Order Number 9010836

Asymptotic cumulative risk and Bayes risk under entropy loss, with applications

Clarke, Bertrand Salem, Ph.D.

University of Illinois at Urbana-Champaign, 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

ASYMPTOTIC CUMULATIVE RISK AND BAYES RISK UNDER ENTROPY LOSS, WITH APPLICATIONS

BY

BERTRAND SALEM CLARKE

B.S., University of Toronto, 1984

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1989.

Urbana, Illinois

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THE GRADUATE COLLEGE

July 27, 1989

WE HEREBY RECOMMEND THAT THE THESIS BY

BERTRAND SALEM CLARKE

ENTITLED ~~ASYMPTOTIC CUMULATIVE RISK AND BAYES RISK~~

UNDER ENTROPY LOSS, WITH APPLICATIONS

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

Andrew R. Barron
Director of Thesis Research

Robert A. Wijsman (for J. Sacks)
Head of Department

Committee on Final Examination†

Andrew R. Barron
Chairperson

Adam T. Martinsek

Zhiliang Ying

Robert A. Wijsman

† Required for doctor's degree but not for master's.

O-517

Abstract

In many areas of application of statistics one has a relevent parametric family of densities and wishes to estimate the density from a random sample. In such cases one can use the family to generate an estimator. We fix a prior and consider the properties of the predictive density as an estiamtor of the density.

We examine the cumulative risk of the estimator and its cumulative Bayes risk under Kullback-Leibler loss. Those two mathematical quantites appear in other contexts with different interpretations. Aside from density estimation, the first quantity occurs in source coding, and hypothesis testing, and the second occurs in source coding, channel coding, and asymptotic convergence of the posterior to a normal. In the first chapter we state our two main results, give some examples, and discuss the applications of our results to those areas.

Our two key results amount to two senses in which the Kullback-Leibler distance between the n-fold product of a distribution in a parametric family and a mixture of such distributions over the parametric family increases as the logarithm of the sample size, provided that in the mixing some mass is assigned near the true distribution. The first is a direct examination of the Kullback-Leibler distance, the second is an examination of the Kullback-Leibler distance after it has again been averaged with respect to the prior. We prove that each is of the form one half the dimension of the parameter times the logarithm of the sample size plus a constant. In both cases the constant is identified.

The key technique for the first result is Laplace integration which gives upper and lower bounds which are asymptotically tight and can be made uniformly good over compact sets in the parameter space. When the parameter space is no longer compact it becomes advantageous to use different techniques in obtaining upper and lower bounds. The convergence holds in an average sense not pointwise uniform. For the upper bound we use an inequality due to Barron (1988) so as to set up an application of the dominated convergence theorem, and for a lower bound we use the fact that the normal has maximal entropy for constrained variance.

Acknowledgements

I wish to express my gratitude to Professor A. R. Barron, my thesis advisor for the last two and a half years. I have greatly appreciated his confidence in me and his enthusiasm. At the same time, I have done my best to learn from his insight and and be instructed by his example. The work contained herein bears the unmistakable impress of his mind and is greatly improved by so doing. Certainly, working under his guidance has been the best part of my graduate career.

Another major influence on me has been Professor J. E. Mittenthal. Our joint work began in 1985, and continues. I am grateful for having had the opportunity to work with him in subjects that I otherwise would never have learned about, and for seeing how theoretical tools are important to problems of real concern.

I would also like to mention Professor R. A. Wijsman: The example he has set for me as a teacher, through the three classes I have taken from him, is my paradigm for how to present statistical ideas properly.

Finally, this work is dedicated to those who discouraged me. Their skepticism motivated me to surpass their expectations, in part resulting in this thesis. I hope that others will be similarly motivated to fight the good fight despite the obstacles they encounter.

Table of Contents

Chapter 1: Applications of the Key Results

1.1 Introduction

Initially this work was motivated by trying to identify the constant term in an asymptotic expansion for the redundancy of a source code. For sample size n, and a d dimensional parameter indexing a family of distributions, Rissanen (1984) gave upper and lower bounds of $(d/2)\log n$, accurate to $o(\log n)$, for the redundancy in a general context. We also narrowed the focus to a parametric setting, but chose to examine the relative entropy distance which represents the redundancy of a Bayes' code. By that extra restriction, and some different techniques, we were able to obtain an expression for the redundancy accurate to o(1). Our result, in this context, is that the redundancy, which we denote $D(P_\theta^n \| M_n)$, where M_n is the mixture with respect to a prior, of the densities in the parametric family of which p_θ is a member, is

$$(d/2)\log n/2\pi e + \log 1/w(\theta) + (1/2)\log\det I(\theta), \qquad (1)$$

plus an error term which goes to zero. We have denoted the prior density by w, and the Fisher information at θ for one observation by $I(\theta)$. There are two hypotheses. One is the finiteness of an expected local supremum of squares of second derivatives of the log-likelihood; the other is that the posterior distribution concentrates on neighborhoods of the true value of the parameter at rate $o(1/\log n)$. An alternate hypothesis involves a restriction on the type of parametrizations which are allowed. We call this concept the soundness of the parametric family. Our bounds are on a more restricted class than were Rissanen's, however they are more accurate. Indeed, we expect that by use of more exacting techniques, such as in Tierney and Kadane (1986), a better expression could be obtained, accurate perhaps to order $O(1/n^2)$.

Although the quantity we examined arose from information theory, it also has a direct statistical interpretation. It is the cumulative risk of the parametric Bayes' estimator under Kullback - Leibler loss, and it is the error exponent of a suitably formulated hypothesis test. The desirability of a more accurate expression stems from this correspondence between cumulative risk and redundancy: one wants to know the risk as well as the cumulative risk. Also, it has been noted by Leonard (1982) and De Groot (1970) that expansions for posterior distributions typically include a $\log n$

term. As the examples of the next section will show our key results fail to account for $1/n$ behavior although work by Haughton (1988) seems to indicate that they are accurate to order $O(1/\sqrt{n})$.

The next step undertaken was to seek conditions under which our approximation, (1), could be made uniformly accurate in the parameter so that we could approximate the integral over the parameter space with respect to a prior density. That mathematical quantity can be interpreted as the Bayes' cumulative risk of the Bayes' density estimator under Kullback-Leibler loss or as the average redundancy of the Bayes' code. This extension approximates those quantities to within o(1). A convenient but weaker version gives O(1) bounds. Thus, we found conditions under which our approximation, (1), to the pointwise redundancy could be averaged with respect to the prior. The resulting quantity, the average redundancy, is mathematically the same as the Shannon mutual information even though they, too, have very different interpretations.

Our result in this case is that the mutual information, $I(\Theta; X^n)$, which is the average of $D(P_\theta^n \| M_n)$ with respect to θ, is

$$(d/2)\log n + H(\Theta) + (1/2)\int w(\theta) \log \det I(\theta) d\theta, \qquad (2)$$

plus an error term which goes to zero with large n. We have used $H(\Theta)$ to denote the entropy of a the parameter as a random variable, and X^n, to denote n repetitions of the experiment. The hypotheses are not quite so neatly summarized because, in our proof that (2) lower bounds $I(\Theta; X^n)$, we used the maximum likelihood estimator, the MLE, requiring it to be consistent and close to the Bayes estimator under squared error loss. Sufficient conditions for those two results to hold can be found in work due to Bickel and Yahav (1969). In the proof that (2) upper bounds $I(\Theta; X^n)$, our key assumption is that the second order Taylor expansion is a uniformly good approximation to the Kullback-Leibler distance between members of the parametric family on the support of the prior, and that our approximation for the Kullback-Leibler number is valid pointwise.

Given a class of Bayes' solutions, which incorporates a minimization, it is natural to ask if one of them achieves the maximal Bayes' risk. Since (2) is an expression for the Bayes' risk which is reasonably accurate, in particular, has a term which is dependent on the prior, we can use some standard reasoning to obtain the minimax estimator, its minimax cumulative risk, and the least favorable prior. In information theoretic terms we have identified a minimax code, and the minimax redundancy.

The rest of this chapter is devoted to examples and applications of the results which are proved later. In the examples we consider the normal family, and a parametric family in exponential form with the natural parametrization and use a result due to Berk (1970) to see that our hypotheses are satisfied. We address density estimation and coding since they were central to the formulation of the problem. Other areas of application we will address are: asymptotic convergence of the posterior to the normal, hypothesis testing and channel capacity. Although we have, for the most part, assumed that the prior has a density with respect to Lebesgue measure, we also consider the case that the prior is discrete, and see that the behavior of the cumulative risk and cumulative Bayes' risk is very different: asymptotically it is a constant dependent only on the prior probability of the true distribution.

Chapter 2 gives a formal proof, under the best hypotheses we could find, of (1), the o(1) asymptotic approximation for the Kullback-Leibler distance between the true density and a mixture of densities with respect to a prior. The technique of proof is to partition the underlying probability space in two ways: one for an upper bound and one for a lower bound. Each partition has two elements: one element which contains those points for which a condition fails and one element which contains those points for which it holds. We show that integration over the set on which either condition fails contributes only a negligible amount. An estimator is introduced for the purpose of proving the result. In earlier versions we used the MLE, or the mode of the posterior. In this version we use one which is a stochastic perturbation of the true value of the parameter. This is not a true estimator since it depends on the estimand. It was introduced by Lehmann (1983) to prove efficiency of Bayes' estimators, but he credits Bickel for the idea. Use of that quantity in this context was first recognized by Barron.

At the end of the chapter we try to show that the hypotheses under which we have proved that (1) closely approximates $D(P_\theta^n \| M_n)$, are not too stringent. This is important because it is not clear when the rate assumption on posterior consistency is satisfied. We introduce the idea of the soundness of a parametrization and show that it is a sufficient condition for the rate assumption. The idea, as explained in Chapter 2, Section 3, is that inferences made from the parametrization should be sound in the sense that a sequence of parameter values converges to a point in the parameter space if and only if the sequence of densities they index converges to the density indexed by the limit point. This ensures that the manifold of densities does not wrap around in any strange ways.

We are able to prove that soundly parameterized families are consistent a posteriori, at the desired rate, by applying the nonparametric work of Kiefer and Wolfowitz (1958), a result due to Schwartz (1965), and a convenient upper bound on the relative entropy due to Barron (1988). Here we restrict attention to random variables taking values in a finite dimensional real space and use the Kolmogorov-Smirnov distence, assuming it to be equivalent to the Euclidean metric on the parameter space. In the last section of Chapter 2, we give another example of the approximation (1), this time for a parametric family of discrete random variables. We do this for the sake of completeness: the other examples are for continuous random variables. We defer this example since it is technically a bit more complicated to work out the approximation without using the theorem. We remark that our results hold for both discrete and continuous random variables.

In Chapter 3 we give a formal proof that the integral of the approximation (1), is a good approximation to the integral of the approximand. We consider two cases: the parameter space is compact, and the parameter space is not compact. In the compact case, we continue the approach taken in Chapter 2 by showing that the approximation for each θ is uniformly accurate over compact sets in the parameter space. This gives general hypotheses, if a bit strong, which we use so as to unambiguously identify the least favorable prior.

In the noncompact case we continue to deal with upper and lower bounds separately. Here we deal with the convergence of the integral itself rather than, as in the compact case, dealing with the convergence of the integrand. Proving that (2) lower bounds $I(\Theta; X^n)$ presents the difficulty of examining the covariance of an estimator, and requiring that it converge to its asymptotic value. One of our hypotheses is that the Bayes' risk under squared error loss be asymptotically $O(1/n)$. This can be inconvenient to verify but is typical of many examples. We expect that the lower bound holds under significantly weaker hypotheses but have not yet proved it. That (2) upper bounds $I(\Theta; X^n)$ is proved without reference to the compactness of the support of the prior. It reduces to the compact case but is different in that it is not pointwise uniform; it, too, deals with the average. It is a limitation of that result that the logarithm of the prior is assumed to be uniformly continuous, for this rules out the normal prior. In the applications sections, we will be drawing our inferences without specifically stating the hypotheses of the results, since they will be identified in the statements later, and this will allow the presentation to be smoother.

1.2 Notation and Statements of Main Results

In the next sections we will be evaluating examples and giving applications of our results. The main quantity which recurs through much of our analysis is the Kullback-Leibler number which when evaluated for distributions P, and Q is

$$D(P \,||\, Q) = \int p(x) \log \frac{p(x)}{q(x)} \lambda(dx),$$

where p, and q are the densities with respect to the dominating measure λ. We take all logarithms to have base e, except in Section 5 where base two is used. Equivalently, we will write the arguments of the Kullback-Leibler number as the densities. Generally, we will assume that a dominated parametric family $\{P_\theta \mid \theta \in \Omega\}$ is given, in which Ω is contained in $\mathbf{R}^d$, and has nonvoid interior. The notation $p(x \mid \theta)$ is used interchangeably with $p_\theta(x)$ to indicate the value of a density. We write $\theta = (\theta_1, \ldots, \theta_d)$ and it will occasionally be convenient to write $D(\theta \,||\, \theta')$ for $D(P_\theta \,||\, P_{\theta'})$. When we think of the parameter θ as a random variable we will denote it by Θ. We assume that Θ has a density with respect to Lebesgue measure and denote that density by w. Often w is called a prior and is used to define a mixture of distributions. The density of the mixture distribution M_n is

$$m_n(x_1, \ldots, x_n) = \int w(\theta) p_\theta(x_1) \cdots p_\theta(x_n) d\theta.$$

Under the true distribution the X_i's are independently and identically distributed, but under the mixture they are only exchangeable. Typically, we will denote the n-fold product of a density or the measure it defines by a superscript n. The n will occasionally be omitted when no ambiguity about the meaning will result. We will often consider a fixed θ_o. When we do we will be assuming that θ_o is in the interior of Ω and that the prior w is continuous almost everywhere, in particular it is continuous, and positive, at θ_o. In addition, we will assume that the prior probability of the boundary of Ω is zero. A sequence of random variables $X_1, \ldots, X_n$, with outcomes $x_1, \ldots, x_n$, will be abbreviated to X^n, and x^n, respectively.

At any point θ in the parameter space, the Fisher information is the $d \times d$ matrix

$$I(\theta) = - \left[E_\theta \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X_1 \mid \theta) \right]_{i,j=1,\ldots,d}.$$

Another concept of information is Shannon's mutual information which we write as

$$I(X; Y) = D(P_{X,Y} \,||\, P_X \times P_Y),$$

for random variables X and Y. We shall be concerned with the case that X is Θ and Y is X^n. For, then we have that

$$I(\Theta; X^n) = \int w(\theta) D(P_\theta^n \| M_n) d\theta,$$

i.e., averaging the Kullback-Leibler number over the parameter space gives the mutual information.

Our two main results are asymptotic expansions for $D(P_\theta^n \| M_n)$ and $I(\Theta; X^n)$. Indeed, because of the averaging relation between them, we hope that, upon integration, a pointwise asymptotic approximation for $D(P_\theta^n \| M_n)$ will give an asymptotic approximation for $I(\Theta; X^n)$. This is in fact what we have shown. First, our pointwise approximation is summarised in the following.

Theorem 1.2.1: *Suppose that $I(\theta_o)$ is positive definite and that there exists a $\xi > 0$ so that for each i,j from 1 to d*

$$E_{\theta_o} \sup_{\{|\theta - \theta_o| < \xi\}} \left| \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x \mid \theta) \right|^2 < \infty, \tag{3}$$

and that the posterior distribution is consistent at rate $o(1/\log n)$. Then

$$D(P_{\theta_o}^n \| M_n) = \frac{d}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \log \det I(\theta_o) + \log \frac{1}{w(\theta_o)} + o(1). \tag{4}$$

We prove the theorem rigorously in Chapter 2.

By posterior convergence at rate $o(1/\log n)$ we mean that for any open set N containing θ_o

$$P_{\theta_o}^n (W(N^c \mid X^n) > \alpha) = o\left(\frac{1}{\log n}\right)$$

for all $\alpha > 0$. Here $W(\cdot \mid X^n)$ is the posterior distribution of Θ given X^n.

We have found various sets of conditions under which integrating the pointwise approximation bounds the mutual information. If the parameter space is compact, denoted K, then we can use moment conditions to show that the error term is uniformly small as a function of θ, so that we can directly integrate with respect to w. When the parameter space is not compact, we use different techniques to obtain lower bounds and upper bounds. We state a version of our result for the compact case. Let $H(X)$ be the entropy of a random variable X,

$$H(X) = \int p(x) \log \frac{1}{p(x)} \lambda(dx),$$

and $N(\theta, \delta)$ denote a neighborhood about θ of radius δ in the Euclidean norm.

Theorem 1.2.2: *Suppose that* $w(\theta)$ *and* $\det I(\theta)$ *are bounded away from zero for* θ *in* K; *for each positive* α *and* δ *we have that*

$$\sup_{\theta \in K} P_\theta(W(N(\theta, \delta)^c \mid X^n) > \alpha) = o(\frac{1}{\log n}), \tag{5}$$

and that for all i *and all* j *there is a* $\delta > 0$ *which satisfies*

$$\sup_{\theta \in K} E_\theta \sup_{\theta' : |\theta - \theta'| < \delta} (\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X_1 \mid \theta'))^4 < \infty. \tag{6}$$

Then we have that (4) is uniformly valid and ,

$$\int w(\theta) D(P_\theta^n \mid\mid M_n) d\theta$$

$$= \frac{d}{2} \log \frac{n}{2\pi e} + \frac{1}{2} \int w(\theta) \log \det I(\theta)\, d\theta + H(\Theta) + o(1). \tag{7}$$

This theorem will be proved rigorously in Chapter 3. We have tacitly assumed that the quantities appearing in (4) and (7) are finite. Note that (6) is a strengthened version of (3), and (5) is a uniform version of the posterior consistency hypothesis of Theorem 1.2.1.

In the sections that follow we will refer back to these two results. Note that the uniformity of (4) is proved only over compact sets. We use this in Sections 4 and 5 in minimaxity arguments, but, elsewhere the extra strength of uniformity is not required. It should be understood, therefore, that outside of those arguments the noncompact case can be included by the same reasoning. The more general validity of (7) is proved in Chapter 3. We have restricted ourselves to the compact case for the sake of exposition.

1.3 Motivational Examples

In the first part of this section we give some examples of Theorem 1.2.1 and in the second part we give an example of Theorem 1.2.2. Where reasonably possible we have evaluated quantities directly so as to see that the result agrees with our theorems. Otherwise we have demonstrated that our theorems give results which are not readily obtainable.

We consider two examples for Theorem 1.2.1. In the first we use a normal parametric family with a normal prior. In this case, we calculate explicitly the same expression as is given by the approximation. After that, we show that the hypotheses of Theorem 1.2.1 are satisfied by any family in exponential form with the natural parametrization.

Let w be a N(0,1) prior for the parameter μ as it appears in a sequence of i.i.d. $N(\mu, 1)$ random variables. We denote the mixture distribution, M_n, with density

$$m_n(x^n) = \int w(\mu)p(x^n \mid \mu)d\mu,$$

and calculate $D(P_\mu^n \mid\mid M_n)$ explicitly. The mixture is

$$m(x^n) = \int_{-\infty}^{\infty} \frac{e^{-\frac{\mu^2}{2}}}{(2\pi)^{(n+1)/2}} e^{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}} d\mu.$$

Expanding the sum in the exponent and completing the square in μ gives

$$-\frac{(n+1)}{2}\mu^2 + \mu n\bar{x} = -\frac{(n+1)}{2}(\mu - \frac{\sum_{i=1}^{n} x_i}{n+1})^2 + \frac{(\sum_{i=1}^{n} x_i)^2}{2(n+1)}.$$

Recognizing the variance as $1/(n+1)$, the mixture is

$$m(x^n) = \frac{e^{-1/2\sum_{i=1}^{n} x_i^2 + (\sum_{i=1}^{n} x_i)^2/2(n+1)}}{\sqrt{n+1}(2\pi)^{n/2}}.$$

Now the Kullback - Leibler distance is

$$D(P_\mu^n \mid\mid M_n) = E_\mu \log \frac{P_\mu^n(X^n)}{M_n(X^n)}$$

$$= \frac{1}{2}\log(n+1) + \frac{1}{2}E_\mu[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n+1} - \sum_{i=1}^{n}(X_i - \mu)^2]$$

$$= \frac{1}{2}\log(n+1) + \frac{n\mu^2}{2(n+1)} - \frac{n}{2(n+1)}. \qquad (8)$$

Theorem 1.2.1 applies since the local supremum condition, (3), is trivial and the rate of convergence is guaranteed by a result due to Berk (1970) which gives an exponential rate. Alternatively, we could argue that, by the conjugacy of the prior we have the required rate of convergence, but that's harder. The approximation

gives

$$D(P_\mu^n \,||\, M_n) = \frac{1}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\mu) + \log\frac{1}{w(\mu)}$$
$$= \frac{1}{2}\log n - \frac{1}{2} + \frac{\mu^2}{2}.$$

We see that the difference between the exact expression and the approximation tends to zero. The log terms are no problem: the difference goes to zero at rate $1/n$ from Taylor series expansions. The remaining terms in (8) tend to $-1/2 + \mu^2/2$ as expected. Note that the difference between the approximation and (8) is of order $1/n$.

Next we look at a more general example: We show that any one parameter exponential family satisfies the hypotheses of Theorem 1.2.1. This actually holds for any finite dimensional exponential family; we assume there is one parameter because the notation is simpler.

Consider the family

$$p(x \mid \eta) = \chi_A e^{\eta T(x) + S(x) + u(\eta)},$$

in which u is the normalizing constant, assumed to be at least twice continuously differentiable and χ_A is the indicator function for the set A. We recall that

$$E_\eta T = -u'(\eta),$$

and that the Fisher information is independent of the data:

$$i_{11}(\eta) = -u''(\eta).$$

Since u is twice continuously differentiable, the expected supremum condition (3) holds. The posterior consistency condition again holds: Berk (1970) showed that an exponential rate holds for families in exponential form.

Now we have a large class of examples to which Theorem 1.2.1 applies. In practice, the result of Theorem 1.2.1 may appear to be more accurate than the direct calculations we have made in certain examples. This will be seen in Chapter 2, Section 4 where, as in the next example, we will implicitly be evaluating only one term, the second, in the right hand side of the decomposition

$$D(P_\theta^n \,||\, M_n) = E_\theta \log\frac{p_\theta(X^n)}{p_{\hat\theta}(X^n)} + E_\theta \log\frac{p_{\hat\theta}(X^n)}{m(X^n)}, \tag{9}$$

since it is that term which captures the logarithmic dependence on the sample size. We evaluate the second term without justifying the steps because the theorem will

guarantee that the result is valid. Despite that, we believe that a justification for how we approximated that term can be given. We remark that the first term tends to $-d/2$, a fact proved in Clarke and Barron (1988), a fact which is plausible since a Taylor expansion about the MLE shows that the term looks like the expected value of $-1/2$ times a Chi-square random variable with d degrees of freedom.

Next we turn to an example in which we are concerned with the average of the earlier approximation. Here, we give an example of Theorem 1.2.2 which assumes compactness of the parameter space. At the end of Chapter 3 we will examine the same distributions as we did for Theorem 1.2.1 and see that, as before, explicit evaluation of mixture densities is difficult. Where they can be directly evaluated, our examples do verify our results.

Consider a $Beta(q,q)$ prior on a sequence of Bernoulli (p) random variables with p as the true value. The prior was chosen not for its conjugacy but rather for the fact that the mixture comes out as a ratio of Beta functions and we can obviously integrate the logarithm of the Fisher information of a Bernoulli against the integrand of the Beta function.

The mixture density is

$$m(x^n) = \frac{B(\sum_{i=1}^{n} x_i + q,\ n - \sum_{i=1}^{n} + q)}{B(q,q)}$$

$$= \frac{\Gamma(\sum_{i=1}^{n} x_i + q)\Gamma(n - \sum_{i=1}^{n} + q)}{\Gamma(n + 2q)B(q,q)}.$$

The Kullback - Leibler number is

$$E_p \log \frac{P_p(X^n)}{m(X^n)} = E_p \log \frac{B(q,q)\Gamma(n+2q)p^{\sum_{i=1}^{n} X_i}(1-p)^{n - \sum_{i=1}^{n} X_i}}{\Gamma(\sum_{i=1}^{n} X_i + q)\Gamma(n - \sum_{i=1}^{n} + q)},$$

which, if we write $\sum_{i}^{n} X_i \sim np$, is approximated by

$$-nH(p) + \log B(q,q) + \log (n + 2q - 1)!$$

$$- \log (np + q - 1)! - \log (n(1-p) + q - 1)!\,.$$

We apply Stirling's formula to the factorials and simplify to obtain

$$\frac{1}{2}\log \frac{n}{2\pi} + \log B(q,q) + (q - 1/2)\log \frac{1}{p(1-p)}.$$

Integration over the parameter space with respect to the Beta(q, q) prior distribution gives the approximation

$$I(\theta, X^n) \sim \frac{1}{2}\log\frac{n}{2\pi} + \log B(q, q) + (q - 1/2)\int_0^1 \frac{p^{q-1}(1-p)^{q-1}}{B(q, q)} \log\frac{1}{p(1-p)}dp.$$

If we apply the formula then we get

$$I(\theta, X^n) \sim \frac{1}{2}\log\frac{n}{2\pi e} + H(Beta(q, q)) + \frac{1}{2}\int_0^1 \frac{p^{q-1}(1-p)^{q-1}}{B(q, q)} \log\frac{1}{p(1-p)}dp.$$

And the entropy of a Beta(q, q) is

$$H(Beta(q, q)) = \log B(q, q) + (q - 1)\int_0^1 \frac{p^{q-1}(1-p)^{q-1}}{B(q, q)} \log\frac{1}{p(1-p)}dp.$$

So our approximation differs from the direct calculation by -1/2, the contribution of the Chi-square term in (9). Our formula is valid since the bound from Theorem 1.2.2 applies: (6) is obviously true and it can be shown that (5) is satisfied by continuity considerations. Strictly speaking, we only have the result on compact subsets of (0, 1) because the Fisher information is unbounded at 0 and at 1. However, this can be overcome by results to be presented in Chapter 3, Sections 2 and 4.

