CONVERGENCE IN TOTAL VARIATION OF PREDICTIVE DISTRIBUTIONS:
FINITE HORIZON

BY

JOSEPH L. LOCKETT

TECHNICAL REPORT NO. 30
OCTOBER 6, 1971

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-24918

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CONVERGENCE IN TOTAL VARIATION OF PREDICTIVE DISTRIBUTIONS:
FINITE HORIZON

by

JOSEPH L. LOCKETT

TECHNICAL REPORT NO. 30
OCTOBER 6, 1971

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT GP-24918

Reproduction in whole or in part is permitted for
any purpose of the United States Government

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>PRELIMINARIES AND NOTATION</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>GENERAL CASE</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>DISCRETE CASE</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>BIBLIOGRAPHY</td>
<td>46</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

A notion of learning from experience is studied in this dissertation. A somewhat informal discussion of the topic has been given by Savage (1954). The concept is intermediate between consistency a posteriori and the merging of opinions in the sense of Blackwell and Dubins (1962).

Assume that $Q$ is a probability measure governing $(X_1, X_2, \ldots)$, absolutely continuous with respect to a probability measure $P$. Further, suppose that there exist regular conditional probabilities $Q_n$ (respectively $P_n$) for the indefinite future $(X_{n+1}, \ldots)$ given the $n$-observation past. Blackwell and Dubins show that almost surely $[Q, P_n - Q_n]$ tends to 0 in total variation. For instance, suppose that two individuals have prior distributions $p$ and $q$ for a parameter $\lambda$ and that, given $\lambda$, $(X_1, X_2, \ldots)$ are independent and identically distributed according to $P$. If $p(A) > 0$ if and only if $q(A) > 0$, it follows that for almost all $\lambda$ the two almost surely $[P = \int P_\lambda^\infty p(d\lambda)]$ will have opinions about every future event which become uniformly close. In particular, for almost all $\lambda$, the difference of their posterior distributions tends to 0 in the sense of total variation with probability 1 $[P_\lambda^\infty]$. Typically, when this occurs both posterior distributions are tending weakly to point mass at $\lambda$, that is,
the (quadratic loss) Bayes estimates of bounded, continuous functions $f$ tend to $f(\lambda)$ simultaneously for all $f$. The latter phenomenon is called consistency a posteriori.

Situations in which consistency implies merging are special. In view of Theorem 1 of Freedman (1963) there are plausible models in which two prior distributions are mutually singular, yet both lead (almost surely) to consistent Bayes estimates for exactly the same set of parameters. Thus the restriction to absolute continuity in the theorem of Blackwell and Dubins is seen to be a severe one. However, when two prior distributions are mutually singular, the resulting posterior distributions are also, so without that restriction no result of the type they describe is possible. Moreover, it should be emphasized that their arguments are too general to permit identification of the exceptional set on which merging does not occur.

Earlier, Doob (1949) obtained a very general result on the consistency of Bayes estimates: roughly speaking, if consistent estimates of a parameter $\lambda$ exist, posterior expectations of a fixed function $f$ tend to $f(\lambda)$ a.s. $[P^\infty_\lambda]$ for almost all values of the parameter under the prior probability. In Chapter 3 an argument similar to Doob's is exploited to extend his conclusions to consistency a posteriori. These results are adequate if, unlike subjective Bayesians, one believes the parameter is chosen by nature according to the prior. Otherwise it is of interest to identify the exceptional set. As is the case with merging, the arguments to which I have alluded are too general to permit that identification.
Freedman has specialized the study of consistency a posteriori to the non-parametric case in which the observations are (independent and identically distributed and) have countable range. In part his work is an extension of the work of Schwartz (1965). He has given (1963) criteria that consistency obtain at particular parameter values (that is, pointwise), and also criteria which render Doob's exceptional set empty. Only in case the range of the data is finite has it been possible without restrictions on the prior to specify precisely the set of parameters on which there is consistency. The exceptional set on which there is not is, in a certain topological sense, usually a "large" null set (Freedman (1965)). Section 5 of the 1963 paper describes other anomalous behavior of posterior distributions. Kiefer (in Watts (1968), p. 140) noted the very difficult nature of the problem of consistency a posteriori and observed that discovery of the essential nature of Bayes estimates might provide the "beginning of a good non-parametric theory".

Fabius (1964) extended some of Freedman's results to the case in which the parameter space is the set of all Borel probabilities on the unit interval. In work mentioned previously, Schwartz (1965) obtained very general pointwise results on the consistency of Bayes estimates of individual functions of parameters.

I have noted that the restriction to absolute continuity which is required to assure the conclusion of Blackwell and Dubins is a severe one, and that once the restriction is removed it is no longer possible to inquire into merging of opinions on the indefinite future. Hence, this dissertation is limited to consideration (for each fixed k) of
conditional distributions of the finite future \((X_{n+1}, \ldots, X_{n+k})\) given \((X_1, \ldots, X_n)\). Also, it appears both difficult and uninteresting to discover whether two opinions merge at false conclusions. Consequently, I study convergence in total variation of one's opinion on the finite future to the true predictive distribution—a phenomenon which may be called "coming to believe the truth about the finite future". Freedman informally introduced the notion.

The fact that consistency a posteriori need not imply convergence, as I mentioned in the opening paragraph, is illustrated in Chapter 3 by an example, and a continuity condition assuring this implication is discussed. In Chapter 4, the study is narrowed to the case of discrete observations. An example (4.1) indicates that, when the prior distribution is also discrete, one comes to believe the truth at an exponential rate. However, if there is no atom of probability on the true parameter, the rate of convergence will be at best algebraic. An elaboration of techniques used by Freedman (1963) is employed to prove a general theorem (4.3) to the effect that under conditions on the prior distribution, convergence occurs, almost surely, at a rate at least \(n^{-1/4+\delta}\), where \(\delta\) is positive and less than \(\frac{1}{4}\), and may depend on the true distribution. In general, the thinner the tail of the distribution the smaller \(\delta\) will be. Examples are given which indicate that, for the geometric distribution, any \(\delta > 0\) will suffice (Example 4.2), but that for a distribution whose tail decays algebraically \((P(|i|) \leq i^{-\alpha} \text{ for some } \alpha > 1)\) a value of \(\delta\) bounded away from 0 is required (Example 4.3). Example 4.4 illustrates the point at which
improvement of technique is required to improve the result of Theorem 4.1. The study of rates in conjunction with consistency a posteriori, let alone opinions on the finite future, is new insofar as it concerns total variation. Berk (1970) has studied related but different problems.

Theorem 4.4 establishes that, almost surely, opinions based on the empirical distribution tend to the truth at the rate \((\log \log n/n)^{1/2}\). Theorem 4.5 is a pointwise result showing that, if the prior distribution puts sufficient probability in neighborhoods of the maximum likelihood estimate, differences of opinions based on the empirical distribution and those derived from the (quadratic loss) Bayes estimate tend to 0 (in total variation) a.s. at the rate \((\log n/n)^{1/2}\).

The theorems of this dissertation refer to posterior and predictive distributions and to functionals of them, but not to expectations of those functionals. The study of rates, then, is of the limits superior of sequences of random variables.
CHAPTER 2

PRELIMINARIES AND NOTATION

Throughout this dissertation, $\mathcal{X}$ denotes a Borel space (Breiman (1968), p. 401) and $\mathcal{I}$ a denumerable set; the $\sigma$-field of Borel subsets of $\mathcal{X}$ (or $\mathcal{I}$) is denoted $\mathcal{A}$. These sets will be the range of observations $(X_1, X_2, \ldots)$, a sequence of $\mathcal{X}$-(or $\mathcal{I}$ as appropriate) valued random variables on the measurable space $(\Omega, \mathcal{F})$. When required, $\mathcal{X}^n$ will denote $\bigtimes_{1}^{n} \mathcal{X}$ for $n \leq \infty$. The family of sub $\delta$-fields of $\mathcal{F}$ induced by $(X_{m}, \ldots, X_{n})$ will be denoted $\left(\mathcal{F}_{n}^{m}\right)$. (When $m = 1$ it will be omitted.) $\mathcal{A}^{m}_{n}$ will denote the associated $\sigma$-field of $\mathcal{X}^n$.

It will be necessary to consider a set of probabilities on the space $(\Omega, \mathcal{F})$. These are indicated as $(P_{\lambda} | \lambda \in \Lambda)$ where $\Lambda$, the index set, is a locally compact, separable metric space, and $\mathcal{B}$ denotes the $\sigma$-field of Borel subsets of $\Lambda$. For each $A \in \mathcal{F}$ the function $P(A) : \lambda \rightarrow P_{\lambda}(A)$ is $\mathcal{B}$-measurable. Usually $(\Omega, \mathcal{F})$ is taken to be $\mathcal{F}^\infty$ with the product Borel structure, and $(X_1, X_2, \ldots)$ are coordinate variables.

A probability on $(\Lambda, \mathcal{B})$ will be denoted $\mu$ and the associated posterior probability given $(X_1, \ldots, X_n)$ will be written $\mu_{n, \omega}$. The function $\mu_{n, \omega}$ on $(\Omega \times \mathcal{B})$ to $[0,1]$ is called a posterior probability given $(X_1(\omega), \ldots, X_n(\omega))$ if
i) for every \( \omega \in \Omega \), the function \( \mu_{n,\omega} : B \rightarrow \mu_{n,\omega}(B) \) is a
probability;

ii) for every \( B \in \mathcal{B} \), the function \( \mu_{n,\omega}(B) : \omega \rightarrow \mu_{n,\omega}(B) \) is
\( \mathcal{F}_n \)-measurable;

iii) for every \( B \in \mathcal{B} \), \( \mu_{n,\omega}(B) \) is a version of \( E(I_B|X_1(\omega),\ldots,X_n(\omega)) \).
(\( I_B \) is the indicator function of the set \( B \). The underlying
probability is \( P = \int P_\lambda \mu(d\lambda) \).)

