LOGICALLY SMOOTH DENSITY ESTIMATION

Andrew R. Barron

LOGICALLY SMOOTH DENSITY ESTIMATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Andrew R. Barron
September 1985

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas M Cover

(Principal Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Abbas El Gamal

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Allen M. Peterson

Approved for the University Committee
on Graduate Studies:

Dean of Graduate Studies & Research

Abstract

A principle for statistical inference is developed using information theory. The motivation is to determine probabilistic laws or theories which provide the shortest complete description of random samples.

Estimates of probability distributions and densities are obtained which provide an asymptotically optimal summary (compression) of the observed data. The data consist of real numbers observed to increasing accuracy. The exact descriptions of the data require two stages: first a candidate distribution P is described using the shortest length computer program for P on a universal computer, then the second stage is the (optimum) Shannon code for the data based on the distribution P. The logically smooth estimates are the distributions which achieve nearly minimum two-stage description length. This approach depends upon both the algorithmic notion of information developed by Kolmogorov and Chaitin and the probabilistic notion of information due to Shannon. Cover, Rissanen, and Sorkin have presented similar estimators.

In this thesis, convergence properties are analyzed in both parametric and non-parametric cases. It is shown that logical smoothing empirically discovers the law for the data. If the true distribution has a finite description, then the estimate is exactly correct for all sufficiently large sample sizes. (For example, from the shortest description of the observed motion of the planets, the true laws of motion are discovered.) If the true distribution is infinitely complex, then a sequence of logically smooth estimates is shown to converge in variation distance.

Logical smoothing may be regarded as a Bayes rule. Nevertheless, the motivation for this approach is the minimization of description length and not any Bayesian inclinations. The proof techniques developed here for logical smoothing apply in general to establish new consistency results for Bayes procedures.

Preface

This September 1985 version differs from the August 1985 version submitted to Stanford University in that Theorems 4.7 and 4.8 are included, the reference list is expanded, and some typographical mistakes are corrected.

Some of the results in this thesis were announced at the 1983 International Symposium on Information Theory. In particular, Theorems 3.3, 3.4, 4.1, and 4.8 were presented (Barron and Cover, 1983).

It is my pleasure to thank my advisor, Professor Thomas M. Cover, for his guidance, encouragement, and support. Professor Cover proposed logical smoothing as an approach to non-parametric inference. His stimulating enthusiam for information theory as a fundamental tool of both engineering and statistics has led us to new contributions in topics as diverse as the asymptotic equipartition property and the central limit theorem.

I am also grateful to my father, Roger L. Barron, a pioneer in the development and application of polynomial learning networks for guidance, control, regression, and pattern recognition. His relentless empirical approach to computer-intensive mathematical modelling convinced me that the number of conceivable models which can be feasibly considered is essentially unlimited. Thus my father's ideas motivated me to develop a theory for unconstrained model selection. It has been a delightful task.

I thank the members of my dissertation reading committee: Professors Cover, Abbas El Gamal, and Allen M. Peterson. Moreover, Professors Cover, El Gamal, Persi Diaconis, and Bruce Lusignan are acknowledged for serving on my oral examination committee.

This work was partially supported by the National Science Foundation under grant ECS 82-11568 at Stanford University.

Andrew R. Barron
Departments of Statistics and
Electrical and Computer Engineering
University of Illinois
Urbana, Illinois 61801

Table of Contents

1. Overview

2. Example

3. Description

4. Inference

Chapter 1. Overview

1.1 Introduction

Let a stochastic process $X_1, X_2, \ldots$ be governed by an unknown probability distribution P^*. We treat the general case that the sequence $\{X_i\}$ is stationary and ergodic as well as the specific case that the X_i are independent and identically distributed (iid). The random variables X_i may be scalar or vector valued. The observed data $\mathbf{X}_n$ consists of the sequence $(X_1, X_2, \ldots, X_n)$ with each X_i revealed to b_n bits accuracy. Estimates of the probability distribution (and its associated mass function or density function) are desired in order to summarize, classify, or predict data.

We show that accurate estimates of the distribution P^* can be obtained from short descriptions of the observed data. Descriptions (or codes) are finite length binary sequences from which the data may be recovered exactly. In particular, we are interested in descriptions which involve candidate distribution estimates P in a natural way.

For any candidate distribution P, there is an exact description of the data $\mathbf{X}_n$ in two stages. First the distribution P is described using $L(P)$ bits. Here $L(P)$ is the length of a binary computer program for P on a universal computer. (A program for a distribution P is an effective procedure for computing the probabilities $P(\mathbf{X}_n)$ to any prescribed accuracy for any $\mathbf{X}_n$ for any n, see Section 3.3. In particular, probabilities may be computed from density functions with simple formuli -- such as the standard normal, the exponential, the Cauchy, and many others. The length of the shortest program for a distribution is a measure of its intrinsic complexity.) Then the second stage is a Shannon code for the data based on the distribution P. The Shannon code assigns a finite length binary sequence from which the data may be recovered, once the distribution P is given. The length of the Shannon code is $\lceil \log 1/P(\mathbf{X}_n) \rceil$ bits (where log is the base 2 logarithm and $\lceil y \rceil$ denotes the least integer not less than y.) Thus the total two-stage description length is $L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil$.

The logically smooth estimator is defined to be a distribution $\hat{P}_n$ which achieves the minimum two-stage description length $\min_P\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil\}$. Logical smoothing achieves the proper balance between low complexity and high likelihood. Redundant or overly-complex estimates are rejected.

Other estimators $\hat{P}$ may also be called logically smooth if the corresponding two-stage description length is not much greater than the minimum. Many popular non-parametric estimators are numerically smooth but not logically smooth -- they do not summarize the data (see section 1.2).

The history of the development of logical smoothing is traced in section 1.2. Key previous work the area is by Cover, Davisson, Rissanen, and Sorkin. In chapter 2, Bernoulli sequences or coin-flips provide a concrete example which illustrates the principles of logical smoothing.

Compression

The data compression properties of Shannon codes are developed in chapter 3. It is shown that $\log 1/P^*(\mathbf{X}_n)$ provides an almost sure lower bound on the length of any description or code for $\mathbf{X}_n$ (Theorem 3.1). We note that the ideal lengths $\lceil \log 1/P^*(\mathbf{X}_n) \rceil$ could be attained exactly only if the true distribution P^* were known. Nevertheless, Shannon codes with respect to mixtures of distributions are shown to achieve the optimum length to first order, even when the true distribution is unknown. Such descriptions are called universal.

In practice, the only Shannon codes that can be implemented are those which are based on distributions that can be described. The theory of computable functions (reviewed in section 3.3) characterizes those distributions that have finite descriptions. The ideal complexity measure $L(P)$ is the length of the shortest computer program that describes the distribution P. The advantage of providing such programs as prefaces to the corresponding Shannon codes should be clear. Instead of coding with a fixed distribution P (agreed to in advance), we are free to code using a distribution $\hat{P}$ which depends on the data. The best such distribution is the one which minimizes the total description length.

Good data compression entails using a distribution estimate which provides an essential summary of the data. Indeed, the minimal program for $\hat{P}_n$ might naturally be called a ***sufficient statistic for description.*** Because, given the first stage, the remaining description may be regarded as conditionally maximally complex. In fact, with probability one, for all large n, the data $\mathbf{X}_n$ is in a set of roughly 2^{nH_n} typical sequences which are all nearly equally likely and which all have complexity near the entropy nH_n. (Here the discrete entropy nH_n is approximately $nh + nb_n$ where h is the differential entropy rate.) Consequently, the code which has length $\lceil \log 1/\hat{P}_n(\mathbf{X}_n) \rceil$ near nH_n amounts to giving the index of $\mathbf{X}_n$ in the set of typical sequences.

In section 3.4 we argue that minimum two-stage descriptions are universal. It is shown that if the true distribution is computable or if it can be approximated by computable distributions in the relative entropy rate sense, then *the minimum two-stage description is an asymptotically optimal compression of the data* (Theorems 3.4 and 3.5). In the iid case it is shown that any density which is less than a computable function with finite integral can be arbitrarily closely approximated by computable densities in the relative entropy sense.

Discovery

We examine the general problem of estimating the law of a stationary ergodic process in chapter 4. It is shown that if the true distribution has a finite description, then ***the logically smooth estimate is exactly correct,*** $\hat{P}_n = P^*$***, for all sufficiently large n, with probability one*** (Theorem 4.1). Thus if the data is governed by a computable law, then this law will eventually be discovered and thereafter never refuted.

For instance, if the true distribution is an exponential distribution with a computable mean (such as 3 or 2/7 or 1/ln2), then the estimate exactly determines both the shape of the distribution and the true value of the parameter.

Since the estimated distribution is exactly correct, the optimal regression, prediction, and classification functions may also be discovered. For example, if the stock market is an ergodic process with a computable distribution, then from the shortest description of stock market data, the optimum sequential investment portfolio may be found so as to maximize the growth rate of capital.

The result that the shortest complete description yields the true distribution is seen to be in agreement with the famous principle of parsimony espoused by the fourteenth century logician William of Ockham, "Explanations should not be multiplied beyond necessity" (Nunquam ponenda est pluralitas sine necessitate, *Commentarium in Sententias* I,27, see Tornay,1938,p.9). The simplest complete explanation is best.

The objective of empirical modeling is to find an accurate parametric family (model) of distributions from the observed data. Parametric families which are indexed by real valued parameters (coefficients) arise naturally from computable distributions by translations, scalings, mixtures, etc. Often the parametric family has a finite description (for instance, the set of translations of the standard normal distribution), even though particular members of the family are infinitely complex (for instance, a normal with a randomly selected mean). Two approachs to empirical modeling are suggested by logical smoothing. The first approach is to describe candidate parametric families and then to describe rational parameter values that achieve nearly maximum likelihood. For smooth families, about $(1/2)\log n^k|\hat{I}|$ bits should be used to describe the parameter estimates, where $\hat{I}$ is the observed Fisher information matrix and k is the dimension of the parameter vector. Roughly $(1/2)\log n$ bits per parameter achieves the right tradeoff between complexity and likelihood. A logically smooth estimated model is a parametric family achieving nearly minimum total description length. It is shown that for (Lebesgue) almost every parameter ***in a smooth family with finite description, the minimum description length based on the true family is within*** $2\log\log n$ ***of the minimum total description length over all families*** (Theorem 4.4). Thus a natural estimator of parametric families is the simplest family among all families achieving total description length within $2\log\log n$ of the minimum.

Another approach to empirical modeling suggested by logical smoothing is to enlarge the collection of candidate distributions to include exchangable or stationary distributions Q which are not necessarily iid or ergodic. The logically smooth estimator is a stationary distribution $\hat{Q}_n$ which achieves $\min_Q\{L(Q) + \lceil \log 1/Q(\mathbf{X}_n)\rceil\}$. Stationary distributions Q are mixtures of ergodic distributions P. There exists a unique prior probability measure v on the space of of ergodic distributions such that $Q(\mathbf{X}_n) = \int P(\mathbf{X}_n)\, dv(P)$ for any $\mathbf{X}_n$. The logically smooth estimate $\hat{Q}_n$ has a corresponding measure $\hat{v}_n$ on distributions. Ordinary parametric families are manifolds $\{P_\theta\}$ in the space of distributions which are traced by varying the parameter vector θ. Let the true distribution P_{θ^*} be in such a manifold, where the true parameter θ^* is selected at random according to a probability measure v^* equivalent to Lebesgue measure. Suppose that the corresponding mixture $Q^* = \int P_\theta\, dv^*(\theta)$ has a finite description. There are many computable parametric families which can model the behavior of samples from Q^*. Thus $\hat{Q}_n = Q^*$ cannot be claimed. Indeed, there may be a simpler family than Q^* that shares the same local behavior. It is shown that *the decomposition of the estimate $\hat{Q}_n$ has a component which concentrates on the true parametric family (manifold).* Indeed, a component of $\hat{v}_n$ is absolutely continuous and assigns strictly positive mass to neighborhoods of the true parameter θ^*, for all sufficiently large n, with probability one (Theorem 4.2).

Density estimation

The focus of the remaining section 4.3 is the non-parametric estimation of probability mass functions and density functions from independent and identically distributed samples. If the true density has a finite description, then (by Theorem 4.1) the logically smooth estimate is exactly correct for all large n. What about densities that are infinitely complex, but can be approximated by densities of finite complexity? We obtain convergence results that hold for any true density that is less than a computable function with finite integral. (For instance, bounded densities with compact support.) The results depend on the accuracy of the samples. Recall that each datum is observed to b_n bits accuracy to the right of the binary point. (For convenience, suppose the X_i are scalar observations; the vector case is handled similarly.) Thus the data is observed to within intervals of width $w_n = 2^{-b_n}$. If the widths w_n are fixed or tend slowly to zero such that $nw_n \to \infty$, then data accumulates in each interval and we essentially have the discrete mass estimation problem. In this case the maximum likelihood estimator (the histogram) converges in variation distance and so does logical smoothing.

On the other hand, if the cell widths w_n tend more rapidly to zero such that nw_n is bounded, then the number of observations in most cells remains bounded. Indeed, many of these small intervals remain empty. In this case the sequence of maximum likelihood estimators is inconsistent (in variation distance). Surprisingly, the logically smooth estimator is consistent in the sense that the sequence of variation distances, on partitions into intervals of width $h_n = 1/n$, converges to zero (Theorem 4.6).

These convergence results suggest a convenient definition of the density estimator $\hat{p}_n$ which is motivated by description length considerations. The data $X_1,X_2,\ldots,X_n$ are first described to log n bits accuracy by minimizing the two-stage description length. This yields estimates $\hat{P}_n(A)$ for the probabilities of each of the intervals A of width $1/n$. Then the remaining $n(b_n-\log n)$ bits are described in a brute force way, i.e. as though they were independent Bernoulli(1/2) random variables. (Indeed, asymptotically these extra bits are independent Bern(1/2) and log n is just the right depth to assure that it does not hurt to describe them in this way.) Equivalently, the data are described using a distribution $\hat{P}_n$ with density $\hat{p}_n$ defined to be conditionally uniform within each cell of width $1/n$ (while retaining the cell probabilities $\hat{P}_n(A)$)). It is shown that *this density estimate $\hat{p}_n$ is consistent in L_1 distance: that is, $\int|\hat{p}_n(x)-p^*(x)|dx \to 0$ with probability one* (Theorem 4.6). This result contrasts sharply with the convergence results for ordinary histograms (Abou-Jaoude,1976) and kernel density estimators (Devroye,1983) where to obtain consistency, the numerical smoothing must extend over a width h_n satisfying $nh_n \to \infty$.

Another simple modification of logical smoothing yields strongly consistent density estimates. The estimate $\hat{p}_n$ is defined to be a density achieving min $\{\ c\ L(P) + \log 1/p(X_1,\ldots,X_n)\ \}$ where the minimum is over all computable iid distributions P with densities p. The factor c is any constant greater than one. This modification yields estimates $\hat{P}_n$ which tend to have less complexity $L(\hat{P}_n)$ than estimates obtained with $c = 1$. The two stage description length based on $\hat{P}_n$ is nearly minimal. It is shown that this density estimate $\hat{p}_n$ is consistent in L_1

distance: $\int \mid p^* - \hat{p}_n \mid \rightarrow 0$ with probability one (Theorem 4.7).

Bayes rule

There are interesting connections between logical smoothing and Bayesian inference. Minimizing the total description length $L(P) + \log 1/P(\mathbf{X}_n)$ is equivalent to maximizing the posterior probability of P given $\mathbf{X}_n$ where this posterior is determined by Bayes rule from the prior probability $2^{-L(P)}$. For the two-stage description to be uniquely decodable, the length function $L(P)$ must satisfy Kraft's inequality $\sum_P 2^{-L(P)} \leq 1$, or equivalently, the prior must be proper. The description length principle provides the rational for having a prior as well as an objective criterion for the choice of prior probabilities. Some subjectivity enters in the choice of computer, but this subjectivity is limited. Indeed, if $L(P)$ is the length function of a ***universal*** computer, then for any other computer with length function $L'(P)$, say, there exists a constant c such that $L'(P) \geq L(P) - c$ for all P (see section 3.3). Similarly, if a statistician has prior opinion specified by some finitely describable mass distribution $v(P)$ then there exists a constant c such that the prior satisfies $v(P) \leq c2^{-L(P)}$ for all P.

We use the proof techniques developed for logical smoothing to establish new consistency results for arbitrary Bayes procedures. Consider the iid case and suppose the prior probability assigns strictly positive mass to relative entropy neighborhoods of the true distribution. The sequence of posterior distributions given real-valued samples $(X_1, X_2, \ldots, X_n)$ is shown to concentrate on the n^{-1}-smoothed variation distance neighborhoods of the true distribution. Counterexamples are constructed showing that concentration within tighter neighborhoods can fail. However, if the prior $v(P)$ is countable and $\sum_P v^{\alpha}(P)$ is finite for some $0 < \alpha < 1$, then the estimate $\hat{P}_n$ which maximizes the posterior probability is consistent in variation distance.

Two aspects of overfit

We often say that an estimated distribution is overfitted if it assigns high likelihood to the data, but does not generalize to new samples similarly drawn. Cross-validation and the criteria of Mallows and Akaike have been proposed (in limited parametric contexts) to try to avoid this overfit (see Barron 1984). The use of consistent estimators is one way to help prevent overfit. Indeed, strong forms of consistency (such as convergence of the relative entropy between true and estimated distributions) imply that with high probability, the likelihood will not drop significantly on new samples. (Unfortunately, models selected by Mallows' and Akaike's criteria are generally inconsistent.)

However, overfit is not synonymous with inconsistency. Overfit also entails using a distribution which is significantly more complex than is required to accurately represent the data. In this sense, kernel density estimates are overfit (though consistent), whereas estimates based on minimum description length properly avoid overfit (and retain the consistency). Although it is intuitively clear that overfit involves excess complexity, the preponderance of previous work has not quantified this aspect.

Three kinds of statisticians

Some rough categories are suggested for statisticians and engineers. Firstly, there are those who limit themselves to the simplest and most thoroughly understood models (for instance, linear parametric models involving the normal distribution). Such models have low complexity (underfit), yet also have relatively low likelihood (unless the true distribution is of the assumed form). Many researchers are in this first category because they lack the creativity to invent new laws, or the persistence to search for models that fit the data, or the willingness to let a computer aid in the search.

Secondly, there are those who recognize the need for accurate nonparametric fits (consistency), but who routinely employ techniques (such as nearest neighbor rules or kernel estimators or Dirichlet process priors) which provide little understanding of the data. Typically, the estimates have limited practical usefulness because all of the data is retained -- rather than summarized. The estimates are overfit.

The third category of scientists are those who employ sufficient insight and computational resources to consider a rich variety of conceivable distributions and find the law which best explains the data. Both the complexity of the law and its fit to the data are accounted for in the search for the shortest complete description of the data. Occasionally, time constraints force the use of the routine techniques, but usually there is ample time to consider a multitude of possibilities and select the best. Diligence is rewarded with discovery.

1.2 History

We begin with a brief discussion of non-parametric density estimation as it relates to logical smoothing. For surveys on the art of density estimation see Cover (1972), Tapia and Thompson (1978), Prakasa Rao (1983), and Devroye and Gyorfi (1985).

Rosenblatt (1956) and Parzen (1962) initiated interest in non-parametric density estimation with their kernel density estimator. This estimate is obtained by convolving the sample distribution with a kernel function. A series of papers culminating with Devroye (1983) show that the kernel density estimator is strongly consistent in L_1 distance for *any* true density on the line if and only if the kernel functions have width parameters h_n satisfying $h_n \to 0$ and $nh_n \to \infty$. However, from the point of view of data compression the kernel density estimator is very poor. Indeed, for most kernels the data locations may be exactly recovered (by deconvolution). The kernel estimate does not summarize the data.

Good and Gaskins (1971) introduced maximum penalized likelihood estimators. Densities are penalized for their "roughness" rather than there descriptive complexity. De Montricher, Tapia, and Thompson (1975) show that a maximum penalized likelihood estimator reduces to a spline with knots at the data points. Again, such an estimate is as complex as the data.