We defer further consideration of examples, such as those with noncompact parameter spaces, until the end of Chapter 3, by which time we will have proved some of the results which are required.

1.4 Applications to Density Estimation

The Kullback-Leibler number has several properties which make it a natural choice as a loss function in the decision theory framework. Chief amongst these are the following: On parametric families it locally approximates squared error loss; it induces convex neighborhoods; it satisfies Pythagorean relations even though it is not a metric; it is nonnegative and equals zero only when its arguments are equal; and, it has a natural interpretation as the redundancy of a source code.

Suppose we are given a parametric family indexed by θ and that θ_o is the true value of the parameter. However, suppose that it is not the parameter 'per se' that interests us. Rather, we are using the parametric family so as to identify the true density which is p_{θ_o}. One natural estimator of $p(x \mid \theta_o)$ at any given x is the

mixture of the densities with respect to the posterior distribution,

$$\hat{p}_n(x; X^n) = \int_\Omega p_\theta(x) w(\theta \mid X^n) d\theta,$$

that is, the posterior mean of $p(x \mid \Theta)$. Observe that this estimator is the predictive density

$$\hat{p}_n(\cdot) = m(\cdot \mid X^n),$$

where $m(\cdot \mid X^n = x^n)$ is the conditional density of X_{n+1} given $X^n = x^n$.

We use the Kullback-Leibler number as the loss function for parametric density estimation and examine the behavior of the cumulative risk. Let δ_k for $k = 0,\ldots,n-1$ be a sequence of density estimators. Each δ_k estimates the density of X_{k+1}, given the data X^k. Here, δ_o is a fixed density function not dependent on the data. When θ_o is true, the risk associated with $\delta_k = \delta_k(X^k)$ is

$$E_{\theta_o} D(P_{\theta_o} \| \delta_k),$$

and we denote the cumulative risk of n uses of an estimator δ_k for $k = 0,\ldots,n-1$ by $C(n, \theta_o, \delta)$. It is the sum of the individual risks:

$$C(n, \theta_o, \delta) = \sum_{k=0}^{n-1} E_{\theta_o} D(P_{\theta_o} \| \delta_k).$$

The sum of the Kullback-Leibler risks is seen to play an important role in some of the applications. Just as the posterior mean of Θ is the Bayes' estimator under squared error loss it turns out that the posterior mean of $p(x \mid \Theta)$ is the Bayes' estimator under relative entropy loss. Adapting a result due to Aitchison (1975), we have the following.

Proposition 1.4.1: *$\hat{p}_n$ is the Bayes' estimator of the density function. The cumulative risk of this estimator is*

$$C(n, \theta, \hat{p}_n) = \sum_{k=0}^{n-1} E_{\theta_o} D(p_{\theta_o} \| \hat{p}_k) = D(P^n_{\theta_o} \| M_n),$$

under the convention that $\hat{p}_0(x) = m_1(x_1)$. Consequently, under the conditions of Theorem 1.2.1, the cumulative risk is approximated by $(d/2)\log n + c$, and the average risk $(1/n)\sum E_{\theta_o} D(p_{\theta_o} \| \hat{p}_k)$ converges to zero at rate $(\log n)/n$.

Proof: The information inequality, $D(p \| q) \geq 0$, with equality if and only if $p = q$, implies that $\hat{p}_n$ is the Bayes' estimator, since, for any other density q, the posterior average of the risk is seen to equal

$$\int_\Omega D(p_\theta \| q) w(\theta \mid X^n) d\theta = \int_\Omega w(\theta \mid X^n) D(p_\theta \| \hat{p}_n)\, d\theta + D(\hat{p}_n \| q).$$

So, we see that the minimum is achieved when the second term is zero, i.e. , when $q = \hat{p}_n$.

By Bayes' rule, $\hat{p}_n$ equals the predictive density, which is

$$m(X_{n+1}=x_{n+1} \mid X^n) = \frac{m_{n+1}(X^n, x_{n+1})}{m_n(X^n)}.$$

Since $\hat{p}_k(x_{k+1}) = m(x_{k+1} \mid X^k)$ is the predictive density, and a sum of logarithms is the logarithm of the product of their arguments, we have, for $m(x^0)$ taken to be identically one that

$$\begin{aligned}\sum_{k=0}^{n-1} E_{\theta_o} D(P_{\theta_o} \| \hat{P}_k) &= \sum_{k=0}^{n-1} E_{\theta_o} \log \frac{p(X_{k+1} \mid \theta_o)}{m(X_{k+1} \mid X^k)} \\ &= E_{\theta_o} \log \prod_{k=0}^{n-1} \frac{p(X_{k+1} \mid \theta_o)}{m(X_{k+1} \mid X^k)} \\ &= E_{\theta_o} \log \frac{p(X^n \mid \theta_o)}{m(X^n)},\end{aligned}$$

which is $D(P^n_{\theta_o} \| M_n)$. □

We remark that under the conditions of Theorem 1.2.1 the individual risk terms $E_{\theta_o} D(P_{\theta_o} \| \hat{P}_n)$ also converge to zero as $n \to 0$. This follows from noting that

$$E_{\theta_o} D(P_{\theta_o} \| \hat{P}_n) = D(P^n_{\theta_o} \| M_n) - D(P^{n-1}_{\theta_o} \| M_{n-1}),$$

and applying Theorem 1.2.1 to each term on the right hand side. Thus, the predictive density is consistent for the true density in expected Kullback-Leibler distance.

Parameter estimation can be regarded as a special case of density estimation in which we restrict the estimator of the density to be of the form $p(x \mid \theta(X^n))$. In the present context we have not restricted the class of estimators in this way. We have used the parametric family as a tool to generate an estimator, relinquishing any information from the family about what the true value of the parameter is. By enlarging the class of estimators we see that in terms of global optimality properties, the Bayes' risk in parametric density estimation lower-bounds the Bayes' risk in parametric estimation:

$$\inf_{\delta} E_w E_\theta D(\theta \| \delta) \geq \inf_{Q} E_w E_\theta D(P_\theta \| Q).$$

Similarly, for the maximin risk we have

$$\sup_{w} \inf_{\delta} \int w(\theta) E_\theta D(\theta \| \delta) d\theta \geq \sup_{w} \inf_{Q} \int w(\theta) E_\theta D(P_\theta \| Q) d\theta,$$

and for the minimax risk we have

$$\inf_{\delta} \sup_{\theta} E_\theta D(\theta \| \delta) \geq \inf_{Q} \sup_{\theta} E_\theta D(P_\theta \| Q) ,$$

where δ is an estimator of the parameter, Q is an estimator of the density and $D(\theta \| \delta) = D(P_\theta \| P_\delta)$ is the relative entropy loss for parameter estimation. The quantity we have approximated therefore gives an asymptotic lower bound on the Bayes' risk of parameter estimation, and in the next result we give implications for the other two global optimality criteria.

Suppose that the support of w is contained in a compact set K and its density is positive there. The minimax cumulative risk is defined to be

$$R_n = R(n, K, \{P_\theta\}) = \inf_{Q_n} \sup_{\theta \in K} D(P_\theta^n \| Q_n),$$

and the maximin cumulative risk is defined to be

$$\begin{aligned} R_n^* = R^*(n, K, \{P_\theta\}) &= \sup_{w} \inf_{Q^n} D(P_\theta^n \| Q^n) \\ &= \sup_{w} D(P_\theta^n \| M_n). \end{aligned}$$

We can easily give an upper bound on the limit of R_n and a lower bound on the limit of R_n^*. This was motivated by the observation that the Bayes' risk of the Bayes' estimator can be rearranged to give

$$\frac{d}{2}\log\frac{n}{2\pi e} + \log c - D(w \| \mu) + o(1),$$

which is minimized when we choose

$$w(\theta) = \frac{\sqrt{\det I(\theta)}}{c},$$

and

$$c = \int_K \sqrt{\det I(\theta)}\, d\theta.$$

This is Jeffreys' prior, see Jeffreys (1967). Jeffreys' prior gives that, asymptotically, the risk is constant as a function of θ, suggesting that Jeffreys' prior is least favorable.

Proposition 1.4.2: *Under the hypotheses of Theorem 1.2.2 we have that*

$$\begin{aligned} \lim_{n \to \infty} \left[R_n - \frac{d}{2}\log n \right] &= \lim_{n \to \infty} \left[R_n^* - \frac{d}{2}\log n \right] \\ &= \frac{d}{2}\log\frac{1}{2\pi e} + \log c. \end{aligned}$$

Therefore, asymptotically, the minimax estimator is the mixture with respect to the Jeffreys' prior which is least favorable.

Proof: We obtain an upper bound on the minimax risk by the uniform result proved in Chapter 3, Section 3. We have that

$$R_n - \frac{d}{2}\log n \leq \sup_{\theta \in K} [\, D(P_\theta^n \,||\, M_n) - \frac{d}{2}\log n \,],$$

and if we choose w to be Jeffreys' prior then the right hand side is upper bounded by the quantity given in the proposition, uniformly in θ, by Theorem 1.2.2. For the second maximin risk we have that

$$R_n^* - \frac{d}{2}\log n \geq \int w(\theta) D(P_\theta \,||\, M_n) d\theta - \frac{d}{2}\log n.$$

So, the averaged result from Theorem 1.2.2 is enough. For Jeffreys' prior we have that the right hand side is lower bounded by the stated quantity. Since the minimax cumulative risk is greater than or equal to to maximin cumulative risk, see Ferguson (1967) pg. 81, we have an upper bound on the greater quantity which is the same as the lower bound on the lesser one, thus proving the proposition. □

If we now consider the individual risks and estimate P_θ by the predictive density then we have risk

$$E_{\theta_o} D(P_{\theta_o} \,||\, \hat{P}_k),$$

which gives cumulative risk

$$D(P_{\theta_o}^n \,||\, M_n) = \sum_{k=0}^{n-1} E_{\theta_o} D(P_{\theta_o} \,||\, \hat{P}_k).$$

Now we see that the cumulative Bayes' risk is

$$\int D(P_\theta^n \,||\, M_n) w(\theta) d\theta = \sum_{k=0}^{n-1} \int E_{\theta_o} D(P_\theta \,||\, \hat{P}_k) w(\theta) d\theta,$$

from which the Bayes' risk of the Bayes' estimator is seen to be

$$\int E_\theta D(P_\theta \,||\, \hat{P}_k) w(\theta) d\theta = o(1).$$

We have now shown that the cumulative risk and the cumulative Bayes' risk increase as $(d/2)\log n$. These results suggest that the "individual risks" behave like $d/(2n)$. This parallels the work of Bickel and Yahav (1969) on parameter estimation. They characterized the almost sure asymptotic behavior of the Bayes posterior risk, to essentially arbitrary precision, in terms of the behavior of the loss function near zero. In the case of squared error loss they obtained behavior of the form

$(1/n)Var(\theta)$. Also, Cencov (1981) obtained $d/(2n)$ as the leading term in an asymptotic expansion for $E\ D(P_{\theta_o} || P_{\hat{\theta}})$, the risk of the MLE $\hat{\theta}$, which was accurate to order $n^{-3/2}$.

To end this section we consider some other convergences which were studied by McCulloch (1986). In particular we will see that the difference between two predictive densities, with respect to different priors, converges to zero. First suppose that the true distribution is M_n, a mixture of independent and identical distributions and that we estimate by another mixture, N_n based on the prior v which has the same support as w. Then the Kullback-Leibler distance between them is

$$D(M_n || N_n) = \int_K D(P_\theta || N_n) w(\theta) d\theta - \int_K D(P_\theta || M_n) w(\theta) d\theta$$
$$= D(w\ || v) + o(1),$$

as n increases by applying Theorem 1.2.2 to each term. If the predictive distribution based on v is denoted by $\hat{Q}_k$ then by direct calculation we have that

$$E_M D(\hat{P}_k || \hat{Q}_k) = D(M_{k+1} || N_{k+1}) - D(M_k || N_k).$$

So, as $k \to \infty$ we see that $E_M D(\hat{P}_k || \hat{Q}_k)$ tends to zero, which means that except for θ in a set of arbitrarily small prior measure, we have $E_\theta D(\hat{P}_k || \hat{Q}_k)$ tending to zero in P_θ probability. We obtain similar behavior for the posteriors:

$$E_M D(w(\cdot | X^n) || v(\cdot | X^n)) = D(w || v) - D(M_n || N_n) = o(1),$$

so, we have that for $k \geq 1$,

$$E_\theta D(w(\cdot | X^n) || v(\cdot | X^n)) \to 0,$$

in the joint probability for X^n and θ. Also, we have that

$$D(w(\cdot | X^n) || v(\cdot | X^n)) \to 0,$$

in the joint probability for X^n and θ. From the recursion relation we see that

$$D(w\ || v) = \sum_{l=0}^{\infty} E_M D(\hat{P}_l || \hat{Q}_l),$$

where, under our convention $E_M\ D(\hat{P}_0 || \hat{Q}_0) = D(M_1 || N_1)$, so the number of times $E_M D(\hat{P}_n || \hat{Q}_n)$ exceeds $1/n$ must have negligible cumulative effect.

The formula we have proved, in Theorem 1.2.1, for the relative entropy assumes that P_{θ_o} is the true density. If the mixture is the true density then estimating with an element of the parametric family is a poor strategy. We see that if the prior v is unitmass at a point θ in the support of w, then the above formula shows

that

$$D(M_n \,||\, P^n_{\theta_o}) = n \int D(\theta \,||\, \theta_o) w(\theta) d\theta - I(\Theta; X^n),$$

that is, the loss increases at rate n no matter what estimator we use, since we know the second term on the right hand side behaves like $(d/2)\log n$.

1.5 Applications to Universal Noiseless Source Coding.

Suppose that X is a discrete random variable whose distribution is in the parametric family $\{P_\theta \mid \theta \in \Omega\}$, and we want to encode a block of data for transmission. It is known that a lower bound on the expected codelength is the entropy of the distribution. Moreover, this entropy bound can be achieved, within one bit, when the distribution is known. Universal codes have expected length near the entropy no matter which member of the parametric family is true. The redundancy of a code is defined to be the difference between its expected length and the entropy.

The problem of providing a noiseless source code for a block of data $X^n=(X_1, \ldots, X_n)$, has been studied extensively, for instance Davisson (1973). Recall that if

$$\phi: \mathbf{X}^n \rightarrow \{0,1\}^*$$

is a uniquely decodeable code with codelengths $l(\phi(X^n))$, where the asterisk indicates the set of all finite length strings of elements of the set, then

$$Q_n(X^n) = 2^{-l(\phi(X^n))}$$

defines a subprobability mass function on $\mathbf{X}^n$, by the Kraft-McMillan inequality. Moreover, for any subprobability mass function $Q_n(X^n)$ for which $-\log Q_n(X^n)$ takes integer values, a uniquely decodeable code exists with those lengths. The redundancy of a code $\Phi = \{\phi(X^n) \mid X^n \in \mathbf{X}^n\}$ is the difference between the expected length of a message under ϕ, and the expected length of a message if we knew the true distribution:

$$R_n(\Phi, P_{\theta_o}) = E[l(\phi(X^n)) - \log(\frac{1}{P_{\theta_o}(X^n)})]$$

$$= E\left[\log\left(\frac{1}{Q_n(X^n)}\right) - \log\left(\frac{1}{P_{\theta_o}(X^n)}\right)\right]$$

$$= D(P_{\theta_o}^n \,||\, Q_n),$$

where the logarithm is taken base 2. Thus the redundancy is the Kullback - Leibler number. We want to choose l so as to minimize the redundancy. Among all sub-probability mass functions Q, the one which minimizes the average of $D(P_{\theta_o}^n \,||\, Q_n)$ with respect to a prior w is the mixture M_n. Thus $D(P_{\theta_o}^n \,||\, M_n)$ is referred to as the redundancy of the Bayes' code. The idealized lengths $\log 1/M_n(X^n)$ may violate the constraint of being integer valued. Nevertheless, the Shannon code based on M_n, i.e., the one with code lengths

$$l(\phi(X^n)) = \left\lceil \log \frac{1}{M_n(X^n)} \right\rceil,$$

has redundancy within 1 bit of $D(P_{\theta_o}^n \,||\, M_n)$.

The concepts of noiseless source coding of discrete data may also be applied to the case of continuous random variables which are arbitrarily finely quantized. In the sense made clear by the following proposition, the relative entropy remains the redundancy for nondiscrete sources. If a noiseless code is specified for every finite quantization of a nondiscrete source, we define the redundancy of that source to be the supremum of the redundancies over all such quantizations.

Proposition 1.5.1: *For a nondiscrete source, the redundancy of the Shannon code based on M_n is $D(P_{\theta_o}^n \,||\, M_n)$, to within one bit. Thus the redundancy of the Bayes code is given asymptotically by*

$$\frac{d}{2}\log\frac{n}{2e\pi} + \frac{1}{2}\log\det I(\theta_o) - \log w(\theta_o),$$

under the conditions of Theorem 1.2.1.

Proof: For any finite partition π, of $\mathbf{X}^n$, we can specify a code book Φ, by use of the Shannon code based on the probability measure restricted to π. For the Shannon code we have an explicit codelength formula:

$$l(\Phi_{n,\pi}(A)) = \left\lceil \log \frac{1}{Q_n(A)} \right\rceil,$$

and the redundancy is:

$$R_{\pi,n}(\Phi_n, P_\theta) = \sum_{A\in\pi} l(\phi_n(A))P_\theta^n(A) - P_\theta^n(A)\log\left(\frac{1}{P_\theta^n(A)}\right).$$

So, to within one bit, the redundancy on the partition is the discrete divergence $\sum_{A \in \pi} P_\theta^n(A) \log P_\theta^n(A)/Q_n(A)$. Taking the supremum over all possible partitions gives $D(P_\theta^n \| Q_n)$, by using a well known theorem, see Kullback, Keegel and Kullback (1980), pg. 6-7. If Q_n is replaced by M_n, then we get the Bayes code, and the result is the asymptotic least upper bound on the redundancy. □

Rissanen (1984) showed that for any code $(d/2)\log n - o(\log n)$ is an asymptotic lower bound on the redundancy for (Lebesgue) almost every θ in the family. Also, he showed that for particular codes based on his minimum description length criterion, a redundancy of order $(d/2)\log n + c_\theta$ is achieved although he did not attempt to optimize the constant. For a discussion of the best constants in Rissanen's framework of two stage codes, see Barron and Cover (1989). The optimum code according to the criteria of minimaxity or minimum average redundancy is not a two stage code of the type originally considered by Rissanen, or by Barron and Cover, rather it is a one stage code based on a mixture M_n, where the choice of prior in the mixture is determined by the criterion. Rissanen (1987) also considers codes based on mixtures, however, he does not identify the constant in the expression of the redundancy.

The most stringent hypothesis in Rissanen (1984) is that the maximum likelihood estimator $\hat{\theta}$ be asymptotically normal. Our hypotheses are about as strong. Sufficient conditions for the asymptotic normality of $\hat{\theta}$ are given by Lehmann (1983) pg. 429-430, and Cramer (1946) pg. 500-501. While we have assumed a bound on the expected supremum of the squares of the second derivatives, both Lehmann and Cramer assume a bound on the expected supremum of the absolute values of the second and third derivatives. We have used a higher moment rather than a higher derivative.

Since we cannot know which member of the parametric family is the true density and we still want to know how well the Shannon code based on M_n performs, we can evaluate the average redundancy. The minimal average redundancy is

$$\int D(P_\theta^n \| M_n)\, w(\theta) d\theta = \inf_Q \int D(P_\theta^n \| Q_n) w(\theta) d\theta,$$

where Q varies over all subprobability mass functions which can be used to generate a code. By definition, it is the Bayes' code which achieves the minimal average redundancy. If we next maximize over possible choices of w then we obtain the maximin redundancy, which we will denote by R_n^*.

Another global optimality criterion is that our coding strategy minimize the maximal redundancy R_n which is

$$R_n = \inf_Q \sup_{\theta \in K} D(P_\theta^n \,||\, Q_n).$$

By definition a minimax code is a code which achieves the minimax redundancy.

Theorem 1.2.2 gives that, essentially, the minimal average redundancy is

$$\int D(P_\theta \,||\, M_n) w(\theta) d\theta = I(\theta; X^n)$$

$$= \frac{d}{2} \log \frac{n}{2\pi e} - D(w \,||\, \mu(K)) + \log c + o(1).$$

in the compact case. Theorems 3.2.1 and 3.4.1 extend that result to the non-compact case. We summarize the implications of our results for minimax and maximin coding in the following proposition.

Proposition 1.5.2: *Under the hypotheses of Theorem 1.2.2 we have that*

$$\lim_{n \to \infty} [\, R_n - \frac{d}{2} \log n \,] = \lim_{n \to \infty} [\, R_n^* - \frac{d}{2} \log n \,]$$

$$= \frac{d}{2} \log \frac{1}{2\pi e} + \log c,$$

where, as in Section 4,

$$c = \int_K \sqrt{\det I(\theta)}\, d\theta.$$

Therefore, the Shannon code based on the mixture with respect to Jeffreys' prior has redundancy within one bit of the minimax redundancy.

Proof: The proof is the same as for Proposition 1.4.2. We have merely changed the physical interpretation of the quantities. □

Davisson and Garcia (1980) used the minimax theorem, see Ferguson (1967) pg. 85, to examine the minimax redundancy in a special case. We were unable to see that in our case the hypotheses of the minimax theorem were satisfied, so we resorted to a direct examination of the quantities.

1.6 Applications to Posterior Convergence

The line of reasoning behind the lower bound on the mutual information is based on the intuition that the Bayes' estimator is asymptotically efficient. Having

proved the result we take the intuition one step further by showing that the posterior distribution for the parameter given the data is asymptotically normal with parameters the posterior mean and posterior variance. To do so it is the upper bound that we need, coupled with the representation used in the lower bound.

The result, asymptotic normality of Bayes' estimators, is not new. Indeed, our proof here is typical in that we have used MLE reasoning to deduce a Bayesian result as in Bickel and Yahav (1969), or Ibragimov and Hasminskii (1980). What makes it interesting is that the mode of convergence, expected Kullback-Leibler distance, is either stronger than other modes of convergence which have been used, or noncomparable with them. Here we assume that the parameter space is d dimensional real space since the normal is supported on a real space and the support of the limiting distribution must include the support of the posterior for the relative entropy to be well defined.

One of the hypotheses of Theorem 3.2.1 is that Theorem 1.2.1 hold for each point in the parameter space. One of its hypotheses was Bayes consistency. Thus we have found an extra set of conditions which imply asymptotic normality.

Proposition 1.6.1: *If we have that*

$$E_M \log \det n \operatorname{cov}(\theta \mid X^n) \to \int w(\theta) \log \det I(\theta)^{-1} \, d\theta, \tag{10}$$

then we have that the posterior distribution conditioned on the data converges to a normal with mean $E(\theta \mid X^n)$ *and covariance* $\operatorname{cov}(\theta \mid X^n)$, *in expected Kullback - Leibler distance, i.e.,*

$$E_M D(P_{\theta \mid X^n} \mid\mid N(E(\theta \mid X^n), \operatorname{cov}(\theta \mid X^n))) \to 0,$$

if and only if the conclusion of Theorem 1.2.2 holds.

Remark: Sufficient conditions for (10) are explored in Chapter 3.