The restriction of \( P_\lambda \) to \( \mathcal{A}_{n+k}^{m+1} \) will (in an abuse of notation)
be written \( P_\lambda^k \). No confusion should result if, when \( k = 1 \) it is
suppressed. The \( k \)-step conditional distribution of the future given
\((X_1,\ldots,X_n)\) is written \( P_n^k \) and is defined as follows:

\[
P_n^k(A) = P(A|\mathcal{F}_n) = \int_{\Lambda} P_\lambda(A) \mu_{n,\omega}(d\lambda) \quad \text{for} \quad A \in \mathcal{A}_{n+k}^{n+1}
\]

(Again, when \( k = 1 \) it is omitted.) The total variation norm \( \|P - Q\| \)
is defined as

\[
\|P - Q\| = \sup_A |P(A) - Q(A)|
\]

where the supremum is taken over the appropriate \( \sigma \)-field. When \( P \) and
\( Q \) are conditional distributions, the norm is a function of \( \omega \).

To the space of Borel probabilities on \( \Lambda \) will be assigned the
relative topology arising from the weak* (\( w^* \)) topology on the space of
Borel probabilities on \( \Lambda^* \), the one-point compactification of \( \Lambda \). Thus,
\( \mu_n \rightarrow \mu \) if and only if
\[
\int f \, d\mu_n \to \int f \, d\mu
\]

for every bounded, continuous function \( f \) on \( \Lambda \) which can be extended so as to be continuous on \( \Lambda^* \). The probability assigning point mass at \( \theta \in \Lambda \) will be written \( \delta_\theta \) and the statement "(\( \theta, \mu \)) is consistent" shall mean that for \( \mathbb{P}_\theta \)-almost all \( \omega \), \( \mu_n, \omega \xrightarrow{w*} \delta_\theta \).

When \( (\theta, \mu) \) is fixed on a discussion and is consistent, I will sometimes say there is consistency \textit{a posteriori}. Throughout this dissertation the symbol \( \theta \) will be reserved for that element of \( \Lambda \) which is the true value of the parameter.

In the material to follow, if \( A \) and \( B \) are sets, the complement of \( A \) is written \( A^c \) and the relative complement of \( A \) in \( B \) (\( B \cap A^c \)) is expressed \( B \setminus A \). Logarithms to base \( b \) are written \( \log_b \). If \( b = e \) it is omitted. By convention \( 0 \log 0 = 0 \).

From time to time the term "eventually" will appear and a statement such as "eventually, almost surely \([P], X_n < c_n\)" should be taken to mean that there is a measurable \( P \)-null set \( F \) and a random variable \( \nu \) defined on \( \Omega \setminus F \) such that, for \( \omega \in \Omega \setminus F \) and \( n \geq \nu(\omega) \) it is the case that \( X_n(\omega) < c_n \).
CHAPTER 3

GENERAL CASE

Results about consistency of Bayes estimates in the most general setting are provided by Doob (1949) and Schwartz (1965). A variant of Doob's result giving consistency in this setting is presented here.

**Theorem 3.1.** Assume that there exists a function \( g : \Lambda \times \mathbb{X}^\infty \rightarrow \Lambda \), measurable with respect to the trace of \( \mathcal{A}_\infty \) in \( \mathcal{B} \times \mathcal{A}_\infty \) such that \( g(\lambda, (x_1, \ldots)) = \lambda \) (a.s. \([P, \lambda]) for each \( \lambda \in \Lambda \). It follows that for \( \mu \)-almost all \( \lambda \in \Lambda \), \((\lambda, \mu)\) is consistent.

**Proof.** Since \( \Lambda^* \) is separable metric it is second countable, and \( C(\Lambda^*) \) the space of bounded continuous functions on \( \Lambda^* \), is separable because \( \Lambda^* \) is compact metric (Dunford and Schwartz, 1958, pp. 21, 340). Let \( \tilde{f}_1, \tilde{f}_2, \ldots \) be dense in \( C(\Lambda^*) \) and let \( f_i = \tilde{f}_i \) restricted to \( \Lambda \). Since each \( f_i \) is bounded it follows from the martingale convergence theorem that \( P(\mathbb{E}(f_i | \mathcal{A}_n) \rightarrow \mathbb{E}(f_i | \text{trace } \mathcal{A}_\infty) \) for all \( i \) \( = 1. \) Now \( \mathbb{E}(f_i(\lambda) | \mathcal{A}_n) = \mathbb{E}(f_i(g(x_1, \ldots)) | \mathcal{A}_n) \) since \( g = \lambda \) a.s. \([P = \int_P \mu(d\lambda)] \). Therefore, since \( g \) is trace \( \mathcal{A}_\infty \) measurable, \( \mathbb{E}(f_i(\lambda) | \mathcal{A}_n) \rightarrow f(g(x_1, \ldots)) \) a.s. \([P]. \) Equivalently, for \( P \)-almost-all \((\lambda, x_1, \ldots), P((\lambda | x_1, \ldots): \mathbb{E}(f_i | \mathcal{A}_n)(\lambda, x_1, \ldots) \rightarrow f_i(\theta) \) for all \( i \) \( = 1. \)
Thus (see Loeve (1963), p. 140), for \( \mu \)-almost-all \( \lambda \),

\[
P_\lambda(x_1, \ldots | \mathcal{A}_n)(\lambda, x_1, \ldots) \rightarrow f_1(\lambda) \quad \text{for all } \ 1) = 1.
\]

\(<>

The condition of Theorem 3.1 is not easy to verify. Therefore the following theorem, also due to Doob, is comforting for practice.

The proof given here is similar in spirit, although not identical, to that of Doob.

**Theorem 3.2.** If

i) \( \Lambda \) is a locally compact, separable metric space;

ii) \( \lambda_1 \neq \lambda_2 \) implies there exists a set \( A \in \mathcal{A} \) such that \( P_{\lambda_1}(A) \neq P_{\lambda_2}(A) \)

then there exists a function \( g : \Lambda \times \mathcal{K}^\infty \rightarrow \Lambda \), measurable with respect to the trace of \( \mathcal{A}_\infty \) in \( \mathcal{B} \times \mathcal{A}_\infty \) such that \( g(\lambda, x_1, \ldots) = \lambda \) a.s.

\[
[P = \int P_\lambda \mu(d\lambda)].
\]

**Proof.** Note first that \( \mathcal{K} \) is a Borel space, so that \( \mathcal{K}^\infty \) is as well. Moreover \( \Lambda^* \) (the one-point compactification of \( \Lambda \)) is a complete, separable, metric space and hence Borel. Because \( \mathcal{K}^\infty \) is Borel there exists a family of sets \( \{A_j\}_{j=1}^\infty \) such that, for each \( \lambda \), \( (P_{\lambda}(A_j))_{j=1}^\infty \) determines \( P_{\lambda}. \) For each fixed \( \lambda \)

\[
(\lim_{n \to \infty} \sum_{i=1}^n I_{[x_i \in A_j]})_{j=1}^\infty = (P_{\lambda}(A_j))_{j=1}^\infty \quad \text{a.s. } [P_{\lambda}];
\]

thus

\[
(\lim_{n \to \infty} \sum_{i=1}^n I_{[x_i \in A_j]})_{j=1}^\infty = (P_{\lambda}(A_j))_{j=1}^\infty \quad \text{a.s. } [P].
\]
Define $Q : A \times \mathcal{F}^\infty \rightarrow [0,1]^\infty$ by

$$Q(x_1, \ldots) = \lim_{n \to \infty} \sum_{i=1}^{n} I_{[x_1 \in A_i]}^{\infty}$$

Note that $Q$ is Borel measurable.

Let $M : A \rightarrow [0,1]^\infty$ have the value at $\lambda$, $(P_{\lambda}(A_j))_{j=1}^{\infty}$. Then by hypothesis $M$ is 1-1 and is Borel. Therefore, according to a theorem of Lusin (see Kuratowski (1958), p. 406) it is bimeasurable.

Let $N$ be defined as $M^{-1}$ on the range of $M$. For $a \in [0,1]^\infty$, $a \not\in \text{Range}(M)$, let $N(a) = \infty$ (the point at $\infty$ of $A^\ast$). Set $g = g(\lambda, x_1, \ldots) = N^\circ Q(\lambda, x_1, \ldots)$. Then $g$ is measurable

$(A \times \mathcal{F}^\infty, \mathcal{B} \times \mathcal{A}_\infty) \rightarrow (A, \mathcal{B})$ and equals the coordinate map $(\lambda, x_1, \ldots) \rightarrow \lambda$ as $[P]$. $hd$

One should recall that the posterior probability measure

$\mu_n, \omega$ as defined is a regular conditional probability for $\lambda$ given $X_1', \ldots, X_n$ so that the conditional expectations $E(f|A_n)$ will be calculated using it. Fabius (1964) has shown that a posterior probability always exists and, moreover, that it can be taken as invariant under all permutations of $\{X_1', \ldots, X_n\}$.

In view of my variant of Doob's theorem and the results of Freedman, it is interesting to inquire whether one will come to believe the truth about the finite future in the presence of consistency a posteriori. That this is not so without restriction may be seen from the following example.
Example 3.1.

Suppose $X$ is a geometric random variable whose probability distribution is indexed by $\lambda \in [0,1]$ as follows:

$$P_{\lambda}(X=i) = \rho(1-\rho)^{i-1} \quad i = 0, 1, 2, \ldots$$

where

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

Suppose that the true distribution of $X$ is $P_{1/2}$ and that the prior distribution of $\lambda$ is uniform, that is,

$$\mu(d\lambda) = d\lambda \quad \text{for } \lambda \in [0,1].$$

In this case the posterior distribution $\mu_n, \omega$ has explicit representation as

$$\mu_n, \omega(\lambda) = \frac{\int_{\frac{1}{2}}^{1} \left(\frac{\lambda}{2}\right)^n \left(1 - \frac{\lambda}{2}\right)^{S_n} d\lambda + \int_{\frac{1}{2}}^{1} \left(\frac{1}{2} + \frac{\lambda}{2}\right)^n \left(\frac{1}{2} - \frac{\lambda}{2}\right)^n d\lambda}{\int_{0}^{1/2} \left(\frac{\lambda}{2}\right)^n \left(1 - \frac{\lambda}{2}\right)^{S_n} d\lambda + \int_{1/2}^{1} \left(\frac{1}{2} + \frac{\lambda}{2}\right)^n \left(\frac{1}{2} - \frac{\lambda}{2}\right)^n d\lambda}$$

where $S_n = \sum_{i=1}^{n} X_i$ (for further discussion of this see Chapter 4).
Let
\[ A' = \{ u : u = \frac{v}{2}, v \in A \cap [0, \frac{1}{2}) \} \]
\[ A'' = \{ u : u = \frac{1}{2} + \frac{v}{2}, v \in A \cap (\frac{1}{2}, 1] \} \]
\[ d = \int_0^{1/4} \rho^n(1-\rho)^n \, d\rho + \int_{3/4}^1 \rho^n(1-\rho)^n \, d\rho \]

Then
\[ \mu_{n,\omega}(A) = d^{-1} \left[ \int_{A'} \rho^n(1-\rho)^n \, d\rho \right]_{A''} \]

It is apparent that, if \( A \) is any neighborhood of \( 1/2 \), \( \mu_{n,\omega}(A) \rightarrow 1 \) (a.s. \( P_{1/2} \)). However since mixtures of geometric random variables are not geometric (for example, no mixture of \( P_{1/4} \) and \( P_{3/4} \) will be \( P_{1/2} \)) it is the case that for no \( k \) does \( \| P_n^k - P_{1/2}^k \| \rightarrow 0 \) almost surely.