There are some popular density estimators which may yield a simple density estimate. For instance, consider the ordinary histogram. Abou-Jaoude (1976) has shown that the histogram is consistent in L_1 distance for any density if and only if the cell widths h_n satisfy $h_n \rightarrow 0$ and $nh_n \rightarrow \infty$. For conservative choices of cell width, the histogram estimates have reasonably short descriptions; however, the structure thus imposed can yield a significant loss in likelihood. Similar statements hold for the density estimates based on an orthogonal series, where the expansion is taken to an appropriate number of terms (see S. Schwartz,1967). The total description length of the data would be nearly minimal only if the series were tailored to the true density. An interesting open question is whether the number of terms (or the cell width of the histogram) which achieves the fastest rates of convergence is also the number which achieves the right tradeoff between complexity and likelihood.

Cover

Cover (1972) introduced a scheme for estimating densities that is similar in spirit to logical smoothing. Briefly, the likelihood $P(\mathbf{X}_n)$ is maximized subject to the complexity constraint $L(P) \leq c_n$ where the constants c_n are chosen such that $c_n \rightarrow \infty$ and $c_n/n \rightarrow 0$. This estimator may be regarded as a special case of the method of sieves (Grenander 1981, Geman and Hwang 1982). The L_1 convergence of Cover's density estimator is readily established using Chernoff tilting arguments(see Theorem 4.7). The complexity of the density estimate is typically near the arbitrary bound c_n. In contrast, logical smoothing automatically determines the right complexity from the data.

A starting point for several ideas used in this thesis is a paper by Cover (1973b). Cover presented a sequence of hypothesis tests for whether a real valued parameter θ (such as a population mean) is in an arbitrarily specified countable set $\{\theta_j\}$ (such as the set of computable real numbers). If the true parameter θ^* is computable then Cover's procedure will estimate θ^* exactly for all large n, with probability one. Briefly, Cover's scheme is to choose the least complex θ_j that lies in a small interval containing the sample mean.

Sufficient statistics for description

A notion of sufficiency of descriptions was introduced by Kolmogorov at the 1974 International Symposium on Information Theory (see Cover 1985). Discrete X are described in two stages, using finite sets S which contain X. The total description length is $L(S) + \log |S|$ where $L(S)$ is the length of the shortest program for a set S and $\log |S|$ is the number of bits required to give the index of X in the set of cardinality $|S|$. (This second stage is a Shannon code with respect to the uniform distribution on S.) A Kolmogorov *minimal sufficient statistic for description* is the program for a set $\hat{S}$ which achieves $\min \{ L(S) : L(S) + \log |S| = L(X) \}$ where $L(X)$ is the length of the shortest program for X. Cover (1973c, 1985) argued that for Bernoulli (θ) sequences $X = (X_1,...,X_n)$ a sufficient statistic for X is a description of $k = \sum_{i=1}^{n} X_i$. The associated program is "S is the set of all $X \in \{0,1\}^n$ such that $\sum X_i = k$." The two stage description length is $\log (k+1) + n\, h(k/n)$ (where h is the binary entropy function) and this length is shown to be nearly minimal. (A candidate for the *minimal* sufficient statistic is a description of the set of sequences with $(1/n)\sum X_i$ in a simple neighborhood of k/n with width given by the standard deviation, see section 2.)

Universal descriptions

Davisson (1973) is responsible for much of the theory of universal source coding. Universal codes are descriptions of data from an unknown distribution such that the codelengths are asymptotically optimal for a large class of possible distributions. Davisson employed Shannon's definition that a sequence of codes is optimal if the average description length (divided by the the sample size) tends to the entropy rate. The excess average description length is called the redundancy. Davisson suggests placing a prior v on the class of distributions and then gives two methods for constructing universal codes. One method is to Shannon code with respect to the mixture of distributions. Davisson recommends the mixture based on the "least favorable" prior achieving the maximin average redundancy (which is also minimax for compact classes, see Davisson and Leon-Garcia,1980).

Davisson (1973,p.790-791) introduced two-stage descriptions as his second method for universal coding. The first stage describes an estimated distribution using the Shannon code with respect to $v^{(n)}$ (a discrete prior constructed from v); the second stage describes the data using the Shannon code with respect to the estimated distribution. Then Davisson (1973, Theorem 7) shows how to obtain universal codes for the class of finite alphabet stationary-ergodic sources using a two-stage code based on conditional histograms. Other literature on universal coding includes Cover (1973a), Elias (1974), Trofimov (1974), Kieffer (1978), Longo and Sgarro (1979), Davisson, McEliece, Pursely and Wallace (1981), Krichevsky and Trofimov (1981), and Davisson (1983).

Rissanen

The minimum description approach to statistical inference was introduced in the context of smooth parametric families by Rissanen (1978,1983). Rissanen assumes that a nested sequence of parametric families is known to accurately model the data (only the parameter dimension and values are free to be estimated) and he determines the asymptotic behavior of the minimum description length. He established that the principal terms of the total description length are $(k/2)$ log n + log $1/P_{\theta_{ML}}(\mathbf{X}_n)$ where k is the dimension of the parameter vector and θ_{ML} is the maximum likelihood estimate. Rissanen (1984a,b) has shown that the minimum average code-length for any sequence of uniquely decodable codes is given by $(k/2)$ log n + E log $1/P_{\theta^*}(\mathbf{X}_n)$ + o(log n), for almost every θ^* in smooth parametric families. (We obtain a strengthening of this result to pointwise rather than average description length, see Theorem 4.3. in Section 4.2) Rissanen suggested the minimum description length as a criterion for model selection. A similar criterion was derived from the Bayesian point of view by G. Schwarz (1978) for Koopman-Darmois or exponential families.

Regression

A natural use of the minimum description length criterion is to select terms in regression models. We minimize the conditional description length of the sequence of dependent variables given the input variables. The dependent variable values (each observed to b bits accuracy) may be described as follows. First a regression function is described which involves k estimated coefficients each determined to (1/2) log n bits accuracy. Also, the maximum likelihood estimate of the residual error variance, which we denote by σ^2_{ML}, is described to (1/2) log n bits accuracy. Then the residual errors are described using a normal distribution with zero mean, variance σ^2_{ML}, and zero correlations. The total description length is roughly $(k+1)/2$ log n + $(n/2)$ log$2\pi e\sigma^2_{ML}$ + nb. (A more refined analysis involves the observed Fisher information in the description length of the parameters.) Hannah (1980) considered (linear) autoregressive and moving average models and established that the minimum description length criterion consistently estimates the parameter dimension, assuming that a bound on the true dimension is known. (Our Theorem 4.2, when specialized to this context, establishes consistency of the estimated dimension, for almost every parameter vector, without requiring bounds on the dimension.)

The minimum description length methodology is eminently suited to situations where there is a multitude of statistical models which feasibly can be considered and where there is considerable variation in the complexity of these models. Non-linear regression and, in particular, polynomial regression involving many variables is a good example. Polynomial learning networks provide a framework for the computer-aided generation and examination of a rich supply (typically hundreds of thousands) of regression models in reasonable time (typically a few seconds on a mainframe computer). For literature on this approach to polynomial regression see Ivakhnenko (1971), R.L. Barron et al.(1975,1984), A.R. Barron (1977,1982,1984), and Farlow (1984).

Sorkin

In the non-parametric context, logical smoothing was independently introduced by Sorkin (1983) and by Barron and Cover (1983, conference presentation). Sorkin suggested minimizing $L(P) + \log 1/P(\mathbf{X}_n)$; however, he did not interpret this as a two-stage description length. Sorkin envisioned that physical, sociological, and economic theories may be discovered using this principle. Indeed, our Theorem 2 shows that discovery of the right theory is the sure consequence of logically smooth inference.

Ockham

The principle of inference of physical laws from simple explanations, has roots in the philosophy of William of Ockham. This fourteenth century logician and natural philosopher stated, "What can be explained with fewer things is vainly explained with more" (*Frustra fit per plura quod potest fieri per pauciora*). In the *Summa Totius Logicae* I,12, Ockham applied this famous razor to defend his theory of knowledge. He maintained that a proposition abstracted by the mind as a summary of observations is not an entity "distinct from the act of understanding, [but rather] it is the act of understanding itself. It is needless to have recourse to many entities when we can get along with fewer ones" (see Tornay,1938,p.9,100-101, Loux,1974,p.74, Boehner,1957,p.xxi).

In Ockham's study of the theory of motion, he employed his "razor" to cut out needless complications in the explanations developed by his predecessors. Concerning the motion of projectiles he declared, "After the separation from the original projector has occurred, the moving body itself is the motion on its own account and not by some absolute force bearing upon it. To sum up: the thing moving and the movement are thoroughly indistinguishable." (*Commentarium in Sententias* II,26, see Tornay,1938,p.171). Thus Ockham anticipates Descartes' law of inertia. Duhem (1909,v.II) traces the influence of Ockham and early Ockhamists upon Copernicus, Galileo, Leonardo de Vinci, and Newton (see also Tornay,1938,p.51-53,165, Moody,1935,p.307-309). For a less grandiose view of Ockham's influence on the history of science, see Goddu (1984).

We restate Ockham's principle in this way: the shortest complete description is the best understanding. For many years, scientists and engineers have been inferring densities or laws from data according to this principle. Data are examined and a simple law is found that fits the data with high likelihood. Consistent inference is the result, provided that the set of distributions considered is sufficiently rich and provided that the statistician properly balances the objectives of simplicity and likelihood.

Chapter 2. Example

Coin Tossing

This section illustrates the principle that short descriptions lead naturally to the inference of probabilities. For a concrete example, we consider sequences of flips of a coin with unknown bias. Coin-flips are regarded as realizations of Bernoulli sequences or binary memoryless sources. Additional discussion of coin tossing and other finite alphabet sources from the description length perspective is found in Rissanen (1984a,b) and in the literature on universal source coding (Cover,1972,1973, Davisson,1973, Davisson, McEliece, Pursley and Wallace,1981, Krichevsky and Trofimov,1981). The analysis here differs from the literature in that pointwise rather than average properties are emphasized.

Let a sequence of zeros and ones $\mathbf{X}_n = (X_1,X_2,...,X_n)$ be the outcomes of n tosses of a biased coin. The sample size n need not large, since the analysis here is valid for all n. The Bernoulli model assumes that the X_i are independent with parameter $\theta = \text{Prob}\{X_i{=}1\}$. The corresponding likelihood is $P_\theta(\mathbf{X}_n) = \theta^S(1{-}\theta)^{n-S}$ where $S = \sum X_i$ is the number of ones. The maximum likelihood estimate $\theta_{ML} = S/n$ is the relative frequency of ones. In terms of information-theoretic quantities, the likelihood may be expressed exactly as

$$P_\theta(\mathbf{X}_n) = 2^{-n(h(\theta_{ML})+D(\theta_{ML}||\theta))}. \tag{2.1}$$

Here h is the binary entropy function, $h(\theta) = \theta \log 1/\theta + (1{-}\theta) \log 1/(1{-}\theta)$, and D is the relative entropy function, $D(\theta_0||\theta) = \theta_0 \log\theta_0/\theta + (1{-}\theta_0) \log(1{-}\theta_0)/(1{-}\theta)$. (An identity similar to (2.1) holds for any finite or countable sample space, not just zeros and ones.) By concavity of the logarithm, the relative entropy $D(\theta_{ML}||\theta)$ is strictly positive unless $\theta = \theta_{ML}$. Consequently, the maximum likelihood satisfies $\log 1/P_{\theta_{ML}}(\mathbf{X}_n) = nh(\theta_{ML})$.

The likelihood is relevant for describing the sequence $\mathbf{X}_n$ because of the Shannon code which has length $\lceil \log 1/P_\theta(\mathbf{X}_n) \rceil$. From equation (2.1), this length is within one bit of $nh(\theta_{ML}) + nD(\theta_{ML}||\theta)$. The data could be described using a Shannon code based on a fixed θ (agreed upon in advance by both the encoder and decoder). Then the codelength exceeds the entropy by the amount $nD(\theta_{ML}||\theta)$. Consequently, this description is highly redundant, unless the guess θ happens to be close to θ_{ML} in the relative entropy sense. (Asymptotically, the redundancy would be small only if we knew or correctly guessed the true θ^*.) Now since $D \geq 0$, with equality only if $\theta = \theta_{ML}$, it appears at first glance that the best Shannon code would use $\theta = \theta_{ML}$ and hence obtain codelength which is within one bit of $nh(\theta_{ML})$. However, for this code to be valid, the parameter θ_{ML} must first be described (otherwise, $\mathbf{X}_n$ could not be uniquely determined from its code). Now about $\log n$ bits suffice to describe S and hence θ_{ML}. So we have a description of the data with total length near $\log n + nh(\theta_{ML})$.

We can do better. Instead of the maximum likelihood estimate θ_{ML}, we use the simplest number $\hat{\theta}$ within a neighborhood of θ_{ML}. As before, this number $\hat{\theta}$ is provided as a prefix to the Shannon code based on $\hat{\theta}$, so that the overall description of the data is uniquely decodable. Since the relative entropy is the amount by which the Shannon codelength exceeds $nh(\theta_{ML})$, we see that the best $\hat{\theta}$ is that number which minimizes the sum of its description length $L(\hat{\theta})$ plus the relative entropy $nD(\theta_{ML}||\hat{\theta})$. If θ_{ML} is sufficiently close to a number θ^* which is easily described (such as 1/2, 4/7, 1/e, or $\pi/4$), then the estimate becomes $\hat{\theta} = \theta^*$ and the minimum total description length is $L(\theta^*) + \lceil \log 1/P_{\theta^*}(\mathbf{X}_n) \rceil$. However, most possible values for θ_{ML} are not close to numbers of low complexity. (Given any n numbers, a proportion of at most 2^{-c} can have descriptions of length less than $\log n - c$. Similarly, given any M distinct intervals, a proportion of at most 2^{-c} can have members with complexity less than $\log M - c$.) For most θ_{ML}, a good $\hat{\theta}$ is found to be a dyadic rational with distance at most $1/\sqrt{nI(\theta_{ML})}$ from θ_{ML}. Here $I(\theta) = 1/(\theta(1-\theta))$ is the Fisher information. Note that $1/\sqrt{nI(\theta_{ML})}$ is the usual estimator of the standard deviation of θ_{ML}. The details showing that the standard deviation provides the right tradeoff between complexity and likelihood will be carried out below.

First we discuss several methods for succinctly describing the sequence $\mathbf{X}_n$. In each case the total description length is nearly

$$\frac{1}{2}\log\frac{nI}{2\pi} + nh(\theta_{ML}). \tag{2.2}$$

Depending on the particular coding scheme, the number I is either the Fisher information at the maximum likelihood, $I = 1/(\theta_{ML}(1-\theta_{ML}))$, or $I = \pi^2$ (in the latter case, $\sqrt{I}$ may be interpreted as the average square root of Fisher informations, $\sqrt{I} = \pi = \int_0^1 1/\sqrt{\theta(1-\theta)}\ d\theta$).

The counting code: The number of ones is described using $\lceil\log(n+1)\rceil$ bits, then $\lceil\log\binom{n}{S}\rceil$ bits are used to give the index of the sequence $\mathbf{X}_n$ within the set of n–sequences having S ones. (Here $\binom{n}{S} = n!/(S!(n-S)!)$ is the number of such sequences.) Thus the total description length is

$$\lceil \log (n+1) \rceil + \lceil \log \binom{n}{S} \rceil. \tag{2.3}$$

From Stirling's formula for the factorial we find that

$$\log \binom{n}{S} \stackrel{\sim}{=} nh(\theta_{ML}) + (1/2) \log (I(\theta_{ML})/2\pi n). \tag{2.4}$$

Indeed $\log\binom{n}{S}$ is less than the right hand side, but not by more than $(\log e)/9$ provided S and $n-S$ are at least one. (This tight bound is credited to Shannon, see Wozencraft and Reiffen,1961,p.72, and follows from the bounds on Stirling's formula due to Robbins,1955.) Thus the total description length in (2.3) is within a small constant (less than three bits) of

$$\frac{1}{2}\log\frac{nI(\theta_{ML})}{2\pi} + nh(\theta_{ML})$$

which is the same as expression (2.2). This counting code may be interpreted as a two-stage code where first $\theta_{ML} = S/n$ is described, and then the data is described using the Shannon code based on the uniform distribution on the set of n–sequences having $n\theta_{ML}$ ones.

The uniform mixture code: Let $Q = \int_0^1 P_\theta d\theta$ be the uniform mixture of Bernoulli distributions. Then $Q(\mathbf{X}_n) = \int_0^1 \theta^S(1-\theta)^{n-S}d\theta = 1/((n+1)\binom{n}{S})$. The same distribution is obtained by uniform selection of the number of ones S within $\{0,1,\ldots,n\}$ and then uniform selection of $\mathbf{X}_n$ within the set of n–sequences having S ones. The Shannon code based on Q has codelength

$$\lceil \log 1/Q(\mathbf{X}_n) \rceil = \lceil \log (n+1) + \log \binom{n}{S} \rceil. \tag{2.5}$$

Note that this description length is within one bit of the counting codelength and (by Stirling's formula) within two bits of expression (2.2).

The Beta(1/2,1/2) mixture code: Let $Q = \int_0^1 P_\theta dv(\theta)$ where v is the Beta(1/2,1/2) distribution with density function $1/(\pi\sqrt{\theta(1-\theta)})$ which is proportional to the square root of the Fisher information. Then $Q(\mathbf{X}_n) = \Gamma(S+1/2)\Gamma(n-S+1/2)/(\pi\Gamma(n+1))$ and from Stirling's formula the Shannon codelength $\lceil \log 1/Q(\mathbf{X}_n) \rceil$ is within a small constant of

$$\frac{1}{2}\log\frac{n\pi}{2} + nh(\theta_{ML}). \tag{2.6}$$

This description length is the same as expression (2.2) with the average interpretation of $\sqrt{I}$. Note that the first term of (2.6) does not depend on θ_{ML}. (Using this fact Krichevsky and Trofimov (1981) showed that (1/2) $\log n$ is the minimax asymptotic redundancy. The Beta(1/2,1/2) distribution is suggested as the least favorable prior.) Comparing π to $\sqrt{I(\theta_{ML})}$, we find that expression (2.6) is a shorter description length when θ_{ML} is near zero or one ($\theta_{ML}(1-\theta_{ML}) < 1/\pi^2$, i.e. $\theta_{ML} < .114$ or $\theta_{ML} > .886$). On the other hand, for intermediate values of θ_{ML}, the uniform mixture code has shorter length.

Two-stage codes for coin-flips: Unlike the mixture codes, the two-stage descriptions explicitly involves inference of the parameter. Here we discuss estimates $\hat{\theta}$ which achieve nearly minimum two-stage description length $L(\theta) + \lceil \log 1/P_\theta(\mathbf{X}_n) \rceil$. From the identity for log-likelihood, $\log 1/P_\theta(\mathbf{X}_n) = nh(\theta_{ML}) + nD(\theta_{ML}||\theta)$, we see that minimizing the two-stage description length is equivalent to minimizing the redundancy $L(\theta) + nD(\theta_{ML}||\theta)$. We let $\hat{\theta}$ be the dyadic rational of lowest denominator in an interval of width δ which contains the maximum likelihood estimator θ_{ML}. Then about $\log^* 1/\delta$ bits are sufficient to describe $\hat{\theta}$. (Here we may set $\log^* x = \lceil \log x + 2 \log \log x \rceil$; a more refined definition is given in section 4.2. In either case, the function $\log^* x$ is within $2 \log \log x$ of the ordinary logarithm function $\log x$). The interval containing θ_{ML} is conveniently chosen to be $A = [\theta_{ML}, \theta_{ML}+\delta]$ if $\theta_{ML} \leq 1/2$, otherwise $A = [\theta_{ML}-\delta, \theta_{ML}]$. By Taylor expansion of the log-likelihood about its maximum, we find that for $\hat{\theta}$ in A, the two stage description length approximately equals $\log^* 1/\delta + \log 1/P_{\theta_{ML}}(\mathbf{X}_n) + (1/2)\delta^2 nI(\theta_{ML}) \log e$. Precise bounds follow from the general fact that the relative entropy is less than the Chi-square distance: in this case, we have $D(\theta_{ML}||\hat{\theta}) \leq (\theta_{ML}-\hat{\theta})^2 I(\hat{\theta}) \log e$. Simple calculus then verifies that (to first order) the minimizing δ is $1/\sqrt{nI(\theta_{ML})}$. Note that for this δ the log-likelihood $\log P_{\hat{\theta}}(\mathbf{X}_n)$ is within a small constant of the maximum. Indeed, the difference from the maximum is $nD(\theta_{ML}||\hat{\theta})$ which is less than $\log e$. Consequently, the two-stage description length based on $\hat{\theta}$ is within a small constant of

$$\log^* \sqrt{nI(\theta_{ML})} + nh(\theta_{ML}). \tag{2.7}$$

This description length is within 2 log log n of expression (2.2).