Proof: Let $\Phi = \Phi_{\theta \mid X^n}$ denote a normal random variable with mean $E(\theta \mid X^n)$ and variance matrix $\operatorname{cov}(\theta \mid X^n)$. By the definition of the mutual information

$$\begin{aligned} I(\Theta; X^n) &= H(\Theta) - H(\Phi \mid X^n) + [H(\Phi \mid X^n) - H(\Theta \mid X^n)] \\ &= H(\Theta) - \frac{1}{2} E_M \log (2\pi e)^d \det \operatorname{cov}(\theta \mid X^n) \\ &\quad + E_M D(w(\cdot \mid X^n) \mid\mid N(E_{w(\cdot \mid X^n)}\theta, \operatorname{cov}_{w(\cdot \mid X^n)}\theta)), \end{aligned}$$

since Φ and $\theta \mid X^n$ have the same first two moments. By rearranging the expression we find that

$$E_M D(w(\cdot \mid X^n) \mid\mid N(E_{w(\cdot \mid X^n)}\theta, \operatorname{cov}_{w(\cdot \mid X^n)}\theta))$$

$$= I(\theta; X^n) - H(\Theta) + \frac{1}{2}E_M \log (2\pi e)^d \det \operatorname{cov}(\theta | X^n),$$

which tends to zero by the assumptions. □

The above proposition demonstrates the equivalence of a tight asymptotic upper bound for $I(\Theta; Xn)$ with posterior normality of the prior.

As a mode of convergence the expected Kullback-Leibler number is quite strong. It dominates both L^1 and Hellinger distance, see Csiszar (1967). Thus we have proved the asymptotic normality of the posterior in the sense that

$$\lim_{n \to \infty} E_{M_n} | w(\theta | X^n) - \phi_{E(\theta | X^n),\, \operatorname{cov}(\theta | X^n)}(\theta) |_1 = 0,$$

which means that except for θ in a set of arbitrarily small measure, the same result holds with expectation defined by p_θ.

It was remarked in Barron (1988) that the tilted prior

$$w^*(\theta) \equiv \frac{w(\theta)e^{-nD_\theta}}{c_n},$$

where the constant c_n is the weight factor which makes w^* integrate to 1 and $D_\theta = D(P_{\theta_o} || P_\theta)$, is an interesting approximation to the posterior distribution $w(\theta | X^n)$ in that it is near normal for θ near θ_o, yet gives the right large deviations approximations, to first order in the exponent, for all θ. He also noted that the chain rule for the Kullback - Leibler number gives

$$D(P^n_{\theta_o} || M^n) + E_{\theta_o} D(w^* || w(\cdot | X^n)) = -\log \int_{\mathbf{R}^m} e^{-nD_\theta} w(\theta) d\theta, \qquad (11)$$

and gave some convergence properties of the terms. Use of our approximations can improve on those results.

Proposition 1.6.2: *If the conditions of Theorem 1.2.2 hold, then*

$$E_\theta D(w^* || w(\cdot | X^n)) \to d/2,$$

uniformly for θ *in a compact set K, and if the conditions of Theorem 1.2.1 hold for each* θ *in* $\mathbf{R}^d$ *then*

$$\liminf_{n \to \infty} \int_{\mathbf{R}^d} w(\theta) E_\theta D(w^* || w(\cdot | X^n)) d\theta \geq d/2.$$

Proof: Equation (11) holds for each θ. By use of Theorem 3.3.1 the first term is exactly characterized and by Lemma 3.3.2 the right hand side is also exactly characterized. The difference is d/2. This proves the first claim. The second claim follows by Fatou's Lemma. □

The $d/2$ discrepancy between the posterior and its approximation arises because the approximation does not depend on the data and so does not track it. The approximation captures the effect of the second term in (9) but not the first. In later work we hope to explore this further.

1.7 An Application to Hypothesis Testing.

It is well known that the likelihood ratio test statistic converges in distribution to 1/2 times a Chi-square random variable with d degrees of freedom. i.e. ,

$$\log \frac{p(X^n \mid \hat{\theta})}{p(X^n \mid \theta_o)} \rightarrow \frac{1}{2}\chi_d^2 ,$$

in law, where $\hat{\theta}$ denotes the maximum likelihood estimate, the MLE, and χ_d^2 is a Chi-square (d) random variable, see Wilks (1962), Chernoff (1954), and it has been proved that its asymptotic expected value is essentially $d/2$, see Clarke and Barron (1988). This result accounts for the missing $-d/2$ in the examples. An analogous result requiring fewer hypotheses can be proved for the statistic $\log m(X^n)/p(X^n \mid \theta)$. We consider a centered version of this statistic obtained by subtracting its mean under the distribution $P_{\theta_o}^n$.

Proposition 1.7.1: *If the assumptions of Theorem 1.2.1 are satisfied, then for X^n distributed according to $P_{\theta_o}^n$,*

$$\log \frac{m_n(X^n)}{p(X^n \mid \theta_o)} + D(P_{\theta_o}^n \mid\mid M_n) \rightarrow \frac{1}{2}(\chi_d^2 - d),$$

in distribution.

Proof: Let $l_n'(\theta) = (1/n)\nabla \log p(x^n \mid \theta)$. We note that by Proposition 2.2.1 and Theorem 2.2.1 the difference of interest has bounds of the following form; valid for a set of P_θ^n probability tending to one as n goes to infinity:

$$-\eta + \frac{n}{2(1+\varepsilon)} l_n'(\theta) I^{-1}(\theta) l_n'(\theta) - \frac{d}{2}$$

$$\leq \log \frac{m(X^n)}{p(X^n \mid \theta)} + D(P_\theta^n \mid\mid M_n)$$

$$\leq \eta + \frac{n}{2(1-\varepsilon)} l_n'(\theta) I^{-1}(\theta) l_n'(\theta) - \frac{d}{2},$$

where η is any small positive number. For X^n distributed according to $P_{\theta_o}^n$ we have that $nl'_n(\theta)I^{-1}(\theta)l'_n(\theta)$ converges in law to a χ_d^2. So, the proposition follows. □

We use that convergence to identify the critical value and the average power of a hypothesis test. Consider testing H: P_{θ_o} versus K: P_θ, $\theta \neq \theta_o$. We constrain the probability of type 1 error to be less than $\alpha_1 \in (0, 1)$, and examine the performance of tests in terms of the probability of type 2 error averaged with respect to a prior density $w(\theta)$ over the class of alternatives K. Let $c(\alpha)$ be the $1-\alpha$ quantile of a centered Chi-square random variable with d degrees of freedom, i.e., $P(\chi_d^2 - E\chi_d^2 > c) = \alpha$. The Bayes' optimal test is defined so as to minimize the average probability of error. By a familiar argument the problem is seen to reduce to a simple versus simple test for $P_{\theta_o}^n$ versus M_n, so the optimal test compares the test statistic $\log m_n(x^n)/p(x^n \mid \theta_o)$ to a critical value $t = t_n(\alpha_1)$. The following proposition shows how to select the critical value in practice. Specifically, Theorem 1.2.1 gives a convenient approximation to it. Moreover, the average power of the test is shown to be related to $D(P_{\theta_o}^n \mid\mid M_n)$.

Proposition 1.7.2: *Under the hypotheses of Theorem 1.2.1, the asymptotic level* α_1 *critical value for the Bayes' test is* $D(P_{\theta_o}^n \mid\mid M_n) - \frac{1}{2}c(\alpha_1)$ *and the optimal average probability of type 2 error is, to within a constant factor dependent only on* α_1,

$$\alpha_2 \doteq e^{-D(P_{\theta_o}^n \mid\mid M_n)}$$

$$\doteq \frac{n^{-\frac{d}{2}}(2\pi e)^{\frac{d}{2}} w(\theta_o)}{\sqrt{\det I(\theta_o)}},$$

in the sense that there exists a bounded interval $[\, L(\alpha_1), U(\alpha_1) \,]$ *such that every test with type 1 error less than or equal to* α_1 *satisfies*

$$\liminf_{n \to \infty} [\, \log \alpha_2 + D(P_{\theta_o}^n \mid\mid M_n) \,] \geq L(\alpha_1), \tag{12}$$

and there exists a test with type 1 error α_1 *for which the upper bound*

$$\limsup_{n \to \infty} [\, \log \alpha_2 + D(P_{\theta_o}^n \mid\mid M_n) \,] \leq U(\alpha_1), \tag{13}$$

holds. The functions L and U can be expressed in terms of $c(\alpha)$.

Remark 1: This extends Stein's lemma, see Chernoff (1956), or Bahadur (1971), for simple versus simple hypotheses, say P_{θ_o} versus P_θ for some $\theta \neq \theta_o$, which asserts that

$$\alpha_2 \doteq e^{-D(P^n_{\theta_o} \,||\, P^n_{\hat{\theta}})}.$$

Remark 2: The classical likelihood ratio test, L.R.T., uses the statistic $\log [\, p(x^n \mid \hat{\theta})/p(x^n \mid \theta_o)]$. Proposition 2.2.1 relates the likelihood ratio test to the Bayes test: since

$$\log \frac{m_n(X^n)}{p(X^n \mid \theta_o)} = \log \frac{p(X^n \mid \hat{\theta})}{p(X^n \mid \theta_o)} + \log \frac{m_n(X^n)}{p(X^n \mid \hat{\theta})}$$

$$\sim \log \frac{p(X^n \mid \hat{\theta})}{p(X^n \mid \theta_o)} + \frac{d}{2} \log \frac{2\pi}{n} + \log \det I(\theta)^{-1},$$

we see that the L.R.T. and the Bayes' test are asymptotically equivalent, a fact which has been previously observed in specific cases. Moreover,

$$2 \log \frac{p(X^n \mid \hat{\theta})}{p(X^n \mid \theta_o)}$$

has an asymptotic Chi-square distribution with d degrees of freedom, see Wilks (1962).

Proof: First we prove the lower bound statement (12). Let $\tilde{C}_n$ be any critical region with $P_{\theta_o}(\tilde{C}_n) \leq \alpha_1$, and let A_n be the "typical set"

$$A_n = \{ x^n \mid \log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} \leq D(P^n_{\theta_o} \,||\, M_n) - \frac{1}{2} c(\alpha) \},$$

where $\alpha > \alpha_1$. Observe that

$$\lim_{n \to \infty} P^n_{\theta_o}(A_n) = \alpha.$$

Then the average probability of type 2 error satisfies

$$\alpha_2 = M_n(\tilde{C}_n^c) \geq M_n(\tilde{C}_n^c \cap A_n) \geq e^{-D(P^n_{\theta_o} \,||\, M_n) + \frac{1}{2} c(\alpha)} P^n_{\theta_o}(\tilde{C}_n^c \cap A_n)$$

$$\geq e^{-D(P^n_{\theta_o} \,||\, M_n) + \frac{1}{2} c(\alpha)} [\, P^n_{\theta_o}(\tilde{C}_n^c) - P^n_{\theta_o}(A_n^c)].$$

Since

$$\lim_{n \to \infty} [\, P^n_{\theta_o}(\tilde{C}_n^c) - P^n_{\theta_o}(A_n^c) \,] = \alpha - \alpha_1 > 0,$$

we may take logarithms to obtain

$$\liminf_{n \to \infty} [\, \log \alpha_2 + D(P^n_{\theta_o} \,||\, M_n) \,] \geq \frac{1}{2} c(\alpha) + \log (\alpha - \alpha_1).$$

where $\alpha \in (\alpha_1, 1)$. Note that c is strictly decreasing in α and ranges from $-E\chi_d^2$ to ∞ and $\log(\alpha - \alpha_1)$ is strictly increasing. It is possible to get an implicit algebraic relation which must be satisfied by the α which maximizes the right hand side. In particular, we chose $\alpha = (\alpha_1 + 1)/2$ so as to get a the lower bound (12).

Now we prove the upper bound, (13). The Bayes' optimal test is of the form reject H if and only if $(X_1, \ldots, X_n) \in C_n$, where C_n is the critical set

$$C_n = \{ x^n \mid \log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} \leq t \}.$$

Choosing

$$t = D(P_{\theta_o}^n \mid\mid M_n) - \frac{c(\alpha_1)}{2},$$

we have that

$$-2[\log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} - D(P_{\theta_o}^n \mid\mid M_n)]$$

converges weakly to a Chi-square random variable with d degrees of freedom. So, the limiting probability of type 1 error is

$$\lim_{n \to \infty} P_{\theta_o}(C_n) = \alpha_1.$$

By Markov's inequality, the average probability of type 2 error satisfies

$$\alpha_2 = M_n(C_n^c) \leq e^{-t} = e^{-D(P_{\theta_o}^n \mid\mid M_n) + \frac{1}{2}c(\alpha_1)}.$$

Thus, taking logs, and rearranging gives

$$\limsup_{n \to \infty} [\log \alpha_2 + D(P_{\theta_o}^n \mid\mid M_n)] \leq \frac{1}{2}c(\alpha_1),$$

so that $c(\alpha_1)/2$ upper bounds the limit superior of the left hand side, thus (13) is proved. □

1.8 The Discrete Case

Up to this point we have assumed that the prior was continuous in the sense of having a density relative to Lebesgue measure. In this section we assume that the prior is discrete, and will shortly add the assumption that the set of points which

have positive mass has no cluster points. We first give an example to motivate our results.

Consider a Bernoulli distribution which puts mass α at $p=u$ and mass $1-\alpha$ at $p=v$ where u,v are parameter values for a Bernoulli (p) random variable. We require that $\alpha \in (0, 1)$ or no mixing occurs. The mixture distribution is

$$m(x^n) = \alpha u^{\Sigma x_i}(1-u)^{n-\Sigma x_i} + (1-\alpha)v^{\Sigma x_i}(1-v)^{n-\Sigma x_i},$$

where summations run from 1 to n. If u is the true value then

$$E_u \log \frac{P_u(X^n)}{m(X^n)} = E_u \log \frac{u^{\Sigma X_i}(1-u)^{n-\Sigma X_i}}{m(X^n)}$$

$$= \log \frac{1}{\alpha} - E_u \log \left(1 + \left(\frac{1-\alpha}{\alpha}\right)\left(\frac{v}{u}\right)^{\Sigma X_i}\left(\frac{1-v}{1-u}\right)^{n-\Sigma X_i}\right).$$

We apply the same approximations as in Chapter 2, Section 4. In that context, the correct answer is given up to the -1/2 term. We use $\sum X_i \sim nu$ and find that the Kullback-Leibler number becomes

$$-\log \alpha - \log \left[1 + \left(\frac{1-\alpha}{\alpha}\right)[(v/u)^u((1-v)/(1-u))^{(1-u)}]^n \right],$$

so the analysis boils down to the behavior of $(v/u)^u((1-v)/(1-u))^{(1-u)}$. This reasoning applies for any $u \neq v$ since

$$(v/u)^u((1-v)/(1-u))^{(1-u)} = e^{-D(u\,||\,v)} < 1,$$

where

$$D(u\,||\,v) = u \log \frac{u}{v} + (1-u)\log \frac{(1-u)}{(1-v)}.$$

Thus, for $u \neq v$, the answer is, asymptotically, $\log 1/\alpha$, a constant independent of n. Indeed, it is the entropy term, $\log 1/w(u)$. We note that for the continuous priors, the $\log n$ term came out of using Laplace's method to deal with exponentially fast concentration at the true value. That cannot work here. Other examples that could be evaluated in a similar fashion gave the same type of answer, dependent on the prior probability of the true distribution but independent of n. Those examples motivated the proof of the following proposition. We have here denoted the parameter by k since we are assuming it is discrete.

In the course of the proof of the next proposition, we will again fall back on the consistency of the MLE. One of Wald's key hypotheses for consistency was that for given k_o there is a sufficiently large r so that we have

$$E_{k_o} \log \sup_{\{k: \|k - k_o\| > r\}} \frac{p(X \mid k)}{p(X \mid k_o)} < \infty. \tag{14}$$

As will be seen it is not Wald's theorem, Wald (1949), which we use but an implication of it due to Wolfowitz (1949), which amounts to a uniform law of large numbers. We denote the prior probability of k by $w(k)$.

Proposition 1.8.1: *Suppose that the distributions with positive mass are distinct, that there is a Kullback-Leibler neighborhood of the true distribution P_{k_o}, of radius $\eta > 0$, which excludes all other distributions in the family, and that Wald's hypothesis (14) is satisfied. Then, as n increases,*

$$D(P_{k_o}^n \| M_n) \to \log \frac{1}{w(k_o)}.$$

Proof: We can rewrite the Kullback - Leibler number as

$$E_{k_o} \log \frac{p_{k_o}(x^n)}{m(x^n)} = \log \frac{1}{w(k_o)} - E_{k_o} \log [1 + \sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(X^n \mid k)}{p(X^n \mid k_o)}]. \tag{15}$$

By using the inequality $-\log(1 + x) \leq 0$ for x positive we have the bound

$$D(P_{k_o}^n \| M_n) \leq \log \frac{1}{w(k_o)}, \tag{16}$$

which we hope is attained in the limit. To get a lower bound it is enough to upper bound the positive quantity

$$E_{k_o} \log [1 + \sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(X^n \mid k)}{p(X^n \mid k_o)}], \tag{17}$$

which appears in (15), by something which shrinks to zero as n increases.

Consider the partition defined by

$$\Omega = \Omega_n = \{x^n: \text{for all } k \neq k_o, \frac{p(x^n \mid k)}{p(x^n \mid k_o)} < e^{-n\lambda}\},$$

and its complement, where $\lambda < \min\{D(k_o \| k): |k_o - k| > \eta\}$. Using the partition, (17) can be written as the sum of two terms. The first is

$$E_{k_o} \chi_\Omega \log [1 + \sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(x^n \mid k)}{p(x^n \mid k_o)}] \leq E_{k_o} \chi_\Omega \log [1 + \frac{(1 - w(k_o))}{w(k_o)} e^{-n\lambda}]$$

$$= P_{k_o}(\Omega) \log [1 + \frac{(1 - w(k_o))}{w(k_o)} e^{-n\lambda}],$$

which clearly tends to zero as n increases. The other term tends to zero also: it is

$$E_{k_o} \chi_{\Omega^c} \log [1 + \sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(x^n \mid k)}{p(x^n \mid k_o)}]$$

$$\leq P_{k_o}(\Omega^c) \log E_{k_o} \frac{\chi_{\Omega^c}}{P_{k_o}(\Omega^c)} [1 + \sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(x^n \mid k)}{p(x^n \mid k_o)}]$$

$$= -P_{k_o}(\Omega^c) \log P_{k_o}(\Omega^c) + P_{k_o}(\Omega^c) \log E_{k_o} \chi_{\Omega^c} [1 + \sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(x^n \mid k)}{p(x^n \mid k_o)}]$$

$$\leq -P_{k_o}(\Omega^c) \log P_{k_o}(\Omega^c) + P_{k_o}(\Omega^c) \log [P_{k_o}(\Omega^c) + \frac{(1 - w(k_o))}{w(k_o)}]. \quad (18)$$

We see that both terms in (18) go to zero:

$$P_{k_o}(\Omega^c) = P_{k_o}(\{x^n: \text{there exists some } k \neq k_o \text{ such that } \frac{p(x^n \mid k)}{p(x^n \mid k_o)} \geq e^{-n\lambda}\})$$

$$\leq P_{k_o}(\{x^n: \sup_{\{k: |k - k_o| > r\}} \frac{p(x^n \mid k)}{p(x^n \mid k_o)} \geq e^{-n\lambda}\})$$

$$+ P_{k_o}(\{x^n: \sup_{\{k: |k - k_o| < r\}} \frac{p(x^n \mid k)}{p(x^n \mid k_o)} \geq e^{-n\lambda}\}). \quad (19)$$

Since there are only finitely many k with $|k - k_o| \leq r$, the second term in (19) goes to zero. The first term in (19) also goes to zero, by Wolfowitz's theorem because we have assumed Wald's hypothesis (14). Now, (18) goes to zero, implying (17) does also. This proves the proposition. □

Thus we have that the Kullback-Leibler number between the true distribution and the mixture of distributions is the logarithm of the reciprocal of the prior probability. This in turn is the same as the Shannon codelength of the true parameter under the prior. This parallels work of Barron (1985) and Barron and Cover (1989) on minimum complexity density estimation where it was proved that if the prior assigns positive mass to the true density then the minimum complexity density estimator converges to the true density with probability 1. Here we conclude that, no matter how long the message, the redundancy is the codelength for the index of the true density, i.e., the mixture behaves like the true density up to a fixed codelength. That the asymptotic formula is independent of n appears to be related to the ability to identify the true distribution unambiguously.

An alternate proof which does not require (14) to be true is possible. It would show that (17) goes to zero by use of Barron and Cover (1989), or Barron (1985) pg. 56. Results there enable us to derive that

$$\sum_{k \neq k_o} \frac{w(k)}{w(k_o)} \frac{p(X^n \mid k)}{p(X^n \mid k_o)}$$

tends to zero, in P_{k_o} probability. By direct calculation its expectation under P_{k_o} is bounded by $1/w(k_o)$. Since the result of adding 1 to it and then taking the logarithm still tends to zero, in probability, we can set up an application of Lemma 3.4.2, so the limit superior of its expectation is less than or equal to zero. Thus, the upper bound (16) is asymptotically tight.

We can easily obtain a result for the average redundancy also.

Proposition 1.8.2: *Assume the entropy of the prior is finite, that the parameter values with positive prior probabilities have no limit points with positive probability and the hypotheses of Proposition 1.8.1 are satisfies for each parameter value. Then*

$$\sum_k D(P_k^n \| M_n) w(k) \to H(K),$$

as n increases.

Remark: The left hand side is the mutual information $I(K; X^n) = H(K) - H(K \mid X^n)$. Thus, an equivalent statement of the proposition is that $H(K \mid X^n)$ goes to zero. When the support of K is a finite set, this convergence is well known in information theory as an application of Fano's inequality, see Blahut (1987) pg. 156.

Proof: We have that for each k

$$0 \leq D(P_k^n \| M_n) \leq \log \frac{1}{w(k)},$$

and that pointwise the quantity in the middle tends to its upper bound, which is integrable with respect to the prior. The proposition follows from the dominated convergence theorem. □

Both the proofs and the results change dramatically from the case of using a continuous prior. Further, we believe an analysis of the implications for source coding, as done in Chapter 1, Section 5, could be carried out.

1.9 A Channel Capacity Interpretation

In this section we briefly give an interpretation of the uniform approximation result from Chapter 3. This will be in terms of what is called the channel capacity, which is the theoretical upper bound on the rate of transmission of data across a communication channel. A channel is basically a conditional distribution which describes the probability distribution of the output received given the input that was sent. The input is an encoded representation of the message. Naturally, we want the output received to be decodable to give the message that was sent; but, it is possible that the transmission was corrupted by background noise, for instance. We assume that a channel is going to be used repeatedly and want a coding scheme which will achieves a rate close to the capacity over repeated uses of the channel. Shannon identified the analytic form of the capacity and showed that any rate up to the capacity was achievable by some coding strategy. We recall that the mutual information between two random variables X and Y is

$$I(X;Y) = \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dxdy,$$

and the capacity of a channel defined by $p(y \mid x)$ is

$$C = \sup_{p(x)} I(X;Y).$$

in which we regard Y as the output and X as the input.

Suppose that we have one broadcaster sending the same encoded message X to each of many receivers $Y_1,\ldots,Y_k$, which are conditionally independent and identically distributed given X. Intuitively, this means that the noise which interferes with the signal received by any one receiver is independent of that received by any other receiver. Thus, the conditional distribution defining the channel is

$$p(y_1,\ldots,y_k \mid x) = \prod_{i=1}^{k} p(y_i \mid x).$$

When a block of coded data $x_1,\ldots,x_n$ is sent, the i^{th} receiver, i between 1 and k, picks up $y_1^i,\ldots,y_n^i$. Suppose the k receivers decode cooperatively, that is they pool their data and then estimate the message sent. Then, the capacity of the resulting channel is

$$C_k = \sup_{p(x)} I(X;Y^k).$$

We can relate the present case to the statistical context by letting X correspond to the parameter and Y^k correspond to the random sample. Thus $p(x)$ takes the role of the density of the prior which we denote by P, $p(y \mid x)$ takes the role of the density of the i.i.d. random variables with corresponding probability denoted by P_x. We denote the Fisher information for the density by $I(x)$ with $x = (x_1,...,x_d)$ varying over a compact set Ω in $\mathbf{R}^d$, and the mixture of the P_x^k's with respect to p we denote by M_k. This correspondence allows us to restate the hypotheses of Theorem 1.2.2 so as to interpret that result in the setting of channels.

Proposition 1.9.1: *Assume that H(X) is finite, the determinant of I(x) is bounded away from zero, that for each positive* α *and each open set* $N(x)$ *containing* x *we have that*

$$\sup_{x \in \Omega} P_x(P(N(x)^c \mid X^k) > \alpha) = o(\frac{1}{\log k}),$$

and that for all i *and all* j *there is a* $\delta > 0$ *which satisfies*

$$\sup_{x \in \Omega} E_x \sup_{x': |x - x'| < \delta} (\frac{\partial^2}{\partial x_i \partial x_j} \log p(Y_1 \mid x'))^4 < \infty,$$

where E_x *denotes the expectation with respect to* $p(y \mid x)$. *Then, we have that*

$$I(X; Y^k) = \frac{d}{2} \log \frac{k}{2\pi e} + H(X) + \int p(x) \log \det I(x) dx + o(1).$$

Remark: The formula is asymptotic in k, the number of receivers, not n the length of the data stream. In this context, the mutual information and the capacity only have interpretations when the length of the data stream is assumed to be large, i.e., over repeated uses of the channel.