It is worthy of note that this example arises from an unnatural parametrization of the family \( P_\lambda \). Indeed in the natural setting no prior belief whatever is attached to the interval \((\frac{1}{4}, \frac{3}{4})\) so that it should not be surprising that one fails to come to believe the truth.

Theorem 3.3 shows that, in the absence of aberrations of the sort just described one can expect to come to believe the truth. In particular in what might be called the non-parametric case in which the distance \( d \) between \( \lambda \) and \( \theta \) satisfies \( d(\lambda, \theta) = \frac{1}{2} \int |\lambda - \theta| \, (dx) \) merging will always proceed from consistency.

In the next quite general theorem, the \( 1 \) in the superscript is intentionally emphasized, as the strong law of large numbers renders the conditions vacuous with \( l \) replaced by \( \infty \).
Theorem 3.3. If $\lambda \to \theta$ implies $\|P^1_\lambda - P^1_\theta\| \to 0$, then $(\theta, \mu)$ consistent implies $\|P^k_n - P^k_\theta\|_{n \to \infty} \to 0$ (a.s. $[P_\theta]$) for any $k$.

Proof. $\|P^k_n - P^k_\theta\| = \sup_{A \in \mathcal{A}_{n+k}} |P^k_n(A) - P^k_\theta(A)|$

$= \sup_{A \in \mathcal{A}_{n+k}} \int_{\Lambda} P^k_\lambda(A) \mu_n,\omega(d\lambda) - P^k_\theta(A)\|

\leq \sup_{A \in \mathcal{A}_{n+k}} \int_{\Lambda} |P^k_\lambda(A) - P^k_\theta(A)| \mu_n,\omega(d\lambda) + \mu_n,\omega(V^c)

for any $V \in \mathcal{B}$. Choose $\varepsilon > 0$. Since $\lambda \to \theta$ implies $\|P^1_\lambda - P^1_\theta\| \to 0$, there exists a neighborhood $V$ of $\theta$ such that $\|P^k_\lambda - P^k_\omega\| < \varepsilon/2k$ for $\lambda \in V$. Since $(\theta, \mu)$ is consistent, there exists an integer $N = N(\omega)$ such that for $n \geq N$, $\mu_n,\omega(V^c) < \varepsilon/2$ (a.s. $[P_\theta]$). According to Lemma 3.1 which follows,

$\|P^k_n - P^k_\theta\| \leq \sup_{\lambda \in V} \sup_{A \in \mathcal{A}_{n+k}} |P^k_\lambda(A) - P^k_\theta(A) + \mu_n,\omega(V^c)|$

$\leq 2k \sup_{\lambda \in V} \|P^1_\lambda - P^1_\theta\| + \mu_n,\omega(V^c)$

$< 2k \frac{\varepsilon}{2k} + \frac{\varepsilon}{2} = \varepsilon$ a.s. $[P_\theta]$ for $n \geq N$
Lemma 3.1. If $P$ and $Q$ are product probabilities on the same measurable space then

$$\|P^k - Q^k\| \leq 2k\|P^1 - Q^1\|$$

where $P^k$ and $Q^k$ are defined on the $k$-fold product measurable space.

Proof. Let $\tau = P + Q$, $p = dP/d\tau$, $q = dQ/d\tau$, and $A^k$ be the positive set of the Jordan-Hahn decomposition of $P^k - Q^k$. Then

$$\|P^k - Q^k\| = |P^k(A^k) - Q^k(A^k)| = \int P^{k-1}(A^k) dP(\omega) - \int Q^{k-1}(A^k) dQ(\omega)$$

where the integral is over the space $\Omega$ on whose measurable subsets $P$ and $Q$ are defined, and $A^k$ is the section of $A^k$ at the point $\omega$.

Thus

$$|P^k(A^k) - Q^k(A^k)| = |P^{k-1}(A^k) - Q^{k-1}(A^k)| dP(\omega) + \int Q^{k-1}(A^k) dP(\omega) - \int Q^{k-1}(A^k) dQ(\omega)$$

$$\leq |\int P^{k-1}(A^k) - Q^{k-1}(A^k) dP'(\omega)| + |\int Q^{k-1}(A^k) dP(\omega) - \int Q^{k-1}(A^k) dQ(\omega)|$$

$$= |\int [P^{k-1}(A^k) - Q^{k-1}(A^k)] dP(\omega)| + |\int Q^{k-1}(A^k) [p(\omega) - q(\omega)] d\tau(\omega)|$$

$$\leq \int |P^{k-1}(A^k) - Q^{k-1}(A^k)| dP(\omega) + \int Q^{k-1}(A^k) |p(\omega) - q(\omega)| d\tau(\omega)$$

$$\leq \|P^{k-1} - Q^{k-1}\| + 2\|P - Q\|$$

Thus $\|P^2 - Q^2\| \leq \|P - Q\| + 2\|P - Q\| \leq 4\|P - Q\|$, and $\|P^k - Q^k\| \leq 2k\|P - Q\|$ by induction. \(\Box\)

15
CHAPTER 4

DISCRETE CASE

An extended examination of the discrete case, that is, the case in which the X's are $\mathcal{Q}$-valued random variables, is made in this section. Throughout, the assumption is made that the condition of Theorem 3.3 is satisfied, hence merging will occur in the presence of consistency. In this discrete case, $\mu_{n,\omega}$ can be expressed explicitly (writing $X_j(\omega)$ for $\{X_j(\omega)\}$) as follows

\[
\mu_{n,\omega}(A) = \frac{\int \prod_{j=1}^{n} P_\lambda(X_j(\omega)) \mu(d\lambda)}{\int \prod_{j=1}^{n} P_\lambda(X_j(\omega)) \mu(d\lambda)} \quad \text{for } A \in \mathcal{B}
\]

whenever the denominator is positive. Otherwise $\mu_{n,\omega}(A) = 0$ for all $A \in \mathcal{B}$. It is clear that, when its denominator is positive, the expression on the right hand side of the above equation satisfies the first two requirements that it be admitted as $\mu_{n,\omega}$, namely, that it be a probability for every $\omega$ and that it be $\mathcal{F}_n$-measurable. The third requirement, that $\mu_{n,\omega}(B)$ be a version of $E(I_B \mid \mathcal{F}_n)$ for each $B \in \mathcal{B}$, is also satisfied. Since each set $F \in \mathcal{F}_n$ is at most a countable union of sets of the form $\{\omega \mid X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n\}$ it is enough to prove the assertion for sets of this form. Now
\[ P_\Lambda(F) = \prod_{j=1}^{n} P_\Lambda(x_j) \] so that

\[ \int_B dP = \int_B \prod_{j=1}^{n} P_\Lambda(x_j) \mu(d\lambda) = \int_B \prod_{j=1}^{n} P_\Lambda(x_j) \mu(d\lambda) = \int_{B} \prod_{j=1}^{n} P_\Lambda(x_j) \mu(d\lambda) \]

Also

\[ \int_F \mu_{n,\omega}(B) dP = \int_F \frac{\int \prod_{j=1}^{n} P_\Lambda(x_j(\omega)) \mu(d\lambda)}{\int \prod_{j=1}^{n} P_\Lambda(x_j(\omega)) \mu(d\lambda)} P_\Lambda(d\omega) \mu(d\lambda) \]

\[ = \int \frac{\int \prod_{j=1}^{n} P_\Lambda(x_j) \mu(d\lambda)}{\int \prod_{j=1}^{n} P_\Lambda(x_j) \mu(d\lambda)} \prod_{j=1}^{n} P_\Lambda(x_j) \mu(d\lambda) \]

since, on \( F, x_j(\omega) = x_j \). Thus

\[ \int_F \mu_{n,\omega}(B) dP = \int \prod_{j=1}^{n} \Lambda(x_j) \mu(d\lambda) \]

and the right hand side of this last equation has already been shown to equal \( \int_B dP \).

When the denominator of (1) is zero, \( \mu_{n,\omega} \) is not a probability.

However, if, for all \( i, P_\theta(i) > 0 \) implies \( \int_{\Lambda} P_\Lambda(i) \mu(d\lambda) > 0 \), the denominator is positive (a.s. \([P_\theta]\)). Since it is only in this case that consistency of the Bayes estimates is possible, it will not be necessary to be concerned with the possibility that for some \( i \) with \( P_\theta(i) > 0, \int_{\Lambda} P_\Lambda(i) \mu(d\lambda) = 0 \), in which case the denominator of (1) is (a.s. \([P_\theta]\)), eventually zero.
The following lemma is of technical interest although it is foreign to the subjective Bayesian spirit.

**Lemma 4.1.** \( P_n^k \) is a regular conditional probability for \((X_{n+1}, \ldots, X_{n+k})\) given \((X_1, \ldots, X_n)\).