Another reasonable bound on the minimum two-stage description length is suggested by the Beta(1/2,1/2) density $1/(\pi\sqrt{\theta(1-\theta)})$. Consider the quantiles of the Beta(1/2,1/2) distribution which partition [0,1] into intervals A_i, $i=1,2,\ldots,\sqrt{n}$ each of probability $1/\sqrt{n}$. Let θ_i be an endpoint of the interval A_i, specifically the endpoint which is closer to 1/2. By the mean value theorem, each A_i has width equal to $\pi\sqrt{\theta(1-\theta)/n} = \pi/\sqrt{nI(\theta)}$ for some θ in A_i. Hence the width of A_i is less than $\pi/\sqrt{nI(\theta_i)}$. Let the estimate $\hat{\theta}$ be the point θ_i corresponding to the interval which contains the maximum likelihood θ_{ML}. This estimate can be described using $\log^*\sqrt{n}$ bits. (If we conditioned on n we could reduce this to $\lceil \log\sqrt{n} \rceil$. However, in order that two-stage descriptions correspond to a fixed prior $2^{-L(P)}$, we use the unconditional description length $L(P)$, rather than the conditional description length $L(P|n)$ -- see section 3.3.) From the Chi-square bound on the relative entropy, we find that $nD(\theta_{ML}||\hat{\theta}) \leq \pi^2 \log e$. Consequently, the two-stage description length corresponding to $\hat{\theta}$ is within a constant of

$$\log^* \sqrt{n} + nh(\theta_{ML}). \tag{2.8}$$

This description length is within 2 log log n of the description length (2.6) for the code based on the Beta(1/2,1/2) mixture.

Note that the two-stage codes tend to have slightly longer description lengths than the mixture codes, but only by a factor bounded by 2 log log n. This behavior is established for general parametric families in section 4.2.

While encoding the data, we may wish to determine which of the five coding schemes mentioned above works best for this data. Then it is necessary to preface the codes with a description of which scheme is being used -- another three bits would suffice. Indeed, if we are armed with a universal computer then coding based upon any computable stationary distribution (or family of distributions) may be considered; we need only preface the code with a description of the distribution.

In the universal coding literature, codes similar to these for coin-tossing have been constructed for any finite alphabet source. We leave the detailed example to develop a general theory.

Chapter 3. Description

3.1 Pointwise Optimality of the Shannon Codelength

In this section we review some basic facts from the theory of source coding for countable alphabets which we need to analyze logical smoothing. Then we present pointwise optimality properties of the Shannon codelength $\log 1/P^*(X)$ which strengthens Shannon's result that the minimum average codelength is the entropy. For any uniquely decodable descriptions of sequences of random data $\mathbf{X}_n$, it is shown that asymptotically $\log 1/P^*(\mathbf{X}_n)$ provides an almost sure lower bound on description length. We require no assumptions whatsoever on the true distribution P^*.

A code is defined to be a 1-1 mapping from a countable alphabet into finite length binary sequences (codewords, descriptions). A ***prefix condition*** code is defined as a code for which no codeword is the prefix of a longer codeword. A code is said to be ***uniquely decodable*** if for any concatenation of codewords, there is no ambiguity as to when one ends and the next begins. Note that prefix condition codes are uniquely decodable. From Leung-Yan-Cheong and Cover (1978, Theorem 2) the addition of roughly $\log l(x)$ bits (to describe the codelength) is enough to make a prefix condition code from any 1-1 code with lengths $l(x)$.

The key result for variable length source coding is the necessary and sufficient condition on the length function $l(x)$ for there to exist a uniquely decodable code. (This condition was obtained first by Kraft,1949, for prefix codes and then by McMillan,1956, for unique decodability.) The Kraft-McMillan theorem states that ***there exists a uniquely decodable code with codeword lengths $l(x)$ if and only if the lengths satisfy $\sum_x 2^{-l(x)} \leq 1$.*** (A simple proof is due to Karush,1961; see Gallager,1968,p.45-49 and p. 514 for the countable case.)

The code construction due to Shannon is this. Suppose we have a length function $l(x)$ which satisfies Kraft's inequality. Let $l_1 \leq l_2 \leq \ldots$ be the ordered lengths and let $j(x)$ be a labeling of of the alphabet such that $l(x) = l_j$. The codeword for x is the binary representation of $\sum_{i=1}^{j-1} 2^{-l_i}$ where the binary expansion is carried out to $l(x)$ bits. Note that longer codewords must increase the sum by at least $2^{-l(x)}$ and hence differ in the first $l(x)$ places from the codeword for x. Thus Shannon's code satisfies the prefix condition and hence it is uniquely decodable.

If $P(x)$ is a probability mass function on the countable alphabet, then define the description length of x given the distribution P to be

$$\lceil \log 1/P(x) \rceil$$

These lengths are naturally defined so as to satisfy Kraft's inequality. The code constructed as above with lengths $\lceil \log 1/P(x) \rceil$ is called the Shannon code for x given P (Shannon 1949, p.29). With slight modification, Shannon codes may be computed from recursive enumerations of the probabilities to increasing accuracy. (The modified code may have length $\lceil \log 1/P(x) \rceil + 1$ for some x, see section 3.3.)

A similar code proposed by Gilbert and Moore (1959) permits the x's to remain in some natural ordering (rather than be reordered in terms of increasing codelengths). The Gilbert-Moore codes have fast algorithms for encoding and decoding and have codelengths given by $\lceil \log 1/P(x) \rceil + 1$. The results that we obtain below for the Shannon codes also hold for these modified codes. The modifications add at most one bit.

The average description length $\sum_x P^*(x) l(x)$ for any uniquely decodable code must exceed the entropy $H(P^*) = \sum_x P^*(x) \log 1/P^*(x)$ (see Gallager 1968,p.50 and p.514). The Shannon code has average description length which is within one bit of the entropy bound, provided $P = P^*$. Otherwise, the Shannon code has additional average description length (redundancy) which is within one bit of the relative entropy $D(P^*||P) = \sum_x P^*(x) \log P^*(x)/P(x)$.

Similar statements hold for sequences of discrete random variables with joint distribution P^*. For instance, if P^* is iid and the variables are observed to fixed accuracy, then the n–sample entropy is n times the single sample entropy. In this case the Shannon code has average length between nH and $nH+1$ (so the one bit of excess length is insignificant). More generally, suppose P^* is stationary and ergodic. By the Shannon-McMillan-Breiman theorem, $\log 1/P^*(\mathbf{X}_n)$ grows at a linear rate: $\log 1/P^*(\mathbf{X}_n) = n(H + o(1))$ almost surely where H is the entropy rate. (Thus the likelihood is nearly constant, $P^*(\mathbf{X}_n) = 2^{-n(H+o(1))}$. In particular, the distribution is nearly uniform over a 1–ϵ support set of typical sequences for which $|\log 1/P^*(\mathbf{X}_n) - nH| < \epsilon$.) If we Shannon code with respect to another distribution P which is iid or Markov then the excess description length is within one bit of the log-likelihood-ratio $\log P^*(\mathbf{X}_n)/P(\mathbf{X}_n)$ which equals $n(D + o(1))$ almost surely, where D is the relative entropy rate of P^* with respect to P. (This result is a generalization of the Shannon-McMillan-Breiman theorem which was recently proved in Barron 1985.)

In what follows we shall treat $\log 1/P^*(\mathbf{X}_n)$ as the ideal description length or measure of information. This we do to determine the pointwise behavior of description lengths rather than just the average. Such results will be useful in latter sections to obtain almost sure consistency results.

Information theorists have not been in agreement as to whether $\log 1/P^*(\mathbf{X}_n)$ rather than the entropy may be treated as the measure of information. Indeed, Kolmogorov (1965) has said, "only the average quantity is a true measure of the information content." Against this view, Lemma 1 is offered below. No description length function, not even the shortest program for $\mathbf{X}_n$ on a universal computer (as suggested by Kolmogorov), can have lengths less than $\log 1/P^*(\mathbf{X}_n)$ by more than a logarithmic factor, except for sequences in a set of probability zero. Moreover, this bound holds with no assumptions whatsoever on the true probability distribution P^*.

It must be understood that the pointwise measure of information $\log 1/P^*(\mathbf{X}_n)$ is with respect to the true distribution. Depending on the sequence $\mathbf{X}_n$, there are many other distributions P (for instance, a point mass at the data $\mathbf{X}_n$) which have Shannon codelength much less than $\log 1/P^*(\mathbf{X}_n)$. However such a "code" is not a valid description of $\mathbf{X}_n$ -- unless it is prefaced with a description of P. The description length principle provides necessary limitations to the principle of maximum likelihood.

Let $\mathbf{X}_n$ be any sequence of random variables taking values in a sequence of countable partitions π_n of an underlying measurable space Ω with probability distribution P^*. Suppose the partitions π_n are refinements: that is, any event $\mathbf{x}_n$ in π_n is a union of events in $\pi_{n'}$ for $n' > n$. Let $\pi_* = \bigcup_n \pi_n$ be the union of the partitions, thus an event $\mathbf{x}$ is in π_* if it is in π_n for some n. (A space and a sequence of partitions that satisfy all the properties we will need is as follows: Ω is the space of infinite sequences $(x_1,x_2,...)$ of real numbers and the partition π_n consists of the cylinder sets corresponding to $(x_1,x_2,...,x_n)$ with each x_i truncated to b_n bits accuracy.)

We argue that $\log 1/P^*(\mathbf{X}_n)$ provides a lower bound on description length in both average and pointwise senses.

Theorem 3.1: Lower bound on complexity.

Let $l(\mathbf{x}_n)$, $\mathbf{x}_n \in \pi_n$, be the length function for any uniquely decodable code on the partition π_n. The description lengths $l(\mathbf{X}_n)$ exceed $\log 1/P^*(\mathbf{X}_n)$, on the average,

$$E\,[\ l(\mathbf{X}_n) - \log 1/P^*(\mathbf{X}_n)\ ] \geq 0, \tag{3.1}$$

and pointwise,

$$l(\mathbf{X}_n) - \log 1/P^*(\mathbf{X}_n) \geq -c_n \tag{3.2}$$

for all n sufficiently large, with probability one. Here c_n is any sequence of constants for which $\sum_n 2^{-c_n} < \infty$ (for instance, $c_n = \log n + 2 \log \log n$).

If the code is uniquely decodable on π_* (the union of the partitions), *then the sequence of pointwise redundancies $l(\mathbf{X}_n) - \log 1/P^*(\mathbf{X}_n)$ is uniformly greater than a random variable $R > -\infty$ (which does not depend on n). Furthermore,*

$$E \inf_n\ (\ l(\mathbf{X}_n) - \log 1/P^*(\mathbf{X}_n)\ ) \geq -\log e. \tag{3.3}$$

Moreover, this inequality (3.3) *also holds for the lengths $l(\mathbf{X}_n) = \lceil \log 1/Q(\mathbf{X}_n) \rceil$ of a Shannon code with respect any sub-probability measure Q on Ω. The pointwise redundancy $l(\mathbf{X}_n) - \log 1/P^*(\mathbf{X}_n)$ is dominated from below.*

Remarks: Any sequence of uniquely decodable codes on π_n can be made uniquely decodable on π_* by adding c_n bits to describe n (where $\sum 2^{-c_n} \leq 1$). The description length of n accounts for the difference in the bounds in (3.2) and (3.3).

The Shannon codes based on a probability measure Q (on Ω) are uniquely decodable on each π_n, but in general are not uniquely decodable on π_* (Kraft's inequality is violated when we sum over n). Nevertheless, a 1-1 code on π_* with lengths $\lceil \log 1/Q(\mathbf{x}) \rceil$ exists by the construction due to Elias (see Cover and King,1978,p.417).

The bound c_n in (3.2) tends to infinity, but at a rate $\log n$ which is much slower than the typical growth rates for $\log 1/P^*(\mathbf{X}_n)$. For instance, if $\mathbf{X}_n$ is $(X_1,X_2,\ldots,X_n)$ truncated to b_n bits accuracy and if the X_i are iid with density p^*, then the growth rate of $\log 1/P^*(\mathbf{X}_n)$ is $nh(p^*) + nb_n$, which is much greater than

$\log n$. (Here h is the differential entropy, $h(p) = \int p(x) \log 1/p(x)\, dx$.)

Lemma 1 motivates the following definitions. The *pointwise redundancy* is defined to be the random variable

$$R_n(\mathbf{X}_n) = l(\mathbf{X}_n) - \log 1/P^*(\mathbf{X}_n). \tag{3.4}$$

Likewise the *average redundancy* is defined to be the expectation $E(R_n)$. We do not require that the entropy $E \log 1/P^*(\mathbf{X}_n)$ be finite, nor that the entropy rate exist. (Compare with Davisson's 1973 definitions of redundancy.)

Proof of Theorem 3.1: Set $Q_n(\mathbf{x}_n) = 2^{-l(\mathbf{x}_n)}$ for $\mathbf{x}_n$ in π_n. Then by Kraft's inequality, Q_n is a sub-probability mass function on π_n. (In the case of a Shannon code with respect to a distribution Q, we have $Q_n(\mathbf{x}_n) \leq Q(\mathbf{x}_n)$ for all $\mathbf{x}_n$.) Note that the pointwise redundancy is the log-likelihood-ratio $R_n(\mathbf{X}_n) = \log P^*(\mathbf{X}_n)/Q_n(\mathbf{X}_n)$. The average redundancy is the relative entropy $E(R_n) = D(P^* || Q_n) = \sum_{\mathbf{x}_n} P^*(\mathbf{x}_n) \log P^*(\mathbf{x}_n)/Q_n(\mathbf{x}_n)$ which is non-negative by the concavity of the logarithm. Thus inequality (3.1) is verified.

Let $A_n = \{ R_n(\mathbf{X}_n) \leq -c\}$ be the event that the redundancy is not greater than $-c$. Then $A_n = \{ P^*(\mathbf{X}_n) \leq 2^{-c} Q_n(\mathbf{X}_n) \}$. Now use Markov's inequality or note directly that $P^*(A_n) = \sum_{A_n} P^*(\mathbf{x}_n) \leq 2^{-c} \sum_{A_n} Q_n(\mathbf{x}_n) \leq 2^{-c}$. Consequently,

$$P^*\{ R_n(\mathbf{X}_n) \leq -c\} \leq 2^{-c}. \tag{3.5}$$

For 2^{-c_n} summable, inequality (3.2) follows by the Borel-Cantelli Lemma.

The proof of (3.3) is similar except that we use $A_n = \{ R_n \leq -c, \min_{k<n} R_k > -c\}$. These events are disjoint with union $A = \{ \inf_n R_n \leq -c\}$. As before we have $P^*(A_n) \leq 2^{-c} \sum_{A_n} Q_n(\mathbf{x}_n)$. Summing over n yields $P^*(A) \leq 2^{-c} \sum_n \sum_{A_n} Q_n(\mathbf{x}_n)$. Note that each $\mathbf{x}$ in π_* appears at most once in this double sum. Thus $P^*(A) \leq 2^{-c} \sum 2^{-l(\mathbf{x})} \leq 2^{-c}$, provided the code is uniquely decodable on π_*. Likewise, for a Shannon code with respect to a distribution Q on Ω, we have $Q_n(\mathbf{x}_n) \leq Q(\mathbf{x}_n)$, and hence $P^*(A) \leq 2^{-c}\sum_n\sum_{A_n} Q(\mathbf{x}_n) = 2^{-c}Q(A) \leq 2^{-c}$. In either case we have shown that

$$P^*\{\ \inf_n R_n(\mathbf{X}_n) \leq -c\ \} \leq 2^{-c}. \tag{3.6}$$

Note that this inequality is a strengthening of (3.5). Define the random variable R to be the infimum of the pointwise redundancies $R = \inf_n R_n(\mathbf{X}_n)$. The expectation ER exceeds $-E(R)^- = -\int_0^\infty P^*\{R \leq -c\}\,dc$, which by (3.5) exceeds $-\int_0^\infty 2^{-c} = -\log e$. Thus the redundancy R_n is dominated from below by a random variable R and this R satisfies $ER \geq -\log e$. So inequality (3.3) is verified and the proof is complete.

Remark: In the above proof we associated mass functions with descriptions by setting $Q_n(\mathbf{x}_n) = 2^{-l(\mathbf{x}_n)}$ on π_n. We should point out that, in general, the sequence of mass functions Q_n need not be compatible with a distribution Q on Ω.

Often, the limit of the redundancies R_n will be infinite. Let Ω_* be the measurable subspace of Ω generated by the sequence of partitions π_n. Suppose we use a Shannon code with respect to a distribution Q. To obtain finite limiting redundancy, it is necessary that a component of Q be absolutely continuous with respect to P^* on Ω_*.

Lemma 3.1: Mutual singularity implies redundancy → ∞.

Let $R_n(\mathbf{X}_n)$ be the pointwise redundancies for a Shannon code based on a sub-probability measure Q on Ω. If Q and P^* are mutually singular on Ω_*, then $\lim_n R(\mathbf{X}_n) = \infty$ with P^*-probability one.

Proof: Note that the redundancy $\lceil \log 1/Q(\mathbf{X}_n) \rceil - \log 1/P^*(\mathbf{X}_n)$ is within one bit of the log-likelihood-ratio $\log P^*(\mathbf{X}_n)/Q(\mathbf{X}_n)$. The sequence of likelihood ratios $Q(\mathbf{X}_n)/P^*(\mathbf{X}_n)$ converges P^*-almost surely to the density ρ of the absolutely continuous component of Q with respect to P^*. (This fact is due to Doob 1953, ch.VII, sec.8 with corrections by Billingsley 1961,p.67). Here ρ is trivially zero if Q and P^* are mutually singular. Thus the limit set of the redundancy is within one bit of log $1/\rho$, which is infinite whenever $\rho = 0$. This completes the proof of Lemma 3.1.

Remark: When a component of Q is absolutely continuous with respect to P^*, this proof provides tight bounds on the limit of the redundancy. For instance, if $Q = \int P dv$ is a mixture of distributions which assigns P^* strictly positive prior probability $v(P^*)$. Then log $1/Q(\mathbf{X}_n)$ is less than log $1/v(P^*)$ + log $1/P^*(\mathbf{X}_n)$. So the redundancy is less than log $1/v(P^*) + 1$, which is finite.

On the other hand, Q and P^* are often mutually singular. For instance, if $Q = \int P dv$ is a mixture of mutually singular distributions on Ω (e.g. stationary, ergodic distributions on infinite sequences) and if the prior probability of P^* is zero (e.g. v is a nonatomic prior), then Q and P^* are mutually singular on Ω and hence (by Lemma 3.1) the redundancy tends almost surely to infinity for any such P^*.

3.2 Pointwise Universal Coding

In chapter 2 we observed that the Shannon code with respect to a single distribution Q, specifically a mixture of Bernoulli distributions, provides universally good descriptions for data drawn from any Bernoulli distribution P^*. We need a general theory of this behavior for P^* iid or stationary and ergodic. Conditions are given on the prior which ensure that the Shannon code with respect to the mixture Q has pointwise redundancy which is $o(n)$ for a large class of true distributions P^*.

Let P^* and P be probability measures on a measurable space. The relative entropy $D(P^*||P)$ is defined as

$$D(P^*||P) = \int p^* \log p^*/p \, d\mu$$

where p^* and p are the density functions (of P^* and P) with respect to any dominating measure μ (such as $\mu = P^*+P$). Note that if P^* has a singular component with respect to P then $D = \infty$. By the concavity of the logarithm, the relative entropy satisfies $D \geq 0$, with equality if and only if $P^* = P$. For the relative entropy restricted to a countable partition π we write

$$D_\pi(P^*||P) = \sum_{A\in\pi} P^*(A) \log P^*(A)/P(A).$$

The relative entropy is approximated by it discrete counterparts in the sense that $D = \sup_\pi D_\pi$ and $D = \lim_n D_{\pi_n}$ for any generating sequence of partitions. Similarly, the variation distance is defined as $||P^*-P|| = \int |p^*-p| d\mu$ or equivalently $||P^*-P|| = \sup_\pi ||P^*-P||_\pi$ where $||P^*-P||_\pi = \sum_{A\in\pi} |P^*(A)-P(A)|$.

