d\theta)}$$

has derivative

$$M_\mu(\cdot) = \frac{N_1(\cdot)}{D(\mu)} - \frac{N_1(\mu)}{D(\mu)^2} D(\cdot)$$

32
where $N_1(\delta) = \int \delta f(x|\theta) \delta(d\theta)$ and $D(\delta) = \int f(x|\theta) \mu(d\theta)$ as before. The norm of $\tilde{N}_\mu$ is computed as follows: let $c = N_1(\mu)/D(\mu)$. This is the Bayes' rule based on $\mu$. Let $g(\theta) = (\theta - c) \hat{f}(x|\theta)$. Define range $g = \sup g - \inf g$. Then

$$||\tilde{N}_\mu|| = \frac{1}{2} \text{ range } g.$$ 

(5) Theorem E assumes a dominated family. In the undominated case, the derivative need not exist in our strong sense. The difficulties can already be seen in the following simple example: take $X$ and $\Theta$ to be the real line mod 10. Let $Q_\delta\{x\} = \frac{1}{4}$ if $x = \Theta \pm 1$; suppose the prior $\mu$ for $\Theta$ has continuous density $f$ on $\Theta$. The posterior for $\Theta$ given $x$ is supported at $x \pm 1$ with mass $\frac{f(x-1)}{f(x-1) + f(x+1)}$ if $\Theta = x-1$ and $\frac{f(x+1)}{f(x-1) + f(x+1)}$ if $\Theta = x+1$.

(B4) \[
T(\mu|\Theta) = \begin{cases} 
\frac{f(x-1)}{f(x-1) + f(x+1)} & \text{if } \Theta = x-1 \\
\frac{f(x+1)}{f(x-1) + f(x+1)} & \text{if } \Theta = x+1 
\end{cases}
\]

The map $T$ is norm continuous at no $x$. To see this, consider a sequence of continuous prior densities $f_n$ converging to $f$ in variation distance but pointwise at no point. More specifically, let $s_n = 1 + \frac{1}{2} + \ldots + \frac{1}{n}$. Let $g_n$ on the line vanish to the left of $s_n - \frac{2}{n}$; increase linearly to the value 1 at $s_n$; decrease linearly to zero at $s_n + \frac{2}{n}$ and vanish to the right of that value.

Wrap $g_n$ around the line mod 10; let $f_n$ be the sum of $f$ and the wrapped $g_n$, normalized to be a density. Clearly $f_n \to f$ in $L_1$, but $f_n(x) \neq f(x)$ for no $x$. The argument is only sketched. Fix a real number $\Theta$ with $0 \leq \Theta < 10$; for any integer $k$, the real number $k + \Theta$ wraps to the same point $\Theta$ in $\Theta$. For infinitely many $n$, for some $k = k_n$, we have $s_n \leq k + \Theta < s_{n+1}$. Then $g_n(k+\Theta) \geq \frac{1}{2}$, because $s_{n+1} - s_n < 1/n$. For such $n$, we have $f_n(\Theta) > [f(\Theta) + \frac{1}{2} \frac{n}{n+2}].$
because \( \int g_n = 2/n \). Since \( g_n \) has only one bump, this can be either at \( x + 1 \) or \( x - 1 \), but not both. Thus, for any \( x \), for infinitely many \( n \), we have both the following relations

\[
\begin{align*}
f_n(x+1) &> [f(x+1) + \frac{1}{x+1} \cdot \frac{n}{n+2}] \\
f_n(x-1) &= f(x-1) \cdot \frac{n}{n+2}.
\end{align*}
\]

So \( f_n \) does not converge pointwise, and, from (84), the map \( T \) is not continuous.

In this example, the posterior is Gateaux differentiable: A derivative can be calculated by considering \( (1-\varepsilon)f + \varepsilon g \) as \( \varepsilon \) tends to zero. The Gateaux derivative exists quite generally, as we will show elsewhere.

(6) In the situation of Theorem E, the same result holds if the weak-star topology is used instead of the norm topology, provided \( f \) is bounded continuous in \( \Theta \).
References


Berger, J. (1982). The robust Bayesian viewpoint. Technical Report No. 82-9,
Department of Statistics, Purdue University.

Math. Statist. 41, 979-990.


38, 1907-1911.

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing infor-


Blackwell, D. and Ramamoorthi, R. V. (1982). A Bayes, but not classically suffi-


Addison Wesley, Reading, Massachusetts.

Breiman, L., LeCam, L. and Schwartz, L. (1964). Consistent estimates and zero-one


25, 617-657.

Dalal, S. R. (1979a). Dirichlet invariant processes and applications to non-
parametric estimation of symmetric distribution functions. Stochastic

Optimizing Methods in Statistics (ed. by J. S. Rustagi), 141-166. Academic
Press, New York.