**Proof.** \( P_n^k(A) \) has been defined as \( \int_A P_n^k(A) \mu_n,\omega(d\lambda) \), so that, by the countable additivity of the indefinite integral, it is a probability on \( \mathcal{A}_{n+1}^{n+k+1} \). It need only be shown that for \( A \in \mathcal{A}_{n+1}^{n+k} \), \( P_n^k(A) \) is a version of \( E(I_A | \mathcal{F}_n) \). Since \( \mu_n,\omega \) is \( \mathcal{F}_n \)-measurable and \( P_n^\lambda(A) \) is (for each \( \lambda \)) a constant it follows, by viewing the integral as the limit of its approximating sums, that it is \( \mathcal{F}_n \)-measurable. It remains to show that for \( B \in \mathcal{F}_n \)

\[
(2) \quad \int_{A \times B} P_n^k(A) P(d\lambda,d\omega) = \int_{A \times B} I_A P(d\lambda,d\omega)
\]

where, as before, \( P = \int \mu(d\lambda) \). Since \( A \) is at most the countable union of sets of the form \((X_{n+1} = y_1, \ldots, X_{n+k} = y_k)\) it suffices that \( A \) be considered such a set. Moreover, for the same reason, \( B \) may be considered to be of the form \((X_1 = x_1, \ldots, X_n = x_n)\). Then the left hand side of (2) is
\[
\int_{\Lambda \times B} P_k^n(\Lambda) \ P(d\lambda, d\omega) \\
= \int_{\Lambda \times \Omega} \left( \prod_{i=1}^{n} P_{\lambda}(y_i) \right) \frac{\prod_{j=1}^{n} P_{\lambda}(x_j) \ \mu(d\lambda)}{\int_{\Lambda} \prod_{j=1}^{n} P_{\lambda}(x_j) \ \mu(dx)} \\
= \int_{\Lambda} \prod_{i=1}^{n} P_{\lambda}(y_i) \ \prod_{j=1}^{n} P_{\lambda}(x_j) \ \mu(d\lambda)
\]

The right hand side of (2) is

\[
\int_{\Lambda \times \Omega} 1_{A \cap B} P(d\lambda, d\omega) = \int_{\Lambda} P_{\lambda}(A \cap B) \ \mu(d\lambda)
\]

where, in an abuse of notation, \(A\) and \(B\) are now viewed as cylinder sets in \(A_\infty\).

\[
\int_{\Lambda} P_{\lambda}(A \cap B) \ \mu(d\lambda) = \int_{\Lambda} \prod_{i=1}^{k} P_{\lambda}(y_i) \ \prod_{j=1}^{n} P_{\lambda}(x_j) \ \mu(d\lambda)
\]

It has been mentioned that Freedman (1963) is able to specify precisely in the discrete case the subset of \(\Lambda\) for which consistency of the Bayes estimates is realized. The following two theorems provide this information. Freedman's setting is non-parametric and he writes \(\lambda(i)\) instead of \(P_{\lambda}(i)\) as indicated below. However, since the hypothesis of Theorem 3.3 is presumed to be satisfied the two statements are equivalent.
Theorem 4.1 (Freedman). Suppose $\theta \in \Lambda$ and $\{i | P_{\theta}(i) > 0\}$ is finite. Then $(\theta, \mu)$ is consistent if and only if $\theta$ is in the topological carrier of $\mu$, that is, the smallest closed set of $\mu$-measure 1.

Theorem 4.2 (Freedman). Let $\theta \in \Lambda$ have

$$H(\theta) = - \sum_{i=1}^{\infty} P_{\theta}(i) \log P_{\theta}(i) < \infty.$$ Let $\mu$ be a probability on $\Lambda$ with the property that for any neighborhood $U$ of $\theta$ and any $\delta > 0$

$$\mu(\lambda | \lambda \in U \text{ and } \sum_{i=1}^{\infty} P_{\theta}(i) \log P_{\lambda}(i) < H(\theta) + \delta) > 0.$$ Then $(\theta, \mu)$ is consistent.

In the investigation of the rate at which one comes to believe the truth about the finite future, it appears that if $\mu(\theta) > 0$, convergence is at an exponential rate, an instance of which is given in Example 4.1. If, however, the prior assigns very little probability to neighborhoods of $\theta$ merging can occur at a very slow rate. Freedman (1963) showed that, in the discrete case, consistency a posteriori is not dependent on local properties of the prior distribution, but occurs (under the appropriate regularity conditions) for every parameter in the topological carrier of the prior. In this case, then, and under the continuity condition of Theorem 3.3, one will almost surely come to believe the truth no matter how unlikely he originally thought it. I conjecture that the convergence may, however, be arbitrarily slow.

Example 4.1.

Suppose $(X_1, X_2, \ldots)$ are i.i.d. with

$$P_{\lambda}(X_1 = 1) = \lambda, \quad P_{\lambda}(X_1 = 0) = 1 - \lambda$$
Here $\Lambda = (0,1]$. The prior distribution $\mu$ is concentrated on two points with $\mu(\theta) = p$, $\mu(\beta) = q = 1-p$. Let $S_n = \sum_{i=1}^{n} X_i$, and suppose that $P_{\theta}$ is the true distribution. Then

$$P_n(X_{n+1} = 1) = \theta \mu_n,\omega(\theta) + \beta \mu_n,\omega(\beta)$$

and it is enough to examine this event (see Lemma 3.1). Moreover, since $\mu_n,\omega(\beta) = 1 - \mu_n,\omega(\theta)$, that $\|P_n - P_\theta\| \to 0$ exponentially, it suffices that $\mu_n,\omega(\theta) \to 1$ at that rate.

$$\mu_n,\omega(\theta) = \frac{\frac{S^2}{\theta^n(1-\theta)}}{\theta^n(p) + \beta^n(1-\theta) n q} = \frac{1}{1 + f_n}$$

where

$$f_n = \left(\frac{S^n}{\theta} \right) \left(\frac{1-\beta}{1-\theta} \right) n q \frac{q}{p}$$

The result follows if $f_n \to 0$ exponentially. But, almost surely,

$$f_n^{1/n} = \left(\frac{S^n}{\theta} \right)^{\theta+o(1)} \left(\frac{1-\beta}{1-\theta} \right)^{1-\theta+o(1)} \left(\frac{q}{p} \right)^{1/n}$$

and it suffices that $f_n^{1/n} < 1$ or that

$$\beta^{\theta+o(1)} (1-\beta)^{1-\theta+o(1)} < \theta^{\theta+o(1)} (1-\theta)^{1-\theta+o(1)} \left(\frac{q}{p} \right)^{1/n}$$

To maximize $\beta^{\alpha}(1-\beta)^{1-\alpha}$ it is clear (from probabilistic considerations regarding the maximum likelihood estimate of the Bernoulli parameter or by differentiating) that one should set $\beta = \alpha$. Moreover the function to be maximized has a unique maximum, and since $\beta^{\alpha}$ is a continuous
function of $\alpha$ for fixed $\beta$, it is the case that, eventually, $r_n^{\alpha n} < 1$.

For the material to follow, a few additional notational conventions are required. As before, $P_\lambda(i)$ will be written for $P_\lambda([i])$, and $R_n^i$ will denote the empirical distribution defined by

$$ R_n^i = n^{-1} \sum_{j=1}^{N_n} I(X_j = i). $$

Let $\mathcal{L}_+ = \{i \mid P_\theta(i) > 0\}$ be enumerated as $\{1, 2, \ldots\}$ and let $\mathcal{L}_n = \{1, 2, \ldots, N_n\}$ where $N_n$ is as defined in the theorems to follow. Let $H_j(\lambda \mid \theta) = \sum_{i=1}^{\mathcal{L}_j} P_\theta(i) [-\log P_\lambda(i)]^j$ and $H_j(\theta) = H_j(\theta \mid \theta)$. When $j = 1$ it will be omitted in the notation. As before $\theta$ is used to represent the true value of the parameter while $\lambda$ is an arbitrary element of $\Lambda$.

The following theorem shows that merging occurs at a rate proportional to $n^{-1/4+\delta}$ for $\delta$ as specified. An example which follows the theorem indicates the possible choice of $\delta$.

**Theorem 4.3.** Suppose $\theta \in \Lambda$ with $H_2(\theta) < \infty$ and $\mu$ a probability on $\mathbb{G}$. For $\delta \in (0, \frac{1}{4})$ define

$$ N_n = \inf \{j \mid - \sum_{i=j+1}^{\infty} P_\theta(i) \log P_\theta(i) < n^{-1/2+2\delta} \} $$

$$ V_n = \left\{ \lambda \mid \sum_{i=1}^{N_n} \frac{P_\theta(i) [P_\theta(i) - P_\lambda(i)]}{\max(P_\theta(i), P_\lambda(i))]^2} < 16n^{-1/2+2\delta} \right\} $$

$$ V_n^* = \{ \lambda \in V_n \mid H(\lambda \mid \theta) < H(\theta) + n^{-1/2+2\delta}, H_2(\lambda \mid \theta) < \nu \} $$

for some $\nu < \infty$.  