The pointwise redundancy $R_n(\mathbf{X}_n)$ of the Shannon code based on a distribution Q is $\lceil \log 1/Q(\mathbf{X}_n) \rceil - \log 1/P^*(\mathbf{X}_n)$ which is within one bit of the log-likelihood ratio. Therefore, the necessary and sufficient condition for the average redundancy per sample to tend to zero, $\lim_n E(R_n)/n = 0$, is that the relative entropy rate equal zero, $\lim (1/n) D_{\pi_n}(P^*||Q) = 0$. Suppose that Q is a mixture of parametrized distributions according to some prior. In this case Davisson (1973,Theorem 4) has an interpretation of the relative entropy rate condition in terms of the mutual information rate between the parameters and the data. For the pointwise redundancy to be $o(n)$, we find weaker and more readily verified conditions on the prior.

Henceforth we assume that the probabilities $P^*,P,etc.$ are defined on a measurable sequence space Ω. In particular, Ω is assumed to be the space of infinite sequences $(x_1,x_2,\ldots)$ with coordinates taking values in a standard Borel space $\mathcal{X}$ (a space which is isomorphic to a Borel subset of the real line). The restriction of a distribution P to the (sigma-field of) events determined by coordinate variables $X_1,X_2,\ldots,X_n$ is denoted by P_n. Priors v are understood to be sub-probability measures on a measurable space of distributions P.

Let $\{\tau_n\}$ be a refining sequence of countable partitions which generate $\mathcal{X}$. The data $\mathbf{X}_n$ take values in a partition π_n of Ω which consists of cylinder sets specified by having the first n coordinates be events in τ_n. The coordinates of the data $\mathbf{X}_n$ are denoted $X_1^{(n)},X_2^{(n)},\ldots,X_n^{(n)}$. We may always take the following case: $\mathcal{X}$ is the real line, τ_n consists of dyadic intervals of width 2^{-b_n}, and the event $X_i^{(n)}$ is the real number X_i observed to b_n bits accuracy.

Consider first the special case that P^* is an iid distribution on Ω. Thus the random variables X_i are independent with identical marginal distribution P_1^*. If P is also iid then the relative entropy satisfies $D(P_n^*||P_n) = nD(P_1^*||P_1)$ and the variation distance satisfies $||P_n^*-P_n|| \leq n||P_1^*-P_1||$.

We establish conditions on the prior which guarantee that the redundancy of the Shannon code with respect to the mixture distribution is reasonably small.

Theorem 3.2: Redundancy = o(n); iid case.

Let $R_n(\mathbf{X}_n)$ be the pointwise redundancies for a Shannon code with respect to a mixture $Q_n = \int P_n v(dP)$ of iid distributions P. Then for the redundancy per sample to tend to zero

$$\lim \frac{R_n(\mathbf{X}_n)}{n} = 0 \quad P^*\text{-almost surely,}$$

it is sufficient that either of the following conditions hold:

(A) *The prior assigns strictly positive mass to the relative entropy sets:*

$$v\{P: D(P_1^*||P_1) < \epsilon\} > 0, \text{ for all } \epsilon>0,$$

or

(B) *The prior assigns enough mass to the variation distance neighborhoods* *in the sense that there exists ϵ_n with $\sum \epsilon_n < \infty$ such that*

$$v\{P\colon n||P_1^*-P_1|| < \epsilon_n\} = e^{-o(n)}.$$

For convergence of $R_n(\mathbf{X}_n)/n$ to zero in probability, the weaker condition that $v\{P\colon n||P_1^*-P_1||<\epsilon\} = e^{-o(n)}$, for all $\epsilon>0$, is sufficient.

Remarks: Similar conditions were introduced by L. Schwartz (1965) for the problem of Bayes consistency. The proof of Theorem 3.2 is based on her analysis (Schwartz,1965,p.22-24). Recent results (Barron 1985) on the domination of log-likelihoods are also required.

The solutions to most problems in information theory have been characterized by the reemergence of information-theoretic quantities such as the entropy, the relative entropy, and the mutual information. In that regard the sufficiency of the relative entropy condition (A) is no surprise. What is surprising is that relative entropy condition is *not* necessary. The redundancy per sample $(1/n)\log(P^*(\mathbf{X}_n)/Q(\mathbf{X}_n))$ can converge to zero almost surely even when the relative entropies $D(P_1^*||P_1)$ are infinite for v-almost every P. It is sufficient that the prior assign enough mass to the set of distributions P with $||P_1^*-P_1||\leq\epsilon/n$. Nevertheless, the relative entropy condition (A) is more readily verifiable. We only need to know that the prior assigns *some* mass to relative entropy neighborhoods, it matters not how much.

Proof of Theorem 3.2: The redundancy $R_n(\mathbf{X}_n)$ is within one bit of the log-likelihood ratio $\log(P^*(\mathbf{X}_n)/Q(\mathbf{X}_n))$. For any $\epsilon>0$ we want $\log(P^*(\mathbf{X}_n)/Q(\mathbf{X}_n)) \leq n\epsilon$ or equivalently $2^{n\epsilon}Q(\mathbf{X}_n)/P^*(\mathbf{X}_n) \geq 1$ for all large n. Now $Q(\mathbf{X}_n)$ is given by the mixture $Q(\mathbf{X}_n) = \int P(\mathbf{X}_n)v(dP)$. Let N be the set $\{ P\colon D(P_1^*||P_1) < \epsilon\}$. Restricting the integral to this set yields

$$2^{n\epsilon}Q(\mathbf{X}_n)/P^*(\mathbf{X}_n)\geq\int_N(2^{n\epsilon}P(\mathbf{X}_n)/P^*(\mathbf{X}_n))v(dP) = \int_N 2^{n(\epsilon-\hat{D})}v(dP) \qquad (3.7)$$

Here $\hat{D}$ is defined by $\hat{D} = (1/n)\log(P^*(\mathbf{X}_n)/P(\mathbf{X}_n))$. Now we use the appropriate law of large numbers to show that $\hat{D}$ tends to the relative entropy $D(P_1^*||P_1)$. Expanding $\hat{D}$ we see that

$$\hat{D} = \frac{1}{n}\sum_{i=1}^{n} \log\, (P^*(X_i^{(n)})/P(X_i^{(n)}))$$

which is the average of n independent terms. For each term i, the sequence $\log(P^*(X_i^{(n)})/P(X_i^{(n)}))$ is dominated (by an integrable random variable) and is almost surely convergent with limit given by $\log\, p^*(X_i)/p(X_i)$ (Barron 1985, Lemma 2). So by the strong law of large numbers for dominated variables,

$$\lim_{n\to\infty} \hat{D} = E \log\, p^*(X_1)/p(X_1) = D(P^*||P) \quad P^*\text{-almost surely.}$$

Thus for each P in N the integrand in (3.7) tends P^* almost surely to infinity. Consequently (by Fubini's theorem) there is a set of sequences with P^* probability one for which $\lim\, (2^{n\epsilon}P(\mathbf{X}_n)/P^*(\mathbf{X}_n)) = \infty$ for v-almost every P in N. Since the integrand is non-negative, Fatou's lemma yields

$$\liminf_n(2^{n\epsilon}Q(\mathbf{X}_n)/P^*(\mathbf{X}_n)) \geq \int_N \liminf_n(2^{n\epsilon}P(\mathbf{X}_n)/P^*(\mathbf{X}_n))v(dP) = \infty,$$

provided $v(N) > 0$. Therefore condition (A) implies that the redundancy satisfies $R_n/n \to 0$ P^*-almost surely.

Now consider the variation distance condition. Let $N_n = \{P: ||P^*-P||_{\pi_n} < \epsilon\}$. We show in general (without restrictions on the distribution) that the redundancy $R_n(\mathbf{X}_n)$ is less than $2 + \log 1/v(N_n)$ except in a set of probability less than ϵ. Then condition (B) implies that $\log 1/v(N_n) = o(n)$, (because $||P^*-P||_{\pi_n} \leq ||P_n^*-P_n|| \leq n||P_1^*-P_1||$ in the iid case).

The redundancy $R_n(\mathbf{X}_n)$ is less than $1 + \log P^*(\mathbf{X}_n)/Q(\mathbf{X}_n)$ which is less than $1 + \log P^*(\mathbf{X}_n)/Q(\mathbf{X}_n|N_n) + \log 1/v(N_n)$ where $Q(\mathbf{X}_n|N_n) = \int_{N_n} P(\mathbf{X}_n)v(dP)/v(N_n)$ is the conditional mixture. Thus it is enough to show that $P^*(\mathbf{X}_n) \leq 2Q(\mathbf{X}_n|N_n)$

except in a set of probability no more than ϵ. Again we use a variant of Markov's inequality. The event $P^*(\mathbf{X}_n) \leq 2Q(\mathbf{X}_n|N_n)$ is the same as the event $P^*(\mathbf{X}_n) \leq 2(P^*(\mathbf{X}_n)-Q(\mathbf{X}_n|N_n))$ which has probability no more than

$$\begin{aligned} 2\textstyle\sum_{\mathbf{x}_n}(P^*(\mathbf{x}_n)-Q(\mathbf{x}_n|N_n))^+ &= ||P^*-Q(\cdot|N_n)||_{\pi_n} \\ &= ||\textstyle\int_{N_n}(P^*-P)v(dP)/v(N_n)\ ||_{\pi_n} \\ &\leq \textstyle\int_{N_n}||P^*-P||_{\pi_n}v(dP)/v(N_n) \\ &< \epsilon. \end{aligned}$$

Thus the redundancy $R_n(\mathbf{X}_n)$ is less than $2 + \log 1/v(N_n)$ except in a set of probability less than ϵ. If $\log 1/v(N_n) = o(n)$ for all $\epsilon>0$, then the redundancy satisfies $R_n(\mathbf{X}_n)/n \to 0$ in probability. If also ϵ_n is summable, then by the Borel-Cantelli Lemma $R_n(\mathbf{X}_n)/n \to 0$ almost surely. This completes the proof of Theorem 3.2.

Remark: In the above proof we invoked the strong law of large numbers to show that the relative entropy density $(1/n)\log P^*(\mathbf{X}_n)/P(\mathbf{X}_n)$ converges almost surely to $D(P_1^*||P_1)$. The assumption was that both P^* and P were iid. The generalization to stationary and ergodic P^* and stationary Markov P was recently obtained in Barron (1985). The upshot is that $(1/n)\log P^*(\mathbf{X}_n)/P(\mathbf{X}_n)$ converges P^*-a.s. to the relative entropy rate

$$\lim_{n\to\infty} \frac{1}{n} \log P^*(\mathbf{X}_n)/P(\mathbf{X}_n) = D^\infty(P^*||P).$$

Here the relative entropy rate is defined as $D^\infty = \lim_n E \log \rho(X_n|X_1,\ldots,X_{n-1})$ where $\rho(X_n|X_1,\ldots,X_{n-1})$ denotes the ratio of conditional densities of P^* with respect P. Equivalently, $D^\infty(P^*||P) = E \log \rho(X_{m+1}|X_1,\ldots,X_m) + I_m$ where m is the Markov order for P and I_m is Shannon's conditional mutual information

$I_m = I(X_1;X_{m+2},X_{m+3},\ldots|X_2,\ldots,X_{m+1})$ (I_m depends only on P^* and is a measure of conditional dependence. If I_m is finite for some m, then I_m decreases to zero as $m\to\infty$.) We see that P^* is closely approximated by a Markov distribution P only if the Markov order m is sufficiently large (that I_m is near zero) and if the relative entropy between the m^{th} order conditional distributions is made sufficiently small. Armed with these results, we extend Theorem 3.2 to more general processes.

Theorem 3.2: Redundancy = o(n); general case.

If P^* is stationary and ergodic and if for each $\epsilon>0$ the prior v assigns strictly positive mass to the set of stationary Markov distributions P for which $D^\infty(P^*||P)<\epsilon$, then the pointwise redundancy $R_n(\mathbf{X}_n)$ for the Shannon code based on $Q_n = \int P_n v(dP)$ satisfies $\lim R_n(\mathbf{X}_n)/n = 0$, P^*–almost surely.

If P^* is any probability measure on Ω, and if the prior v satisfies $v\{P:||P^*-P||_{\pi_n}< \epsilon_n\} = e^{-o(n)}$ for a summable sequence ϵ_n, then the pointwise redundancy satsifies $\lim R_n(\mathbf{X}_n)/n = 0$, P^*–almost surely.

3.3 Computability and Complexity

In the preceding two sections we treated a code as a 1-1 (and uniquely decodable) mapping from a countable alphabet into finite length binary sequences. However, practical codes must satisfy an additional requirement. There must be effective procedures for encoding and decoding. (The necessity of computable codes is noted in Elias 1975.) Consider Shannon's code construction which we discussed in section 4. Encoding and decoding algorithms are readily designed, provided there is a recursive enumeration of the length function. In the case of the Shannon code with respect to a probability distribution P, we require that P be computable. Previously, computable distributions have been introduced in the context of complexity based definitions of randomness by Martin-Lof (1966), Schnorr (1977), and Schnorr and Fuchs (1977). In this section we review the necessary theory of computability and complexity.

Let $\{0,1\}^*$ denote the set of finite length sequences of zeros and ones (including the empty string Λ). Let M be a partial recursive function from $\{0,1\}^*$ into a countable set $\{x\}$. (Here $\{x\}$ may be $\{0,1\}^*$, or the integers, or π_* -- the union of the partitions in section 3.2.) The binary strings in the domain of M are called programs. Of special interest are the partial recursive functions with domains that satisfy the prefix condition: no program is the prefix of another. The *algorithmic complexity* of a symbol x with respect to M is defined as follows:

$$L_M(x) = \text{length of the shortest program } \phi \text{ such that } M(\phi) = x,$$

if there is no such program then $L_M(x) = \infty$.

There exists a *universal* complexity measure L_U, that is, there is a partial recursive function U such that for any other partial recursive function M, there is a constant C_M such that

$$L_U(x) \leq L_M(x) + C_M \text{ for all } x. \tag{3.8}$$

The definition of algorithmic complexity and the proof of universality is due to Kolmogorov (1965). Similar formulations independently appeared in Solomonoff (1964) and Chaitin (1966,1969).

The notion of partial recursive functions was first defined by Kleene (1936) who showed that a function is partial recursive if and only if it is computable by a Turing machine. A Turing machine is a simple model for digital computation developed by Turing (1936). The key property that Turing established is the existence of a universal Turing machine U. The universal machine can simulate the computation of any other Turing machine M. By providing a code for M of length C_M as a prefix to the programs of M, the universal machine is seen to satisfy property (3.8) and hence it provides the universal measure of complexity.

"Turing's analysis exposes the reasons why no computation procedure, in the sense of an unambiguous set of instructions which a human being could follow without using any imagination, is likely to be one which a Turing machine could not carry out." (Crossley et.al.,1972). Thus Church (1936) was led to his Thesis that no computational procedure will be considered as an algorithm unless it can be presented as a Turing machine. Independently, Post (1936,1943) developed an equivalent model of computability. Since then numerous other classes of reasonable computing machines have been proposed, but none has been found that is not Turing-computable (see Robinson,1950, Markov,1954, Chomsky,1959, Minsky,1967, Rogers,1967, Hopcroft and Ullman,1969). Even modern high-level computer languages, such as Pascal, may be used as a basis for the exposition of computability theory (see Kfoury, Moll, and Arbib,1982). With some effort, we could use any of these models of computation to define the algorithmic complexity of probability distributions. For definiteness, we use modified Turing machines similar to the machines proposed by Chaitin (1975,1976).

Our Turing machines may be visualized as follows. There are three tapes (a program tape, a work tape, and an output tape), ten commands, and a table indexed by finitely many states. The program tape contains a binary program ϕ of finite length. Initially, the leftmost bit is in position to be read by the machine. The *read* command shifts the program tape one bit to the left. The work tape and output tape are initially all blank have no length constraints. The commands to ***shift right, shift left, write 0, write 1***, and ***write blank*** all apply to the work tape. The ***output 0, output 1***, and ***output comma*** commands write a zero, one, or comma and shift the output tape one bit to left. The tenth possible command is to ***halt***. The table determines the specific function implemented by the Turing machine. The entries in the table give the next state and the next command as a function of the current state and the contents of the current positions of the input and work tapes. Each such table defines a different machine M. (There are $(10k)^{6k}$ Turing machines with k states.) The first state in the table indicates the initial state. The execution then proceeds according to the above prescription, one step at a time. At any time t the output tape displays a finite set of strings in $\{0,1\}^*$ with the strings delineated by commas.

let $M^t(\phi)$ be the finite set of strings output by the computer M up to time t when its program is ϕ. Let $M(\phi) = \bigcup_t M^t(\phi)$. Thus $M(\phi)$ is the output of the computer M when given the program ϕ. If M halts on the program ϕ, then the output $M(\phi)$ is necessarily a finite set of strings in $\{0,1\}^*$. If M does not halt on ϕ, then $M(\phi)$ denotes the possibly infinite set of strings ***recursively enumerated*** by M with program ϕ. If the machine does not read all of the program ϕ or if it shifts past the end of the program, then ϕ is not a valid program and $M(\phi)$ is undefined. Note that by construction, the programs in the domain of M satisfies the prefix condition. These properties were observed by Chaitin (1976).

Chaitin (1975,Theorem 2.2, 1976,Theorem 1) established the existence of a universal Turing machine U with a prefix domain. The *Chaitin complexity* of a set A of strings is defined as

$$L(A) = L_U(A) = \begin{cases} \text{length of shortest program } \phi \text{ such that } U(\phi) = A \\ \infty, \text{ if there is no such program.} \end{cases}$$

The ***recursively enumerable*** sets are defined to be those possibly infinite sets A such that $L(A) < \infty$. As Chaitin (1976) noted, this is equivalent to the standard definition of recursive enumeration.

The Chaitin complexity is a universal complexity measure (Chaitin,1976, Theorem 1). For any other prefix domain computer M, there exists a constant C_M such that

$$L(A) \leq L_M(A) + C_M \text{ for all recursively enumerable sets } A. \tag{3.9}$$

This universality is established in the standard way. The Turing machine U is of the form given above (three tapes, ten commands, finitely many states) and its state table is constructed so that U can simulate any other such Turing machine. Any program for U begins with a prefix condition code for the table defining some other machine M and then follows with a program for M. Corresponding to the shortest program for A on M we have a program on U of length $L_M(A) + C_M$. Consequently (3.9) holds and U is universal.

For finite sets A, we define $L^t(A)$ to be the length of the shortest program for A that runs in time less than or equal to t. Thus

$$L^t(A) = \min \{\text{length}(\phi)\text{: } U^t(\phi) = A\}.$$

Note that $L^t(A)$ is a monotone function which decreases to $L(A)$ as $t \to \infty$. The computation of a set A is said to be ***feasible*** in time t if $L^t(A) < \infty$. Thus $L^t(\cdot)$ measures the complexity of feasible computations. However, unlike $L(\cdot)$, the time-constrained complexity is not universal. The speed of computations depends on the choice of computer.

For sets A which consist of a single finite length string, Chaitin (1975) established the dramatic relationship $L(A) \approx \log 1/P(A)$ where $P(A)$ is the probability that a universal prefix domain machine computes A when it is given a random program (Bernoulli(1/2) sequence). Here $\approx$ means to within an additive constant. However, Chaitin (1976) also showed that no such relationship holds for sets A which contain one infinite string or more than one finite string.