Proof: This result follows from noting that

$$I(X; Y^k) = \int p(x)p(y^k \mid x) \log \frac{p(x)p(y^k \mid x)}{p(x)m(y^k)} dy^k dx$$

is of the same form as the averaged redundancy

$$\int w(\theta) D(P_\theta^n \mid\mid M_n) d\theta = \int w(\theta)p(x^n \mid \theta) \log \frac{w(\theta)p(x^n \mid \theta)}{w(\theta)m(x^n)} dx^n \, d\theta,$$

under the correspondence in notation already defined and then using Theorem 1.2.2. □

We also obtain an asymptotic form for the capacity, by translating Proposition 1.4.2 into the present notation.

Proposition 1.9.2: *Under the hypotheses of Proposition 1.9.1 we have that the capacity of k receivers, C_k is*

$$C_k = \frac{d}{2} \log \frac{k}{2\pi e} + c,$$

where $c = \int \sqrt{\det I(x)}\, dx$.

Proof: It is clear that by the same mathematics as in Section 5 the proposition is true. □

So, the capacity increases as the logarithm of the number of receivers which means that for large k there are coding schemes which achieve rates of transmission, over repeated uses of the channel, arbitrarily close to $(d/2) \log (k/2\pi e) + c$.

Chapter 2: The Cumulative Risk

2.1 Intuition

In this chapter we will characterize the asymptotic behavior of the Kullback-Leibler distance between the n-fold product of a given, true, member of a parametrized family of densities and a mixture of products of such densities with respect to a continuous prior which assigns positive mass to each open set which contains the true value of the parameter. We will show that the distance increases with the logarithm of the sample size plus a constant which we identity. The logarithmic form comes from the fact that a continuous mixture of densities is used and includes densities very close to the true one. If the mixture excluded a neighborhood of the true density we would expect the behavior of the quantity to be of the order of the sample size like the case of two independent distributions. Earlier, in Chapter 1, Section 8, we saw that if positive mass is assigned to the true density then the quantity converges to a constant.

First we give the intuition behind the main result of this chapter. The intuition can be formalized into a proof; however, the desired result is true more generally than the intuition suggests. In particular, we will use the MLE for θ, denoted $\hat{\theta}$ in the outline below, but a different estimator will be used in the proof to follow. Also, the intuition will account for the lost -1/2 in the example of the Beta prior on the Bernoulli in Chapter 1, Section 3 and in the exponential prior on the Poisson which we will examine at the end of this chapter.

We consider a parametrized family of distributions $\{P_\theta, \theta \in \Omega\}$ on a measurable space, with $\Omega \subset \mathbf{R}^d$, and assume that $X^n = X_1,\ldots,X_n$ are i.i.d. with respect to the distribution P_{θ_o}. Let $w(\theta)$ be a prior density for θ with respect to Lebesgue measure, and M_n the mixture of distributions with respect to w, with density m_n. We identify the asymptotic behavior of the Kullback-Leibler number $D(P^n_{\theta_o} \,||\, M_n)$ to o(1) accuracy. The result is

$$D(P^n_{\theta_o} \,||\, M_n) = \frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\theta_o) + \log\frac{1}{w(\theta_o)} + o(1). \qquad (1)$$

Equation (1) is an improvement on earlier similar expansions. Krichevsky and Trofimov (1981) identified the $(d/2)\log n$ term in the Bernoulli case, and Rissanen

(1984) gave general conditions for bounds on the redundancy of a code which were of a similar form. Rissanen (1983) used an expansion similar to ours, but it had different terms. Schwarz (1978) used a $(d/2)\log n$ penalty term in a penalized likelihood criterion for the number of parameters in a model selection context, and he proved a Bayes' optimality property of the criterion. In future work, we hope to demonstrate the relevance of the results here to some model selection considerations from Barron & Barron (1988), and from Haughton (1988).

The expansion can be conjectured from recalling equation (9) from Chapter 1, which was the decomposition

$$D(P^n_{\theta_o} \,||\, M_n) = E_{\theta_o}\log\frac{p(X^n \mid \hat{\theta})}{m(X^n)} + E_{\theta_o}\log\frac{p(X^n \mid \theta_o)}{p(X^n \mid \hat{\theta})}. \tag{2}$$

The first term on the right is a modification of the posterior distribution. Walker (1967) showed that the standardized posterior can be well approximated by a $N(\hat{\theta}, I(\hat{\theta}))$ under suitable technical conditions. His proof basically used Laplace integration on the mixture density. Since we are approximating $\log [\, p(x^n \mid \theta)/m(x^n)\,]$ in expectation, not just in probability, the Laplace integration introduces new difficulties. However, they can be dealt with so the technique can be adapted to give

$$\frac{d}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log\det I(\theta_o) + \log\frac{1}{w(\theta_o)},$$

as an approximation of the first term in (2).

The second term in (2) looks as though we should use a second order Taylor expansion of $\log p(x^n \mid \theta)$ about $\hat{\theta}$. Such a result was formulated by Wilks (1962), Chernoff (1954), and Wald (1943), but they only proved convergence to a Chi-square random variable in distribution, whereas we are concerned with an expected value. By the second order Taylor expansion, we have

$$E_{\theta_o}\log\frac{p(X^n \mid \theta_o)}{p(X^n \mid \hat{\theta})} = -1/2 E_{\theta_o}\sqrt{n}\,(\theta_o - \hat{\theta})^t I^*(\theta^*)\sqrt{n}\,(\theta_o - \hat{\theta}),$$

where $I^*(\theta^*)$ is the empirical Fisher information matrix evaluated at a point θ^* on the line segment joining θ_o and $\hat{\theta}$. Using a first order Taylor expansion on $\nabla \log p(x^n \mid \theta)$, about $\hat{\theta}$, we obtain an expression for $S_n = (1/\sqrt{n})\sum_{i=1}^{n} \nabla \log p(X_i \mid \theta_o)$, so that we can re-express the second term of (2) as

$$\frac{-1}{2}\,[E_{\theta_o} S_n^t I^{-1}(\theta_o) S_n + E_{\theta_o} S_n^t (A_n - I(\theta_o)^{-1}) S_n 1_{\Omega_\varepsilon}],$$

where Ω_ε is essentially the set on which A_n converges to $I(\theta_o)^{-1}$. The first term tends to $-d/2$ since $E_{\theta_o} S_n^i S_n = I(\theta_o)$ and the first term inside the bracket is seen to be d by a familiar calculation. Thus approximating by use of (2) gives

$$D(P_{\theta_o}^n \,||\, M_n) = \frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\theta_o) + \log\frac{1}{w(\theta_o)}, \tag{3}$$

which is of the same form as (1). The validity of (3) can be established, as shown in Clarke and Barron (1988); however, the introduction of the MLE in the approximation requires additional assumptions to guarantee its consistency. The approach we give below avoids such assumptions.

We recall that in Chapter 1 Section 3, a result due to Berk (1970) was used to obtain the posterior consistency at rate $o(1/\log n)$. In parametric families which are not of exponential form, Berk's conditions can be difficult to verify. Accordingly, we sought sufficient conditions for posterior convergence which would be easy to verify, and give the desired rate of $o(1/\log n)$. Using the nonparametric work of Kiefer and Wolfowitz (1958), we found such conditions involving a criterion which we call the soundness of a parametric family. Essentially, a family is soundly parametrized if and only if the mapping from the parameter space with the Euclidean topology into the collection of distributions is a homeomorphism onto its image under the relative topology induced by the restriction of a suitable distance measure. We have restricted our attention to random variables taking values in a finite dimensional real space and used the Kolmogorov-Smirnov distance. Our result, Theorem 2.3.1, is a restatement of Theorem 1.2.1 which gives the same conclusion. It assumes only soundness and the expected supremum condition, see equation (3) of Chapter 1.

2.2 The Main Theorem

In the approximation we are seeking, the only quantities which appear are the Fisher information and the prior density at the true value. This suggests that, ideally, the only conditions which should be introduced are those which will control them. The behavior of the Fisher information can be controlled, for present purposes, by assuming that the expected values of local suprema of the squares of the second derivatives of the log density are finite, i.e., there exists a $\xi > 0$ so that for

each i,j from 1 to d

$$E_{\theta_o} \sup_{\{|\theta - \theta_o| < \xi\}} \left| \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X_1 | \theta) \right|^2 < \infty. \tag{4}$$

Throughout this chapter, it is assumed that w is continuous and positive at θ_o and that the Fisher information $I(\theta_o)$ is positive definite. Equation (4) implies the first derivative condition,

$$E_{\theta_o} \left| \frac{\partial}{\partial\theta_k} \log p(X_1 | \theta_o) \right|^2 < \infty,$$

is satisfied so the two definitions of Fisher information, one in terms of expected second derivatives, the other in terms of products of first derivatives are equivalent, see Lehmann, (1983, Lemma 2.6.1).

Formally, the theorem we will prove in this chapter is the following.

Theorem 2.2.1: *Let the family* $\{ P_\theta \}$ *satisfy the local supremum condition (4). Assume that for the parameter value* θ_o, $I(\theta_o)$ *is positive definite and that* $w(\theta_o) > 0$. *Then we have the upper bound*

$$\limsup_{n \to \infty} \left[D(P^n_{\theta_o} \,||\, M_n) - \frac{d}{2} \log \frac{n}{2\pi e} - \log \frac{1}{w(\theta_o)} - \frac{1}{2} \log \det I(\theta_o) \right] \leq 0. \tag{5}$$

If, in addition, we have that the posterior distribution is consistent with rate $o(1/\log n)$, *then*

$$\liminf_{n \to \infty} \left[D(P^n_{\theta_o} \,||\, M_n) - \frac{d}{2} \log \frac{n}{2\pi e} - \log \frac{1}{w(\theta_o)} - \frac{1}{2} \log \det I(\theta_o) \right] \geq 0. \tag{6}$$

Clearly, expansion (1) only holds when both hypotheses are satisfied.

We will first prove the theorem, and then give conditions which ensure the rate of posterior convergence we are assuming. By posterior convergence at rate $o(1/\log n)$, we mean that, for any open set N containing θ_o,

$$P^n_{\theta_o}(W(N^c | X^n) > \alpha) = o\left(\frac{1}{\log n}\right),$$

for all $\alpha > 0$. Here, $W(\cdot | X^n)$ is the posterior distribution of Θ given X^n. In particular we will give sufficient conditions for posterior convergence at rate $O(\frac{\log n}{n})$.

To prove the theorem we will use a proposition which gives upper and lower bounds on the integrand of $D(P^n_{\theta_o} \,||\, M_n)$ on certain sets which have high probability and will permit tight bounds. We introduce the following notation. Let

$$N(\theta_o, \delta) = \{\theta : |\theta - \theta_o| \leq \delta\},$$

where the inner product defining the norm is with respect to $I(\theta_o)$. For $0 < \varepsilon < 1$ and $\delta > 0$ define

$$A_n(\delta, \varepsilon, \theta_o) = \{ \int_{N(\theta_o, \delta)^c} p(x^n \mid \theta) w(\theta) d\theta \tag{7}$$

$$\leq \varepsilon \int_{N(\theta_o, \delta)} p(x^n \mid \theta) w(\theta) d\theta \},$$

$$B_n(\delta, \varepsilon, \theta_o) = \{(1 - \varepsilon)(\theta - \theta_o)^t I(\theta_o)(\theta - \theta_o) \} \leq (\theta - \theta_o)^t I^*(\theta_t)(\theta - \theta_o) \tag{8}$$

$$\leq (1 + \varepsilon)(\theta - \theta_o)^t I(\theta_o)(\theta - \theta_o)$$

$$\text{for all } \theta \in N(\theta_o, \delta) \text{ and all } t \in [0, 1] \},$$

where $\theta_t = t(\theta - \theta_o) + \theta_o$, and $I^*(\theta)$ is the empirical Fisher information at θ. Also, let

$$C_n(\delta, \theta_o) = \{ l_n'(\theta_o)^t I^{-1}(\theta_o) l_n'(\theta_o) \leq \delta^2 \}, \tag{9}$$

where we have denoted the average score function by

$$l_n'(\theta_o) = \frac{1}{n} \nabla \log p(x^n \mid \theta_o).$$

The set A_n contains those points x^n for which the posterior probability of the neighborhood N is at least $1/(1 + \varepsilon)$; the set B_n allows us to bound an empirical estimate of the Fisher information by its true value; and, the set C_n is the set where the second moment of the random variable $\hat{\theta}$ below is well behaved. We bound the behavior of the prior by the modulus of continuity of its logarithm on a neighborhood of the true value:

$$\rho(\delta, \theta_o) = \sup_{\theta' \in N(\theta_o, \delta)} \left| \log \frac{w(\theta')}{w(\theta_o)} \right|,$$

and the analog to the MLE which we will use is

$$\hat{\theta} = \theta_o + I^{-1}(\theta_o) l_n'(\theta_o),$$

a stochastic perturbation about the true value of the parameter.

We record a handy identity which will be used in the proof of the proposition. Let $u = \theta_o + 1/(1 - \varepsilon)\,(\hat{\theta} - \theta_o)$. Then, by completing the square (add and subtract $(1/(1 - \varepsilon)^2) l_n'^t(\theta_o) I(\theta_o)^{-1} l'_n(\theta_o)$) we have that

$$(\theta' - \theta_o)^t l'_n(\theta_o) - \frac{1}{2}(1 - \varepsilon)(\theta' - \theta_o)^t I(\theta_o)(\theta' - \theta_o)$$

$$= -\frac{(1 - \varepsilon)}{2}(\theta' - u)^t I(\theta_o)(\theta' - u) + \frac{1}{2(1 - \varepsilon)} l_n'(\theta_o)^t I^{-1}(\theta_o)' l_n'(\theta_o). \tag{10}$$

Next we state and prove tight upper and lower bounds on the density ratio. We will use a second order Taylor expansion about θ_o, then apply the formula from completing the square, and finally recognize a normal integral.

Proposition 2.2.1: *On the set $A_n \cap B_n$ we have the upper bound*

$$\frac{m(x^n)}{p(x^n \mid \theta_o)} \leq (1+\varepsilon) w(\theta_o) e^{\rho(\delta,\theta_o)} e^{\frac{n}{2(1-\varepsilon)} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o)} (2\pi)^{d/2}$$

$$\times \mid n(1-\varepsilon) I(\theta_o) \mid^{-1/2}. \tag{11}$$

On $B_n \cap C_n$ we have the lower bound

$$\frac{m(x^n)}{p(x^n \mid \theta_o)} \geq w(\theta_o) e^{-\rho(\delta,\theta_o)} e^{\frac{n}{2(1+\varepsilon)} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o)} (2\pi)^{d/2}$$

$$\times \mid n(1+\varepsilon) I(\theta_o) \mid^{-1/2} (1 - 2^{d/2} e^{-\varepsilon^2 n \delta^2/8}). \tag{12}$$

Proof: In both cases we apply Laplace integration to the mixture density. For the upper bound, (11), we have, by restriction to A_n, see (7), and then to B_n, see (8), that

$$\frac{m(x^n)}{p(x^n \mid \theta_o)} \leq (1+\varepsilon) \int_{N(\theta_o,\delta)} \frac{p(x^n \mid \theta')}{p(x^n \mid \theta_o)} w(\theta') d\theta'$$

$$= (1+\varepsilon) \int_{N(\theta_o,\delta)} e^{n(\theta'-\theta_o)' l_n'(\theta_o) - \frac{n}{2}(\theta'-\theta_o)' \hat{I}(\theta_o)(\theta'-\theta_o)} w(\theta') d\theta'$$

$$\leq (1+\varepsilon) w(\theta_o) e^{\rho(\delta,\varepsilon)} \int_{N(\theta_o,\delta)} e^{n(\theta'-\theta_o)' l_n'(\theta_o) - \frac{n}{2}(1-\varepsilon)(\theta'-\theta_o)' I(\theta_o)(\theta'-\theta_o)} d\theta'$$

$$= (1+\varepsilon) w(\theta_o) e^{\rho(\delta,\varepsilon)} e^{\frac{n}{2(1-\varepsilon)} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o)}$$

$$\int_{N(\theta_o,\delta)} e^{-\frac{n}{2}(1-\varepsilon)(\theta'-u)' I(\theta_o)(\theta'-u)} d\theta'$$

$$= (1+\varepsilon) w(\theta_o) e^{\rho(\delta,\varepsilon)} e^{\frac{n}{2(1-\varepsilon)} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o)} (2\pi)^{d/2} \mid n(1-\varepsilon) I(\theta_o) \mid^{-1/2},$$

where we have used (10) so as to pull out the exponential factor.

For the lower bound, (12), we have

$$\frac{m(x^n)}{p(x^n \mid \theta_o)} \geq \int_{N(\theta_o,\, \delta)} \frac{p(x^n \mid \theta')}{p(x^n \mid \theta_o)} w(\theta') d\theta'$$

$$= \int_{N(\theta_o,\, \delta)} e^{n(\theta' - \theta_o)' l_n'(\theta_o) - \frac{n}{2}(\theta' - \theta_o) I^*(\tilde{\theta}_t)(\theta' - \theta_o)} w(\theta') d\theta',$$

where $\tilde{\theta}_t \in < \theta', \theta_o >$. Again, we use the identity stated above, (10), but we now replace $(1 - \varepsilon)$ with $(1 + \varepsilon)$ and let $u = \theta_o + 1/(1 + \varepsilon)(\hat{\theta} - \theta_o)$. Because of the restriction to B_n, see (8), we can continue the inequality

$$\geq w(\theta_o) e^{-\rho(\delta,\, \theta_o)} \int_{N(\theta_o,\, \delta)} e^{n(\theta' - \theta_o)' l_n'(\theta_o) - \frac{n}{2}(1 + \varepsilon)(\theta' - \theta_o)' I(\theta_o)(\theta' - \theta_o)} d\theta'$$

$$= w(\theta_o) e^{-\rho(\delta,\, \theta_o)} e^{\frac{n}{2(1 + \varepsilon)} l_n'(\theta_o)' I^{-1}(\theta_o) l_n'(\theta_o)}$$

$$\times \int_{N(\theta_o,\, \delta)} e^{-(1 + \varepsilon)\frac{n}{2}(\theta' - u)' I(\theta_o)(\theta' - u)} d\theta'$$

$$= w(\theta_o) e^{-\rho(\delta,\, \theta_o)} e^{\frac{n}{2(1 + \varepsilon)} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o)} \left[\int_{\mathbf{R}^d} e^{-(1 + \varepsilon)\frac{n}{2}(\theta' - u)' I(\theta_o)(\theta' - u)} d\theta \right.$$

$$\left. - \int_{N(\theta_o,\, \delta)^c} e^{-(1 + \varepsilon)\frac{n}{2}(\theta' - u)' I(\theta_o)(\theta' - u)} d\theta' \right]. \qquad (13)$$

Since we have restricted to C_n, see (9), and the inner product is with respect to $I(\theta_o)$ we have that by writing $\alpha = 1/(1 + \varepsilon)$ and using the definition of u and $\hat{\theta}$ that

$$\begin{aligned} \mid \theta' - u \mid &= \mid \theta' - \theta_o - \alpha(\hat{\theta} - \theta_o) \mid \\ &= \mid \theta' - \theta_o - \alpha I^{-1}(\theta_o) l_n'(\theta_o) \mid \\ &\geq \mid \theta' - \theta_o \mid - \alpha \mid I^{-1}(\theta_o) l_n'(\theta_o) \mid \\ &\geq \delta - \alpha l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o) \\ &\geq (1 - \alpha)\delta = \frac{\varepsilon\delta}{(1 + \varepsilon)}. \end{aligned}$$

Consequently, in the second integral of (13), the integrand is not greater than

$$e^{-n\varepsilon^2\delta^2/4(1 + \varepsilon)} \cdot e^{-(1 + \varepsilon)n \mid \theta' - u \mid^2/4}.$$

So, expanding the domain of integration in the second integral of (13), and rearranging, we have the lower bound (12). □

We now have some control over the logarithm of the mixture density over the true density. The integrand of the Kullback-Leibler number approximated by the theorem uses the reciprocal of that density ratio. Thus, when obtaining upper bounds on the Kullback-Leibler number we will be concerned with the probability of $B_n \cap C_n$; and, when obtaining lower bounds, the probability of $A_n \cap B_n$ will be important. It will come out in the course of the proof that we require the probabilities of the complements of those sets to decrease at a fast enough rate. Any rate faster than $1/\log n$ is enough; however, we have found it convenient to use $1/n$ for B_n^c and C_n^c. We have assumed a suitable rate for A_n^c, deferring a result which will give sufficient conditions for that rate.

Before launching into a proof of the theorem, we describe the bounds that we will use on the probabilities. The expected supremum condition (4) controls both the probability of B_n^c and of C_n^c since it guarantees that certain second moments exist. That will mean, in particular, that the variance of

$$\sup_{\theta': |\theta_o - \theta'| < \delta} \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_1 | \theta'),$$

and of

$$\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_1 | \theta_o),$$

and the expectation of

$$\frac{\partial}{\partial\theta_j} \log p(x | \theta') \frac{\partial}{\partial\theta_k} \log p(X_1 | \theta_o),$$

are finite for any choice of i,j from 1 to d. The finiteness of those expectations will be used with the elementary result, based on Chebyshev's inequality, that for i.i.d. outcomes of a random variable X with finite variance under probability P,

$$P(|\overline{X} - E_P X_1| > \varepsilon) \leq \frac{1}{n\varepsilon^2} E_P\, n\, (\overline{X} - E_P X_1)^2 1_{\{|\overline{X} - E_P X_1| > \varepsilon\}},$$

where

$$E_P\, n\, (\overline{X} - E_P X_1)^2 1_{\{|\overline{X} - E_P X_1| > \varepsilon\}} \to 0.$$

The first part of the proof will be taken up by getting suitable bounds; then we will actually prove the approximation (1), by proving (5) and (6).

Proof of Theorem 2.2.1: From the posterior consistency assumption we have that for each $\delta > 0$, and each $\varepsilon > 0$,

$$P_n(A_n^c(\delta, \varepsilon, \theta)) = o(\frac{1}{\log n}),$$

by use of the relation

$$W(N(\theta_o, \delta)^c \mid X^n) = \frac{1}{1 + \frac{\int_{N(\theta_o, \delta)} p(x^n \mid \theta) w(\theta) d\theta}{\int_{N(\theta_o, \delta)^c} p(x^n \mid \theta) w(\theta) d\theta}}.$$

From Chebyshev's inequality, we will obtain bounds of the form

$$P_{\theta_o}(B_n^c(\delta, \varepsilon, \theta_o)) \leq \frac{c_1(\theta_o, n, \varepsilon, \delta)}{n\varepsilon'^2}, \tag{14}$$

where ε' is a function of ε, and δ which is positive and tends to zero as ε and δ tend to zero. The function c_1 tends to zero, as n increases, for any fixed δ and ε. By Markov's inequality we will show that we have

$$P_{\theta_o}(C_n^c(\delta, \theta_o)) \leq \frac{c_2(\theta_o, n, \delta)}{n\delta^2}, \tag{15}$$

where c_2 tends to zero, as n increases, for any fixed δ. We proceed with proving that c_1 and c_2 exist as we want.