22
If
\[ \sum_{n=1}^{\infty} N_n n^{-1-\delta} < \infty \]
\[ \sum_{i=1}^{\infty} P_{\theta}(i)^{-1} m_i^{-1-\delta} < \infty \] where \( m_i = \min\{n \mid i \in \mathcal{O}_n\} \)

then (a.s. \([P_{\theta}]\)) eventually
\[ \|P_n^k - P_{\theta}^k\| \leq 30kn^{-1/4+\delta} + \mu(V_n^{-1}) \exp(-n^{1/2+2\delta}) \]

**Proof.**
\[
\begin{align*}
\|P_n^k - P_{\theta}^k\| &= \sup_{\lambda \in \mathcal{A}_{n+1}} \int \left| P_n^k(A) \mu_n,\omega(d\lambda) - P_{\theta}^k(A) \right| \\
&\leq \sup_{\lambda \in \mathcal{A}_{n+1}} \int \left| P_n^k(A) - P_{\theta}^k(A) \right| \mu_n,\omega(d\lambda) \\
&\leq \sup_{\lambda \in \mathcal{A}_{n+1}} \int \left| P_n^k(A) - P_{\theta}^k(A) \right| \mu_n,\omega(V_n^c) + \mu_n,\omega(V_n) \\
&\leq \sup_{\lambda \in \mathcal{A}_{n+1}} \|P_n^k - P_{\theta}^k\| + \mu_n,\omega(V_n^c)
\end{align*}
\]

(The last step is a very poor approximation as may be seen in Example 4.4.) By Lemma 3.1

(3) \[ \|P_n^k - P_{\theta}^k\| < 2k\|P_n^k - P_{\theta}\| \leq 2k \left[ \sum_{i=1}^{N} |P_n^k(i) - P_{\theta}(i)| + \sum_{i=N+1}^{\infty} (P_n^k(i) + P_{\theta}(i)) \right] \]

By the defining condition for \( N_n \)
\[
(4) \quad r(n) = \sum_{i=N_n+1}^{\infty} P_\theta(i) < n^{-1/2+25}
\]

for \( n \) such that \( i > N_n \) implies \( -\log P_\theta(i) > 1 \). Let

\[
J_n = \{ i \in \mathcal{J}_n \mid P_\lambda(i) > P_\theta(i) \} \quad \text{and} \quad K_n = \mathcal{J}_n \setminus J_n
\]

where the dependence on \( \lambda \) has been suppressed on the notation. It follows, for \( \lambda \in \mathcal{V}_n \), that

\[
\sum_{i \in J_n} \frac{P_\theta(i)}{P_\lambda(i)} (P_\lambda(i) - P_\theta(i))^2 + \sum_{i \in K_n} P_\theta^{-1}(i)(P_\theta(i) - P_\lambda(i))^2 < 16n^{-1/2+25}
\]

A fortiori each of the sums on the left hand side is also bounded by \( 16n^{-1/2+25} \) whence

\[
(5) \quad \sum_{i \in K_n} (P_\theta(i) - P_\lambda(i)) < 4n^{-1/4+5}
\]

This is so because \( \sum b_i^2 \) is maximized, subject to the constraint \( \sum a_i b_i^2 \leq c \) (\( a_i > 0, b_i \geq 0 \) for all \( i \)) when \( b_i = 0 \) for \( i \neq j \), \( b_j = (c/a_j)^{1/2} \) whence \( j \) satisfies \( a_j = \min_1 a_i \) provided such a minimum exists. In the present case \( a_i = P_\theta^{-1}(i) \) so that there is such a minimum. Now

\[
(6) \quad \sum_{J_n} P_\lambda(i) - P_\theta(i) \leq \sum_{K_n} P_\theta(i) - P_\lambda(i) + \sum_{\mathcal{J}_n \setminus \mathcal{O}_n} P_\theta(i)
\]

\[
\leq 4n^{-1/4+5} + n^{-1/2+25}
\]

\[
< 5n^{-1/4+5} \quad \text{eventually.}
\]
Note that $\sum_{\lambda \in V_n} P^\lambda(i) \geq \sum_{\lambda \in V_n} P^\lambda(\lambda(i)).$ Thus

$$\sup_{\lambda \in V_n} \sum_{\lambda \in V_n} P^\lambda(i) \leq 1 - \inf_{\lambda \in V_n} \sum_{\lambda \in V_n} P^\lambda(i)$$

$$\leq 1 - \inf_{\lambda \in V_n} \left[ \sum_{\lambda \in V_n} P^\lambda(i) + \sum_{\lambda \in V_n} P^\lambda(i) - 4n^{-1/4+\delta} \right]$$

$$\leq r(n) + 4n^{-1/4+\delta}$$

$$< 5n^{-1/4+\delta} \quad \text{eventually}$$

Finally, in view of the bounds obtained on (4) through (7), one obtains from (3) that

$$(8) \quad \sup_{\lambda \in V_n} \|P^\lambda - P^\lambda\| < 30kn^{-1/4+\delta} \quad \text{eventually}.$$
Define

\[ S_n = \sum_{i=1}^{N} \frac{(n_i - nP_\theta(i))^2}{nP_\theta(i)} \]

Clearly

\[ S_n \geq \sum_{i=1}^{N} P_\theta(i) \left( \frac{n_i}{n} - P_\theta(i) \right)^2 \]

\[ \frac{n_i}{n} - \max(P_\theta(i), \frac{n_i}{n}) \]

so that, if eventually \( S_n < 16n^{-1/2+2\delta} \) it follows that the empirical probability distribution \( R_n \) defined by \( R_n(i) = n_i/n \) is eventually inside \( V_n \) (a.s. \([P_\theta])\). Straightforward, but tedious, calculation yields

\[ \text{Var}(S_n) \leq 2N_n n^{-2} + n^{-3} \sum_{i=1}^{N} [P_\theta(i)]^{-1} [1-P_\theta(i)] \]

Invoking the Borel-Cantelli lemma and the Chebyshev inequality one obtains \( P_\theta(|S_n - ES_n| > n^{-1/2+2\delta} \text{ i.o.}) = 0 \) provided

\[ \sum_{n=1}^{\infty} n^{-1-4\delta} \text{Var} S_n < \infty \]

Thus, provided (9) is satisfied, it must be the case that (a.s. \([P_\theta])\)

\[ S_n < n^{-1/2+2\delta} + n^{-\frac{1}{\delta}} \sum_{i=1}^{N} (1 - P_\theta(i)) \]

(The second term on the right hand side is \( o(n^{-1/2+2\delta}) \) by i.) For

(9) it is enough that \( \sum_{n} N_n n^{-1-4\delta} < \infty \), which is i) and that

\[ \sum_{n=1}^{\infty} n^{-2-4\delta} \sum_{i=1}^{N} (P_\theta(i))^{-1} \sum_{i=1}^{n} (P_\theta(i))^{-1} < \infty \]. If \( m_i = \min(n_i | i \in \mathcal{C}_n) \)
\[
\sum_{n=1}^{\infty} n^{-2+\delta} \sum_{i=1}^{N} (P_{\theta}(i))^{-1} \leq c \sum_{i=1}^{\infty} (P_{\theta}(i))^{-1} (m_{i}^{-1+\delta} + \frac{m_{i}}{m_{i}^2})
\]

where \( c \) depends only on \( \delta \). Thus ii) insures the result that \( R_{n} \in V_{n} \) eventually (a.s. \([P_{\theta}]\)). By discarding a null set, suppose that \( R_{n} \in V_{n} \).

The function \(-H(\cdot \mid R_{n}) : \lambda \mapsto \sum_{i=1}^{n} \frac{1}{n} \log P_{\lambda}(i)\) is concave with maximum at \( R_{n} \). To verify concavity compute (writing \( \lambda(i) \) for \( P_{\lambda}(i) \))

\[
H(\alpha \lambda_{1} + (1-\alpha) \lambda_{2} \mid p) = \sum_{i=1}^{\infty} p(i) \left[ -\log(\alpha \lambda_{1}(i) + (1-\alpha) \lambda_{2}(i)) \right]
\leq \sum_{i=1}^{\infty} p(i) \left[ -\alpha \log \lambda_{1}(i) - (1-\alpha) \log \lambda_{2}(i) \right]
\]

\[
= \alpha H(\lambda_{1} \mid p) + (1-\alpha) H(\lambda_{2} \mid p)
\]

where the inequality follows from Jensen's inequality and the concavity of the log. To verify that \( H(\theta) \leq H(\lambda \mid \theta) \) note that if for some \( i \), \( \lambda(i) = 0 \) while \( \theta(i) > 0 \) there is nothing to prove. Otherwise

\[
\sum_{i=1}^{\infty} \theta(i) \left[ \log \theta(i) - \log \lambda(i) \right] = \sum_{i=1}^{\infty} \theta(i) \log \frac{\lambda(i)}{\theta(i)}
\leq -\log E_{\theta}(\frac{\lambda}{\theta}) \text{ where } E_{\theta} \text{ denotes expectation}
\leq -\log E_{\theta}(\frac{\lambda}{\theta}) = 0
\]

again by Jensen.
Because of the concavity of $-H(\cdot \mid R_n)$ it follows that

$$\sup_{\lambda \in \partial V_n} \sum_{i=1}^{N_n} \frac{n_i}{n} \log P_\lambda(i) = \max_{\lambda \in \partial V_n} \sum_{i=1}^{N_n} \frac{n_i}{n} \log P_\lambda(i)$$

where $\partial C$ denotes the boundary of $C$.

For $\lambda \in \partial V_n$, since $\log(1+x) \leq x - \frac{x^2}{2(1+x)^2}$ (where $x^+ = \max(x,0)$)

$$\sum_{i=1}^{N_n} P_\theta(i) \log P_\lambda(i) - \sum_{i=1}^{N_n} P_\theta(i) \log P_\theta(i)$$

$$= \sum_{i=1}^{N_n} P_\theta(i) \log \left[ 1 + \frac{P_\lambda(i) - P_\theta(i)}{P_\theta(i)} \right]$$

$$\leq \sum_{i=1}^{N_n} P_\lambda(i) - \sum_{i=1}^{N_n} P_\theta(i) - \frac{1}{2} \sum_{i=1}^{N_n} \frac{P_\theta(i)[P_\lambda(i) - P_\theta(i)]^2}{\max(P_\theta(i), P_\lambda(i))^2}$$

$$\leq 1 - (1 - n^{-1/2+2\delta}) - 8n^{-1/2+2\delta}$$

$$\leq 7n^{-1/2+2\delta}$$

Further

$$\max_{\lambda \in \partial V_n} \sum_{i=1}^{N_n} P_\theta(i) \log P_\lambda(i) \leq -H(\theta) + n^{-1/2+2\delta} - 7n^{-1/2+2\delta}$$

and by Lemma 4.2 which follows, eventually (a.s. $[P_\theta]$)

$$\max_{\lambda \in \partial V_n} \sum_{i=1}^{N_n} \frac{n_i}{n} \log P_\lambda(i) \leq -H(\theta) - 4n^{-1/2+2\delta}$$

Thus eventually (a.s. $[P_\theta]$)

28
(10) \[ \mu_n, \omega_n(V_n^c) \leq D_{n}^{-1} \text{exp}[n(-H(\theta) - 3n^{-1/2+2\delta})] \]