Now we build up the necessary definitions of computable functions which take on integer, rational, or real values. Let $\mathbf{N}$ be the set of positive integers $\{1,2,...\}$. Let $\overline{x}$ denote the binary representation of integers x. An integer-valued function $f(x)$, $x \in \mathbf{N}$ is said to be a computable function if its graph $\{\overline{x}, \overline{f}(x)\}$ is a recursively

enumerable set. The graph is the set of all ordered pairs $(\bar{x}, \bar{f}(x))$. The complexity of a function f is defined to be the Chaitin complexity of its graph:

$$L(f) = L(A) \text{ for } A = \{\bar{x}, \bar{f}(x)\}. \tag{3.10}$$

Similarly, if $\{x\}$ is any other countable space (such as the rationals), then label the x's with a binary string $\bar{x}$ (a 1-1 coding of $\{x\}$ into $\{0,1\}^*$). Then the complexity of a rational-valued function f on $\{x\}$ is also defined as in (3.10).

Consider a real-valued function f on a countable space $\{x\}$. Let $g(x,a)$ be a rational-valued function on $\{x\} \times \mathbf{N}$ which approximates f to all accuracies a:

$$|g(x,a) - f(x)| \leq 2^{-a} \text{ for all } x, a. \tag{3.11}$$

Then the complexity of real-valued function f is defined to be the complexity of the simplest rational-valued function satisfing (3.11):

$$L(f) = \min\{L(g)\text{: } g \text{ satisfies (3.11)}\}. \tag{3.12}$$

A real-valued function f is said to be computable if $L(f) < \infty$.

Finally, the complexity of a probability measure P on a measurable space Ω is defined to be the complexity of the real-valued function $\log 1/P$ restricted to the countable space $\pi_* = \bigcup_n \pi_n$, where π_n is a sequence of countable partitions which generate Ω. From (3.10) and (3.12), we see that the complexity $L(P)$ is the length of a program for a (prefix domain universal computer) which recursively enumerates (a binary string representation of) the infinite set $\{\mathbf{x}, a, P^a(\mathbf{x})\}$, where $P^a(\mathbf{x}) = g(\mathbf{x},a)$ is the simplest rational-valued function which approximates $P(\mathbf{x})$ in the sense that

$$|\log 1/P(\mathbf{x}) - \log 1/P^a(\mathbf{x})| < 2^{-a} \tag{3.13}$$

for all $\mathbf{x}$ in π_*, and all accuracies a. If there is such a recursive enumeration, then the probability distribution P is said to be computable. Note that for large a, inequality (3.13) amounts to requiring that $P^a(\mathbf{x})$ be close to $P(\mathbf{x})$ to within a multiplicative factor of 1 ± 2^{-a}.

Shannon codes revisited

Recall from section 4, that for any length function $l(x)$ (on a countable space) which satisfies Kraft's inequality, there is a Shannon code given by the first $l(x)$ bits in the binary representation of $\sum 2^{-l(x')}$ where the sum is over all x' that precede x in a list ordered in terms of increasing length $l(x)$. Moreover, Shannon codes with lengths $l_P(x) = \lceil \log 1/P(x) \rceil$ were defined based on probability measures P (on the countable space $\{x\}$). Here we discuss how to obtain Shannon codes from a recursive enumeration of the set $\{x, a, P^a(x)\}$ where P^a satisfies (3.13).

Note that if the real number $\log 1/P(x)$ is not an integer, then there exists an accuracy a sufficiently large that

$$\lceil \log 1/P(x) \rceil = \lceil \log 1/P^a(x) \rceil .$$

However, if $\log 1/P(x)$ is an integer, then the sequence of approximating intervals (defined in (3.13)) always straddle the integer. Hence the integer part is never resolved. We circumvent this difficulty by slightly modifying the Shannon codelength. Simply fix an accuracy $a \geq 1$ and define $\tilde{P}(x)$ to be the corresponding lower bound on the probability. Specifically, define

$$\tilde{P}(x) = P^a(x)(1\text{-}2^{-a}). \tag{3.14}$$

Then Shannon code with respect to $\tilde{P}$. The codelengths are

$$l_{\tilde{P}}(x) = \lceil \log 1/\tilde{P}(x) \rceil . \tag{3.15}$$

These codelengths exceed $\lceil \log 1/P(x) \rceil$ by at most one bit.

It remains to show that we can recursively enumerate the set $\{x, l_{\tilde{P}}(x)\}$ in order of increasing codelength. We use the fact that the probabilities sum to one. To find the shortest codelength, proceed as follows. Pick x,a arbitrarily and let $\epsilon = P^a(x)(1-2^{-a})$. Run the enumeration of $\{x,a,P^a(x)\}$ while summing the approximate probabilities (replacing the less accurate terms with more accurate ones as we proceed) until the sum of the probabilities is determined to be greater than $1-\epsilon$. This implies that the x with the shortest codelength is in the finite set of x's that have been seen so far. For these x's compute the lengths $l_P(x)$ and find the shortest. Thus we have the first $(x,l_P(x))$ in the new enumeration. In general the algorithm subtracts the (approximate) probabilities corresponding to the lengths enumerated so far and then finds the next largest codelength, etc.

Two-stage programs

We note that there exists a constant C such that for every computable probability distributions P on a countable set $\{x\}$, and for every x, there is a computer program for the data x with length

$$C + L(P) + l_P(x). \tag{3.16}$$

Here $L(P)$ is the length of a program which recursive enumerates the probability distribution (to arbitrary accuracies) and $l_P(x) = \lceil \log 1/\tilde{P}(x) \rceil$ is the length of the Shannon code for the data x. (By the prefix condition, these may be concatenated without ambiguity.) The constant C is the number of bits of a preface which indicates what should be done with the remaining bits of the program. It instructs the computer that it should read the remaining bits, then run the recursive enumeration, while summing the $2^{-\text{codelengths}}$ for x's in order of increasing codelength, until the binary representation of the sum equals the Shannon codeword, then output the corresponding data x and halt.

Henceforth we ignore the constant C and the *tilde* in equations (3.15) and (3.16) -- our estimates are not effected. We still refer to $L(P) + \lceil \log 1/P(x) \rceil$ as a description length: note that Kraft's inequality is still satisfied. We do not drop the integer part notation $\lceil\ \rceil$. It serves as a constant reminder that we are directly minimizing integer description lengths. (The interpretations of $\min\{L(P) + \lceil \log 1/P(x) \rceil\}$ as a modified likelihood principle or as a Bayes rule are mathematically important and statistically revealing but motivationally secondary to the description length interpretation.)

Toward practicality

Given data X, let $\hat{P}$ achieve $\min\{L(P) + \lceil \log 1/P(X) \rceil\}$. Can a minimizing distribution $\hat{P}$ be found in finite time? In computable time? The answers are yes to the former but an unfortunate no the the latter.

Probabilities $P(X)$ are computed by running a recursive enumeration of P until the triple $(X, a, P^a(X))$ is output for a given accuracy a in an specified format. Of course not all programs are recusive enumerations of sub-probability distributions. Some programs will never halt or never output an acceptable triple for X. Some programs will output a probability $P^a(X)$ that is not consistent with previously enumerated probabilities (e.g. the total mass assigned might be greater than one.) If the inconsistency occurs for x's already enumerated, then such programs can be ignored. However, if as yet unseen outputs are inconsistent, then we could be fooled into nonsense estimators. A solution to this problem is to append to all programs a routine of fixed length which modifies the outputs of unacceptable programs to force consistency with the probabilities enumerated thus far. Acceptable outputs are left untouched.

The search for the minimizing distribution $\hat{P}$ proceeds as follows. List the computer programs in order of increasing length. Given data X begin to execute all the programs in a diagonal fashion. As soon as a program (of length $L(P)$) produces an acceptable output $P(X)$, we need not check any program of length longer than $L(P) + \lceil \log 1/P(X) \rceil$. This bound is refined as more programs are considered. At time t the distribution with shortest $L(P) + \lceil \log 1/P(X) \rceil$ is denoted by $\hat{P}^t$. Now at some (finite but uncomputable) time $t(X)$ a program achieving $\min_P \{ L(P) + \lceil \log 1/P(X) \rceil \}$ halts and hence $\hat{P}^t = \hat{P}$ for all time $t \geq t(X)$. Thus the logically smooth estimate is found, but we never know when.

3.4 Logical Data Compression

Here we present universal coding properties of logical smoothing. The pointwise redundancy is shown to be $o(n)$ for any true distribution which can be approximated by computable distributions in the relative entropy sense. Furthermore, we show that iid distributions with densities are approximated accurately (in the relative entropy sense) by computables (unless the density has peaks or tails undominated by computable functions). Thus data drawn from any reasonable density is described with asymptotically negligible redundancy.

First note that if the true distribution P^* is computable, then there is a two-stage description using $L(P^*) + \lceil \log 1/P^*(\mathbf{X}_n) \rceil$ bits. Thus for a computable distribution P^*, the pointwise redundancy of the minimum two-stage description is bounded by the constant $L(P^*) + 1$. In section 4.2, it will be shown that for parametric families, the pointwise redundancy is of order $\log n$. Here we treat redundancy for the general non-computable case.

We assume (as in section 3.2) that Ω is a space of infinite sequences with coordinates in a standard Borel space $\mathcal{X}$. The n-sample of data $\mathbf{X}_n$ takes values in a partition π_n of Ω which consists of cylinder sets specified by having the first n coordinates be events in refining partitions τ_n which generate $\mathcal{X}$. For instance, $\mathcal{X}$ may be the real line and τ_n the partition into dyadic intervals of width 2^{-b_n}. Then the data $\mathbf{X}_n$ in π_n are determined by real numbers $X_1, X_2, \ldots, X_n$ observed to b_n bits accuracy.

Let $\Gamma = \{P\}$ be a countable set of computable probability distributions on Ω. This Γ is the list of candidate estimates. For each distribution P in Γ, let $L(P)$ denote the length of a prefix domain computer program for P. For distributions P not in Γ, set $L(P) = \infty$.

Let $B(\mathbf{x}_n) = \min\{L(P) + \lceil \log 1/P(\mathbf{x}_n) \rceil : P \in \Gamma\}$ be the minimum two-stage description length for $\mathbf{x}_n$ in π_n. By definition, the pointwise redundancy is given by $R_n(\mathbf{x}_n) = B(\mathbf{x}_n) - \log 1/P^*(\mathbf{x}_n)$.

A stationary ergodic distribution P^* on Ω is said to be approximated by computable distributions in the relative entropy rate sense if for any $\epsilon > 0$, there exits a computable stationary Markov distribution P in Γ such that $D^\infty(P^*||P) < \epsilon$. This is equivalent to $\inf_{P \in \Gamma} D^\infty(P^*||P) = 0$. (The relative entropy rate D^∞ was defined in section 3.2.)

Using Theorem 3.2 on the redundancy of mixture codes, we establish the following Theorem. Recall that the pointwise redundancy is the number of excess bits beyond the ideal length $\log 1/P^*(\mathbf{X}_n)$.

Theorem 3.3: Redundancy = o(n) for two-stage descriptions.

If P^* is a stationary ergodic distribution which can be approximated by computable distributions in the relative entropy rate sense, then the pointwise redundancy $R_n(\mathbf{X}_n) = \min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil\} - \log 1/P^*(\mathbf{X}_n)$ of the minimum two-stage description satisfies

$$\lim \frac{R_n(\mathbf{X}_n)}{n} = 0 \quad P^*\text{-almost surely.}$$

Proof: Define the sub-probability measure $Q = \sum 2^{-2L(P)} P$. Note that $Q(\mathbf{X}_n) = \sum 2^{-2L(P)} P(\mathbf{X}_n)$ may be regarded as an average of $2^{-L(P)} P(\mathbf{X}_n)$ with weights $2^{-L(P)}$. The average does not exceed the maximum. Thus $Q(\mathbf{X}_n) \leq \max_P \{2^{-L(P)} P(\mathbf{X}_n)\}$. Taking logarithms yields

$$\min_P\{L(P) + \log 1/P(\mathbf{X}_n)\} \leq \log 1/Q(\mathbf{X}_n).$$

Note that the left side of this inequality is within one bit of the minimum two-stage description length. Hence the Shannon code with respect the distribution Q provides an upper bound on the minimum two-stage description. But by Theorem 3.2, the redundancy using Q is $o(n)$. Therefore, the redundancy of the minimum two-stage description is also $o(n)$.

A more direct proof is as follows. For any $\epsilon>0$, let P be such that $L(P)$ is finite and $D^\infty(P^*||P) <\epsilon$. Then the redundancy per sample satisfies

$$\frac{R_n(\mathbf{X}_n)}{n} \leq \frac{L(P)+1}{n} + \frac{1}{n} \log \frac{P^*(\mathbf{X}_n)}{P(\mathbf{X}_n)}$$

The first term in the upper bound tends to zero and the second term tends to $D^\infty(P^*||P)$ almost surely as $n\rightarrow\infty$. Thus $\limsup R_n/n \leq D^\infty<\epsilon$. Since ϵ is arbitrarily small, $\lim R_n/n = 0$ almost surely. This completes the proof of Theorem 3.3.

Remarks: We define the closure in the relative entropy sense of the set Γ of candidate distributions to be

$$\overline{\Gamma} = \{\ P^*\!:\ \inf_{P\in\Gamma} D^\infty(P^*||P) = 0\ \}.$$

Thus $\overline{\Gamma}$ is the set of all distributions P^* that can be approximated by computable distributions in the relative entropy sense. Theorem 3.3 shows that the minimum two-stage description is asymptotically optimal (to first order) for any distribution P^* in $\overline{\Gamma}$.

Can every distribution be approximated by computables in the relative entropy sense or are there distributions which remain unapproachable? Consider the iid case and restrict attention to probability density functions (with respect to Lebesgue measure on the real line). Let Γ be the set of densities p with computable distribution functions. (If a density function is computable then so is the distribution, but the converse need not hold.) Recall that the relative entropy between densities is defined as $D(p^*||p) = \int p^* \log p^*/p$.

Theorem 3.4: Approximating densities.

A density function p^* is approximated by density functions with computable distributions in the relative entropy sense, i.e., $\inf_p D(p^*||p) = 0$, *if and only if there exists a density function p with a computable distribution for which the relative entropy $D(p^*||p)$ is finite. Thus*

$$\inf_{p \in \Gamma} D(p^*||p) = \begin{cases} 0 & \text{if } D(p^*||p) < \infty \text{ for some } p \\ \infty & \text{otherwise.} \end{cases}$$

In particular, if p^* is less than a computable function with finite integral, then $\inf_p D(p^*||p) = 0$.

Remarks: Thus $\bar{\Gamma}$ includes all bounded densities with compact support, all bounded densities with tails which decrease faster than some computable and integrable function, and unbounded densities with computable and integrable bounds on the peaks.

Proof of Theorem 3.4:

The second claim is proved as follows. Let $q(x)$ be a computable, integrable function for which $p^*(x) < q(x)$ for almost all x and let $c = \int q(x)dx$. Then $p(x) = q(x)/c$ is a computable probability density function. Moreover, the density ratio $p^*(x)/p(x)$ is bounded by the constant c. Hence p is a density for which the relative entropy $D(p^*||p)$ is finite. (In fact, $D(p^*||p)$ is less than $\log c$).

Now we show that finite D implies the existence of computable distributions for which D is arbitrarily small. Let $p_0(x)$ be a density function for which the relative entropy $D(p^*||p_0)$ is finite and the distribution P_0 is computable. Using a step function $\psi(x)$ which takes on rational values on a finite set of intervals with rational endpoints, we construct a density $p(x) = p_0(x)2^{\psi(x)}/c$, where $c = \int p_0(x)2^{\psi(x)}dx$, for which the relative entropy $D(p^*||p)$ is as small as desired. Then the distribution P satisfies $P(A) = P_0(A)2^r/c$ for any set A in an interval for which $\psi(x) = r$. Clearly the distribution P is computable.

Consider the relative entropy $D(p^*||p)$. Let $h(x) = \log p^*(x)/p_0(x)$ be the log-likelihood-ratio. The relative entropy satisfies

$$\begin{aligned} D(p^*||p) &= E \log \frac{p^*}{p} \\ &= E \log \frac{p^* c}{p_0\, 2^{\psi}} \\ &= E\,(h-\psi) + \log c \\ &\leq E\,|h-\psi| + \log c. \end{aligned}$$

To make the relative entropy small, we design the step function ψ not only to approximate h in the L_1 sense, but also to have $c = \int p_0\, 2^{\psi}$ not much bigger than one. The proof uses standard approximation techniques from real analysis.

Given any integer $k>1$, let $n> \log k$ be sufficiently large that the integrals $\int_{|x|>n} p_0$, $\int_{|x|>n} |h|\,dP^*$, and $\int_{|h|>n} |h|\,dP^*$ are each less than $1/k$. Let $\delta>0$ be sufficiently small that if B is any set with Lebesgue measure $m(B)<\delta$ then $P^*(B) < 1/nk$ and $P_0(B) < 2^{-n}/k$ (such a δ exists by the absolute continuity of the distributions P^* and P_0).

We define a function $\bar{h}$ which is a truncated version of h. For $|x|>n$, set $h(x)=0$. For $|x|\leq n$, let the function $\bar{h}(x)$ be $h(x)$ clipped at $\pm n$, that is $\bar{h} = h$ for $|h|\leq n$, $\bar{h} = n$ for $h\geq n$, and $\bar{h} = -n$ for $h\leq -n$. Note that $\bar{h}$ approximates h in the sense that $\int |h-\bar{h}|\,dP^* \leq \int_{|x|>n} |h|\,dP^* + 2\int_{|h|>n} |h|\,dP^* < 3/k$.

Let $\phi(x) = \sum_{i=1}^{N} r_i I_{A_i}(x)$ be a simple function such that $\phi(x) \le \bar{h}(x)+1/k$ for $|x| \le n$. Specifically, set $N=2nk$, $r_i=(i-1)/k - n$, and $A_i = \{x : r_i \le \bar{h} < r_i+1/k, |x| \le n\}$ for $i=1,2,\ldots,N$. For each set A_i, let R_i be a disjoint union of intervals with rational endpoints such that the Lebesgue measure of the symmetric difference $A_i \Delta R_i$ is less than δ/N (such approximating sets R_i exist by the measurability of the A_i). Set $\psi(x)$ equal to the step function $\sum_{i=1}^{N} r_i I_{R_i}(x)$ clipped at $\pm n$. Then ψ is a computable function agreeing with ϕ except in a set B with measure $m(B) \le m(\bigcup_i (A_i \Delta R_i)) < \delta$. Furthermore, ψ approximates $\bar{h}$ in the sense that $\int |\bar{h} - \psi| dP^* \le 1/k + 2n\int_B dP^* < 3/k$. Moreover, ψ approximates h, since by the triangle inequality $\int |h - \psi| dP^* < \int |h - \bar{h}| dP^* + \int |\bar{h} - \psi| dP^*$, which is less than $6/k$.

Now we examine $c = \int p_0\, 2^{\psi}$. Note that by construction the function $p_0(x)2^{\psi}(x)$ is less than or equal to the probability density $p^*(x) = p_0(x)2^{h(x)}$ except possibly for x satisfying $|x|>n$, $h(x)<-n$, or x in the set B. Thus the integral $\int p_0\, 2^{\psi}$ is bounded by $\int p^* + \int_{|x|>n} p_0 + \int_{h<-n} p_0\, 2^{-n} + \int_B p_0 2^n$ which is less than $1 + 1/k + 2^{-n} + 2^n P_0(B) < 1 + 3/k$. Therefore

$$\begin{aligned} D(p^* \| p) &\le \int |h - \psi| dP^* + \log c \\ &< \frac{6}{k} + \log(1 + \frac{3}{k}) \\ &< \frac{12}{k}. \end{aligned}$$

Taking k to be as large as we wish, we have a computable distribution P with density p that is arbitrarily close to the uncomputable density p^* in the relative entropy sense. This completes the proof of Theorem 3.4.

We conclude that the redundancy of the minimum two-stage description is asymptotically negligible for any density function p^* (except for those unusual densities which are infinitely far from every computable density). Therefore, the minimum two-stage description is a universal code for data compression.

Chapter 4. Inference

Thus far, the focus of this thesis has been discussion of the asymptotics of the lengths of descriptions. Now we bring statistical considerations to the forefront. We ask, "Are the distributions which minimize the description length accurate estimates of the true distribution?" Affirmative answers are provided in the computable, parametric, and non-parametric cases. We shall see that information-theoretic principles are crucial to addressing the questions of statistical inference as well as the questions of data compression.