To show the existence of c_1 it is enough to examine sets of the form

$$P_{\theta_o}(\sup_{t:\, 0 \leq t \leq 1} \mid i^*_{j,k}(\theta_t) - i_{j,k}(\theta_o) \mid < \varepsilon')$$

This is suggested by noting that $B_n(\delta, \varepsilon, \theta_o)$ can be written as

$$B_n(\delta, \varepsilon, \theta_o) = \{\varepsilon < \frac{\xi^t I^{-1/2}(\theta_o)(I^*(\theta_t) - I(\theta_o))I^{-1/2}(\theta_o)\xi}{\xi^t \xi} < \varepsilon\},$$

where $\xi = I^{1/2}(\theta_o)(\theta - \theta_o)$ varies over the set $I^{1/2}(\theta_o)N(0, \delta)$, and t varies from zero to one. Without loss of generality we can assume that the norm of ξ is one, since the normalizing factors cancel. Now we see that if the largest of the absolute values of the eigenvalues of

$$I^*(\theta_t) - I(\theta_o)$$

is small enough, then the desired inequality defining $B_n(\delta, \varepsilon, \theta_o)$ is satisfied. Taking the largest absolute eigenvalue gives a norm. By the finite dimensionality of the matrix space it is equivalent to any other norm. We choose the norm which takes the maximum of the entries. To show that the convergence of the empirical Fisher information to the true Fisher information holds in that norm it is enough to show that each entry of $I^*(\theta_t) - I(\theta_o)$ tends to zero. With that in mind we choose

$\varepsilon' = \varepsilon'(\varepsilon, \delta)$ so that when each entry is less than ε' we know that the inequality in B_n is satisfied. Thus it is enough obtain an upper bound of the required form for

$$P_{\theta_o}(\sup_{\theta \in N_\delta} | i^*_{j,k}(\theta) - i_{j,k}(\theta_o) | > \varepsilon'),$$

since there are finitely many entries. It is upper bounded by adding and subtracting $i^*_{j,k}(\theta_o)$ to get

$$P_{\theta_o}(\sup_{|\theta_o - \theta'| < \delta} | i^*_{j,k}(\theta') - i^*_{j,k}(\theta_o) | > \frac{\varepsilon'}{2}) + P_{\theta_o}(| i^*_{j,k}(\theta_o) - i_{j,k}(\theta_o) | > \frac{\varepsilon'}{2}). \quad (16)$$

By Chebyshev the second term in (16) is upper bounded by

$$\frac{4}{n\varepsilon'^2} E_{\theta_o} 1_{\{| i^*_{j,k}(\theta_o) - i_{j,k}(\theta_o)| > \frac{\varepsilon'}{2}\}} (\sqrt{n}\ (\ i^*_{j,k}(\theta_o) - i_{j,k}(\theta_o)\ )\ )^2. \quad (17)$$

For the first term in (16), we choose δ so small that

$$E \sup_{|\theta_o - \theta'| < \delta} | \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X \mid \theta') - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X \mid \theta_o) | < \frac{\varepsilon'}{4}$$

and set up another application of Chebyshev's inequality. Let

$$Y_m = \sup_{|\theta_o - \theta'| < \delta} | \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_m \mid \theta') - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_m \mid \theta_o) |$$

and let

$$V(\theta_o) = \{| \bar{Y} - E_{\theta_o} Y_1 | > \frac{\varepsilon'}{4}\}.$$

Now, the first term in (16) is upper bounded by

$$P_{\theta_o}(| \bar{Y} - E_{\theta_o} Y_1 | > \frac{\varepsilon'}{4}) \leq \frac{16}{n\varepsilon'^2} E_{\theta_o} 1_{V(\theta_o)} (\sqrt{n}\ (\bar{Y} - E_{\theta_o} Y_1)^2. \quad (18)$$

Adding the bounds (17) and (18) for the terms of (16), we see that we have an expression for c_1 of the form desired for (14): c_1 decreases to zero, albeit slowly, because of the presence of the indicator function which tends to zero and multiplies a uniformly integrable function. Here we have used the result that convergence in distribution, with convergence of the expected absolute mean, implies uniform integrability.

Similarly from Markov's inequality we can identify an expression for c_2 for use in (15):

$$P_{\theta_o}(C_n^c) \leq \frac{1}{n\delta} E_{\theta_o} 1_{C_n^c}\ n\ l'(\theta_o) I(\theta_o)^{-1} l'(\theta_o).$$

Again, the expectation goes to zero as n increases since the convergence in distribution and constancy of the expected value of $l'(\theta_o)^t I(\theta_o)^{-1} l'(\theta_o)$ guarantees the required uniform integrability.

We will sandwich the desired quantity, $D(P^n_{\theta_o} \,||\, M_n)$, between upper and lower bounds which will both converge to the same expression. By definition

$$D(P^n_{\theta_o} \,||\, M_n) = \int_{R^n} p(x^n \,|\, \theta_o) \log \frac{p(x^n \,|\, \theta_o)}{m_n(x^n)} \lambda(dx^n).$$

To get the bounds we will use two decompositions of the integral, each a sum of two terms. The bound on the probability of B_n^c will be used in both the upper bound and the lower bound because it appears in both parts, (11) and (12), of Proposition 2.2.1. However, it will be seen that the required rate of decrease to zero on the relevant probabilities is $o(1/\log n)$ for the lower bound, but is of the form c_n/n where $c_n \to 0$ as $n \to \infty$ for the upper bound.

For the lower bound (6), our first decomposition is:

$$D(P^n_{\theta_o} \,|\, M_n) = \int_{A_n \cap B_n} p(x^n \,|\, \theta_o) \log \frac{p(x^n \,|\, \theta_o)}{m_n(x^n)} \lambda(dx^n) \tag{19}$$

$$+ \int_{(A_n \cap B_n)^c} p(x^n \,|\, \theta_o) \log \frac{p(x^n \,|\, \theta_o)}{m_n(x^n)} \lambda(dx^n). \tag{20}$$

Expression (20) represents the error part which we hope is small. For it we will use a Jensen's inequality argument to show that given any $\eta > 0$ expression (20) is greater than or equal to $-\eta$, for all large n. Indeed, (20) equals

$$-P((A_n \cap B_n)^c \,|\, \theta_o) \int_{(A_n \cap B_n)^c} \frac{p(x^n \,|\, \theta_o)}{P((A_n \cap B_n)^c \,|\, \theta_o)} \log \frac{m_n(x^n)}{p(x^n \,|\, \theta_o)} \lambda(dx^n)$$

$$\geq -P((A_n \cap B_n)^c \,|\, \theta) \log \int_{(A_n \cap B_n)^c} \frac{m_n(x^n)}{P((A_n \cap B_n)^c \,|\, \theta_o)} \lambda(dx^n)$$

$$= -P((A_n \cap B_n)^c \,|\, \theta_o) \log \frac{M_n((A_n \cap B_n)^c)}{P((A_n \cap B_n)^c \,|\, \theta_o)}$$

$$\geq P((A_n \cap B_n)^c \,|\, \theta_o) \log P((A_n \cap B_n)^c \,|\, \theta_o)$$

$$\geq -\eta,$$

for n large enough, since $P((A_n \cap B_n)^c \mid \theta_o) \to 0$. For the lower bound on (19) we use the upper bound part of Proposition 2.2.1, equation (11). Now the lower bound is

$$D(P^n_{\theta_o} \mid\mid M_n) \geq -E_{\theta_o} 1_{A_n \cap B_n} \log [(1+\varepsilon) e^{\rho(\delta,\theta_o)} e^{\frac{n}{2(1+\varepsilon)} l'_n(\theta_o) I(\theta_o)^{-1} l'_n(\theta_o)}$$
$$\times (2\pi)^{d/2} \mid n(1+\varepsilon) I(\theta_o) \mid^{-1/2}] - \eta,$$

so we have that

$$D(P^n_{\theta_o} \mid\mid M_n) - \frac{d}{2}\log n \geq P_{\theta_o}(A_n \cap B_n)[\frac{d}{2}\log(1+\varepsilon) - \log(1+\varepsilon) e^{\rho(\delta,\theta_o)}]$$
$$+ P_{\theta_o}(A_n \cap B_n)[\frac{d}{2}\log\frac{1}{2\pi} + \log\frac{1}{w(\theta_o)} + \frac{1}{2}\log\det I(\theta_o)]$$
$$- \frac{n}{2(1-\varepsilon)} E_{\theta_o} 1_{A_n \cap B_n} l'^t_n(\theta_o) I(\theta_o)^{-1} l'_n(\theta_o) - \eta$$
$$+ \frac{d}{2}(\log n)\,[P_{\theta_o}(A_n \cap B_n) - 1]. \qquad (21)$$

The limit of the right hand side exists as n increases, and is a function of ε, η, δ, and θ_o, which we now identify. To see that the last term of (21) gives zero we use the fact that $P_{\theta_o}(A_n \cap B_n) - 1 \geq o(1/\log n)$. To identify the limit of the expectation in (21), let

$$z_n = \sqrt{n} I(\theta_o)^{-1/2} l'_n(\theta),$$

then $z^t z$ is uniformly integrable. In fact, it converges in distribution to a χ^2_d random variable in P_{θ_o} distribution, with convergent, indeed constant, expected value:

$$E_{\theta_o} z^t z = \operatorname{tr} E_{\theta_o} z z^t = \operatorname{tr} E_{\theta_o} n l'_n(\theta) l'_n(\theta_o)^t I(\theta_o)^{-1} = \operatorname{tr} I(\theta_o)^{-1} I(\theta_o) = d.$$

Thus $1_{A_n \cap B_n} z^t z$ is uniformly integrable too, and is convergent to χ^2_d also, so it has the same asymptotic expected value, d.

By taking the limit as n goes to infinity in the last lower bound (21), we have the following inequality:

$$\liminf_{n \to \infty} [\, D(P^n_{\theta_o} \mid\mid M_n) - \frac{d}{2}\log n \,]$$
$$\geq \frac{d}{2}\log(1+\varepsilon) - \log(1+\varepsilon) e^{\rho(\delta,\theta_o)} - \eta$$
$$+ \frac{d}{2}\log\frac{1}{2\pi} + \log\frac{1}{w(\theta_o)} + \frac{1}{2}\log\det I(\theta_o) - \frac{d}{2(1-\varepsilon)}. \qquad (22)$$

Now we let ε, η, and δ decrease so that the first part of (22) is seen to be zero and the second part gives the constants claimed by expression (6).

It remains to obtain the upper bound (5). We use a slightly different decomposition:

$$D(P_{\theta_o}^n \| M_n) = \int_{B_n \cap C_n} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} \lambda(dx^n) \tag{23}$$

$$+ \int_{(B_n \cap C_n)^c} p(x^n \mid \theta_o) \log \frac{p(x^n \mid \theta_o)}{m_n(x^n)} \lambda(dx^n). \tag{24}$$

In this case term (24) is the error which we hope is small. If we invert the argument of the log, restrict the domain of integration in the definition of m_n, and rewrite the inner integrand then we have an upper bound on (24) which is of the form:

$$-\int_{(B_n \cap C_n)^c} p(x^n \mid \theta_o) \log \int_{\{\theta : \|\theta - \theta_o\| \le \delta\}} e^{-\log \frac{p(x^n \mid \theta_o)}{p(x^n \mid \theta)}} w(\theta) d\theta \lambda(dx^n).$$

Since θ is restricted to a neighbourhood about θ_o, we can use a Taylor expansion of $\log p(x^n \mid \theta)$:

$$\log p(x^n \mid \theta) - \log p(x^n \mid \theta_o) = (\theta - \theta_o)^t S_n(\tilde{\theta}),$$

where $S_n(\theta) = \nabla \log p(X^n \mid \theta)$. Now, the last integral is less than or equal to

$$-\int_{(B_n \cap C_n)^c} p(x^n \mid \theta_o) \log \int_{\{\theta : \|\theta - \theta_o\| \le \delta\}} e^{-\sup_{\theta', \tilde{\theta} \in N(\theta, \delta)} (\theta' - \theta)^t S_n(\tilde{\theta})} w(\theta) d\theta \lambda(dx^n)$$

$$\le -P((B_n \cap C_n)^c \mid \theta_o) \log W(\{\theta : \|\theta - \theta_o\| \le \delta\})$$

$$+ E_{\theta_o} \chi_{(B_n \cap C_n)^c} \sum_{i=1}^{n} \sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X_i \mid \tilde{\theta}), \tag{25}$$

an upper bound for (24). By consistency the first term in (25) is no problem. For the second term in (25), we add and subtract a convenient quantity. The result is

$$E_{\theta_o} \chi_{(B_n \cap C_n)^c} \sum_{i=1}^{n} \Big[\sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X_i \mid \tilde{\theta})$$

$$- E_{\theta_o} \sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X_i \mid \tilde{\theta}) \Big]$$

$$+ n P_{\theta_o}((B_n \cap C_n)^c) E_{\theta_o} \sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X \mid \tilde{\theta}). \tag{26}$$

The last term in (26) is upper bounded by using

$$P_{\theta_o}((B_n \cap C_n)^c) \leq P_{\theta_o}(B_n^c) + P_{\theta_o}(C_n^c),$$

and then the bounds on $P_{\theta_o}(B_n^c)$, and $P_{\theta_o}(C_n^c)$, that were derived earlier. For fixed values of ε and δ, we have that c_1 and c_2, as in (14) and (15), tend to zero. Thus we have upper bounds on $P_{\theta_o}(B_n^c)$ and $P_{\theta_o}(C_n^c)$, which tend to zero as n increases, for fixed ε and δ, even when multiplied by the factor n. So, the last term of (26) tends to zero as n increases.

The first term in (26) is upper-bounded by use of the Cauchy-Shwartz inequality so as to recognize a variance term: that upper bound is

$$\sqrt{E_{\theta_o}\chi_{(B_n \cap C_n)^c}} \sqrt{Var_{\theta_o}\left(\sum_{i=1}^{n} \sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X_i \mid \tilde{\theta})\right)}.$$

The union of events bound in the first factor gives an upper bound on the last expression:

$$\leq \sqrt{c_1/n + c_2/n} \sqrt{n\, Var_{\theta_o}\left(\sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X_1 \mid \theta)\right)}$$

$$= \sqrt{c_1 + c_2} \sqrt{Var_{\theta_o}\left(\sup_{\theta', \tilde{\theta} \in N(\theta_o, \delta)} (\theta' - \theta_o)^t \nabla \log p(X_1 \mid \theta)\right)}.$$

Again, we use the fact that c_1 and c_2, from (14) and (15), go to zero as n increases, for fixed values of ε and δ.

Having controlled the error term adequately, we deal with term (28) by the lower bound (12), given in Proposition 2.2.1. Thus we have an upper bound on the relative entropy:

$$D(P_{\theta_o}^n \mid\mid M_n) \leq -E_{\theta_o} 1_{B_n \cap C_n} \log \left[w(\theta_o) e^{-\rho(\delta, \theta_o)} e^{\frac{n}{2(1-\varepsilon)} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o)} \right.$$

$$\left. \times (2\pi)^{d/2} \mid n(1-\varepsilon) I(\theta_o) \mid^{-1/2} (1 - 2^{d/2} e^{-(1-\varepsilon)n\delta/8}) \right] + \eta$$

$$= P_{\theta_o}(B_n \cap C_n) \left[\frac{d}{2} \log(1 - \varepsilon) - \rho(\delta, \theta_o) + \log (1 - 2^{d/2} e^{-(1-\varepsilon)n\delta/8}) \right]$$

$$+ P_{\theta_o}(B_n \cap C_n) \left[\frac{d}{2} \log \frac{n}{2\pi} + \frac{1}{2} \log \det I(\theta_o) + \log \frac{1}{w(\theta_o)} \right]$$

$$- \frac{n}{2(1-\varepsilon)} E_{\theta_o} 1_{B_n \cap C_n} l_n'(\theta_o) I^{-1}(\theta_o) l_n'(\theta_o) + \eta. \qquad (27)$$

Now reasoning similar to that used to conclude the proof of the lower bound (6) gives the upper bound (5): From (27), form the upper bound inequality analogous to (22), and then let $n \to \infty$. After that, let ε, δ, and η go to zero. □

2.3 Posterior Consistency

Bayes' consistency has two different definitions. One, which we use here and call posterior consistency, is that the posterior distribution converge to a degenerate distribution at the true parameter value. The other, which we do not use here, is that a Bayes' estimator resulting from minimizing the Bayes' risk in a decision theory framework converges in probability to its true value. Sufficient conditions for posterior consistency usually assume one of two forms: the Wald style assumptions as used, for instance, by Le Cam (1953), see Theorem 5b; the other is the hypothesis testing approach of Schwartz (1965). Formally, by posterior consistency, with rate $O(f(n))$, we mean that, when θ_o is taken to be true then, for every $\alpha > 0$ and $\delta > 0$,

$$P_{\theta_o}(W(\{\theta: |\theta - \theta_o| > \delta | X^n\}) > \alpha) \leq c(\alpha, \delta) f(n).$$

where $c(\alpha, \delta)$ is a constant and $f(n) \to 0$ as $n \to \infty$. Posterior consistency with rate $o(f(n))$ is defined similarly. We have thus far assumed that we have a suitable rate for the convergence of the posterior to a degenerate distribution at the true parameter value. In this section we will prove that posterior consistency with the rate assumption we have made is a consequence of the parametrization provided it satisfies a certain condition.

By analogy with the use of the term in mathematical logic, we call that condition the soundness of the parametric family. Specifically, a parametric family is sound if and only if it satisfies the topological condition

$$\theta' \to \theta \Leftrightarrow P_{\theta'} \to P_\theta,$$

where convergence in the parameter space is in the Euclidean metric and the appropriate mode of convergence in the set of probabilities will be denoted by d. Soundness forces parameter estimation to correspond to probability estimation.

We assume that the random variables take values in a k-dimensional real space, and we choose d to be the Kolmogorov-Smirnov distance, the L^∞ norm on the distribution functions. Non-Euclidean spaces for the X_i can also be handled with a suitable choice for the mode of convergence d. More generally we could choose

$$d_S(P_{\theta'}, P_\theta) = \sup_{A \in S} |P_{\theta'}(A) - P_\theta(A)|$$

as the mode of convergence, where S is a Vapnik-Chervonenkis collection of sets, see Vapnik and Chervonenkis (1971). Any choice of S amounts to requiring that the empirical probabilities of sets in S converge uniformly to their true values. Clearly, d is a special case of d_S.

The key requirement on any choice of distance measure d is that it define neighborhoods which admit uniformly consistent tests for hypotheses of the form

$$H_o: \theta = \theta_o \quad \textit{versus} \quad H_1: |\theta - \theta_o| > \delta,$$

for some positive δ. It is known that the Kolmogorov-Smirnov distance has that property, as do the distances of Vapnik and Chervonenkis, and the variation distance for the discrete case, see Barron (1989), and Hoeffding and Wolfowitz (1958).

Soundness makes it impossible for a family to 'fold back on itself' because the only members of the family which are close to a given P_θ are those P_θ's for which θ' is close to θ. That is, no member of the family can be realized as the limit of other members of the family unless the corresponding parameters converge also. This is nontrivial particularly when taking a limit along a sequence of parameter values which has no limit in $\mathbf{R}^d$. The point here is that the unique value of θ can be obtained by convergence of estimates within the family. Schwartz (1965) gave examples of unsound parametrizations which had anomalous properties.

As a consequence of the results of Schwartz (1965), it can be shown that if the family is sound, and if $D(P_{\theta_o} \| P_\theta)$ is continuous at $\theta = \theta_o$ then the posterior distribution is consistent. To obtain a rate of posterior convergence, we will add the assumption that $D(P_{\theta_o} \| P_\theta)$ admits the second order Taylor expansion

$$D(P_{\theta_o} \| P_\theta) = \frac{1}{2}(\theta - \theta_o)^t I(\theta_o)(\theta - \theta_o) + o(\|\theta_o - \theta\|^2), \tag{28}$$

We remark that the local supremum condition (4) is sufficient for the Taylor expansion assumption. There are two key hypotheses in Theorem 2.2.1: condition (4) and the posterior consistency assumption. The latter condition is generally not easy to verify directly, so we present a result which gives sufficient conditions for it to be true.

Theorem 2.3.1: *If a parametric family for random variables taking values in* $\mathbf{R}^k$ *is soundly parametrized and the Taylor expansion (28) holds, then posterior consistency with rate* $O(\frac{\log n}{n})$ *is satisfied.*

The intuitive connection between soundness and posterior convergence is that soundness makes it impossible for any density more than δ away from the true density to mimic its behavior. In the exponential family case of Chapter 1, Section 3, we were able to cite results due to Berk which guaranteed posterior consistency. However the result above seems more generally valid. After proving the theorem, we suggest how that analysis can be carried over to random variables not taking values in a real space.

Theorem 2.3.1 falls into the hypothesis testing approach to posterior consistency. It merely states sufficient conditions for the hypothesis test to exist for a distance which, when restricted to a parametric family, is equivalent to the Euclidean metric. We will prove the existence of a test which has type 2 error uniformly upper-bounded by e^{-nr} for some positive r. Such tests are called uniformly exponentially consistent (UEC). The test we identify also has type 1 error which is decreasing exponentially fast as a function of n.

The next two propositions amount to a proof of Theorem 2.3.1. We divide the proof of the theorem in two pieces so as to isolate the part which uses the reality of the random variables. Later, when we consider generalizing, it will only be necessary to examine Proposition 2.3.1, which is the following.

Proposition 2.3.1: *Suppose the family* $\{P_\theta\}$ *is sound under the* L^∞ *distance measure denoted d, then, for any* $\delta > 0$, *there exists a UEC hypothesis test of* $\theta = \theta_o$ *versus* $|\theta - \theta_o| > \delta$.

Proof: By soundness, given $\delta > 0$ there exists an $\varepsilon > 0$ such that $|\theta - \theta_o| > \delta$ implies $d(P_\theta, P_{\theta_o}) > \varepsilon$. If we have a UEC test of

$$\text{H: } P = P_{\theta_o} \textit{ versus } \text{K: } P \in \{Q \mid d(Q, P_{\theta_o}) > \varepsilon\},$$

then we have a UEC test of

$$\text{H: } \theta = \theta_o \textit{ versus } \text{K: } \theta \in \{\theta' \mid |\theta' - \theta_o| > \delta\}.$$

Now all that remains is to identify a UEC test for the nonparametric hypothesis test. Let $\hat{P}_n$ denote the empirical distribution, choose $0 < \varepsilon' < \varepsilon$ and let

$$C_n = \{x^n \mid d(\hat{P}_n, P_o) > \varepsilon'\}$$

be the critical region. By the Kiefer-Wolfowitz theorem (1958) we have that

$$P_o(C_n) \leq 2e^{-n\varepsilon'^2/8},$$

for $P_o = P_{\theta_o}$, and for any choice Q in the alternative we want to show that

$$Q(C_n^c) = Q(d(\hat{P}_n, P_o) \leq \varepsilon')$$

is exponentially small. From the triangle inequality, we have that, for X^n in C_n^c,

$$\begin{aligned} \varepsilon \leq\ & d(P_o, Q) \leq d(P, \hat{P}_n) + d(\hat{P}_n, Q) \\ & \leq \varepsilon' + d(\hat{P}_n, Q). \end{aligned}$$

So we have a lower bound on how likely it is for the empirical distribution to remain a finite distance away from the true distribution, when the true distribution is in the alternative. By the Kiefer - Wolfowitz theorem we have

$$\begin{aligned} Q(C_n^c) \leq\ & Q(d(\hat{P}, Q) \geq \varepsilon - \varepsilon') \\ & \leq 2e^{-n(\varepsilon - \varepsilon')^2/8}, \end{aligned}$$

independently of Q in the alternative. □

The second proposition uses the conclusion of the first proposition as its hypothesis and obtains a Bayes' consistency result which is stronger than what is actually required for the theorem.

Proposition 2.3.2: *Suppose that the prior density is continuous and positive at* θ_o, *and that* $D(P_{\theta_o} \| P_\theta)$ *admits the second order Taylor expansion (28) where* $I(\theta_o)$ *is positive definite. Then, if there exists a UEC test of* $\theta = \theta_o$ *versus* $\theta \in N(\theta_o, \delta)^c$ *there is an* $r > 0$ *so that*

$$P_{\theta_o}(\int_{N(\theta_o, \delta)} w(\theta)p(x^n \mid \theta)d\theta < e^{nr} \int_{N(\theta_o, \delta)^c} w(\theta)p(x^n \mid \theta)d\theta) = O(\frac{\log n}{n}),$$

and, consequently,

$$P_{\theta_o}(W(N(\theta_o, \delta \mid X^n)^c > 2e^{-nr}) = O(\frac{\log n}{n}).$$

Proof: To make use of the existence of a UEC test we will first want to show that for every $r' > 0$ the probability of the set

$$\tilde{U}_n^c = \{\int_{N(\theta_o, \delta)} w(\theta)p(x^n \mid \theta)d\theta < e^{-nr'}p(x^n \mid \theta_o)\}$$

is $O((\log n)/n)$. It is equivalent to show that

$$U_n^c = \{\frac{\int_{N(\theta_o, \delta)} w(\theta)p(x^n \mid \theta)d\theta}{W(N(\theta_o, \delta))} < e^{-nr'}p(x^n \mid \theta_o)\} \tag{29}$$

has probability bounded by O((log n) /n). That change is convenient since the left hand side of the inequality in (29) can be recognized as

$$m(x^n \mid N(\theta_o, \delta)) = \frac{\int_{N(\theta_o, \delta)} w(\theta)p(x^n \mid \theta)d\theta}{W(N(\theta_o, \delta))},$$

a mixture of distributions with respect to a different prior. By Markov's inequality we first note that

$$P_{\theta_o}(U_n^c) \le P_{\theta_o}\left(\frac{1}{n} \left|\log \frac{p(x^n \mid \theta_o) W(N(\theta_o, \delta))}{\int_{N(\theta_o, \delta)} w(\theta) p(x^n \mid \theta) d\theta}\right| > r'\right)$$

$$\le \frac{1}{nr'} E_{\theta_o} \left|\log \frac{p(x^n \mid \theta_o)}{\int_{N(\theta_o, \delta)} w(\theta) p(x^n \mid \theta) d\theta}\right|$$

$$\le \frac{1}{nr'}\left(D(P_{\theta_o}^n \| M^n(\cdot \mid N(\theta_o, \delta))) + 2e^{-1}\right), \tag{30}$$

where we have used the fact that the negative part of the integrand in Kullback - Leibler number is always bounded below by e^{-1}, since $x \log x \ge -e^{-1}$. It is enough to upper bound the Kullback-Leibler number in (30) by $O(\log n)$. We recall from Barron (1987) that

$$D(P_{\theta_o}^n \| M^n(\cdot \mid N(\theta_o, \delta))) = E_{\theta_o} \log \frac{p(X^n \mid \theta_o) W(N(\theta_o, \delta))}{\int_{N(\theta_o, \delta)} w(\theta) p(x^n \mid \theta) d\theta}$$

$$\le n\varepsilon^2 - \log \frac{W(V_\varepsilon)}{W(N(\theta_o, \delta))}, \tag{31}$$

where

$$V_\varepsilon = \{\theta \mid D(P_{\theta_o} \| P_\theta) < \varepsilon^2\}.$$

By the second order Taylor expansion of $D(P_{\theta_o} \| P_\theta)$ we see that there exists $\tau > 0$ so that for all small $\varepsilon > 0$,

$$V_\varepsilon \supset B(\theta_o, \tau \varepsilon^2),$$

and by continuity of the prior density there exists $0 < v < w(\theta_o)$ and $c' > 0$, such that

$$W(V_\varepsilon) = W(B(\theta_o, \tau \varepsilon)) = \int_{B(\theta_o, \tau \varepsilon)} w(\theta) d\theta$$

$$\ge (w(\theta_o) - v) c' \varepsilon^d.$$

Now we have that (31) gives

$$D(P_{\theta_o}^n \| M^n(\cdot \mid N(\theta_o, \delta))) \le n\varepsilon^2 - d \log \varepsilon + c, \tag{32}$$

where c is a constant. As a function of ε the right hand side of (32) is minimized by $\varepsilon_n = \sqrt{d/2n}$ for which choice we have

$$D(P^n_{\theta_o} \,||\, M^n(\cdot \mid N(\theta_o, \delta))) \leq \frac{d}{2}\log n + c.$$

Now we can upper bound (30), the result of Markov's inequality, by:

$$P_{\theta_o}(U_n^c) \leq \frac{1}{nr'}(\frac{d}{2}\log n + c + 2e^{-1}), \tag{33}$$

which is clearly $O((\log n)/n)$.