Similarly,

\[ D_{n} \mu_n, \omega_n(V_n) = \int_{V_n} \prod_{j=1}^{n} P_{\lambda}(X_j(\omega)) \mu(d\lambda) \]

\[ \geq \inf_{\lambda \in \mathcal{V}_n^*} \exp[n \sum_{i=1}^{n} \frac{1}{n} \log P_{\lambda}(i)] \mu(V_n^*) \]

Because \( H_2(\theta) < \infty \), Lemma 4.2 implies the existence of a positive constant (depending on \( \lambda \)) \( c_\lambda \) for which

\[ \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \frac{1}{n} \log P_{\lambda}(i) + H(\lambda | \theta)}{(\log \log n/n)^{1/2}} > -c_\lambda \quad \text{a.s.} \ [P_\theta] \]

Thus, eventually (a.s. \([P_\theta]\)) for \( \lambda \in \mathcal{V}_n^* \)

\[ \sum_{i=1}^{n} \frac{1}{n} \log P_{\lambda}(i) > -H(\lambda | \theta) - c_\lambda \left( \frac{\log \log n}{n} \right)^{1/2} \]

\[ > -H(\theta) - n^{-1/2+2\delta} - n^{-1/2+2\delta} \]

Eventually, then, (a.s. \([P_\theta]\))

(11) \[ \mu_n, \omega_n(V_n) \geq Dn^{-1} \mu(V_n^*) \exp[n(-H(\theta) - 2n^{-1/2+2\delta})] \]

Combining (10) and (11) yields

\[ \frac{\mu_n, \omega_n(V_n^c)}{\mu_n, \omega_n(V_n)} \leq \mu(V_n^*)^{-1} \exp(-n^{1/2+2\delta}) \quad \text{a.s.} \ [P_\theta] \]

29
Lemma 4.2. If $H_2(\lambda|\theta) < \infty$ then

$$\lim \frac{\sum_{i=1}^{\infty} \frac{n_i}{n} \log P_{\lambda}(i) + H(\lambda|\theta)}{(\log \log n/n)^{1/2}} < c$$

Proof.

$$\sum_{i=1}^{\infty} \frac{n_i}{n} \log P_{\lambda}(i) = n^{-1} \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} I_{(X_j = i)} \log P_{\lambda}(i) = n^{-1} \sum_{j=1}^{n} Y_j$$

where

$$Y_j = \sum_{i=1}^{\infty} I_{(X_j = i)} \log P_{\lambda}(i)$$

Now $E_\theta Y_j = \sum_{i=1}^{\infty} P_\theta(i) \log P_\lambda(i) = -H(\lambda|\theta)$. Since $E_\theta Y_j^2 = H_2(\lambda|\theta) < \infty$ and the $Y_j$'s are i.i.d. it follows by the version of the law of the iterated logarithm due to Hartman and Wintner (1941) that

$$\lim \frac{\sum_{j=1}^{n} Y_j + nH(\lambda|\theta)}{(2n\sigma^2 \log \log n/n)^{1/2}} \leq 1$$

where $\sigma^2 = \text{Var} Y_j$

Since $\log \log n > \log \log n/n^{1/2}$ for $n > \sigma^2$ the result follows with $c \geq \sigma \sqrt{2}$.

The following examples show the application of Theorem 4.3.

Example 4.2.

Suppose $X_1, X_2, \ldots$ is a sequence of geometric random variables, where
\[ P_\zeta(X_j = i) = \zeta r^{i-1} \quad \text{and} \quad r = 1 - \zeta \]

Let \( M_n = -\log_b k_n + \log_b (-\log_b k_n) \) where \( \varphi = 1 - \theta \), \( b = \varphi^{-1} \) and \( k_n = (2c_2)^{-1} n^{-1/2+2\delta} \) for \( \delta > 0 \) and \( c_2 = \log b \). (Recall that \( \log = \log_e \).) Although \( M_n \) as defined does not satisfy the condition of Theorem 4.3 defining \( N_n \) in that there may be smaller integers \( j \) for which \( \sum_{j+1}^{\infty} P_\theta(i) \log P_\theta(i) < n^{-1/2+2\delta} \), nothing is lost by using the proposed definition. Now

\[
\sum_{j+1}^{\infty} P_\zeta(i) \log P_\zeta(i) = \sum_{j+1}^{\infty} \zeta r^{i-1} \log(\zeta r^{i-1})
\]
\[
= \zeta (\log \zeta) \sum_{j+1}^{\infty} r^{i-1} + \zeta (\log r) \sum_{j+1}^{\infty} (i-1) r^{i-1}
\]
\[
= \zeta \log \zeta \frac{r^j}{1-r} + \zeta \log r \left( \frac{jr^j}{1-r} + \frac{r^{j+1}}{(1-r)^2} \right)
\]
\[
= r^j (\log \zeta + \frac{r \log r}{\zeta} + j \log r)
\]
\[
= r^j (-c_1 - c_{2j})
\]

where

\[ c_1 = -\log \zeta - \frac{r \log r}{\zeta} \quad \text{and} \quad c_2 = -\log r \]

If \( M_n \) is as proposed

\[
- \sum_{M_n+1}^{\infty} P_\theta(i) \log P_\theta(i) = [c_1 + c_2(-\log_b k_n + \log_b(-\log_b k_n))] k_n(-\log_b k_n)^{-1}
\]
\[
\leq k_n c_2 [1 + (c_1 + \log_b(-\log_b k_n)(-\log_b k_n)^{-1})]
\]
\[
< n^{-1/2+2\delta}
\]
which is as desired. Conditions i) and ii) of the theorem are satisfied since

\[ \sum_{n=1}^{\infty} M_n n^{-1-4\delta} < \sum n^{-1-4\delta} \log_b n < \infty \]

\[ \sum_{i=1}^{\infty} P_\theta(i) m_i^{-1-4\delta} \leq c \sum_{i=1}^{\infty} b b_i 2i(-1-4\delta) < \infty \]

where ii) follows from \( P_\theta(i) = \theta \phi^{i-1} = \frac{\theta}{\phi} b^{-i} \) and \( M_n < \frac{1}{2} \log_b n \), thus \( m_i \geq b^{2i} \). Examine now the conditions defining \( V_n \) and \( V^*_n \).

If \( d(i) = 1 - (P_\zeta(i)/P_\theta(i)) \) then

\[ T_n(\zeta) = \sum_{i=1}^{M_n} \frac{P_\theta(i)[P_\zeta(i) - P_\theta(i)]^2}{\max(P_\zeta(i), P_\theta(i))^2} \leq \sum_{i=1}^{\infty} P_\theta(i) d^2(i) \]

\[ = \sum_{i=1}^{\infty} \theta \phi^{i-1}[1 - (1 + \frac{d}{\theta})(1 - \frac{d}{\phi})] \text{ where } d = \zeta - \theta \]

\[ = \sum_{i=1}^{\infty} \theta \phi^{i-1}[(i-1)\frac{d}{\phi} - \frac{d^2}{\theta}] + o(d^2) \]

from Taylor's theorem. Computing the sum, one finds

\[ T_n(\zeta) \leq d^2 \left[ \frac{\theta + \phi^2}{\theta^2 + \phi^2} \right] + o(d^2) \]

so that if \( |\zeta - \theta| < \frac{4((\theta^2 \phi^2)/(\theta^2 + \phi^2))^{1/2} n^{-1/4+\delta}}{1/4+\delta} \), \( \zeta \in V_n \). Note that, for \( x \geq -1/2 \), \( x(1-x) < \log(1+x) \). Thus
\[
H(\xi | \theta) - H(\theta) = - \sum_{i=1}^{8} p_{\theta}(i) \log \frac{p_{\xi}(i)}{p_{\theta}(i)}
= - \sum_{i=1}^{8} p_{\theta}(i) \log(1 - d(i))
\leq - \sum_{i=1}^{8} p_{\theta}(i) [-d(i)(1 + d(i))]
= - \sum_{i=1}^{8} \theta_{\phi}^{i-1}(\frac{\theta_{\xi}}{\theta_{\phi}} - 1)(2 - \frac{\xi_{i-1}}{\theta_{\phi}})
= \frac{12d}{\theta} + o(d)
\]

If \(|\xi - \theta| < (\theta/12) n^{-1/2+2\delta}\) then \(\xi \in V_n\). Clearly \(H(\xi | \theta) < \infty\) for \(\xi \in (0,1)\). Thus if the symmetric interval of half-width \((\theta/12) n^{-1/2+2\delta}\) is called \(W_n\) and if \(\mu(W_n)^{-1} \exp(-n^{-1/2+2\delta})\) is \(o(n^{-1/4+\delta})\) it follows that \(\|P_n - P_\xi\| = O(n^{-1/4+\delta})\) a.s.

One should note that, in the preceding example the results followed for any \(8 \in (0, 1/4)\). That this is not always the case may be seen in the following example where the true distribution has a fatter tail than the geometric.

**Example 4.3.**

Suppose \(P_\theta(i) = k_\alpha i^{-\alpha}\).

Then \(M_n = n^{2\alpha-2}\) satisfies \(- \sum_{i=M_n+1}^{1-3\delta} p_{\theta}(i) \log p_{\theta}(i) < n^{-1/2+2\delta}\).

The conditions of Theorem 4.3 require
1) \[ \sum_{n=1}^{\infty} M_n n^{-1-4\delta} < \infty \] which requires \[ \frac{1-3\delta}{2\alpha-2} - 1 - 4\delta < -1, \text{ or } \delta > \frac{1}{8\alpha-5}; \]

ii) \[ \sum_{i=1}^{\infty} P_\theta(1)^{-1} m_1^{-1-4\delta} < \infty \text{ or } \sum_{i=1}^{\infty} i^{\alpha+1} \frac{(2\alpha-2)}{1-3\delta} (-1-4\delta) < \infty \] which requires \( \alpha - (1 + 4\delta) \frac{2\alpha-2}{1-3\delta} < -1, \delta > \frac{3-\alpha}{11\alpha-5} \) which is satisfied if \( \delta \) is as required for i).

Note that, if \( \alpha = 2, \delta > 1/11 \) so that convergence at the rate \( n^{-1/4+1/11+\varepsilon} \) for any \( \varepsilon > 0 \) is attainable. It is of interest that, in this case \( \sum P_\theta(1)^{1/2} = \infty \) so that Theorem 4.4 does not apply.