Throughout this chapter we assume that the measurable space Ω, on which the distributions are defined, is a space of infinite sequences $(x_1,x_2,\ldots)$ with coordinates taking values in a standard Borel space X. (Ω is endowed with the product sigma-field.) Moreover, it is assumed that the random data $\mathbf{X}_n$ takes values in refining partitions π_n which generate the sequence space Ω. (For instance, the partition π_n of Ω may consist of cylinder sets with the first n coordinates given by events in refining partitions τ_n which generate X. Furthermore, the space X may be the real line and the partitions τ_n may consist of dyadic intervals of width 2^{-b_n}. In which case, the data $\mathbf{X}_n$ consist of real numbers $X_1,X_2,\ldots,X_n$ observed to b_n bits accuracy.) The true probability distribution (or law) P^* on Ω is assumed in all cases to be stationary and ergodic.

4.1 Computable Laws

The true distribution P^* is here assumed to be computable (see section 3.3 for a definition of computable distributions). We note that every distribution that has been effectively described by statisticians is indeed computable. Moreover, for effectively described parametric families, the distribution is computable whenever the parameter values are computable. (On the other hand, distributions with randomly selected parameters are not computable with probability one, but we shall handle that case in the next section.) Examples of computable distributions include the uniform, the Gamma, the Student t, Pearson's distributions, and many others. Indeed, the number of computable distributions is countably infinite.

Let Γ denote a countable collection of candidate distributions P. We assume that each P in Γ is a computable sub-probability measure on Ω and that each P is stationary and ergodic. In the ideal case, Γ consists of all such computable distributions. Or Γ may be restricted to a list of distributions which would be reasonable in a specific case. For instance, Γ may be all computable iid distributions with densities.

For each distribution P in Γ, there is a description length $L(P)$ which is the length of a prefix condition code or computer program for P. Then using the Shannon codes, there are two-stage descriptions of the data $\mathbf{X}_n$ with total length $L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil$. The estimated distribution $\hat{P}_n$ is a distribution which achieves $\min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil : P \in \Gamma\}$ -- the minimum two-stage description.

The most important property satisfied by our estimator is the following.

Theorem 4.1: Empirical Revelations.

If the true distribution P^* is a computable distribution (in Γ), then the estimate $\hat{P}_n$ is exactly correct,

$$\hat{P}_n \equiv P^*,$$

for all sufficiently large n, with P^* probability one.

At the risk of multiplying explanations beyond necessity, we give three different proofs of this theorem. In this way we demonstrate various tools for examining the minimum description length. The first proof is perhaps the simplest. The second proof has a Bayesian viewpoint. The third proof avoids the use of martingale theory in the iid case.

First Proof: Consider the Shannon code based on the distribution $Q = \sum 2^{-L(P)}P$, where the sum is over all P in Γ which are not identical to P^*. (By Kraft's inequality $\sum 2^{-L(P)} \leq 1$, so the mixture Q is a valid sub-probability distribution.) Since each stationary, ergodic distribution P (which is not equal to P^*) must be mutually singular with respect to P^*, the mixture distribution Q is also mutually singular with respect to P^* on the sequence space Ω. Hence, by Lemma 3.1, the redundancy $\lceil \log 1/Q(\mathbf{X}_n) \rceil - \log 1/P^*(\mathbf{X}_n)$ tends almost surely to infinity. Thus for all n sufficiently large,

$$L(P^*) + \lceil \log 1/P^*(\mathbf{X}_n) \rceil < \lceil \log 1/Q(\mathbf{X}_n) \rceil$$

$$\leq \min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil : P \neq P^*\}$$

Here the second inequality follows from the fact that the sum $Q(\mathbf{X}_n) = \sum 2^{-L(P)}P(\mathbf{X}_n)$ exceeds its maximum term. Thus the two-stage description using the true distribution P^* is uniformly shorter than all other two-stage descriptions. Therefore, the minimizing distribution $\hat{P}_n$ is the true P^*. This completes the first proof.

Second Proof of Theorem 4.1: This proof is substantially due to Doob (1949). Regard $2^{-L(P)}$ as a prior probability on distributions and consider $\text{Prob}\{P{=}P^*|\mathbf{X}_n\}$, the conditional probability that the random distribution P equals P^* given the data $\mathbf{X}_n$. This conditional probability is equal to the ratio of the joint likelihood $2^{-L(P^*)}P^*(\mathbf{X}_n)$ to the marginal likelihood $2^{-L(P^*)}P^*(\mathbf{X}_n) + Q(\mathbf{X}_n)$. (Here Q is the mixture defined above.) With respect to the joint distribution of P and $\{X_1,X_2,...\}$, the conditional probability $\text{Prob}\{P{=}P^*|\mathbf{X}_n\}$ is a martingale which converges almost surely to the indicator function that P equals P^*. Hence, with respect to P^*, it converges almost surely to one. But

$\text{Prob}\{P{=}P^*|\mathbf{X}_n\} > 1/2$ is equivalent to $2^{-L(P^*)}P^*(\mathbf{X}_n) > Q(\mathbf{X}_n)$. Hence, for all sufficiently large n, the joint likelihood at P^* is greater than the sum of all the other joint likelihoods. Taking logarithms and rounding up to integer lengths, we have that

$$L(P^*) + \lceil \log 1/P^*(\mathbf{X}_n) \rceil \leq \min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil : P{\neq}P^*\}$$

Thus P^* achieves the minimum description length. To see that eventually P^* is the unique minimizing distribution, let n be large enough that $\text{Prob}\{P{=}P^*|\mathbf{X}_n\}$ is greater than 2/3. Then the joint likelihood $2^{-L(P^*)}P^*(\mathbf{X}_n)$ is at least two times as large as all other joint likelihoods. So the description length using P^* is at least one bit less than all others. Therefore, $\hat{P}_n$ equals P^* and this completes the second proof of Theorem 4.1.

Third Proof of Theorem 4.1: This proof avoids martingale theory in the iid case. Instead, we use the monotone convergence theorem and the strong law of large numbers. Note that $\hat{P}_n = P^*$ for all large n, with probability one, if and only if the probability of error $P^*\{\hat{P}_n \neq P^* \text{ for some } n{>}k\}$ decreases to zero as $k{\to}\infty$. By the union of events bound, this probability of error is less than $\sum_P P^*(A_k(P))$, where the sum is over all P not equal to P^*. Here $A_k(P)$ is the event that for some $n{>}k$ the description lengths satisfy $L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil \leq L(P^*) + \lceil \log 1/P^*(\mathbf{X}_n) \rceil$, which happens only if the log-likelihood-ratio satisfies $\log P^*(\mathbf{X}_n)/P(\mathbf{X}_n) \leq -L(P){+}c$. (Here $c = L(P^*){+}1$.) But these log-likelihood-ratios tend almost surely to infinity (either by martingale properties as in Lemma 3.1, which holds in general, or by $\log P^*(\mathbf{X}_n)/P(\mathbf{X}_n) = n(D + o(1))$ with $D{>}0$, which holds for P iid or Markov -- see equation (3.8)). Consequently, the probabilities $P^*(A_k(P))$ decrease to zero as $k{\to}\infty$. Moreover, as shown in the proof of Theorem 3.1, the probabilities $P^*(A_k(P))$ are dominated by $P^*(A_0(P)) \leq 2^{-L(P)+c}$ which is summable in P. Thus, by the monotone convergence theorem, $\lim_k \sum_P P^*(A_k(P)) = \sum_P \lim_k P^*(A_k(P)) = 0$. Therefore, except for sequences in a set of zero probability, $\hat{P}_n = P^*$ for all large n. This completes the third proof of Theorem 4.1.

4.2 Parametric Laws

One aim of statistical inference is to determine suitable models or parametric families of distributions to explain data. We do not assume that a correct model is known. Instead, we consider every conceivable parametric family (in a countable list of candidate families) so as to find the family which yields the best description of the data.

What happens if the true distribution is a member of one of the families on the list? We show that although the true family is not unique, the estimator identifies a simple family which contains the true distribution.

Let $\mathbf{P}$ denote the space of all stationary and ergodic probability distributions on Ω. We regard parametric families $\{P_\theta : \theta \in \Theta\}$ as manifolds in the space $\mathbf{P}$. A parameter space Θ is a subset of $\boldsymbol{R}^k$ for some dimension k, i.e., each θ is a parameter vector with real-valued coordinates.

We suppose that the true distribution P^* is in some parametric family $\{P_\theta : \theta \in \Theta\}$. Thus $P^* = P_{\theta^*}$ where θ^* is the true parameter vector. The results we obtain hold for all θ^* except those in a set of zero measure with respect to a prior measure v^*. (The prior v^* may be equivalent to Lebesgue measure on Θ). The property that makes possible the discovery of parametric families is the assumption that the mixture distribution $Q^* = \int P_\theta dv^*(\theta)$ is computable. Note that Q^* may be computable even if P_{θ^*} is not.

In general, priors v are probability measures on $\mathbf{P}$, the space of stationary, ergodic distributions. (We need that $\mathbf{P}$ is a measurable space; indeed, $\mathbf{P}$ inherits the sigma-field of Borel sets from Ω in the usual way, i.e. ergodic distributions correspond to sets of infinite sequences which have the appropriate limiting relative frequences of events.) Parametric families correspond to priors which are concentrated on a manifold. In particular, a prior v on the parameter space Θ of a family $\{P_\theta\}$ is also to be regarded as a prior on the space $\mathbf{P}$. Such a prior assigns zero probability to the set of distributions not in $\{P_\theta\}$. Mixture distributions are expressed as $Q^* = \int P dv^*(P)$ as well as $Q^* = \int P_\theta dv^*(\theta)$. The global view of priors permits comparison of the sets of distributions on which they concentrate.

In trying to estimate a parametric family, we are confronted with the following difficulty. There are many parametric families which intersect a given family $\{P_\theta\}$ at P_{θ^*}. Indeed, the points of intersection might even include a whole segment of the manifold, $\{P_\theta : \theta \in N\}$, where N is a neighborhood of θ^*. For example, suppose the true distribution is iid Gaussian with unknown mean θ^* and unit variance. As an alternative family consider $\{P_\theta : \theta \in \boldsymbol{R}\}$ where, for all θ outside [0,1], the distribution is Cauchy with location θ, but for θ in [0,1], the distribution is Gaussian. If θ^* happens to be in [0,1], these families are statistically indistinguishable. Nevertheless, logic dictates that the simplest family which intersects P^* is preferred.

One way to use parametric families to describe data is to try the two-stage descriptions based on mixture distributions $Q = \int P dv$ where we are free to try a multitude of priors v corresponding to various families. (Latter we will discuss two-stage descriptions where the first stage involves both a description of the family and a description of a parameter value.)

The mixture distributions are remarkably general. Indeed, any stationary distribution Q on Ω has an ergodic decomposition $Q = \int P dv$ for some prior v (Oxtoby 1952). Thus a stationary distribution Q may be regarded as the (unconditional) distribution on infinite sequences that obtains when first a stationary, ergodic distribution P is chosen at random according to a prior v and then (conditionally) the data sequence is drawn according to P. Mixtures of stationary, ergodic distributions are stationary but not ergodic (unless the prior v is degenerate).

Our inference of parametric families proceeds as follows. We try the two-stage descriptions based on stationary distributions Q and determine a stationary distribution $\hat{Q}_n$ which achieves the minimum. Corresponding to $\hat{Q}_n$ is a prior $\hat{v}_n$ such that $\hat{Q}_n = \int P d\hat{v}_n$. This $\hat{v}_n$ may be regarded as an estimated prior on distributions which we hope is partially concentrated on the true manifold.

In particular, let Γ be any countable collection of stationary distributions Q on Ω. For each Q, let $L(Q)$ denote the length of a prefix condition code or computer program for Q. Ideally, Γ consists of all computable stationary distributions and $L(Q)$ denotes the length of the shortest program for Q, but we are not restricted to this case. For each Q in Γ there is a two-stage description of the data $\mathbf{X}_n$ with length $L(Q) + \lceil \log 1/Q(\mathbf{X}_n) \rceil$. We let $\hat{Q}_n$ be a distribution achieving the minimum two-stage description length $\min\{L(Q) + \lceil \log 1/Q(\mathbf{X}_n) \rceil\}$. The estimate $\hat{Q}_n$ differs from the $\hat{P}_n$ obtained in the previous section because we are no longer restricted to ergodicity. Let $\hat{Q}_n = \int P d\hat{v}_n$ be the ergodic decomposition of the estimate $\hat{Q}_n$.

For any prior v, let $v = v^{ac} + v^s$ denote the decomposition into components v^{ac} and v^s which are (respectively) absolutely continuous and singular with respect to v^*. The absolutely continuous component v^{ac} is concentrated with v^* on the manifold (for if B is the set of distributions not on the manifold, then $v^*(B) = 0$ implies $v^{ac}(B) = 0$). Likewise, $Q^{ac} = \int P dv^{ac}$ and $Q^s = \int P dv^s$ are the absolutely continuous and the singular components of $Q = \int P dv$ with respect to $Q^* = \int P dv^*$. Because of the high dimensionality of the space of all ergodic distributions, most priors v are singular (i.e. v^{ac} is zero) with respect to priors which are supported on low dimensional manifolds. Surprisingly, our estimated prior $\hat{v}_n$ is not singular with respect to v^*.

Ideally, we would like for the estimated prior $\hat{v}_n$ to be equivalent to (mutually absolutely continuous with) the prior v^* on the true manifold $\{P_\theta\}$ in a neighborhood of θ^*. For then $\hat{v}_n$ and v^* would agree as to which distributions are possible and impossible in this neighborhood. We show that these properties nearly hold for large n.

Theorem 4.2: Finding simple but accurate families.

Suppose that $Q^* = \int P dv^*$ *is a computable distribution on the list* Γ *and that the prior* v^* *is concentrated on a manifold* $\{P_\theta\}$. *Let* $\hat{Q}_n = \int P d\hat{v}_n$ *be a distribution which achieves the minimum two-stage description length. Then the estimated prior* $\hat{v}_n$ *has an component* $\hat{v}_n^{ac}$ *on the manifold which is absolutely continuous with respect to* v^* *and which assigns strictly positive mass to* $\{P_\theta : \theta \in N\}$ *for all sufficiently large n, with* P_{θ^*} *probability one, for any neighborhood N of* θ^*, for v^* *almost every* θ^*.

Moreover, the singular component is asymptotically negligible in the sense that $\lim \hat{Q}_n^s(\mathbf{X}_n)/\hat{Q}_n(\mathbf{X}_n) = 0$ *with probability one. Furthermore, the estimated posterior distribution* $\hat{v}_n(B \mid \mathbf{X}_n) = \int_B P(\mathbf{X}_n) d\hat{v}_n / \int P(\mathbf{X}_n) d\hat{v}_n$ *asymptotically concentrates on neighborhoods of* θ^*, *in the sense that* $\lim \hat{v}_n(B \mid \mathbf{X}_n) = 1$, *with* P_{θ^*} *probability one, where* $B = \{P_\theta : \theta \in N\}$ *for any neighborhood N of* θ^*, *for* v^* *almost every* θ^*.

Proof: Since $\hat{Q}_n$ achieves the minimum two-stage description length we have that $L(\hat{Q}_n) + \lceil \log 1/\hat{Q}_n(\mathbf{X}_n) \rceil$ does not exceed $L(Q^*) + \lceil \log 1/Q^*(\mathbf{X}_n) \rceil$ and hence

$$\hat{Q}_n(\mathbf{X}_n)2^{-L(\hat{Q}_n)} > Q^*(\mathbf{X}_n)2^{-c} \tag{4.1}$$

where $c = L(Q^*) + 1$. We show that the right side of (4.1) is significantly greater than $\hat{Q}_n^s(\mathbf{X}_n)2^{-L(\hat{Q}_n)}$. Indeed, by mutual singularity on Ω, the ratio $(\sum Q^s(\mathbf{X}_n)2^{-L(Q)})/\ Q^*(\mathbf{X}_n)$ tends to zero with Q^* probability one. (Here the sum is over all Q in Γ and Q^s denotes the singular component with respect to Q^*.) Now the sum exceeds the $\hat{Q}_n$ term. Therefore,

$$\lim_n \frac{\hat{Q}_n^s(\mathbf{X}_n)2^{-L(\hat{Q}_n)}}{Q^*(\mathbf{X}_n)} = 0$$

and from inequality (4.1),

$$\lim_n \frac{\hat{Q}_n^s(\mathbf{X}_n)}{\hat{Q}_n(\mathbf{X}_n)} = 0$$

with Q^* probability one and hence with P_{θ^*} probability one, for v^* almost every θ^*. Hence, for all large n, the estimated prior $\hat{v}_n$ (and its absolutely continuous component $\hat{v}_n^{ac}$) assign positive probability to the manifold on which v^* concentrates. It remains to examine the behavior in neighborhoods of θ^*.

Consider a fixed set N in the true parameter space Θ and condition on the event that the random θ^* is in N. Then the distribution for the data is $Q_*^N = \int_B P dv^*/v^*(B)$ where $B = \{P_\theta : \theta \in N\}$. For each Q, let $Q^{N^c} = \int_{B^c} P dv$ be the distribution obtained by removing the component on the set $B = \{P_\theta : \theta \in N\}$. Then Q_*^N is mutually singular with respect to each Q^{N^c} and hence the ratio $(\sum Q^{N^c}(\mathbf{X}_n)2^{-L(Q)})/\ Q_*^N(\mathbf{X}_n)$ tends to zero with Q_*^N probability one. So as before we have

$$\lim_n \frac{\hat{Q}_n^{N^c}(\mathbf{X}_n)2^{-L(\hat{Q}_n)}}{Q^*(\mathbf{X}_n)} = 0$$

and from inequality (4.1),

$$\lim_n \frac{\hat{Q}_n^{N^c}(\mathbf{X}_n)}{\hat{Q}_n(\mathbf{X}_n)} = 0 \tag{4.2}$$

with Q_*^N probability one, and hence with P_{θ^*} probability one for v^* almost every θ^* in N.

By assumption, the parameter space Θ for the true family is a subset of the Euclidean space $\boldsymbol{R}^k$ for some k. Therefore there exists a countable basis $\{N\}$, that is, a countable collection of sets such that for any parameter θ^*, there is a sequence of neighborhoods N_j in $\{N\}$ which contain θ^* and have diameter decreasing to zero as $j \rightarrow \infty$.

Consequently, from (4.2), the event $\bigcap_N \lim_n (\hat{Q}_n^{N^c}(\mathbf{X}_n)/\hat{Q}_n(\mathbf{X}_n)) \ I_{\{\theta^* \in N\}} = 0$ has Q^* probability one and hence P_{θ^*} probability one for v^* almost every θ^*. Therefore, letting N be any neighborhood of θ^*, we have that $\lim \hat{Q}_n^{N^c}(\mathbf{X}_n)/\hat{Q}_n(\mathbf{X}_n) = 0$ or equivalently $\lim \hat{v}_n(B^c|\mathbf{X}_n) = 0$ with P_{θ^*} probability one, for v^* almost every θ^*. Thus the estimated prior $\hat{v}_n$ is asymptotically concentrated on the true family in neighborhoods on the true parameter. This completes the proof of Theorem 4.2.

Remarks: Theorem 4.2 may be used to obtain a stronger conclusion in certain contexts. For instance, suppose the list of candidate families $\{P_\theta\}$ is reduced to a set of families that have no non-trivial intersection. Then the true family is discovered for all large n, with probability one, for almost every parameter value.

The mixture distributions may be used to obtain lower bounds on the minimum two-stage description length $\min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil\}$ where the minimum is over stationary and ergodic distributions P. If the prior v^* is continuous, i.e., it contains no mass points, then the mixture $Q^* = \int P dv^*$ is mutually singular (on Ω) with respect to each P. Hence the following result is obtained as a special case of either Lemma 3.1 or Theorem 4.2.

Lemma 4.1: The continuous mixture bound on two-stage descriptions.

Let $Q^* = \int P dv^*$ *denote the mixture distribution with respect to a prior* v^* *which contains no mass points. Then the minimum two-stage description length* $B(\mathbf{X}_n) = \min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil\}$ *exceeds the Shannon codelength* $\lceil \log 1/Q^*(\mathbf{X}_n) \rceil$ *for all large n, with* P^* *probability one, for* v^* *almost every* P^*. *Indeed the redundancy* $B(\mathbf{X}_n) - \log 1/Q^*(\mathbf{X}_n)$ *tends to infinity with probability one.*

This result is especially useful for smooth parametric families for which the mixture likelihoods $Q(\mathbf{X}_n) = \int P_\theta(\mathbf{X}_n) dv(\theta)$ can be evaluated exactly or at least closely approximated.