At last we use the hypothesis on the existence of a UEC test. By an argument due to Schwartz (1965, p. 22, in the proof of Theorem 6.1) the existence of a UEC test implies the existence of r_o, $r_1 > 0$ so that

$$P_{\theta_o}(p(x^n \mid \theta_o) \leq e^{nr_o}\textstyle\int_{N(\theta_o, \delta)^c} w(\theta)p(x^n \mid \theta)d\theta \,) \leq e^{-nr_1}. \tag{34}$$

Now we can obtain a bound on the probability of concern. Let $r \in (0, r_o)$ and set $r' = r_o - r$. Then, by use of U_n to set up (33) and (34), we have that

$$\begin{aligned}
&P_{\theta_o}\{\textstyle\int_{N(\theta_o, \delta)} w(\theta)p(x^n \mid \theta)d\theta < e^{nr}\int_{N(\theta_o, \delta)^c} w(\theta)p(x^n \mid \theta)d\theta \,\} \\
&\quad \leq P_{\theta_o}(U_n \cap \{\textstyle\int_{N(\theta_o, \delta)} w(\theta)p(x^n \mid \theta)d\theta < e^{nr}\int_{N(\theta_o, \delta)^c} w(\theta)p(x^n \mid \theta)d\theta\}) + P_{\theta_o}(U_n^c) \\
&\quad \leq P_{\theta_o}(p(x^n \mid \theta_o) < e^{n(r + r')}\textstyle\int_{N(\theta_o, \delta)^c} w(\theta)p(x^n \mid \theta)d\theta) + P_{\theta_o}(U_n^c) \\
&\quad \leq O(\frac{\log n}{n}) + e^{-nr_1} = (\frac{\log n}{n}),
\end{aligned}$$

which gives the desired result. □

The above two propositions use mild hypotheses to guarantee posterior consistency at a good rate, for random variables taking values in a real space. Here, the key assumption was soundness.

Other conditions for posterior consistency at rate $O((\log n)/n)$ have been given. The familiar conditions of Wald (1949) are sufficient. This can be seen by verifying that a uniformly consistent test exists in this case, or more directly by showing that equation (34) follows from the conclusion of Wolfowitz (1949). See also Strasser (1981) and Le Cam (1953).

In some cases, however, Wald's condition that

$$E_{\theta_o} \sup_{|\theta| > r} \log \frac{p(x \mid \theta)}{p(x \mid \theta_o)} < \infty,$$

for r large enough is not satisfied or hard to verify. We find the soundness condition to be more fundamental, and in some cases easier to verify.

Because of the Kiefer-Wolfowitz theorem we have restricted our attention to random variables taking values in finite dimensional real spaces. To handle random variables taking values in separable spaces which are not necessarily real, a distance function more general than the Kolmogorov-Smirnov distance is required. A natural choice might be the Prohorov metric. A topic for future investigation is whether a UEC test always exists against the complement of a Prohorov neighborhood of a distribution. The results of Vapnik and Chervonenkis (1971) may be useful in such an investigation.

One of their results generalizes the Kiefer-Wolfowitz theorem as follows. Let $m^S(n)$ denote the maximum number of subsamples that can be induced on X^n by sets in S. Then

$$P_{\theta_o}(d_S(\hat{P}_n, P_{\theta_o}) > \varepsilon) \leq 4m^S(2n)e^{\frac{-n\varepsilon^2}{8}},$$

where $n > 2/\varepsilon^2$. The Vapnik-Chervonenkis condition on the collection S of sets, or a sequence of such collections S_n, is that $m^{S_n}(n)$ grow at a sub-exponential rate.

Another sequence of distances such that a sequence of UEC tests exists for the hypothesis test

$$P_{\theta_o} \quad \textit{versus} \quad \{Q \mid d_{S_n}(P_{\theta_o}, Q) > \varepsilon\},$$

is obtained by taking S_n to be the field generated by any partition with cardinality of order $O(n)$ as in Barron (1989). The sequence S_n does not necessarily satisfy the Vapnik-Chervonenkis conditions and test statistics other than $d_{S_n}(P_{\theta_o}, \hat{P}_n)$ are obtained which have the desired properties. This result is used in Barron (1988) to formulate general conditions for the convergence of posterior distributions.

2.4 A Further Example

We conclude this chapter by presenting one more example for the sake of examining the discrete case explicitly. This will demonstrate that the theorem is a lot easier to use than direct approximation.

Consider an exponential prior for the Poisson distribution with parameter λ and true value λ_o. The full model for n outcomes $x_1 = k_1,\ldots,x_n = k_n$ is

$$p_\lambda(k^n) = \frac{e^{-\lambda}\lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda}}{k_1!\ldots k_n!}.$$

The mixture density is

$$m(k^n) = \frac{\Gamma(\sum_{i=1}^{n} k_i + 1)}{k_1!\ldots k_n!(n+1)^{\sum_{i=1}^{n} k_i + 1}}.$$

We will use Stirling's formula on the Gamma function and the approximation that $\sum_{i=1}^{n} k_i \sim n\lambda_o$. The relative entropy is

$$E_{\lambda_o} \log \frac{P_{\lambda_o}(k^n)}{m(k^n)} = E_{\lambda_o} \log \frac{e^{-n\lambda_o}\lambda_o^{\sum_{i=1}^{n} k_i}(n+1)^{\sum_{i=1}^{n} k_i + 1}}{\Gamma(\sum_{i=1}^{n} k_i + 1)}$$

$$- \log \frac{e^{-n\lambda_o}\lambda_o^{n\lambda_o}(n+1)^{n\lambda_o + 1}}{(n\lambda_o)!}$$

$$= -n\lambda_o + n\lambda_o \log \lambda_o + (n\lambda_o + 1)\log(n+1)$$

$$- \log (\sqrt{2\pi n\lambda_o}(n\lambda_o)^{n\lambda_o}e^{-n\lambda_o}).$$

Simplifying the last expression gives

$$\frac{1}{2}\log\frac{n+1}{2\pi\lambda_o} + \lambda_o,$$

which is off by -1/2 from the result of the theorem: it is

$$\frac{1}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log I(\lambda_o) + \log\frac{1}{e^{-\lambda_o}}.$$

Note that the error of 1/2 comes from the fact that we have not evaluated the Chi-square term in (2).

To complete this example we consider a slight modification of the last joint density. We change the prior by relocating the exponential at $a > 0$. The full model is now

$$e^{-(\lambda - a)}\lambda^{\sum_{i=1}^{n} k_i}\frac{e^{-n\lambda}}{k_1!\ldots k_n!}.$$

The mixture density is

$$m(k^n) = \int_a^\infty e^{-(\lambda - a)}\lambda^{\sum_{i=1}^{n} k_i}\frac{e^{-n\lambda}}{k_1!\ldots k_n!}d\lambda,$$

which can be transformed to give an incomplete Gamma function by a change of variables. Incomplete Gamma functions are difficult to evaluate. Fortunately, for $\lambda > a$, the theorem still applies so we have

$$D(P_\lambda^n \,||\, M_n) = \frac{1}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\frac{1}{\lambda} + \log\frac{1}{e^{-(\lambda - a)}} + o(1),$$

since the Fisher information and the value of the prior are easy to compute.

Chapter 3: The Bayes' Cumulative Risk

3.1 Introduction

In this chapter we extend the work of Chapter 2 so as to obtain an asymptotic expression for the Bayes' cumulative risk of the Bayes' estimator. The Bayes' risk is the result of integrating the risk with respect to the prior and the risk of the Bayes' estimator was approximated in the last chapter. If the prior concentrates on a compact set K then a uniformly good approximation to the risk should integrate to give a good approximation of the Bayes' risk. That is, we hope

$$\int_K D(P_\theta^n \,||\, M_n)\, w(\theta) d\theta$$

$$= \frac{d}{2}\log\frac{n}{2\pi e} + H(\Theta) + \frac{1}{2}\int w(\theta) \log \det I(\theta)\, d\theta + \int o(1) w(\theta) d\theta, \qquad (1)$$

in which $H(\Theta)$ is the entropy of the prior, and $o(1)$ is a function of n and θ which satisfies

$$\lim_{n \to \infty} \sup_{\theta \in K} | o(1) | = 0,$$

so that the right hand side of (1) is our candidate approximation, provided each term is finite. The Bernoulli case with $Beta(q,q)$ prior is an example in which the terms are finite and the identity is valid.

One approach to the problem would be to integrate the decomposition from equation (2) of Chapter 2:

$$E_w D(P_\theta^n \,||\, M_n) = E_w E_{p_\theta} \log \frac{p(X^n \mid \theta)}{p(X^n \mid \hat{\theta})} + E_w E_{p_\theta} \log \frac{p(X^n \mid \hat{\theta})}{m(X^n)},$$

and prove the following observations to be true for compact parameter spaces.

1) The MLE $\hat{\theta}$ is uniformly close to its true value at the desired rate:

$$\sup_{\theta \in K} P_\theta(| \theta - \hat{\theta} | > \varepsilon) = o(1/n).$$

2) For each $\theta \in K$ we have that

$$2 \log \frac{p(X^n \mid \hat{\theta})}{p(X^n \mid \theta)} \to \chi_d^2,$$

in P_θ distribution, uniformly, with expected values matching as well, in the limit.

3) The Laplace integration to approximate the integrand of the second term can be done uniformly.

4) The errors introduced by 2) and 3) tend to zero uniformly.

Pointwise versions of these conclusions were shown in Clarke and Barron (1988), and we believe that the uniform versions can be established with techniques similar to those used there. However, the conditions required to use the MLE would be more stringent than those which emerge from making the proof given in Chapter 2 hold uniformly.

The uniformization of either proof for the approximation of $D(P_\theta^n \,||\, M_n)$ breaks down when the support of the prior is not compact. Aside from the technical difficulty of quantities which, for the sake of the proof must be finite, yet go to infinity on non-compact spaces - suprema of expected squared logs of density ratios in the MLE type proof, suprema of Fisher information being infinite in either, for example - we suspect that uniformity is not the right criterion because those parameters which correspond to regions of small prior density should make a relatively smaller contribution to the approximation.

That was the motivation for considering the information theoretic aspects of the problem. One can recognize the cumulative Bayes' risk as the Kullback - Leibler distance between the joint distribution for (Θ, X^n) and the product of the marginal distributions for Θ and X^n. This is the Shannon mutual information, which we denote $I(\Theta; X^n)$. Explicitly, we have

$$\int_K D(P_\theta^n \,||\, M_n) w(\theta) d\theta = \int_K \int_{\mathbf{R}^n} p_\theta(x^n) \log \frac{p_\theta(x^n)}{m_n(x^n)} dx^n w(\theta) d\theta$$

$$= D(P_{\Theta, X^n} \,||\, P_\Theta \times P_{X^n})$$

$$= I(\Theta; X^n)$$

$$= H(\Theta) - H(\Theta \,|\, X^n).$$

We have denoted the conditional entropy of Θ given X^n by

$$H(\Theta | X^n) = \int H(\Theta \,|\, X^n = x^n) m(x^n) \, \lambda(dx^n)$$

$$= \int m(x^n) \int p(\theta \mid x^n) \log \frac{1}{p(\theta \mid x^n)} \, d\theta \lambda(dx^n).$$

We have seen that the entropy $H(\Theta)$appeared in our candidate approximation. The other terms in (1) are seen to approximate the conditional entropy $H(\Theta \mid X^n)$. We find that it is easier in the non-compact case to examine this conditional entropy than to uniformize the approximation to $D(P_\theta^n \mid\mid M_n)$. We shall evaluate a suitable lower bound for $I(\Theta; X^n)$ by maximizing an entropy. To upper bound $I(\Theta; X^n)$ we will use bounds on $D(P_\theta^n \mid\mid M_n)$.

3.2 An Upper Bound

In this section we convert a pointwise version of Theorem 2.2.1 into an averaged version by use of the dominated convergence theorem. The key idea is to use a uniform version of the same inequality as was used in the proof of Proposition 2.3.2. We assume that the second order Taylor expansion of the Kullback-Leibler distance between two elements of the parametric family uniformly upper bounds that Kullback-Leibler number. It turns out that we must also make a hypothesis which essentially forces the logarithm of the prior to be uniformly continuous.

If the parameter space is compact then the proof simplifies: the hypotheses are not needed since continuous functions on compact sets are uniformly continuous there and the existence of two continuous derivatives will guarantee that the second order Taylor series is a good approximation.

Theorem 3.2.1: *Assume that* $H(\Theta)$ *is finite,*

$$\int \mid \log \det I(\theta) \mid w(\theta) d\theta < \infty,$$

and that the hypotheses of Theorem 2.2.1 hold for each θ *in the support of the prior. Assume that there is a positive* δ *and a constant c so that for all* θ', θ *in the support of the prior with*

$$D(\theta \mid\mid \theta') \leq \delta,$$

we have that

$$D(P_\theta \mid\mid P_{\theta'}) \leq \frac{c}{2}(\theta - \theta')^t I(\theta)(\theta - \theta'),$$

and that

$$w(\theta') \geq \frac{w(\theta)}{c}.$$

Then we have the upper bound

$$\limsup_{n \to \infty} [\, I(\Theta; X^n) - \frac{d}{2}\log\frac{n}{2\pi e} - \frac{1}{2}\int \log\det I(\theta) w(\theta) d\theta - H(\Theta)] \leq 0.$$

Proof: From an inequality due to Barron (1987) we have that

$$D(P_\theta^n \,||\, M_n) \leq n\varepsilon - \log W(\{\theta' \,|\, D(\theta \,||\, \theta') < \varepsilon\}).$$

The prior probability of the neighborhood of θ can be lower bounded. We let $\varepsilon = \varepsilon_n = d/n$ and assume n is so large that $d/n < \delta$. Now we have

$$\begin{aligned} W(\{\theta' \,|\, D(\theta \,||\, \theta') < \varepsilon_n\}) &\geq W(\{\theta' \,|\, c(\theta' - \theta)^t I(\theta)(\theta' - \theta) < \varepsilon_n\}) \\ &\geq \frac{w(\theta)}{c} \int_{\{c(\theta' - \theta)' I(\theta)(\theta' - \theta) < \varepsilon_n\}} d\theta' \\ &= \frac{w(\theta)}{c'} \det I(\theta)^{-1/2} \left(\frac{d}{nc}\right)^{d/2}, \end{aligned}$$

where c' is the result of absorbing the volume of the unit ball in d dimensions into c. Now we have that

$$D(P_\theta^n \,||\, M_n) - \frac{d}{2}\log n \leq d - \log w(\theta) + \log c' + \frac{1}{2}\log\det I(\theta) - \frac{d}{2}\log\frac{d}{c}.$$

in which the upper bound is independent of n and integrable with respect to θ. Now we can apply the dominated convergence theorem to see that by the pointwise convergence from Theorem 2.2.1 we have

$$\begin{aligned} \limsup_{n \to \infty} I(\Theta, X^n) - \frac{d}{2}\log n &\leq \int \limsup_{n \to \infty} [\, D(P_\theta^n \,||\, M_n) - \frac{d}{2}\log n \,] w(\theta) d\theta \\ &= \int [\frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{1}{w(\theta)} + \frac{1}{2}\log\det I(\theta)]\, w(\theta) d\theta \,. \,\square \end{aligned}$$

An improved version of that result would eliminate the above condition on the prior, and assume that the set on which the Taylor expansion failed to be a good upper bound had probability decreasing to zero fast enough that its contribution to the mutual information tended to zero. This means that the probability of the set would have to be decreasing at least at rate $o(1/\log n)$, a sort of prior consistency. This would mean that c would depend on n also, and increase slowly. Then one would decompose the mutual information into two parts: an integral over the good set in the parameter space and an integral over its complement which would go to

zero.

Another approach is to use an identity due to Barron (1988) which was used earlier. We recall that from Chapter 1, Section 6,

$$D(P^n_{\theta_o} \,||\, M^n) + E_{\theta_o} D(w^* \,||\, w(\cdot \,|\, X^n)) = -\log \int_{R^m} e^{-nD_\theta} w(\theta) d\theta, \qquad (2)$$

where $D_\theta = D(\theta_o \,||\, \theta)$. An upper bound on the first term of (2) can be derived from the pointwise convergence of $E_\theta D(w^* \,||\, w(\cdot \,|\, X^n))$ to d/2 and being able to perform a uniform Laplace integration on the right hand side. The uniformization of the Laplace integration will require conditions on D_θ also. Although they will be different from those of Theorem 3.2.1 it is not clear whether they are weaker or easier to verify.

Priors, such as the normal, which tail off quickly generally do not have a uniformly continuous logarithm. Intuitively, from looking at the identity (2), they are the ones which assign most mass near the maximum. Indeed, as the examples suggest, an upper bound of the desired form for their asymptotic behavior exists and is not contained in the above result.

We therefore present an upper bound which holds for an appropriate sequence of compact sets and gives the expected approximation on that sequence. We hope eventually to prove that the sequence of complements gives an asymptotically negligible contribution. The hypotheses will not require the uniform continuity of the logarithm of the prior, or that the Taylor expansion provide a uniform upper bound. However, the result is weaker also: the approximation is only along a sequence.

Since we will be integrating the identity (2) with respect to the prior we must first cut down to the set on which our approximation will be valid. Two types of constraints are required to define the domain. Let

$$U_n = \{\, \theta \mid \text{for } (\theta - \theta')^t I(\theta)(\theta - \theta') < \delta_n, \text{ we have}$$

$$1)\ D(\theta \,||\, \theta') \le \frac{(1 + \varepsilon_n)}{2}(\theta - \theta')^t I(\theta)(\theta - \theta'), \text{ and}$$

$$2)\ |w(\theta) - w(\theta')| \le \xi_n,\ w(\theta) \ge 2\xi_n \,\},$$

where the sequences ξ_n, ε_n, and δ_n are positive and thought of as decreasing. The key hypothesis is on how quickly U_n increases to the support of the prior.

Proposition 3.2.1: *Assume the desired approximation makes sense: $H(\Theta)$ is finite and*

$$\int |\log \det I(\theta)| w(\theta) d\theta < \infty,$$

that the hypotheses of Theorem 2.2.1 hold for each θ *in the support of the prior. Assume that we can choose the sequences* ξ_n, ε_n, *and* δ_n *to be positive and tending to zero so that*

$$n\delta_n \to \infty,$$

and

$$W(U_n^c) = o\left(\frac{1}{\log n}\right).$$

Then we have the upper bound

$$\limsup_{n \to \infty} \left[\int_{U_n} w(\theta) D(P_\theta^n \| M_n)\, d\theta - \frac{d}{2} \log \frac{n}{2\pi e} \right.$$

$$\left. - \frac{1}{2} \int w(\theta) \log \det I(\theta)\, d\theta - H(\Theta) \right] \leq 0.$$

Remark: If we assume that $W(U_n^c) = o(1/n)$ then the integral over U_n^c is asymptotically negligible by a calculation using Jensen's inequality. However, we have been unable to find an example which satisfies that rate assumption. We suspect that one can prove $W(\{w(\theta) \leq a_n\})$ is essentially of order a_n since we are evaluating the prior measure of a set which is typically defined by upper bounding the parameter itself.

Proof: Integrating (2) over U_n gives

$$\int_{U_n} w(\theta) D(P_\theta \| M_n) d\theta = -\int_{U_n} \log \left[\int_{R^d} e^{-nD(\theta \| \theta')} w(\theta') d\theta' \right] w(\theta) d\theta$$

$$- \int_{U_n} w(\theta) E_\theta D(w^* \| w(\cdot \mid X^n)) d\theta.$$

By Lemma 3.3.3 the integrand of the second term converges to $d/2$ pointwise and is positive. Applying Fatou's lemma we see that the limit inferior of its integral is bounded below by $d/2$ also.

For the first term we use Laplace integration uniformly. On U_n we have that

$$\int_{R^d} e^{-nD(\theta \| \theta')} w(\theta') d\theta' \geq \int_{\theta' : |\theta - \theta'| < \delta_n} e^{-nD(\theta \| \theta')} w(\theta') d\theta'$$

$$\geq \int_{\theta' : |\theta - \theta'| < \delta_n} e^{-n\frac{(1+\varepsilon_n)}{2}(\theta - \theta')' I(\theta)(\theta - \theta')} w(\theta') d\theta'$$

$$\geq (w(\theta) - \xi_n) \left[\int_{R^d} e^{-n\frac{(1+\varepsilon_n)}{2}(\theta - \theta')' I(\theta)(\theta - \theta')} d\theta' \right.$$

$$\left. - \int_{\theta' : |\theta - \theta'| \geq \delta_n} e^{-n\frac{(1+\varepsilon_n)}{2}(\theta - \theta')' I(\theta)(\theta - \theta')} d\theta' \right]$$

$$\geq w(\theta)(2\pi)^{d/2}|nI(\theta)|^{-1/2}(1 - \frac{\xi_n}{w(\theta)})(1 + \varepsilon)^{-d/2}$$

$$\times (1 - e^{-n\delta_n}2^{d/2}).$$

Integrating the negative logarithm of that over U_n gives

$$\int_{U_n} w(\theta) D(P_\theta \| M_n) d\theta = \int_{U_n} w(\theta) \left[\log\frac{1}{w(\theta)} + \frac{d}{2}\log\frac{1}{2\pi} + \frac{1}{2}\log\det I(\theta) \right.$$

$$\left. + \frac{d}{2}\log n + \log \frac{(1 + \varepsilon_n)^{d/2}}{(1 - e^{-n\delta_n/4}2^{d/2})} - \log\left(1 - \frac{\xi_n}{w(\theta)}\right) \right] d\theta.$$

As U_n increases the first three terms in the integrand, with the $-d/2$, tend to the constants

$$H(\Theta) + \frac{d}{2}\log\frac{1}{2\pi e} + \frac{1}{2}\int w(\theta) \log\det I(\theta)\, d\theta.$$

The fourth term of the integrand is $W(U_n)(d/2)\log n$ which is asymptotically equivalent to $(d/2)\log n$ by the rate assumption on the prior probabilities. The other two terms go to zero as n increases: the fifth because $n\delta \to \infty$ and $\varepsilon_n \to 0$ and the sixth by the dominated convergence theorem because

$$\chi_{U_n} \log\left(1 - \frac{\xi_n}{w(\theta)}\right) \to 0,$$

almost everywhere with respect to w, and $0 \leq \xi_n/w(\theta) \leq 1/2$. Thus the limit superior is as claimed. □

The sets U_n force uniformity by requiring a lower bound on the prior and by requiring that the second order Taylor expansion be good on a sequence of sets which is increasing fast enough. In this result, it is as if we have considered a sequence of priors with compact support which tend to the true prior. For, the θ's have been weighted with either their true relative weight or with zero. In the theorem before, they were all weighted with their true relative weights. At the end of this chapter we will give an example to show that the rate assumption on U_n can be satisfied.