It is clear that the rate of convergence which is obtained in Theorem 4.3 is far from the rate \( (\log \log n/n)^{1/2} \) which I suppose to be possible, at least under appropriate restrictions on the prior distribution. In the following example, examination is made of the approximations used in the proof of Theorem 4.3 in the simplest possible case--that of Bernoulli random variables to discover the point at which the better rate is lost.

Example 4.4.

Suppose \( A = [0,1] \) and that \( P_\lambda(1) = \lambda, P_\lambda(0) = 1-\lambda \). The prior \( \mu \) is uniform so that \( \mu(d\lambda) = d\lambda \). Then

\[ \|P_n - P_\theta\| = \left| \int_0^1 \lambda \mu_{n,\omega} (d\lambda) - \theta \right| = \left| \frac{8}{n+2} - \theta \right| \]
where \( S_n = \sum_{i=1}^{n} X_i \). By the law of the iterated logarithm, the quantity decreases to 0 (a.s.) eventually as \((\log \log n/n)^{1/2}\). Theorem 4.3 in the case gives only the result that (a.s.) eventually

\[
\| P_n - P_\theta \| \leq 30 n^{-1/4+\delta} + \mu(V_n^*)^{-1} \exp(-n^{1/2+2\delta})
\]

Now

\[
V_n = \left\{ \lambda \left| \sum_{i=1}^{N_n} \frac{P_\theta(i) [P_\Lambda(i) - P_\theta(i)]^2}{[\max(P_\theta(i), P_\Lambda(i))]^2} < 16n^{-1/2+2\delta} \right\}
\]

\[
\cup \left\{ \lambda \left| \sum \frac{[P_\theta(i) - P_\theta(i)]^2}{P_\theta(i)} < 16n^{-1/2+2\delta} \right\}
\]

The condition thus reduces to requiring

\[
(\lambda-\theta)^2 < 16n^{-1/2+2\delta} \theta(1-\theta)
\]

\( V_n^* \) is the subset of \( V_n \) for which \( H(\lambda|\theta) - H(\theta) < n^{-1/2+2\delta} \). But

\[
H(\lambda|\theta) - H(\theta) = \theta \log \frac{\theta}{\Lambda} + (1-\theta) \log \frac{1-\theta}{1-\lambda}
\]

\[
\leq -\theta \log(1 + \frac{\varepsilon}{\theta}) - (1-\theta) \log(1 - \frac{\varepsilon}{1-\theta})
\]

where \( \varepsilon = \lambda-\theta \). Since for \( x > -1/2, \log 1 + x > x(1-x) \), \( H(\lambda|\theta) - H(\theta) < \varepsilon^2 [\theta(1-\theta)]^{-1} \), so that \( \mu(V_n^*) \geq \text{const.} \, n^{-1/4+\delta} \) and the last term of the bound for \( \| P_n - P_\theta \| \) vanishes rapidly.

The first approximation in the proof of Theorem 4.3 is
\[ |\int_{\Lambda} P_{\lambda}^{k}(A) \, \mu_{n,\omega}(d\lambda) - P_{\theta}(A)| \leq \int_{\Lambda} |P_{\lambda}^{k}(A) - P_{\theta}(A)| \, \mu_{n,\omega}(d\lambda). \]

In the present example, the right hand side of the above inequality is

\[ \int_{0}^{1} |\lambda - \theta| \, \mu_{n,\omega}(d\lambda) \leq \left[ \int_{0}^{1} (\lambda - \theta)^2 \, \mu_{n,\nu}(d\lambda) \right]^{1/2} \]

Observe that

\[ \int_{0}^{1} \lambda^2 \, \mu_{n,\omega}(d\lambda) = \frac{(S+1)(S+2)}{(n+2)(n+3)} = \frac{S}{n} + o(1) \]

so that

\[ \int_{0}^{1} |\lambda - \theta| \, \mu_{n,\omega}(d\lambda) \leq \left| \frac{S}{n} - \theta \right| + o(1) \]

and nothing has been lost in this approximation. The next step however involves estimating \( \int V_{n} |P_{\lambda}^{k}(A) - P_{\theta}(A)| \, \mu_{n,\omega}(d\lambda) \) by the supremum of the integrand on \( V_{n} \), which, in this example, is the supremum of \( |\lambda - \theta| \) on \( V_{n} \). As has previously been calculated, here is the Achilles heel for \( \sup_{\lambda \in V_{n}} |\lambda - \theta| \) can be bounded only by \( \frac{4n^{-1/4} + 8(\theta(1-\theta))^{1/2}}{n} \).

The difficulties encountered in the proof of Theorem 4.3 arise principally from the fact that it is not possible to examine directly the quantity in which one is interested—namely \( \sum |\int P_{\lambda}(1) \, \mu_{n,\omega}(d\lambda) - P_{\theta}(1)| \).

Accordingly it is necessary to make use of the empirical distribution \( R_{n} = \{ n_{i}/n \} \), ensuring that the results obtained can not be better than corresponding results obtainable for the maximum likelihood estimate.
The following two theorems make this detour explicit. The result of Theorem 4.4 is the best possible, and is useful in its own right. Theorem 4.5, the result of using Laplace's method for an infinite number of coordinates, is not a theorem about probability—it is essentially an arithmetic result—but its application depends entirely on probabilistic arguments.

**Theorem 4.4.**

\[
\lim \frac{\|P^n_\theta - R_n\|}{(2 \log \log n/n)^{1/2}} \leq \sum_{i=1}^{\infty} [P_\theta(i)(1 - P_\theta(i))]^{1/2} \quad \text{a.s.}
\]

where \( R_n \) is the empirical distribution defined by \( R_n(i) = \frac{n_i}{n} \).

**Proof.** Note that if \( \sum P_\theta(i)^{1/2} = \infty \) the theorem is true trivially. Suppose then that \( \sum P_\theta(i)^{1/2} < \infty \).

Consider the case \( P_\theta(i) = 0 \) for \( i \notin \{1, \ldots, N\} \) for some \( N < \infty \). Then the weak form of the law of the iterated logarithm for Bernoulli trials gives

\[
\lim \frac{|n_i - nP_\theta(i)|}{(2n \psi(i) \log \log n)^{1/2}} = 1 \quad \text{a.s.} \quad [P_\theta] \text{ where } \psi(i) = P_\theta(i)(1-P_\theta(i))
\]

Since, for each \( n \)

\[
\|P^n_\theta - R_n\| \leq \sum_{i=1}^{N} \left| \frac{n_i}{n} - P_\theta(i) \right|
\]

it follows that
\[
\lim \frac{\|P_\theta - R\|}{(2 \log \log n/n)^{1/2}} \leq \sum_{i=1}^{N} \psi(i)^{1/2} \quad \text{a.s.}
\]

In the case in which \(\mathcal{J}_+\) is denumerable it is still the case that, for fixed \(i\),

\[
\lim \frac{|P_\theta(i) - \frac{n_i}{n}|}{(2 \log \log n/n)^{1/2}} = \psi(i)^{1/2} \quad \text{a.s.}
\]

and thus that

\[
\sum_{i=1}^{\infty} \lim \frac{|P_\theta(i) - \frac{n_i}{n}|}{(2 \log \log n/n)^{1/2}} = \sum_{i=1}^{\infty} \psi(i)^{1/2} \quad \text{a.s.}
\]

The result will follow if

\[(12) \quad \lim \sum_{i=1}^{\infty} \frac{|P_\theta(i) - \frac{n_i}{n}|}{(\log \log n/n)^{1/2}} \leq \sum_{i=1}^{\infty} \lim \frac{|P_\theta(i) - \frac{n_i}{n}|}{(\log \log n/n)^{1/2}} \]

Since, for each \(m\), and any sequence of sequences \(\{a_{mn}\}_{m=1}^{\infty}\) and any \(N\)

\[
\sup_{n \geq N} \sum_{m} a_{mn} \leq \sum_{m} \sup_{n \geq N} a_{mn}
\]

it follows that

\[
\lim \sum_{i=1}^{\infty} \frac{|P_\theta(i) - \frac{n_i}{n}|}{(\log \log n/n)^{1/2}} \leq \lim_{m \to \infty} \sum_{i=1}^{\infty} \sup_{n \geq m} \frac{|P_\theta(i) - \frac{n_i}{n}|}{(\log \log n/n)^{1/2}}
\]

Then (12) will follow if

38
\[
\lim \sum_{m}^{\infty} \sup_{n \geq m} \frac{|P_{\theta}(i) - \frac{n_{i}}{n}|}{(\log \log n/n)^{1/2}} \leq \sum_{i=1}^{\infty} \lim_{m} \frac{|P_{\theta}(i) - \frac{n_{i}}{n}|}{(\log \log n/n)^{1/2}}
\]

Writing
\[
d_{m} \text{ for } \sup_{n \geq m} \frac{|P_{\theta}(i) - \frac{n_{i}}{n}|}{(\log \log n/n)^{1/2}}
\]

it is required that

(13) \[
\lim \sum_{m}^{\infty} d_{m} \leq \sum_{i=1}^{\infty} \lim_{m} d_{m}
\]

Consider the auxiliary random variables \(Z_{j}^{i}\) where for each \(i\), \((Z_{1}^{i}, Z_{2}^{i}, \ldots)\) are i.i.d. with

\[
P_{\theta} \left\{ Z_{j}^{i} = \left( \frac{1 - P_{\theta}(i)}{P_{\theta}(i)} \right)^{1/2} \right\} = P_{\theta}(i)
\]

\[
P_{\theta} \left\{ Z_{j}^{i} = -\left( \frac{P_{\theta}(i)}{1 - P_{\theta}(i)} \right)^{1/2} \right\} = 1 - P_{\theta}(i)
\]