Description length for smooth families

In this section we show that if $\{P_\theta\}$ is a computable family which satisfies appropriate smoothness conditions then there is a two-stage description of the data $\mathbf{X}_n$ with length within 2 log log n bits of $(k/2) \log n + \log 1/P_{\theta_{ML}}(\mathbf{X}_n)$, where θ_{ML} is the maximum likelihood parameter estimate and k is the dimension of the parameter vector. Moreover, this same length $(k/2) \log n + \log 1/P_{\theta_{ML}}(\mathbf{X}_n)$ is asymptotically within a constant of the Shannon codelength $\log 1/Q(\mathbf{X}_n)$ with respect to the mixture distributions $Q = \int P_\theta v(\theta) d\theta$ for continuous prior density functions $v(\theta)$. But by Lemma 4.1 (also Theorem 3.1), $\log 1/Q(\mathbf{X}_n)$ is an asymptotic lower bound on the minimum description length, for data drawn from P_{θ^*} for (Lebesgue) almost every θ^*. Consequently, it is shown that $(k/2) \log n + \log 1/P_{\theta_{ML}}(\mathbf{X}_n)$ is an asymptotically achieved lower bound on the minimum description length.

Example

Suppose the distribution on sequences in Ω is iid according to the Gaussian distribution with density function $p_\theta(x) = (1/\sqrt{2\pi})\exp\{-(x-\theta)^2/2\sigma^2\}$. Here θ is the mean or location parameter and σ^2 is the variance or scale parameter. In this example we regard σ^2 as a known computable number, whereas θ is unknown (and likely to be uncomputable). For the sequence of real-valued variables $X_1,X_2,\ldots,X_n$, the joint density function is given by

$$p_\theta(X_1,\ldots,X_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\sum(X_i-\theta)^2/2\sigma^2}$$

$$= p_{\theta_{ML}}(X_1,\ldots,X_n)\, e^{-n(\theta-\theta_{ML})^2/2\sigma^2}$$

where $\theta_{ML} = (1/n)\sum_{i=1}^n X_i$ is the maximum likelihood parameter estimate. (For more general families a similar result holds by a second order Taylor expansion, in which case $1/\sigma^2$ is replaced with the Fisher information.)

Let $v(\theta)$ be a prior density function on the parameter θ. (We will need v to be continuous and strictly positive.) Consider the mixture distribution $Q = \int P_\theta v(\theta)\,d\theta$. This mixture has a joint density function given by

$$q(X_1,\ldots,X_n) = \int p_\theta(X_1,\ldots,X_n) v(\theta)\,d\theta$$

$$= p_{\theta_{ML}}(X_1,\ldots,X_n) \int e^{-n(\theta-\theta_{ML})^2/2\sigma^2} v(\theta)\,d\theta.$$

Note that the integrand is insignificant except in a neighborhood of θ_{ML} (within which $v(\theta)$ will be nearly constant). Therefore the following approximation is to be expected,

$$q(X_1,\ldots,X_n) \approx p_{\theta_{ML}}(X_1,\ldots,X_n) v(\theta_{ML}) \int e^{-n(\theta-\theta_{ML})^2/2\sigma^2} d\theta$$

$$= p_{\theta_{ML}}(X_1,\ldots,X_n) v(\theta_{ML}) (\frac{2\pi\sigma^2}{n})^{1/2} \tag{4.3}$$

Indeed, this approximation is a special case of Laplace's method of integration which is quite accurate in general for approximating mixture densities (see Lemmas 4.2 and 4.3, below).

Suppose that the observed data is $\mathbf{X}_n = (X_1^b, X_2^b, ..., X_n^b)$ where X_i^b is X_i observed to $b = b_n$ bits accuracy. It will be shown (Theorem 4.3>) that if $b_n \rightarrow \infty$ then the approximations for densities also hold for the probabilities, i.e.,

$$P_\theta(\mathbf{X}_n) \approx P_{\theta_{ML}}(\mathbf{X}_n) e^{-n(\theta-\theta_{ML})^2/2\sigma^2} \tag{4.4}$$

and

$$Q(\mathbf{X}_n) \approx P_{\theta_{ML}}(\mathbf{X}_n) v(\theta_{ML}) (\frac{2\pi\sigma^2}{n})^{1/2}. \tag{4.5}$$

The length of the Shannon code with respect to Q is now approximated. Taking logarithms in (4.5) yields

$$\lceil \log \frac{1}{Q(\mathbf{X}_n)} \rceil \approx \frac{1}{2} \log \frac{n}{2\pi\sigma^2} + \log \frac{1}{v(\theta_{ML})} + \log \frac{1}{P_{\theta_{ML}}(\mathbf{X}_n)}. \tag{4.6}$$

If in place of θ_{ML} we use any $\hat{\theta}$ which differs from θ_{ML} by no more than $\sigma/\sqrt{n}$, then by the expansion (4.4), the approximation (4.6) remains accurate to within (1/2) log e bits.

Surprisingly, the terms in the expansion (4.6) are suggestive of a "two-stage" description of the data. A distribution $P_{\hat{\theta}}$ is described as follows. First the family (in this case iid Gaussian) and the prior v are described using $L(\mathit{Gauss.}) + L(v)$ bits (these lengths should be added to both sides of (4.6)). Then using the prior v we describe the parameter value $\hat{\theta}$ which is the maximum likelihood estimate given to $k_n = \lceil \log \sqrt{n}/\sigma \rceil$ bits accuracy. Roughly, the term $\log 1/v(\theta_{ML})$ corresponds to the bits to the left of the binary point and the term $k_n = 1/2 \log n/\sigma^2$ corresponds to the bits to the right of the binary point. In particular, let N be the dyadic interval of width 2^{-c} which contains θ_{ML} where the constant c is chosen such that $\log 1/v(\theta)$ does not significantly vary within such intervals (for non-Gaussian cases we will also need to assume that N is small enough that the Fisher information does not vary significantly within N). This interval N is described using about $\log 1/(v(\theta_{ML})2^{-c}) = c + \log 1/v(\theta_{ML})$ bits for the Shannon code based on the prior v. The remaining $(k_n - c)$ bits of $\hat{\theta}$ may then be given, once the integer $k_n = \log\sqrt{n}/\sigma$ is described. Now integers k are exactly described using $\log^* k = \lceil \log k + 2 \log\log k \rceil$ bits for the Shannon code with respect to the mass function $P(k) = 1/k(\log k)^2$ for $k>2$. Finally the fact that $\hat{\theta}$ is within $\sigma/\sqrt{n}$ of the maximum likelihood estimate (and the expansion as in (4.3)) implies that the Shannon code with respect to $P_{\hat{\theta}}$ has codelength $\lceil \log 1/P_{\hat{\theta}}(\mathbf{X}_n) \rceil$ near $\log 1/P_{\theta_{ML}}(\mathbf{X}_n)$. Putting it all together, we have a two-stage description of the data with length within a term of order $\log\log\log\sqrt{n}$ of the following

$$\log \frac{1}{v(\theta_{ML})} + \log \frac{\sqrt{n}}{\sigma} + \log\log \frac{\sqrt{n}}{\sigma} + \log \frac{1}{P_{\theta_{ML}}(\mathbf{X}_n)}. \tag{4.7}$$

At first the iterated logarithm may seem to be more accurate an expansion than necessary. But note that without the $\log\log\sqrt{n}/\sigma$ term, the two-stage description length as in (4.7) would be within a constant of the Shannon codelength based on the continuous mixture as in (4.6), which would violate Lemma 4.1. The unexpected dividend of our analysis is that we can assert that the two-stage description length in (4.7) is within $\log\log\sqrt{n}/\sigma$ of the minimum two-stage description length over *all* families. Thus we have presented a distribution which achieves nearly minimum description length. This is a surprising result because we know that to find the exact minimum would require unbounded computational resources.

Now we present the tools needed to handle smooth families in general. The following Lemma is patterned after a result in Tierney and Kadane (1984, Theorem 1). The proof is fairly standard and hence omitted. Additional useful background is in Polya and Szego (1972, Pt.II, problem 201 and solution), and in

De Bruijn (1958, chapter 4). The function $g_n(\theta)$ in the Lemma will be used to denote log-likelihood-ratios in the succeeding development. Similarly, $g(\theta)$ will be the relative entropy and $G(\theta)$ will be the Fisher information matrix.

Lemma 4.2: Laplace's method of integration.

Let $g_n(\theta)$, $g(\theta)$ and $v(\theta)$ be functions on $\boldsymbol{R}^k$. Suppose the following conditions are satisfied.

(i) *$g_n(\theta)$ and $g(\theta)$ are twice continuously differentiable in a neighborhood of θ^*. The matrices of second partial derivatives are denoted $G_n(\theta)$ and $G(\theta)$ respectively.*

(ii) *$g(\theta)$ has a unique minimum at θ^* and* $\det G(\theta^*) > 0$.

(iii) $\lim_n g_n(\theta^*) = g(\theta^*)$.

(iv) *$G_n(\theta) \rightarrow G(\theta)$ uniformly on a neighborhood of θ^*.*

(v) *For any $\delta > 0$, there exists $c(\delta) > 0$ such that if $|\theta-\theta^*| \geq \delta$, then $g_n(\theta) \geq g(\theta^*) + c(\delta) > 0$, for all large n.*

(vi) *$v(\theta)$ is continuous, integrable, and $v(\theta^*) > 0$.*

Then for all large n, $g_n(\theta)$ has a unique minimum at θ_{ML} and $\theta_{ML} \rightarrow \theta^*$ as $n \rightarrow \infty$. The function $g_n(\theta)$ satisfies

$$g_n(\theta) = g_n(\theta_{ML}) + \frac{1}{2}(\theta-\theta_{ML})^T \hat{G}\, (\theta-\theta_{ML})(1+\epsilon_n) \tag{4.8}$$

where $\hat{G} = G(\theta_{ML})$ and the factor ϵ_n tends to zero as $\theta-\theta_{ML} \rightarrow 0$ and $n \rightarrow \infty$. Furthermore,

$$\int e^{-ng_n(\theta)} v(\theta)\, d\theta = e^{-ng_n(\theta_{ML})} v(\theta_{ML}) \left(\frac{1}{n^k \det \hat{G}} \right)^{1/2} (2\pi)^{k/2}\, (1 + o(1)) \tag{4.9}$$

where $o(1) \rightarrow 0$ as $n \rightarrow \infty$.

We note that $\lim_n G_n(\theta_{ML}) = \lim_n G(\theta_{ML}) = G(\theta^*)$. Therefore, the approximations remain valid with either $G_n(\theta_{ML})$ or $G(\theta^*)$ in place of $\hat{G}$. Also θ_{ML} may be replaced with any $\hat{\theta}$ for which $|\hat{\theta} - \theta_{ML}| = o(1/\sqrt{n})$.

Now we provide conditions on parametric density functions that ensure the applicability of Laplace's method. Let $\{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^K$, be a parametric family of iid distributions with density functions $p_\theta(x)$. Let $X_1, X_2, \ldots$ be independent random variables with common density function $p^*(x)$. (The true distribution P^* need not be in the family.) The observed data is assumed to be of the form $\mathbf{X}_n = (X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)})$ where $X_i^{(n)}$ is a small interval containing X_i. In particular, $X_i^{(n)}$ is in a countable partition τ_n of the line. We require that the maximum width of intervals in τ_n tend to zero as $n \to \infty$.

Let $g(\theta) = E \log p^*(X)/p_\theta(X)$ be the relative entropy which we will need to be finite for some open set of parameters θ. The assumptions will also ensure that $g(\theta)$ is twice continuously differentiable. We let $G(\theta)$ be the corresponding matrix of second partial derivatives $(\partial^2/\partial\theta_j\theta_k)g(\theta)$.

Theorem 4.3: Approximations for smooth likelihoods.

Suppose the following conditions are satisfied by a parametric family of densities $p_\theta(x)$.

(1) The derivatives $(\partial/\partial\theta_j)p_\theta(x)$ and $(\partial^2/\partial\theta_j\theta_k)p_\theta(x)$ exist and are continuous on the set $\{(\theta,x) : |\theta - \theta^*| \leq \delta\}$ for some $\delta > 0$.

(2) $g(\theta)$ has a unique minimum at θ^* and $\det G(\theta^*) > 0$.

(3) The derivatives $(\partial^2/\partial\theta_j\theta_k)\log 1/p_\theta(x)$ and $((\partial/\partial\theta_j)\log 1/p_\theta(x))^2$ are locally dominated in the sense that there exists $\delta > 0$ and $c(x)$ with finite $Ec(X)$ for which $|(\partial^2/\partial\theta_j\theta_k)\log 1/p_\theta(x')|$ and $((\partial/\partial\theta_j)\log 1/p_\theta(x'))^2$ are less than

$c(x)$ for all $|\theta-\theta^*| \leq \delta$ and $|x' - x| \leq \delta$.

(4) $E \log \dfrac{p^*(X)}{\sup_\theta p_\theta(X)} > -\infty$.

(5) $\lim_{|\theta|\to\infty} p_\theta(x) = 0$ *for all x.*

(6) $v(\theta)$ is continuous, integrable, and $v(\theta^*) > 0$.

Then with probability one, there exists a unique θ_{ML} maximizing the likelihood $P_\theta(\mathbf{X}_n)$ for all large n. The maximum likelihood estimate θ_{ML} is consistent, i.e., $\lim_n \theta_{ML} = \theta^*$ *almost surely. Let $\hat{G} = G(\theta_{ML})$. The log-likelihood satisfies*

$$\log \frac{1}{P_\theta(\mathbf{X}_n)} = \log \frac{1}{P_{\theta_{ML}}(\mathbf{X}_n)} + \frac{n}{2}(\theta-\theta_{ML})^T \hat{G}\,(\theta-\theta_{ML})(1+\epsilon_n) \qquad (4.10)$$

where $\epsilon_n \to 0$ almost surely as $|\theta-\theta_{ML}| \to 0$ and $n \to \infty$. Furthermore, the mixture distribution $Q = \int P_\theta v(\theta) d\theta$ satisfies

$$Q(\mathbf{X}_n) = P_{\theta_{ML}}(\mathbf{X}_n) v(\theta_{ML}) \left(\frac{(2\pi)^K}{n^K \det \hat{G}} \right)^{1/2} (1 + o(1)). \qquad (4.11)$$

where $o(1) \to 0$ almost surely as $n\to\infty$.

Remarks: The approximations remain valid with $\hat{G} = G(\theta_{ML})$ replaced by the Fisher information $G(\theta^*)$ or the observed Fisher information $G_n(\theta_{ML})$ (the matrix with entries $(1/n)(\partial^2/\partial\theta_j\theta_k)\log 1/P_\theta(\mathbf{X}_n)$). Note that by (4.10), if $\hat{\theta} = \theta_{ML} + O(1/\sqrt{n})$ then the log-likelihood at $\hat{\theta}$ differs from the maximum by at most a constant.

Proof: We show that the assumptions (1)-(6) imply the satisfaction of the conditions (i)-(vi) of Lemma 4.2 with g_n given by $g_n(\theta) = (1/n) \log P^*(\mathbf{X}_n)/P_\theta(\mathbf{X}_n)$. The verification of conditions (i), (ii), and (vi) is immediate. Condition (iii) holds by the strong law of large numbers for dominated variables applied to

$g_n(\theta^*) = (1/n) \sum \log P^*(X_i^{(n)})/P_{\theta^*}(X_i^{(n)})$. (The domination of the variables $\log P^*(X_i^{(n)})/P_{\theta^*}(X_i^{(n)})$ for each fixed index i is a consequence of Barron 1985, Lemma 2.)

The verification of condition (iv) is the most difficult step. We need to show that the matrix $G_n(\theta)$ with entries $(1/n)(\partial^2/\partial\theta_j\theta_k)\log 1/P_\theta(\mathbf{X}_n)$ $=(1/n)\sum_{i=1}^n(\partial^2/\partial\theta_j\theta_k)\log 1/P_\theta(X_i^{(n)})$ converges to $G(\theta)$ uniformly on a neighborhood of θ^*. First consider the densities. From calculus we have

$$\frac{\partial}{\partial\theta_j}\log\frac{1}{p_\theta(x)} = \frac{-1}{p_\theta(x)}\frac{\partial}{\partial\theta_j}p_\theta(x)$$

and

$$\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{p_\theta(x)} = \left(\frac{-1}{p_\theta(X)}\frac{\partial^2}{\partial\theta_j\theta_k}p_\theta(x)\right) + \left(\frac{1}{p_\theta(X)}\frac{\partial}{\partial\theta_j}p_\theta(x)\right)\left(\frac{1}{p_\theta(X)}\frac{\partial}{\partial\theta_k}p_\theta(x)\right)$$

Similarly we calculate the derivatives of $\log 1/P_\theta(A)$ for any compact interval A. The first derivatives are given by

$$\frac{\partial}{\partial\theta_j}\log\frac{1}{P_\theta(A)} = \frac{-1}{P_\theta(A)}\frac{\partial}{\partial\theta_j}P_\theta(A)$$

$$= \frac{-1}{P_\theta(A)}\int_A\frac{\partial}{\partial\theta_j}p_\theta(x)\,dx$$

and second derivatives are given by

$$\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(A)} = \left(\frac{-1}{P_\theta(A)}\int_A\frac{\partial^2}{\partial\theta_j\theta_k}p_\theta(x)\,dx\right) \tag{4.12}$$

$$+\left(\frac{1}{P_\theta(A)}\int_A\frac{\partial}{\partial\theta_j}p_\theta(x)\,dx\right)\left(\frac{1}{P_\theta(A)}\int_A\frac{\partial}{\partial\theta_k}p_\theta(x)\,dx\right)$$

Here we employed exchanges in the order of differentiation and integration (which are valid since the derivatives are continuous and hence uniformly contintous on the compact set $\{(\theta,x) : |\theta-\theta^*|\leq\delta,\ x\in A\}$). Now we simplify expression

(4.12) as follows

$$\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(A)} = \int_A\left(\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{p_\theta(x)}\right)\frac{p_\theta(x)dx}{P_\theta(A)} - R_{jk} \tag{4.13}$$

where the remainder term R_{jk} is given by

$$R_{jk} = \int_A\left(\frac{\partial}{\partial\theta_j}\log\frac{1}{p_\theta(X)}\right)\left(\frac{\partial}{\partial\theta_k}\log\frac{1}{p_\theta(X)}\right)\frac{p_\theta(x)dx}{P_\theta(A)}$$

$$- \int_A\left(\frac{\partial}{\partial\theta_j}\log\frac{1}{p_\theta(X)}\right)\frac{p_\theta(x)dx}{P_\theta(A)}\int_A\left(\frac{\partial}{\partial\theta_k}\log\frac{1}{p_\theta(X)}\right)\frac{p_\theta(x)dx}{P_\theta(A)}. \tag{4.14}$$

Examining the terms in (4.13) and (4.14) we see that the integrals are all finite (from assumption 3). Let $A = X^{(n)}$ be a sequence of sets decreasing to X. By a standard generalization of the Lebesgue density theorem (as in Ash, 1972, p.78), each of the (normalized) integrals on $X^{(n)}$ tends almost surely to the integrand at X. Consequently,

$$\lim_n \frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(X^{(n)})} = \frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{p_\theta(X)}$$

with P_θ probability one and hence P^* probability one (since finite relative entropy $g(\theta) = D(P^*||P_\theta)$ implies that P^* is absolutely continuous with respect to P_θ). Note also that the terms in (4.13) and (4.14) involve averages with respect P_θ conditioned on the set A. Now averages are less than suprema, so we have

$$\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(A)} \le \sup_{x\in A}\left|\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{p_\theta(x)}\right| + 2\sup_{x\in A}\left|\frac{\partial}{\partial\theta_j}\log\frac{1}{p_\theta(x)}\right| \sup_{x\in A}\left|\frac{\partial}{\partial\theta_k}\log\frac{1}{p_\theta(x)}\right|$$

Consider the interval $X^{(n)}$ in τ_n that contains the random variable X. Let N be large enough that the maximum width interval in τ_n is less than δ for all $n \ge N$. Then, $X^{(n)}$ is a subset of $\{x|x-X|\le\delta\}$. Therefore,

$$E\sup_{\substack{n\ge N\\ |\theta-\theta^*|\le\delta}}\left|\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(X^{(n)})}\right|$$

$$\leq E \sup_{\substack{|x-X|\leq\delta \\ |\theta-\theta^*|\leq\delta}} \left|\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{p_\theta(x)}\right| + 2E\left(\sup_{\substack{|x-X|\leq\delta \\ |\theta-\theta^*|\leq\delta}} |\frac{\partial}{\partial\theta_j}\log\frac{1}{p_\theta(x)}| \sup_{\substack{|x-X|\leq\delta \\ |\theta-\theta^*|\leq\delta}} |\frac{\partial}{\partial\theta_k}\log\frac{1}{p_\theta(x)}|\right)$$

$$\leq E\sup|\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{p_\theta(x)}| + 2\left\{E\sup(\frac{\partial}{\partial\theta_j}\log\frac{1}{p_\theta(x)})^2\ E\sup(\frac{\partial}{\partial\theta_k}\log\frac{1}{p_\theta(x)})^2\right\}^{1/2}$$

$$\leq 3Ec(X) < \infty.$$

Thus $(\partial^2/\partial\theta_j\theta_k)\log 1/P_\theta(X_i^{(n)})$ is locally dominated.