3.3 Bounds in the Compact Case

From chapter two we recall the sets $A_n(\delta, \varepsilon, \theta)$, $B_n(\delta, \varepsilon, \theta)$, and $C_n(\delta, \varepsilon, \theta)$, see expressions (7), (8), and (9). They were introduced so that tight upper and lower bounds on the log density ratio between a member of the parametric family and the mixture could be found, and the complements of the sets on which those bounds were valid had probability decreasing to zero fast enough. Rather than assessing convergences under a fixed θ_o we consider an arbitrary element of the parameter space since we want our results to be valid over a compact set. Under uniform hypotheses it is clear that, pointwise, the statement of Proposition 2.2.1 is clearly valid, for all θ as is Theorem 2.2.1. It is only the rates of decrease of probabilities which may not hold uniformly in θ, the n which is large enough for one value of θ may not be large enough for another. The slightly stronger assumptions we have introduced below control the rates so that the formula is valid uniformly for θ in a given compact set K.

Since we are extending a proof which has already been given we will not go over all the details, we will merely show that the terms which arise can be controlled. The first step in obtaining that control is the following lemma which gives a rate of decrease for the quantities appearing in Chapter 2 in the proof of Theorem 2.2.1. We assume a rate of decrease on a set and make moment assumptions so that the rate of decrease of the expected value of a quantity of interest is given in terms of the rate on the set.

Lemma 3.3.1: *Consider two sequences of random variables X_i, Y_i which are i.i.d. and satisfy X_i is independent of Y_j if and only if $i \neq j$. Let U_n be a set such that*

$$P_\theta(U_n) \leq c(\theta) f(n),$$

where c is continuous on K. We assume that for each $\theta \in K$, X_i and Y_j are mean zero with finite variance and that

$$E_\theta X_1^2 Y_1^2 < \infty.$$

Then using an overbar to denote a sample average, we have that there is a bounded function c_1 on K so

$$nE_\theta 1_{U_n} \overline{X}\ \overline{Y} \leq c_1(\theta)\sqrt{f(n)}.$$

Proof: By the Cauchy - Schwartz inequality and the hypotheses we have

$$nE_\theta 1_{U_n} \overline{X}\ \overline{Y} = \frac{1}{n} E_\theta \sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j 1_{U_n}$$

$$\leq \frac{1}{n} P_\theta(U_n)^{1/2} \sqrt{E_\theta(\sum_{i=1}^n X_i \sum_{j=1}^n Y_j)^2}$$

$$= \frac{\sqrt{c(\theta) f(n)}}{n} \sqrt{E_\theta(\sum_{i=1}^n X_i \sum_{j=1}^n Y_j \sum_{k=1}^n X_k \sum_{l=1}^n Y_l)}.$$

We can expand the second factor and see that most of the terms are zero. The ones that aren't give the upper bound

$$\frac{\sqrt{c(\theta) f(n)}}{n} \sqrt{nE_\theta X_1^2 Y_1^2 + n^2 E_\theta X_1^2 E_\theta Y_1^2 + 2n^2 E_\theta (X_1 Y_1)^2}$$

$$\leq \sqrt{c(\theta)} \sqrt{3E_\theta X_1^2 Y_1^2 + (1/n)\, E_\theta X_1^2 E_\theta Y_1^2} \sqrt{f(n)},$$

which gives the stated result. □

We will apply the lemma to each of the three quantities which must be controlled for the approximation of Theorem 2.2.1 to be valid on K. Our result gives upper bounds as well as lower bounds. To a certain extent the upper bounds duplicate the result of the last section; however, here we have a result uniform in θ, rather than one which is true only after the expectation with respect to θ has been taken. We state our result as the following.

Theorem 3.3.1 *Suppose that* $w(\theta)$ *and* $\det I(\theta)$ *are bounded away from zero on* K, *and that for each* θ *there is an* $\varepsilon > 0$ *so that for any* $\alpha > 0$ *we have*

$$\sup_{\tilde{\theta} \in B(\theta, \varepsilon)} P_{\tilde{\theta}}(W(N(\tilde{\theta}, \delta)^c \mid X^n) > \alpha) = o(\frac{1}{\log n}). \tag{3}$$

Assume that for all k *and all* l *there is a* $\delta > 0$ *which satisfes*

$$\sup_{\tilde{\theta} \in B(\theta, \varepsilon)} E_{\tilde{\theta}} \sup_{\theta' : |\tilde{\theta} - \theta'| < \delta} (\frac{\partial^2}{\partial \theta_k \partial \theta_l} \log p(x_1 \mid \theta'))^4 < \infty. \tag{4}$$

Then for any compact set K *we have that*

$$\lim_{n \to \infty} \sup_{\tilde{\theta} \in K} \left| D(P_{\tilde{\theta}}^n \mid\mid M_n) - \frac{d}{2} \log \frac{n}{2\pi e} - \frac{1}{2} \log \det I(\tilde{\theta}) - \log \frac{1}{w(\tilde{\theta})} \right| = 0.$$

Proof: We first show the conclusion for the case of a small compact set contained in one $B(\theta, \varepsilon)$ and then extend. Note that the expected local supremum on the fourth moment of the second derivative implies that the first derivative is controlled, that is we also have for each i

$$\sup_{\tilde{\theta} \in B(\theta, \varepsilon)} E_{\tilde{\theta}} (\frac{\partial}{\partial \theta_i} \log p(x_1 \mid \tilde{\theta}))^4 < \infty.$$

Examination of the proof of Theorem 2.2.1 shows that the upper and lower bounds on $D(P_{\tilde{\theta}}^n \| M_n)$ depend as follows. For the lower bound we require that

$$P_{\tilde{\theta}}((A \cap B)^c) = o(\frac{1}{\log n}), \tag{5}$$

uniformly in $\tilde{\theta}$ and for the upper bound we require that

$$P_{\tilde{\theta}}(B^c), P_{\tilde{\theta}}(C^c) = o(\frac{1}{n}). \tag{6}$$

General conditions under which the assumption on A_n can be met can be derived by extending the reasoning in Chapter 2, in the section on posterior consistency. However, in the present context it is enough to show that (5) and (6) are satisfied. We extend the pointwise results of Chapter 2, Section 2. There, we identified three quantities which had to tend to zero if the theorem were to be true. We do that here and prove the desired convergences.

Given the stated assumptions we will see that use of Lemma 3.3.1 implies

$$\lim_{n \to \infty} \sup_{\tilde{\theta} \in B(\theta, \xi)} E_{\tilde{\theta}}(1_U) \, (\sqrt{n} \, (\, i^*_{j,k}(\tilde{\theta}) - i_{j,k}(\tilde{\theta}) \,) \,)^2 = 0, \tag{7}$$

$$\lim_{n \to \infty} \sup_{\tilde{\theta} \in B(\theta, \xi)} E_{\tilde{\theta}}(1_V) \, (\, \sqrt{n} \, (\sup_{|\tilde{\theta} - \theta'| < \delta} | \, i^*_{j,k}(\theta') - i^*_{j,k}(\tilde{\theta}) | \tag{8}$$

$$- E_{\tilde{\theta}} \sup_{|\tilde{\theta} - \theta'| < \delta} | \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log p(X \mid \theta') - \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log p(X \mid \tilde{\theta}) \,) \,)^2 = 0,$$

and

$$\lim_{n \to \infty} \sup_{\tilde{\theta} \in B(\theta, \xi)} E_{\tilde{\theta}} 1_{C_n^c} \, n \, l'(\tilde{\theta}) I(\tilde{\theta})^{-1} l'(\tilde{\theta}) = 0, \tag{9}$$

where U is the event

$$U = \{ | \, i^*_{j,k}(\tilde{\theta}) - i_{j,k}(\tilde{\theta}) | > \frac{\varepsilon}{2} \},$$

and V is as in Chapter 2, Section 2. Expressions (7), (8), and (9) are the quantities which must be made uniformly close to zero for θ in K which is stronger than the statement that pointwise in θ the limits are zero. For case (9), the easiest, we have that

$$E_{\tilde{\theta}} 1_{C_n^c} \, n \, l'(\tilde{\theta}) I(\tilde{\theta})^{-1} l'(\tilde{\theta}) = tr \, I^{-1}(\tilde{\theta}) n E_{\tilde{\theta}} 1_{C_n^c} l'(\tilde{\theta}) l'^t(\tilde{\theta}),$$

and the (i,j) entry of the matrix is

$$n E_{\tilde{\theta}} 1_{C_n^c} \frac{1}{n} \frac{\partial}{\partial \theta_i} \log p(x^n \mid \tilde{\theta}) \frac{1}{n} \frac{\partial}{\partial \theta_j} \log p(x^n \mid \tilde{\theta}).$$

Since fourth moments are controlled, and from Chapter 2, equation (15), we have that

$$P_{\tilde{\theta}}(C_n^c) = O(\frac{1}{n}),$$

Lemma 3.3.1 gives that

$$nE_{\tilde{\theta}} 1_{C_n^c} l'(\tilde{\theta}) l'^t(\tilde{\theta}) = O(\frac{1}{\sqrt{n}})$$

which clearly goes to zero uniformly for $\tilde{\theta}$ in $B(\theta, \varepsilon)$, by continuity, so (9) holds.

For case (7) we consider the centered random variables

$$X_i = Y_i = \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_i \mid \tilde{\theta}) - E_{\tilde{\theta}} \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_1 \mid \tilde{\theta}).$$

By assumption their fourth moments are uniformly controlled, and by Chapter 2, expression (17), we have that

$$P_{\tilde{\theta}}(U) = O(\frac{1}{n}),$$

so, by Lemma 3.3.1, we conclude that (7) holds.

Similarly, for case (8), we consider the centered random variables

$$X_i = Y_i = \sup_{|\tilde{\theta} - \theta'| < \delta} \left| \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_i \mid \theta') - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_i \mid \tilde{\theta}) \right|$$

$$- E_{\tilde{\theta}} \sup_{|\tilde{\theta} - \theta'| < \delta} \left| \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_i \mid \theta') - \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(x_i \mid \tilde{\theta}) \right|.$$

Since the fourth moments are uniformly controlled by assumption, and from Chapter 2 expression (18) we have that

$$P_{\tilde{\theta}}(V) = O(\frac{1}{n}),$$

Lemma 3.3.1 implies that (8) is satisfied.

Now, for an arbitrary compact set K, choose a finite open cover $B_i = B(\theta_i, \xi_i)$ where i runs over the finite index set so that both supremal conditions, (3) and (4), are satisfied on each open set in the covering. On each B_i we have the desired bound holding uniformly in θ. Taking the maximum or minimum of finitely many such bounds goes to zero as well. □

We remark that continuity in θ of the expectations appearing in (7), (8) and (9) is not enough. We had to rule out the possibility that the supremum as a function of

n remains bounded away from zero.

Corollary 3.3.1: *Under the conditions of Theorem 3.3.1, we obtain the desired approximation for the Bayes' risk of the Bayes' estimator:*

$$I(\Theta, X^n) = \frac{d}{2}\log\frac{n}{2\pi e} + H(w) + \frac{1}{2}\int (\log\det I(\theta))\, w(\theta)d\theta + o(1).$$

Proof: Since the supremum in the conclusion of Theorem 3.3.1 tends to zero, so also does the average with respect to $w(\theta)$. The corollary follows. □

Remark: If all we wanted was an approximation to the mutual information then we did not really need the uniformity. All we needed was the pointwise result and then conditions sufficient to take the limit of the integrals. We used the extra strength so that Jeffreys' prior could be identified. Indeed, this result is implied by Theorem 3.2.1 and Theorem 3.4.1, which was tacitly assumed in the Beta-Bernoulli example in Chapter 1, Section 3.

We next turn to a weaker version of a similar result, from an information theoretic viewpoint, by use of the identity (2). We will get an upper bound which differs from the asymptotically accurate one by $d/2$. This will be easy given the following. We recall the definition of the tilted prior:

$$w^*(\theta) = \frac{w(\theta)e^{-nD(\theta_o \| \theta)}}{c_n},$$

where c_n is the normalizing constant.

Lemma 3.3.2: *If $D(P_{\theta_o} \| P_\theta)$ admits the second order Taylor expansion*

$$D(P_{\theta_o} \| P_\theta) = \frac{1}{2}(\theta - \theta_o)^t I(\theta_o)(\theta - \theta_o) + o(\|\theta - \theta_o\|^2),$$

with second derivative equal to the Fisher information matrix, then the normalizing constant is

$$c_n = (1 + o(1))(2\pi)^{d/2} w(\theta)\sqrt{\det(nI(\theta_o))^{-1}},$$

and the error term o(1) can be made uniformly small over compact sets in the parameter space.

Proof: We apply Laplace integration to

$$c = \int_{R^m} e^{-nD(\theta\|\theta')} w(\theta')d\theta'.$$

For fixed θ, we note that $-D(\theta\|\theta')$ has a maximum at $\theta' = \theta$. Thus,

$$\int_{R^m} e^{-nD(\theta\|\theta')} w(\theta')d\theta' \sim w(\theta)\det\left[\frac{n}{2\pi}\frac{\partial^2}{\partial\theta'^2} D(\theta\|\theta')\,|_{\theta' = \theta}\right]^{-\frac{1}{2}}.$$

The second derivative of D is just the Fisher information matrix, so rearranging the last expression gives the approximation pointwise. The uniformity follows from the continuity. □

If we turn (2) into an inequality by removing the posterior term, then integrating out θ with respect to w gives an upper bound on the mutual information:

$$I(\Theta, X^n) \leq -\int_K \left[\log \int_K e^{-nD(\theta \| \theta')} w(\theta') d\theta' \right] w(\theta) d\theta$$

$$\leq \log \frac{n}{2\pi} + \int \frac{1}{2} \log \det I(\theta) w(\theta) d\theta + H(\Theta) + \int w(\theta) o(1) d\theta,$$

which, in the limit, gives the claimed approximation up to $d/2$. The following lemma allows us to recover the $d/2$.

Lemma 3.3.3: *Assume that the hypotheses of Theorem 2.2.1 hold for each* θ. *Then pointwise in* θ *we have*

$$E_\theta D(w^* \| w(\cdot | X^n)) \to d/2.$$

Proof: Recalling that the local supremum condition, equation (4) of Chapter 2, implies that the Taylor expansion of Lemma 3.2.2 is valid, we use that result and Theorem 2.2.1 to identify the behavior of two of the terms in equation (2). This allows us to solve for the third, which gives the result in the lemma. □

By the uniformity of Theorem 3.3.1 and of Lemma 3.2.2 we know that the pointwise convergence is actually uniform for θ in compact sets.

3.4 Lower Bound in the Noncompact Case

In this section we give conditions under which an asymptotic lower bound of the desired form holds without assuming compactness. We will be concerned with the behavior of various quantities under the mixture distribution for X^n, and under joint distribution for Θ and $X_1, \ldots, X_n, \ldots$ in order to obtain a lower bound. The intuition behind Theorem 3.4.1 was suggested by Bernardo (1979), and Hartigan (1983), but they offered no proof. It is here that we use the maximum entropy argument.

In the proof of the theorem to follow we will want to upper bound the expectation of the log determinant of a conditional variance. A result we need is that n times the posterior variance converges to the inverse Fisher information matrix. To obtain the result we tried to directly apply Laplace integration to the posterior variance as in Chapter 2 but we have not been able to control the error term. We therefore used a different technique motivated by the work of Bickel and Yahav (1969). In their paper they used consistency of the MLE in order to prove the asymptotic normality of the standardized posterior, in a strong mode of convergence. It is unnatural to use the MLE in proving that the variance of the Bayes' estimator is what it should be, however, we are as yet unable to obtain the result by any other method.

Proposition 3.4.1: *We make the following assumptions:*

1) The parameter has a finite second moment:

$$\int \theta^t \theta w(\theta) d\theta < \infty.$$

2) For each θ *there is an* $\varepsilon = \varepsilon(\theta) > 0$ *so that the expected suprema of the second derivatives is finite:*

$$E_\theta \sup_{|\theta - \psi| < \varepsilon} \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X_1 | \psi) \right| < \infty.$$

3) The Fisher information is positive definite for each θ.

4) For each θ *there is a* $\rho = \rho(\theta)$ *large enough that*

$$E_\theta \sup_{\psi: |\theta - \psi| > \rho} \log \frac{p(X_1 | \psi)}{p(X_1 | \theta)} < 0.$$

5) For each θ_o *and for any* $\delta > 0$ *small enough we have that for each* θ

$$E_{\theta_o} \log \frac{p(X_1 | \theta_o)}{\sup_{\theta': |\theta - \theta'| < \delta} p(X_1 | \theta')} < \infty.$$

6) For each x, *as* $||\theta||$ *increases,* $p(x | \theta) \to 0$.

Then we have for each θ_o *that*

$$n \operatorname{cov}(\theta | X^n) \to I^{-1}(\theta_o),$$

in P_{θ_o} *probability.*

Remark: Hypotheses 4), 5), and 6) are originally due to Wald (1949). In work due to Bickel and Yahav (1969) only one of Wald's assumptions was used, a mistake

since in the context of their work, that requires 4) to hold for arbitrarily small values of ρ, see pg. 263, equation (2.30). It can be shown that as ρ shrinks to zero the expression in 4) tends to a positive value, so their condition cannot be satisfied. Their result is corrected by use of all three Waldean assumptions: Their Lemma 2.6 is then true by Wald's original technique of covering $B(\theta, \varepsilon)^c$ by the union of $B(\theta, \rho)^c$, where ρ satisfies 4), with finitely many small balls B_i each satisfying 5). The mistake was noted by A. R. Barron.

It can be seen that there is hope of significantly weakening the hypotheses by using a Laplace integration argument on n cov$(\theta | X^n)$, thereby removing the Wald conditions 4), 5), and 6), for the consistency of the MLE. Thus Proposition 3.4.1 would play a role in the lower bound here analogous to the role played by equation (11) of Chapter 2, from Proposition 2.2.1 in Theorem 2.2.1.

Proof of Proposition 3.4.1: We first derive a modified form of a result due to Bickel and Yahav (1969). The modification is that the standardized posterior located at the Bayes' estimate of the parameter, rather than the MLE, is asymptotically normal in an L^2 sense. Here we tacitly assume a version of the MLE restricted to a small open set about the true value of the parameter as justified by Lemma 2.1 in Bickel and Yahav (1969). Fix a value of θ, which we shall take as being true. Let $\tilde{\theta}$ denote the posterior mean, the Bayes' estimator under squared error loss, and $\hat{\theta}$ denote the MLE. We denote the normal density with mean 0 and covariance matrix the inverse Fisher information by ϕ.

First we show that $\sqrt{n}(\hat{\theta} - \tilde{\theta}) \to 0$ in P_θ probability: let $v = \sqrt{n}(\theta - \hat{\theta})$ and $u = v + \sqrt{n}(\tilde{\theta} - \hat{\theta})$ then

$$\begin{aligned}\sqrt{n}(\tilde{\theta} - \hat{\theta}) &= E[\sqrt{n}(\theta - \hat{\theta}) | X^n] \\ &= \int \sqrt{n}(\theta - \hat{\theta})\, w(\theta | X^n) d\theta \\ &= \int v \left[\frac{w(\hat{\theta} + v/\sqrt{n} | X^n)}{n^{d/2}} - \phi(v) \right] dv\end{aligned}$$

since ϕ has mean zero. The last expression tends to zero in P_θ probability by Theorem 2.2 of Bickel and Yahav (1969).

Next, we have that:

$$\begin{aligned}&\int v\, v^t \left| \frac{w(\tilde{\theta} + v/\sqrt{n} | X^n)}{n^{d/2}} - \phi(v) \right| dv \\ &= \int v\, v^t \left| \frac{w(\hat{\theta} + \frac{\sqrt{n}(\tilde{\theta} - \hat{\theta}) + v}{\sqrt{n}} | X^n)}{n^{d/2}} - \phi(v) \right| dv\end{aligned}$$

$$= \int (u - \sqrt{n}(\tilde{\theta} - \hat{\theta}))\,(u - \sqrt{n}(\tilde{\theta} - \hat{\theta})^t$$

$$\left| \frac{w(\hat{\theta} + u/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi(u - \sqrt{n}(\tilde{\theta} - \hat{\theta})) \right| du,$$

where we have used the same change of variables. We take an upper bound by adding and subtracting $\phi(u)$ inside the absolute value bars and applying the triangle inequality. We get two terms. One is

$$\int (u - \sqrt{n}(\tilde{\theta} - \hat{\theta}))\,(u - \sqrt{n}(\tilde{\theta} - \hat{\theta})^t \left| \frac{w(\hat{\theta} + u/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi(u) \right| du,$$

which tends to zero by use of Theorem 2.2 from Bickel and Yahav (1969) and the fact that $\sqrt{n}(\tilde{\theta} - \hat{\theta})$ tends to zero with P_θ probability 1, which we have just derived. The other term is

$$\int (u - \sqrt{n}(\tilde{\theta} - \hat{\theta}))(u - \sqrt{n}(\tilde{\theta} - \hat{\theta})^t \left| \phi(u) - \phi(u - \sqrt{n}(\tilde{\theta} - \hat{\theta}) \right| du,$$

which goes to zero by similar reasoning.

We use this last result to prove the proposition. If we let $s = \sqrt{n}(\theta - E(\theta \mid X^n))$, then we have

$$n \operatorname{cov}(\theta \mid X^n) = \int n(\theta - E(\theta \mid X^n))^t(\theta - E(\theta \mid X^n))w(\theta \mid X^n)d\theta$$

$$= \int s\, s^t \left[\frac{w(\tilde{\theta} + s/\sqrt{n} \mid X^n)}{n^{d/2}} - \phi(s) \right] ds + I^{-1}(\theta).$$

The first term goes to zero by the earlier calculation. The second term is exactly what we want. □

The proposition gives conditions under which we have

$$\log \det n \operatorname{cov}(\theta \mid X^n) \to - \log \det I(\theta_o),$$

in P_{θ_o} probability for each θ_o. That is the quantity which will appear in the course of proving the main theorem of this section, a tight, o(1), asymptotic lower bound on the mutual information between Θ and X^n. We next give a lemma which will be used in conjunction with the proposition.

Lemma 3.4.1: *Suppose that for each* θ *we have that*

$$f_n(X^\infty) \to f(\theta),$$

in $P_{X^\infty \mid \theta}$. *Then the convergence holds in the joint measure:*

$$f_n(X^\infty) - f(\theta) \to 0,$$

in P_{Θ, X^∞}.

Proof: Let $\varepsilon > 0$ and note:

$$P_{\Theta, X^\infty}(\mid f_n(X^\infty) - f(\theta) \mid > \varepsilon) = \int_{R^d} w(\theta) P_{X^\infty \mid \theta}\{ \mid f_n(X^\infty) - f(\theta) \mid > \varepsilon\} d\theta,$$

which goes to zero by the dominated convergence theorem. □

The next idea we introduce so as to obtain a lower bound is a one sided version of uniform integrability. Following Chow and Teicher (1978) we say that a sequence of random variables Y_n is uniformly integrable from above if and only if its positive part is uniformly integrable. Equivalent to uniform integrability from above is the condition

$$\lim_{r \to \infty} \sup_n E\ Y_n\ 1_{\{Y_n > r\}} = 0.$$

As with uniform integrability, uniform integrability from above interacts nicely with inequalities. Specifically, if $X_n \geq Y_n$ for each n and X_n is uniformly integrable from above then Y_n is uniformly integrable from above. This follows from noting that X_n is uniformly integrable from above if and only if X_n^+ is uniformly integrable. But, $X_n^+ > Y_n^+$ so Y_n^+ is uniformly integrable, which is equivalent to the uniform integrability from above of Y_n.

We only use uniform integrability from above since obtaining a lower bound on $I(\Theta, X^n)$ will require us to upper bound the conditional entropy term which arises in its definition.

We next prove two quick lemmas for which an explicit statement will be convenient. The first gives sufficient conditions which we will use to show that the quantity of interest is uniformly integrable from above. It is modeled on the proof in Billingsley (1986), pg. 348.

Lemma 3.4.2: *If a sequence of positive random variables* Y_n *satisfies*

$$\sup_n E\ Y_n < \infty,$$

then $Z_n = \log Y_n$ *is uniformly integrable from above.*

Proof: Let g be the exponential function, $g(r) = e^r$. Then, for $r > 1$, the function re^{-r} is decreasing and consequently we have the inequalities

$$0 \leq \sup_n EZ_n 1_{\{Z_n > r\}} = \sup_n E\ g(Z_n) \frac{Z_n 1_{\{Z_n > r\}}}{g(Z_n)}$$

$$\leq \frac{r}{g(r)} \sup_n Eg(Z_n).$$

By assumption the expectation on the right is finite and $r/g(r)$ converges to zero as $r \to \infty$, so the lemma is proved. □

The next lemma uses uniform integrability from above to identify how a limit of expectations is related to the expectation of the limit.