For each \(i\), \(E_{\theta} Z_{j}^{i} = 0, E_{\theta} (Z_{j}^{i})^{2} = 1\). Let \(S_{n}^{i} = \sum_{j=1}^{n} Z_{j}^{i}\), then

\[
|n_{i} - nP_{\theta}(i)| = \left| \sum_{j=1}^{n} [I_{\{X_{j}=i\}} - P_{\theta}(i)] \right|
\]

\[
= \psi(i)^{1/2} \left| \sum_{j=1}^{n} Z_{j}^{i} \right|
\]

\[
= \psi(i)^{1/2} \left| S_{n}^{i} \right|
\]
By a theorem of Siegmund (1969, p. 530) if $EZ = 0$, and $EZ^2 < \infty$ then
\[ E\left[ \sup_n \frac{|S_n|}{(n \log \log n)^{1/2}} \right] < \infty. \] Since each of the $Z_i$ has variance 1, it must be that there exists $u$ independent of $i$ such that
\[ E\left[ \sup_n \frac{|S_n^i|}{n (n \log \log n)^{1/2}} \right] \leq u < \infty. \] A similar manipulation may be seen in the lemma (11) of Olshen and Siegmund (1971). One concludes that
\[ \widetilde{g} = \sup_i \sup_n \frac{|S_n^i|}{n (n \log \log n)^{1/2}} < \infty \quad \text{a.s.} \quad [P_0] \]
Let $g = \tilde{g}$ if $\tilde{g} < \infty$, $g = 0$ otherwise and define $f_i = \psi(i)^{1/2}g$. It is clear that $d_i^{1/2} \leq f_i \quad \text{(a.s.) and that}$
\[ \sum_{i=1}^{\infty} f_i = g \sum_{i=1}^{\infty} \psi(i)^{1/2} < \infty \]
by supposition; (13) follows by the dominated convergence theorem. $\triangleright$

**Theorem 4.5.** Suppose
\[ N_n = \min \left\{ j \mid \max \left[ \sum_{j+1}^{\infty} \frac{n_i}{n}, - \sum_{j+1}^{\infty} \frac{n_i}{n} \log \frac{n_i}{n} \right] < c^2 \frac{\log n}{n} \right\} \quad \text{for some} \quad c > 0 \]
Let
\[ W_n = \left\{ \lambda \mid \sum_{i=1}^{n} \frac{n_i}{n} \left( P_{\lambda}(i) - \frac{n_i}{n} \right) < 9c^2 \frac{\log n}{n} \right\} \]
\[ W^*_n = \left\{ \lambda \in W_n \mid \sum_{i=1}^{n} \frac{n_i}{n} \log P_{\lambda}(i) > \sum_{i=1}^{\infty} \frac{n_i}{n} \log \frac{n_i}{n} - c^2 \frac{\log n}{n} \right\} \]

40
Then

\[ \|P_n^k - R_n^k\| < 24\alpha k^{1/2} \frac{\log n}{n} + \mu(W_n) n^{-1} n^{-2} \]

Proof.

\[ (14) \|P_n^k - R_n^k\| = \sup_{\Lambda \subset A_{n+1}} \sum_{\lambda \in \Lambda} \left| P_{n,\lambda}(A) - R_n(A) \right| \]

\[ \leq 2k \sup_{\lambda \in W_n} \sum_{i=1}^n \left| P_{\lambda}(i) - \frac{n_i}{n} \right| + \sup_{\lambda \in W_n} \sum_{i \in J_n} \left( P_{\lambda}(i) + \frac{n_i}{n} \right) \]

\[ + \mu_n, \omega(W_n^c) \]

by Lemma 3.1. Let

\[ J_n = \left\{ i \in \mathcal{I}_n \mid P_{\lambda}(i) > \frac{n_i}{n} \right\}, \quad K_n = \frac{J_n}{J_n} \]

By an argument developed in the proof of Theorem 4.3, for \( \lambda \in W_n \)

\[ \sum_{i \in K_n} \left( \frac{n_i}{n} - P_{\lambda}(i) \right) < 3c \left( \frac{\log n}{n} \right)^{1/2} \]

thus

\[ \sum_{i \in J_n} \left( P_{\lambda}(i) - \frac{n_i}{n} \right) < 3c \left( \frac{\log n}{n} \right)^{1/2} + c^2 \frac{\log n}{n} \]

\[ < 4c \left( \frac{\log n}{n} \right)^{1/2} \]

eventually.

Whence

\[ (15) \sup_{\lambda} \sum_{i \in J_n} \left| P_{\lambda}(i) - \frac{n_i}{n} \right| < 7c \left( \frac{\log n}{n} \right)^{1/2} \]
Similarly

\[
\sup_{W_n} \sum_{\lambda \in \mathcal{D}_n} P_\lambda(i) \leq 1 - \inf_{W_n} \sum_{\lambda \in \mathcal{D}_n} P_\lambda(i)
\]

\[
\leq 1 - \left[ \sum_{j_n} \frac{n_j}{n} + \sum_{k_n} \frac{n_k}{n} - 3c \left( \frac{\log n}{n} \right)^{1/2} \right]
\]

\[
\leq 4c \left( \frac{\log n}{n} \right)^{1/2} \text{ eventually.}
\]

Next, let

\[
D_n = \int_{\Lambda} \prod_{j=1}^{n} P_\lambda(X_j) \, \mu(d\lambda) = \int_{\Lambda} \left( \exp \sum_{i=1}^{\infty} n_i \log P_\lambda(i) \right) \mu(d\lambda)
\]

By arguments developed in detail in Theorem 4.3

\[
D_n \mu_n(\omega^0_n) \leq \exp(n \sup_{\lambda \in \mathcal{W}_n} \sum_{i=1}^{N_n} \frac{n_i}{n} \log P_\lambda(i))
\]

and

\[
\sum_{i=1}^{N_n} \frac{n_i}{n} \log \frac{P_\lambda(i)}{n_i/n} \leq \sum_{i=1}^{N_n} \frac{n_i}{n} \left[ \frac{P_\lambda(i) - \frac{n_i}{n}}{\frac{n_i}{n}} - \frac{1}{2} \left( \frac{P_\lambda(i) - \frac{n_i}{n}}{\max(P_\lambda(i), \frac{n_i}{n})} \right)^2 \right]
\]

\[
\leq 1 - (1 - e^2 \frac{\log n}{n}) - \frac{1}{2} \cdot 9e^2 \frac{\log n}{n}
\]

\[
\leq -3e^2 \frac{\log n}{n} \text{ for } \lambda \in \partial W_n
\]

Thus

\[
\sup_{\lambda \in \mathcal{W}^c_n} \sum_{i=1}^{N_n} \frac{n_i}{n} \log \lambda(i) \leq -H(R_n) - 2e^2 \frac{\log n}{n}
\]

whence
\[
D_n \mu_n, \omega_n (W^c_n) \leq \exp[-nH(R_n) - 2e^2 \frac{\log n}{n}]
\]

Similarly, using the definition of \(W^*_n\)

\[
D_n \mu_n, \omega_n (W^*_n) \geq \mu(W^*_n) \exp[-nH(R_n) - e^2 \frac{\log n}{n}]
\]

Thus

\[(17) \quad \mu_n, \omega_n (W^c_n) \leq \mu(W^*_n)^{-1} - e^2\]

Combining (15)-(17) in (14) yields the result. 

One should note that \((\log n)/n\) in the conditions of Theorem 4.5 can be replaced by any other quantity with a consequent change in the factor multiplying the quantity \(\mu(W^*_n)^{-1}\). The choice of \((\log n)/n\) was made to retain at least a negative power of \(n\) as this factor.

In assessing these results, one is impelled to consider the nature of the neighborhoods \(V_n\) and \(W_n\). Recall, for example, that

\[
W_n = \left\{ \lambda \mid \sum_{i=1}^{N} \frac{n_i}{n} \frac{(P^{(i)}(\lambda) - \frac{n_i}{n})^2}{\max(P^{(i)}(\lambda), \frac{n_i}{n})} < 9e^2 \frac{\log n}{n} \right\}
\]

A (slightly) smaller neighborhood has

\[(18) \quad \sum_{T_n} \frac{(P^{(i)}(\lambda) - \frac{n_i}{n})^2}{\frac{n_i}{n}} < 9e^2 \frac{\log n}{n}\]

43
where \( T_n = \{ 1 \leq i \leq N_n | n_i > 0 \} \). For ease of notation, let

\[
(P_\lambda(i) - \frac{n_i}{n}) = \varepsilon_i \quad \text{and} \quad 9c^2 \frac{\log n}{n} = d. \quad \text{(Temporarily suppress dependence on } n) \text{.}
\]

The neighborhood (13) is then represented as

\[
\sum_i \frac{n_i \varepsilon_i}{n_i^2} < d. \quad \text{If one replaces } \frac{n_i}{n} \text{ by } P_\theta(i) \text{ and changes coordinates by letting } \delta_i = \varepsilon_i P_\theta(i)^{1/2}, \text{ this neighborhood may be represented as an } N_n \text{-sphere of radius } d^{1/2} : \sum_i \delta_i^2 < d.
\]

Since the volume of such a sphere is proportional to

\[
d_n^{n/2} = (9c^2 \frac{\log n}{n})^{n/2}
\]

it is clear that, if the prior probability has no atom at } \theta \text{ and if } N_n \to \infty, \text{ the expression } \mu(W_n^{*})^{-1} n^{-c^2} \text{ may grow very quickly. If, however, } N_n \text{ is bounded above (as, for example, when } \mathcal{L}_+ \text{ is finite) or if } \mu \text{ is concentrated on certain arcs in } \Lambda, \text{ one may expect to obtain a usable result from the theorem. Arcs amenable to computation are homeomorphic images of common parametric distributions, for example, Poisson, geometric (see Example 4.2) and negative binomial probabilities.}

\[\text{It should be noted that the starred neighborhoods can be expected to afford little or no additional difficulty since, for example, in the case of } W_n^{*}, \text{ the additional requirement is that}\]

[44]
\[
\sum_{1}^{n} \frac{n_i}{n} \log \frac{\lambda}{\lambda(i)} > \sum_{1}^{\infty} \frac{n_i}{n} \log \frac{n_i}{n} - \epsilon^2 \log \frac{n}{n}
\]

or

\[
\sum_{1}^{n} n_i \log \frac{n \lambda(i)}{n_i} > -\epsilon^2 \log n
\]

or

\[
\prod_{1}^{n} \left( \frac{n \lambda(i)}{n_i} \right)^{n_i} > n^{-\epsilon^2}
\]
BIBLIOGRAPHY