Lemma 3.4.3: *If Y_n is uniformly integrable from above and converges in probability to a random variable Z, then*

$$\limsup_{n \to \infty} E\ Y_n \leq E\ Z.$$

Proof: The proof is easy: Write

$$EY_n = EY_n 1_{\{Y_n \leq r\}} + EY_n 1_{\{Y_n > r\}}.$$

For fixed r, the limit superior of the first term is bounded by $E\ Z\ 1_{\{Z \leq r\}}$ since the random variables $Y_n\ 1_{\{Y_n < r\}}$ are bounded above. For r large enough, the second term is finite by the uniform integrability from above. As r increases we have the desired result. □

Now we state and prove the key result of this section. Here, $\tilde{\theta} = E(\theta \mid X^n)$ and the operator E by itself means expectation with respect to the joint distribution.

Theorem 3.4.1: *Assume the hypotheses of Proposition 3.4.1, that*

$$\limsup_{n \to \infty} n\ E(\theta_i - \tilde{\theta}_i)^2 < \infty,$$

and that

$$\int |\log \det I(\theta)|\ w(\theta) d\theta < \infty.$$

Then we have that

$$\liminf_{n \to \infty} [I(\Theta, X^n) - \frac{d}{2} \log \frac{n}{2\pi e} - \frac{1}{2} \int \log \det I(\theta) d\theta - H(\Theta)] \geq 0.$$

Remark 1: We have written the proof of the lower bound so that it is clearly seen that the extra assumptions are used to identify the constants so as to get the o(1) convergence. Later, we will see that weaker conditions will give O(1) or coarser bounds.

Remark 2: If there is any estimator with Bayes' risk of order $O(1/n)$ then the Bayes' estimator has risk of the same order, since its risk is minimal.

Proof of Theorem 3.4.1: By definition we have that

$$
\begin{aligned}
I(\theta;X^n) &= H(\theta) - H(\theta \mid X^n) \\
&= H(\theta) - \int_{R^n} H(\theta \mid X^n = x^n) m(x^n) \lambda(dx^n) \\
&= H(\theta) - \int_{R^n} H(\theta - \tilde{\theta} \mid X^n = x^n) m(x^n) \lambda(dx^n) \\
&\geq H(\theta) - \frac{1}{2} \int_{R^n} m(x^n) \log [(2\pi e)^d \det E_{w(\cdot \mid x^n)} n(\theta - \tilde{\theta})(\theta - \tilde{\theta})^t] \lambda(dx^n) \\
&= H(\theta) + \frac{d}{2} \log \frac{n}{2\pi e} \\
&\quad - \frac{1}{2} \int_{R^n} \log \det E_{w(\cdot \mid x^n)} \sqrt{n}(\theta - \tilde{\theta}) \sqrt{n}(\theta - \tilde{\theta})^t m(x^n) \lambda(dx^n), \qquad (10)
\end{aligned}
$$

where the inequality comes from the fact that the normal achieves the maximal entropy under a variance constraint.

We will show that $\log \det n \operatorname{cov}(\theta \mid X^n)$ is uniformly integrable from above with respect to the mixture by bounding it with a sum of functions each of which is uniformly integrable from above. An inequality which we will use is due to Hadamard, see Samelson (1974) pg. 228, and is that for any positive definite matrix K with diagonal entries k_{ii},

$$\det K \leq \prod_{i=1}^{d} k_{ii}.$$

Consequently,

$$\log \det K \leq \sum_{i=1}^{d} \log k_{ii} \leq \sum_{i=1}^{d} (k_{ii} - 1).$$

That inequality means we have the following bounds:

$$
\begin{aligned}
\log \det [\, n \operatorname{cov}(\theta \mid X^n) \,] &\leq \sum_{i=1}^{d} \log [\, n \, Var(\theta_i \mid X^n) \,] \\
&\leq \sum_{i=1}^{d} n \, Var(\theta_i \mid X^n) - d.
\end{aligned}
$$

By assumption,

$$\sup_{n,i} E_{M_n} E_{\theta \mid X^n} n(\theta_i - \tilde{\theta}_i)^2 < \infty,$$

so by Lemma 3.4.2 we have that

$$\log E_{\theta \mid X^n} n(\theta_i - \tilde{\theta}_i)^2$$

is uniformly integrable from above. Since uniform integrability from above interacts

nicely with inequalities as described before we can conclude that

$$\log \det [\, n \ \mathrm{cov}(\theta \mid X^n) \,]$$

is uniformly integrable from above, and therefore so is

$$\log \det [\, n \ \mathrm{cov}(\theta \mid X^n) \,] + \log \det I(\theta).$$

By Proposition 3.4.1 we have that

$$\log \det n \ [\, \mathrm{cov}(\theta \mid X^n) \,] + \log \det I(\theta) \to 0,$$

in $P_{X^n \mid \theta}$ probability, for each θ in the support of w, and therefore, by Lemma 3.4.1, in the joint probability of (Θ, X^∞). Now, by Lemma 3.4.3,

$$\limsup_{n \to \infty} E_m \, [\, \log \det n \mathrm{cov}(\theta \mid X^n) \,] \leq \int \log \det I^{-1}(\theta) w(\theta) d\theta.$$

Now, from inequality (10), we have that

$$\liminf_{n \to \infty} [\, I(\Theta; X^n) - H(\Theta) - \frac{d}{2} \log \frac{n}{2\pi e} \,] \geq - \limsup_{n \to \infty} E_{M_n} \, [\, \log \det n \mathrm{cov}(\theta \mid X^n) \,]$$

$$= \int w(\theta) \log \det I(\theta) \, d\theta,$$

which proves the theorem. □

We remark that there is a proof of Hadamard's inequality based on the entropy of normal random variables due to Cover and El Gamal (1983).

This lower bound is of the desired form. The hypotheses are not all that restrictive, although there are a lot of them. They were used only so that the constants could be identified. Given that the integral term may be extremely large if the prior assigns mass to sets close to those points where the Fisher information is zero or infinity, it is helpful to know the constants, for they will in part determine how large n must be for the asymptotic $(d/2)\log n$ behavior to dominate. Despite that, there may be cases where a less accurate approximation will suffice. That allows for simplifications. An application of Jensen's inequality, and a matrix inequality for determinants gives

$$I(\Theta, X^n) \geq H(\Theta) + \frac{d}{2} \log \frac{n}{2\pi e} - \frac{1}{2} \log \det E_{\theta, X^n} \{ n(\theta - \hat{\theta})(\theta - \hat{\theta})^t \},$$

where $\hat{\theta}$ is any estimator of θ. If we choose the Bayes' estimator under the generalized squared error loss in the expectation, then by Lemma 4.5.2 in Lehmann (1983) the right hand side is as large as it can be. If $\hat{\theta}$ is the Bayes' estimator and is Bayes' efficient in the sense of

$$E_{\theta, X^n} \, n(\theta - \hat{\theta})(\theta - \hat{\theta})^t \to \int I(\theta)^{-1} w(\theta) d\theta,$$

then we have a lower bound which is accurate apart from constants. If we were to use a different estimator we might choose $\hat{\theta}$ to be the MLE. Then a familiar Taylor expansion argument shows that $\sqrt{n}(\theta - \hat{\theta})$ converges in distribution to a random vector with covariance $I^{-1}(\theta)$. However showing that the sequence of expected values $E\ n(\theta - \hat{\theta})(\theta - \hat{\theta})^t$ converges is problematic.

One reason why the argument is difficult is that we want to identify a limit in a situation where it is not enough to know what happens at the limit. We are concerned with how that limit is approached. An example illustrates the point. The asymptotic variance for an estimator may exist and the estimator may even be efficient in the sense of the asymptotic distribution having variance which achieves the Cramer - Rao lower bound, however, for every finite n the variance may be infinite. Consider estimating the Fisher information in the Poisson (λ) distribution where $\lambda > 0$, a noncompact parameter space. It is $1/\lambda$ and the MLE for it is the reciprocal of $\hat{\lambda}$ the MLE for λ, which is 0, with positive probability for each n, although the probability tends to zero so that efficiency in the sense of convergence in distribution to $N(0, I(\theta)^{-1})$ still obtains. In our case we must have the variances for finite n converging to the variance of the asymptotic distribution.

3.5 Examples Continued

In this section we give some examples of the results proved in this chapter for the case of noncompact parameter spaces. Recall that from Section 2 we have a tight upper bound for some choices of prior and an upper bound which is weaker but more generally valid. From Section 4 we have a tight lower bound which is valid whenever n times the posterior variance tends to the inverse Fisher information in probability. The latter is true under conditions which include consistency for the MLE.

We remark that it is difficult to find examples which can be approximated directly, let alone evaluated explicitly. The normal is the exception: we can easily take the expectation of the expression derived in Chapter 1 Section 3, with respect to the same prior and get the same result as in our theorem. Alternatively, we can use some information-theoretic reasoning. The average $\overline{X}$ is a sufficient statistic for the

parameter, and $(\Theta, \bar{X})$ is jointly normally distributed, so by direct calculation

$$I(\Theta; X^n) = I(\Theta; \bar{X}) = \frac{1}{2}\log(1 + n). \tag{11}$$

This answer is familiar from channel capacity calculations in information theory in the case of a Gaussian channel. Note that the $2\pi e$ does not appear since we have used a normal prior rather than the least favorable Jeffreys' prior. The normal has moments of all orders, satisfies local supremum conditions of all orders in all derivatives, and satisfies the Wald consistency hypotheses. Also the average is an unbiased estimator of the mean which upper bounds the Bayes' risk of the Bayes' estimator under squared error loss and decreases like $1/n$. Thus, our lower bound, Theorem 3.4.1, applies and gives $(1/2)\log n$ equivalent to the answer above.

The hypotheses of our upper bound, Theorem 3.2.1, do not hold. The Taylor expansion assumption is certainly valid since the second order Taylor expansion is one half the square distance between the parameters which is the Kullback - Leibler distance. The problem is that the logarithm of the normal prior is not uniformly continuous. If we changed the prior to a standard exponential then we would know that exact upper and lower bounds hold from Theorems 3.2.1 and 3.4.1, and the approximation would then be

$$\frac{1}{2}\log\frac{n}{2\pi e} + 1 + \frac{1}{2}\int_0^\infty e^{-\mu} \log\det I(\mu)\, d\mu,$$

which simplifies to $1/2\log(ne/2\pi)$. In this case, though, the calculations for a direct approximation are quite difficult to carry out. We will return to the normal prior case shortly; for the moment we see that the desired approximation is a lower bound:

$$\frac{1}{2}\log\frac{n}{2\pi e} + H(N(0, 1)) + \frac{1}{2}\int_0^\infty \log\det I(\mu) w(\mu) d\mu$$

$$= \frac{1}{2}\log n, \tag{12}$$

since the entropy of the normal is $(1/2)\log 2\pi e$. Equation (12) is a tight upper bound also on the sequence U_n of compact sets as will be seen. From equation (11), of course, we know that the desired upper bound (12) is exact.

The difficulty in finding an example on a noncompact parameter space which one can approximate without using the theorem arises because the integral defining the mixture becomes intractable and the approximation to it must be valid on the whole parameter space. The hypothesis that the determinant of the Fisher

information be nonzero is necessary for the above calculations. However, it is worth noting that if we had an expansion accurate to order $1/n$ then in cases where the argument of the logarithm is of the form $n+1$ the contribution of the 1 might weaken that hypothesis.

Next we turn to a general class of examples, densities which are of exponential form with a one dimensional parameter. We work in the natural parameterization, as in Chapter 1, Section 3. Thus we consider the family

$$p(x \mid \eta) = \chi_A e^{\eta T(x) + S(x) + u(\eta)},$$

in which u is the normalizing constant, assumed to be at least twice continuously differentiable and χ_A is the indicator function for the set A. We recall that

$$E_\eta T = -u'(\eta),$$

and assume that a prior has been chosen so that the log determinant of the Fisher information, $-u''(\eta)$, has finite expectation under the prior and that the second moment of the parameter is finite.

For the lower bound on $I(\Theta; X^n)$ we note that the hypotheses of Proposition 3.4.1 are all satisfied: the exponential family has all moments of all orders of all derivatives finite; and assumptions 4) and 5) can be verified also. In this case there is no need since we know that the MLE is the average of the $T(X_i)$ and is consistent by the law of large numbers. The other hypotheses for the lower bound hold since the average of the $T(X_i)$ is unbiased and has finite variance, implying that the Bayes' risk of the Bayes' estimator is of order $1/n$. So, the lower bound holds.

For the upper bound, we require that the logarithm of the prior be uniformly continuous, have finite entropy, give finite expectation to the log Fisher information. From the Chapter 1, Section 3 we already know that Theorem 2.2.1 holds pointwise. If the family is written as before then

$$D(\eta \mid\mid \eta') = -u'(\eta)(\eta - \eta') + u(\eta) - u(\eta'),$$

which we require be upper bounded by a quadratic in $\eta - \eta'$ on a strip around the line $\eta = \eta'$. We are unable to identify good conditions to impose on u in order to get a satisfactory upper bound.

Thus, all we have is the lower bound:

$$I(\Theta; X^n) \geq \frac{1}{2}\log\frac{n}{2\pi e} + H(\Theta) + \frac{1}{2}\int \log\left(-u''(\eta)\right) w(\eta)d\eta + o(1).$$

Finally, in parallel to Chapter 2, Section 4, we consider a sequence of i.i.d. Poisson (λ) random variables k_i. We cannot use the exponential prior because it puts mass on neighborhoods of zero where the Fisher information tends to infinity. Instead we consider the location exponential with parameter a fixed and positive, that is we know that the true value parameter is greater than a.

First, we check the assumptions of the lower bound result. The Bayes' risk is of order $1/n$ since the average is unbiased; the Fisher information is $1/\lambda$ and the integral of $\log \lambda$ with respect to the prior is finite. Except for 4), the hypotheses of Proposition 3.4.1 are clearly satisfied since the log density is

$$k_i \log \lambda - \lambda - \log k_i!\,,$$

For 4) in Proposition 3.4.1, the expected supremum is

$$\sup_{\mu:\, |\lambda - \mu| > \rho} \lambda - \mu + \lambda \log \frac{\mu}{\lambda},$$

which tends to $-\infty$ as ρ increases. So the lower bound holds.

No upper bound can be stated because we have been unable to verify the Taylor expansion property. In the proofs of upper bounds here we have basically been expecting the probability to pile up where $D(P_\lambda \| P_\mu)$ is smallest, which is the line $\lambda = \mu$. We then want a neighborhood of that line whose thickness either does not shrink to zero, as λ gets large, or shrinks to zero at the same rate as the quadratic $c(\lambda - \mu)^2$ does, on which the Taylor expansion of the relative entropy between P_λ and P_μ uniformly bounds it from above.

It is worth noting that there are cases in which the diameter of a neighborhood about the line $\lambda = \mu$ can shrink to zero in relative entropy distance even when the parameters are bounded away from each other. We note that the relative entropy between a Poisson(λ) and a Poisson (μ) is

$$D(\lambda \| \mu) = \lambda \log \frac{\lambda}{\mu} + \lambda - \mu,$$

and, if we choose $\lambda_n = n$ and $\mu_n = n + 1$, then $D(\lambda_n \| \mu_n) \to 0$ even though $\mu_n - \lambda_n = 1$.

Finally, we return to the case of the standard normal prior, which we denote by $w(\theta)$, on normal random variables, which we considered in Chapter 1, Section 3. Since one of the hypotheses of Theorem 3.2.1 is not satisfied, we apply Proposition 3.2.1 on the sequence of compact sets U_n which in our case are of the form

$$U_n = \{\theta : \text{ for } (\theta - \theta')^t I(\theta)(\theta - \theta_o) < \delta_n \text{ we have that}$$

$$| w(\theta) - w(\theta') | \leq \xi_n, \text{ and } w(\theta) \geq 2\xi_n \}.$$

The Fisher information is constant so the inner product is with respect to a constant times the identity matrix which is the Euclidean distance which we denote $|\cdot|$. By the mean value theorem we have that there is an M such that

$$| w(\theta) - w(\theta') | \leq M | (\theta - \theta') | \leq M \delta_n .$$

If we choose $\xi_n = \delta_n M = 1/\sqrt{n}$, then

$$n\delta_n^2 \to \infty ,$$

and to show that $W(U_n^c) = o(1/\log n)$ it is enough to verify the rate condition

$$W(\{\theta: w(\theta) \leq \frac{1}{\sqrt{n}}\}) = o(\frac{1}{\log n}),$$

since the continuity condition is automatic. We use the familiar inequality, see Van Trees (1968) pg. 138, that for the standard normal $W(\{\theta > c\}) \leq (1/2)\, e^{-c^2/2}$, so we have that

$$W(\{\theta: \frac{e^{-\theta^2/2}}{\sqrt{2\pi}} \leq 1/\sqrt{n}\}) = 2W(\{\theta: \theta > \sqrt{\log \frac{n}{2\pi}}\})$$

$$\leq e^{-(1/2)\log \frac{n}{2\pi}}$$

$$= \sqrt{\frac{2\pi}{n}},$$

which is clearly $o(1/\log n)$. Now, we have an upper bound which is identical to the lower bound derived earlier but is only valid on a sequence of sets. However, we know that, by (11), $I(\Theta; X^n) = \frac{1}{2}\log(n+1)$, for the normal prior on normal random variables so there must be a result which accounts for it.

3.6 Conclusions

We have identified the underlying mathematical behavior of two quantities which admit diverse physical and statistical interpretations. These include cumulative risk, redundancy, and hypothesis testing for $D(P_\theta^n \| M_n)$ and cumulative Bayes'

risk, average redundancy, channel capacity and posterior convergence for $I(\Theta; X^n)$.

The underlying form in each case is, asymptotically, $(d/2)\log n + c$, and the forms of c have been identified.

References

Aitchison, J. (1975). "Goodness of prediction fit." *Biometrika* (62): 547-554.

Bahadur, R. R. (1971). "Some limit theorems in statistics." in *Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics. Philadelphia.

Barron, A. R. (1985). *Logically Smooth Density Estimation.* Ph. D. thesis, Stanford University.

Barron, A. R. (1987). "Are Bayes rules consistent in information?" in Cover T. M. and Gopinath B., Eds. *Problems in Communications and Computation.* Springer - Verlag. New York.

Barron, A. R. (1988). "The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions." University of Illinois Technical Report # 7.

Barron, A. R. (1989). " Uniformly powerful goodness of fit tests." *Annals of Statistics* (17): 107-124.

Barron, A. R. & Barron, R. L. (1988). Statistical Learning Networks. 1988 Symposium on the Interface: Statistics and Computing Science. Reston, Virginia.

Barron, A. R. & Cover, T. M. (1989). "Minimum complexity density estimation." Submitted to *I.E.E.E. Transactions on Information Theory.*

Berk, R. H. (1970). "Consistency a posteriori." *Annals of Statistics* (41): 894-906.

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society Series B* (41): 113-147.

Bickel, P. & Yahav, J. A. (1969). "Some contributions to the asymptotic theory of Bayes solutions." *Z. Wahrscheinlichkeitstheorie verw. Geb.* (11): 257-276.

Billingsley, P. (1986). *Probability and Measure.* John Wiley and Sons. New York.

Blahut, R. E. (1987). *Principles and Practice of Information Theory.* Addison - Wesley. Reading.

Cencov, N. N. (1981). *Statistical Decision Rules and Optimal Inference.* American Mathematical Society. Providence.

Chernoff, H. (1954). "On the distribution of the likelihood ratio." *Annals of Mathematical Statistics* (25): 573-578.

Chernoff, H. (1956). "Large sample theory: parametric case." *Annals of Mathematical Statististics* (27): 1-22.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory Independence Interchangeability and Martingales.* Springer-Verlag. New York.

Chung, K. L. (1974). *A Course in Probability Theory.* Academic Press. New York.

Clarke, B. & Barron, A. R. (1989). "Information theoretic asymptotics of Bayes methods." University of Illinois Technical Report # 26.

Cover, T. & El Gamal, A. (1983). "An information theoretic proof of Hadamard's inequality." *I.E.E.E. Transactions on Information Theory* (29): 930-931.

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton. Princeton.

Csiszar, I. (1967). "Information-type measures of difference of probability distributions and individual observations." *Studia Sciences Mathematica Hungarica* (2): 299-318.

Davisson, L. D. (1973). "Universal noiseless coding." *I.E.E.E. Transactions on Information Theory* (19): 783-795.

Davisson, L. D. & Leon - Garcia, A. (1980). "A source matching approach to finding minimax codes." *I.E.E.E. Transactions on Information Theory* (2): 166-174.

De Bruijn, N. G. (1958). *Asymptotic Methods in Analysis.* Dover. New York.

De Groot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill. New York.

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press. New York.

Hartigan, J. A. (1983). *Bayes Theory.* Springer-Verlag. New York.

Haughton, D. (1988). "On the choice of a model to fit data from an exponential family." *Annals of Statistics* (16): 342-355.

Hoeffding, W. & Wolfowitz, J. (1958). "Distinguishability of sets of distributions." *Annals of Mathematical Statistics* (29): 700-718.

Ibragimov I. A. & Hasminskii R. Z. (1980). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag. New York.

Jeffreys, H. (1967). *Theory of Probability.* Oxford. New York.

Kiefer, J. & Wolfowitz, J. (1958). "On the deviations of the empiric distribution function of vector chance variables." *Transactions of the American Mathematical Society* (87): 173-186.

Krichevsky, R. E. & Trofimov, V. K. (1981). "The performance of universal encoding." *I.E.E.E. Transactions on Information Theory* (27): 199-207.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley. New York.

Kullback, S., Keegel, J. C., & Kullback, J. H. (1980). *Topics in Statistical Information Theory.* Springer-Verlag. Berlin.

Le Cam, L., (1953). "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates." in Neyman J., Loeve M., and Struve O. Eds. *University of California Publications in Statistics, Volume 1.* Cambridge University Press. London.

Lehmann, E. L. (1983). *Theory of Point Estimation.* Wiley. New York.

Leonard T. (1982). Comment on "A simple predictive density function." *Journal of the American Statistical Association* (77): 657-658.

McCulloch, R. E. (1986). "Information asymptotics and inequalities for posterior and predictive distributions." Submitted to *Canadian Journal of Statistics.*

Rissanen, J. (1984). "Universal coding, information, prediction, and estimation." *I.E.E.E. Transactions on Information Theory* (30): 629-636.

Rissanen, J. (1983). "A universal prior for integers and estimation by minimum description length." *Annals of Statistics* (11): 416-431.

Rissanen, J. (1987). " Stochastic complexity." *Journal of the Royal Statistical Society, Series B* (49): 223-239.

Samelson, H. (1974). *An Introduction to Linear Algebra.* John Wiley and Sons. New York.

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics* (6): 461- 464.

Schwartz, Lorraine (1965). "On Bayes consistency." *Z. Wahrscheinlichkeitstheorie* (4): 10-26.

Strasser, H. (1981). "Consistency of maximum likelihood and Bayes' estimates." *Annals of Statistics* (9): 1107-1113.

Stigler, S. M. (1986). "Laplace's 1774 memoir on inverse probability." *Statistical Science* (1): 359-378.

Tierney, L. & Kadane, J. (1984). " Accurate approximations for posterior moments and marginal densities." University of Minnesota Technical Report # 431.

Tierney, L. & Kadane, J. (1986). " Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association* (81): 82-86.

Van Trees, H. (1968). *Detection, Estimation and Modulation Theory.* Wiley. New York.

Vapnik, V. N. & Chervonenkis, A. (1971). "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability and Its Applications* (16): 264-280.

Wald, A. (1943). "Tests of statistical hypotheses concerning several parameters when the number of observations is large." *Transactions of the American Mathematical Society* (54): 426-482.

Wald, A. (1949). "Note on the consistency of the maximum likelihood estimate." *Annals of Mathematical Statistics* (20): 595-601.

Walker, A. M. (1967). "On the asymptotic behaviour of posterior distributions." *Journal of the Royal Statistical Society, Series B* (31): 80-88.

Wilks, S. S. (1962). *Mathematical Statistics.* Wiley. New York.

Wolfowitz, J. (1949). "On Wald's proof of the consistency of the maximum likelihood estimate." *Annals of Mathematical Statistics* (20): 601-602.

Vita

Bertrand S. Clarke was born 29 July 1963 in Toronto, Canada. He entered University College at the University of Toronto in September of 1980 and graduated May of 1984 in pure mathematics with a minor in statistics. He began his graduate studies in the department of mathematics at the University of Illinois at Urbana - Champaign in August of 1984 and remained there for three and one half years. After that he transferred to the department of statistics where he remained until summer 1989.

While in the department of mathematics he was a teaching assistant for two years. Also he was a research assistant under Dr. J. Mittenthal in the department of cell and structural biology. This led to a publication entitled "An Optimality Criterion in Epimorphic Regeneration," which appeared December 1988, in the Journal of Mathematical Biology. Since 1987 he has been a research assistant under Dr. A. Barron, under whose guidance he completed the requirements for a doctorate in statistics. He has three papers co - authored with Dr. A. Barron, one is University of Illinois Technical Report # 26, one which will appear in the Transactions on Information Theory, and another which is due for submission in the coming months.