For convenience set $G_{jk}(\theta) = E\,(\partial^2/\partial\theta_j\theta_k)\log 1/p_\theta(X)$. (Note that $G_{jk}(\theta) = E\,(\partial^2/\partial\theta_j\theta_k)\log p^*(X)/p_\theta(X) \quad = (\partial^2/\partial\theta_j\theta_k)\,E\log p^*(X)/p_\theta(X)$. The exchange of expectation and derivative is valid by application of the mean value and dominated convergence theorems.) Given any $\epsilon>0$ and any θ satisfying $|\theta-\theta^*|\leq\delta/2$, the monotone convergence theorem shows that there exists $\delta(\theta)>0$ sufficiently small and $N(\theta)$ sufficiently large that

$$E \sup_{\substack{n\geq N(\theta) \\ |\theta'-\theta|\leq\delta(\theta)}} \frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_{\theta'}(X^{(n)})} \leq G_{jk}(\theta)+\epsilon$$

and

$$\sup_{|\theta'-\theta|\leq\delta(\theta)} |G_{jk}(\theta')-G_{jk}(\theta)| \leq \epsilon.$$

Now $\{\theta : |\theta-\theta^*|\leq\delta/2\}$ is a compact set covered by $\bigcup_\theta\{\theta' : |\theta'-\theta|\leq\delta(\theta)\}$. Therefore there exists a finite subcovering $\bigcup_{m=1}^M\{\theta : |\theta-\theta^m|\leq\delta(\theta^m)\}$. Then by the strong law of large numbers,

$$\limsup_n \sup_{|\theta-\theta^*|<\delta/2}\left(\frac{1}{n}\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(\mathbf{X}_n)} - G_{jk}(\theta)\right)$$

$$\leq \limsup_n \max_{1\leq m\leq M}\sup_{|\theta-\theta^m|\leq\delta(\theta^m)}\left(\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\theta_j\theta_k}\log\frac{1}{P_\theta(X_i^{(n)})} - G_{jk}(\theta)\right)$$

$$\leq \max_{m} \limsup_{n} \frac{1}{n} \sum_{i=1}^{n} \sup_{\substack{n' \geq N(\theta^m) \\ |\theta-\theta^m| \leq \delta(\theta^m)}} \left(\frac{\partial^2}{\partial \theta_j \theta_k} \log \frac{1}{P_\theta(X_i^{(n')})} - G_{jk}(\theta) \right) .$$

$$= \max_{m} E \sup_{\substack{n \geq N(\theta^m) \\ |\theta-\theta^m| \leq \delta(\theta^m)}} \left(\frac{\partial^2}{\partial \theta_j \theta_k} \log \frac{1}{P_\theta(X_i^{(n)})} - G_{jk}(\theta) \right) \quad \text{(with probability one)}$$

$$< 2\epsilon.$$

Therefore the positive part of $(1/n)(\partial^2/\partial\theta_j\theta_k) \log 1/P_\theta(\mathbf{X}_n) - G_{jk}(\theta)$ converges to zero uniformly in a neighborhood of θ^* with probability one. Similarly the negative part converges uniformly. Thus condition (iv) is verified. Finally condition (v) follows from assumptions (4) and (5) as in Wald's proof of the consistency of the maximum likelihood estimate (Wald 1949, extended by Wolfowitz 1949). Since the conditions for Lemma 4.2 are now verified, the results (4.10) and (4.11) follow by Laplace's method. This completes the proof of Theorem 4.3.

A computable parametric family $\{P_\theta : \theta \in \Theta\}$, $\Theta \in R^k$ is here defined as a family of distributions for which there exists a recursive enumeration of the set $\{\theta, x, a, P_\theta^a(x)\}$ for all rational θ in Θ, all X in π_*, (the union of the partitions π_n), and all accuracies a. Let $\tilde{\Theta}$ be the set of rational parameters in Θ. We assume that Θ is the closure of $\tilde{\Theta}$. The likelihood $P_\theta(\mathbf{X}_n)$ is assumed to be continuous on $\tilde{\Theta}$ for each x, so that it extends uniquely to all θ in Θ.

The results of this section are summarized by the following theorem.

Theorem 4.4: Description length for smooth families.

Let $\{P_\theta : \theta \in \Theta\}$ be a computable parametric family of iid distributions with density functions $p_\theta(x)$ which satisfy the smoothness assumptions of Theorem 4.3. Assume also that the prior density $v(\theta)$ is computable and let $Q = \int P_\theta v(\theta) d\theta$ be the mixture distribution. Set

$$L_n = \frac{1}{2} \log \frac{n^k \det \hat{G}}{(2\pi)^k} + \log \frac{1}{v(\theta_{ML})} + \log \frac{1}{P_{\theta_{ML}}(\mathbf{X}_n)}$$

Then the minimum two-stage description length $\min_P\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil\}$ *is between* L_n *and* $L_n + 2 \log\log n + c$ *for all large* n*, with* P_{θ^*} *probability one, for almost every* θ^*.

Proof: This theorem is an immediate consequence of the following 3 results.

(a) There exists a two-stage description of $\mathbf{X}_n$ with length less than

$$L_n + 2 \log\log n + c.$$

(b) The Shannon code with respect to the mixture Q has length

$$\log 1/Q(\mathbf{X}_n) = L_n + o(1).$$

(c) The minimum two-stage description length exceeds $\log 1/Q(\mathbf{X}_n)$ with P_{θ^*} probability one, for almost every parameter θ^*.

Recall that (b) is from Theorem 4.3 and that (c) is from Lemma 4.1. It remains to establish (a). The proof parallels the argument given for the univariate case (see also Rissanen 1983). First a constant number of bits are used to describe the family and the prior. Then $\log 1/v(\theta_{ML}) + kc$ bits are used to describe a small cube with sides of width 2^{-c} which contains θ_{ML} and within which $v(\theta)$ and $G(\theta)$ (the Fisher information) do not significantly vary. Fixing a representative θ in the cube, we consider the orthogonalization $G(\theta) = U^T \lambda U$ where U is an orthonormal matrix and λ is the diagonal matrix of eigenvalues. The components $\tilde{\theta}_j$ of the rotated parameter vector $\tilde{\theta} = U\theta_{ML}$ are then described to $k_j = \log 1/\Delta_j$ bits accuracy, resulting in a loss in log-likelihood of at most $(n/2)\sum_j \lambda_j \Delta_j^2$. We set $\Delta_j = 1/\sqrt{n\lambda_j}$ or equivalently $k_j = (1/2)\log n\lambda_j$. The number of these small cells is $\prod_{j=1}^k (1/\Delta_j) = n^{k/2}(\det G)^{1/2}$. To give the index of the cell containing θ^* (and hence to describe a simple parameter vector $\hat{\theta}$ in this cell) requires at most $\log n^{k/2}(\det G)^{1/2} + 2 \log\log n^{k/2}(\det G)^{1/2}$ bits. Consequently the length of the description of both $P_{\hat{\theta}}$ and $\mathbf{X}_n$ is within a constant of $L_n + 2 \log\log n$. This completes the proof of Theorem 4.4.

4.3 Non-parametric Laws

In this section we establish convergence of the logically smooth estimate $\hat{P}_n$ to nearly any true distribution P^*. The emphasis will be on the iid case with density function p^* and on obtaining consistency as strong or nearly as strong as convergence of the variation distance. (The variation distance is the L_1 distance between the densities.) Our starting point is the work of L. Schwartz (1965). She showed that if there exist uniformly consistent tests, then Bayes rules are consistent. Our essential contribution is the discovery of some uniformly consistent tests in the non-parametric context.

Recall the set-up of section (3.2). Stationary and ergodic distributions P are defined on a space Ω of infinite sequences $(x_1,x_2,...)$ with coordinates in a standard space X. If P is a distribution on Ω, we denote by P_n the distribution for $(X_1,X_2,\ldots,X_n)$. The random data $\mathbf{X}_n$ takes values in a sequence of refining partitions π_n which generates Ω. In particular we assume that the partition π_n consists of cylinder sets with the first n coordinates in the product of refining partitions τ_n, where the sequence τ_n generates X. Thus if $(X_1,X_2,...)$ is a random sequence in Ω, then at time n, the observed data are $\mathbf{X}_n = (X_1^{(n)},X_2^{(n)},\ldots,X_n^{(n)})$ where $X_i^{(n)}$ is the cell in τ_n which contains X_i.

In section 3.2 we examined the redundancy $\lceil \log 1/Q(\mathbf{X}_n) \rceil - \log 1/P^*(\mathbf{X}_n)$ of the Shannon code based on $Q_n = \int_N P_n\, v(dP)$ where v is a prior on a measurable space of distributions and $N = \{P{:}D^\infty(P^*||P) < \epsilon$, for P iid or Markov$\}$. We showed that if $v(N) > 0$ then the redundancy is of order $o(n)$. Thus if if N_n is any sequence of (possibly shrinking) neighborhoods which contain N, then the redundancy of the Shannon code based on the mixture $Q_n = \int_{N_n} P_n\, v(dP)$ is also of order $o(n)$. Equivalently,

$$\int_{N_n} P(\mathbf{X}_n) v(dP) > P^*(\mathbf{X}_n) 2^{-o(n)} \text{ almost surely.} \tag{4.16}$$

On the other hand, we would expect to have greater redundancy if the Shannon code were based on $Q_n = \int_{B_n} P_n\, v(dP)$ where $B_n = N_n^c$ is the set of distributions outside the neighborhood N_n.

We show that a sufficient condition for large redundancy is the existence of a uniformly consistent sequence of tests of the hypothesis $P = P^*$ versus the composite hypothesis $P \in B_n$. We consider tests of the form: decide $P = P^*$ if and only if $\mathbf{X}_n$ is in an acceptance region A_n (where A_n is a set in the field generated by π_n). The probability of error is $P^*(A_n^c)$ if $P = P^*$ or $P(A_n)$ if P is in B_n. Randomized tests are also permitted. In this case the acceptance region $A_n(S_n)$ depends on a random variable S_n which is independent of $\mathbf{X}_n$ (and independent of P). Let P_{S_n} denote the distribution of S_n. The (average) probability of error is then given by $P\{\mathbf{X}_n \in A_n(S_n)\} = \int P(A_n(s))P_{S_n}(ds)$ for P in B_n and $P^*\{\mathbf{X}_n \in A_n^c(S_n)\} = \int P^*(A_n^c(s))P_{S_n}(ds)$ for $P = P^*$. Note that randomized tests include non-randomized tests as a degenerate case.

A sequence of tests is said to be consistent for P^* versus B_n if the probability of error tends to zero for each P. It is *uniformly* consistent if there exists $r_n \to \infty$ such that

$$P^*\{\mathbf{X}_n \in A_n^c(S_n)\} \leq 2^{-r_n} \quad \text{and} \quad \sup_{P \in B_n} P\{\mathbf{X}_n \in A_n(S_n)\} \leq 2^{-r_n} \tag{4.17}$$

The sequence of tests is uniformly *exponentially* consistent if (4.17) holds for $r_n = n\epsilon$ for some $\epsilon > 0$. We remark that in the iid case with fixed $B_n = B$ (not depending on n), L. Schwartz (1965) showed that if a test is uniformly consistent then it is uniformly exponentially consistent. In that context, she also obtained results analogous to the following Lemma.

Lemma 4.3: On uniformly consistent tests and redundancy

If there exists a uniformly consistent test of $P = P^*$ *versus* $P \in B_n$, *then the redundancy* $R_n(\mathbf{X}_n) = \lceil \log 1/Q_n(\mathbf{X}_n) \rceil - \log 1/P^*(\mathbf{X}_n)$ *of the Shannon code based on* $Q_n = \int_{B_n} P_n v(dP)$ *satisfies* $R_n(\mathbf{X}_n) \geq r_n - c_n$ *except for* $\mathbf{X}_n$ *in a set of* P^* *probability less than* 2^{-c_n+1} *for any* c_n *less than* r_n. *In particular, if the test is uniformly exponentially consistent then the per sample redundancy is almost surely bounded away from zero, i.e., for some* $\epsilon>0$, $R_n(\mathbf{X}_n)/n > \epsilon$ *for all large n. Equivalently,*

$$\int_{B_n} P(\mathbf{X}_n)\, v(dP) \leq P^*(\mathbf{X}_n)\, 2^{-n\epsilon} \tag{4.18}$$

for all large n, with P^* *probability one.*

Proof: Consider the event $\{R_n(\mathbf{X}_n) \leq r_n - c_n\}$ that the redundancy is not greater than $r_n - c_n$. Note that this event is contained in $E_n = \{P^*(\mathbf{X}_n) \leq Q_n(\mathbf{X}_n) 2^{r_n - c_n}\}$. Clearly, the set E_n has the property that when it is intersected with any A_n (in the field generated by π_n), the probability satisfies $P^*(E_n \cap A_n) \leq Q_n(E_n \cap A_n) 2^{r_n - c_n}$. We find that

$$\begin{aligned} P^*\{R_n(\mathbf{X}_n) \leq r_n - c_n\} &\leq P^*(E_n) \\ &\leq P^*(E_n \cap A_n) + P^*(A_n^c) \\ &\leq Q_n(A_n) 2^{r_n - c_n} + P^*(A_n^c) \\ &= \int_{B_n} P(A_n) v(dP)\, 2^{r_n - c_n} + P^*(A_n^c). \end{aligned}$$

Conditioning on $S_n =$ s and applying this inequality with $A_n(s)$ yields

$$P^*\{R_n(\mathbf{X}_n) \leq r_n - c_n\} \leq \int_{B_n} P(A_n(s)) v(dP)\, 2^{r_n - c_n} + P^*(A_n^c(s)).$$

Note that the left side does not depend on s. Averaging the right side with respect to P_{S_n} yields

$$P^*\{R_n(\mathbf{X}_n) \leq r_n - c_n\} \leq \int_{B_n} P\{\mathbf{X}_n \in A_n(S_n)\} v(dP)\, 2^{r_n - c_n} + P^*\{\mathbf{X}_n \in A_n^c(S_n)\}$$

$$\leq 2^{-r_n}\, 2^{r_n - c_n} + 2^{-r_n}$$

$$< \; 2^{-c_n} + 2^{-c_n}.$$

Thus $R_n(\mathbf{X}_n)$ is greater than $r_n - c_n$ except in a set of probability less than 2^{-c_n+1}. In particular if the test is uniformly exponentially consistent, then $r_n = 2n\epsilon$ for some $\epsilon > 0$. For $c_n = n\epsilon$ we find that (4.18) holds except in a set of probability less than $2^{-n\epsilon}$. Hence (4.18) holds eventually almost surely by the Borel-Cantelli lemma. This completes the proof of Lemma 4.3.

Remarks: Thus if there exists a uniformly exponentially consistent sequence of tests of P^* versus the complement of (possibly shrinking) neighborhoods N_n, then the codelengths based on mixtures of distributions in N_n are asymptotically less than the codelengths based on mixtures in $B_n = N_n^c$. Moreover, the difference $\log 1/\int_{N_n} P(\mathbf{X}_n) v(dP) - \log 1/\int_{N_n^c} P(\mathbf{X}_n) v(dP)$ tends to infinity as $n \to \infty$, with probability one. Indeed, from (4.16) and (4.18) we find that the ratio $\int_{N_n} P(\mathbf{X}_n)\, dv / \int_{N_n^c} P(\mathbf{X}_n)\, dv$ tends to infinity. Therefore, the posterior distribution is asymptotically concentrated on the neighborhoods N_n, in the sense that

$$v(N_n|\mathbf{X}_n) = \frac{\int_{N_n} P(\mathbf{X}_n) v(dP)}{\int P(\mathbf{X}_n) v(dP)} \to 1 \quad P^* \text{ almost surely} \tag{4.19}$$

Likewise for a countable prior which assigns mass $v(P) = 2^{-L(P)}$ to a list $\Gamma = \{P\}$ of distributions, we have that all two-stage description lengths $L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil$ based on distributions P in $B_n = N_n^c$ are uniformly greater than $\log 1/P^*(\mathbf{X}_n) + n\epsilon$ for all large n. Indeed, these two-stage description lengths are all greater than $\log 1/\sum_{P \in B_n} P(\mathbf{X}_n) v(P)$ (since a sum is greater than its maximal term), which by Lemma (4.3) is greater than $\log 1/P^*(\mathbf{X}_n) + n\epsilon$. On the other hand if $\inf_{P\in N_n} D^{\infty}(P^*\|P) = 0$, then by Theorem 3.3 (Section 3.4), for any $0<\delta<\epsilon$ the minimum two-stage description length is less than $\log 1/P^*(\mathbf{X}_n) + n\delta$ for all large n with probability one. Thus any distribution $\hat{P}_n$ which achieves the minimum two-stage description length will be in the neighborhood N_n for all large n. Note also that any distribution having two-stage description length within $n(\epsilon-\delta)$ of the minimum will also be in N_n. Thus we have established the following Lemma.

Lemma 4.4: Consistent tests imply consistent estimates.

Suppose there exists a uniformly exponentially consistent sequence of tests of the hypothesis P^* versus the composite hypothesis $P \in N_n^c$. Suppose also that P^* is approximated by distributions on the countable list Γ in the relative entropy rate sense. Then there exists $\epsilon>0$ sufficiently small, such that for all large n, if $\hat{P}_n$ is any estimated distribution with two-stage description length within $n\epsilon$ of the minimum $\min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil : P \in \Gamma\}$, *then the estimate $\hat{P}_n$ is in the set N_n.*

Now we restrict attention to the iid case. To avoid unnecessary notation we use P (rather than P_1) to denote a marginal distribution on X as well as to denote the distribution on the sequence space Ω. (The distinction should be clear from the context.) Recall from section 3.2 that the relative entropy between distribution restricted to a countable partition β of X is given by $D_\beta(P\|P^*) = \sum_{A\in\beta} P(A) \log P(A)/P^*(A)$. Similarly the variation distance on β is $\|P^*-P\|_\beta = \sum_{A\in\beta} |P^*(A)-P(A)|$. Csiszár (1967) and Kullback (1967) established the inequality $D \geq d^2/2$ between the relative entropy D and the variation distance d.

Consider the β_n-variation neighborhoods given by $N_n = \{P : \|P^*-P\|_{\beta_n}<\epsilon\}$ where β_n is a partition of the space X. If β_n is a sequence of refining partitions which generates X, then the sequence of neighborhoods N_n shrink to the variation distance neighborhood $\{P: ||P^*-P|| < \epsilon\}$, which in turn includes the set $N = \{P: D(P^*||P) < \epsilon^2/2\}$. Uniformly consistent tests do not exist if we use variation distance or relative entropy neighborhoods. We show that for natural choices of β_n, uniformly consistent tests do exist for P^* versus N_n^c.

Note that there are two sorts of partitions involved in our analysis: the partitions τ_n within which our observations $X_i^{(n)}$ live and the partitions β_n corresponding to the hypothesis tests under consideration. We assume that our observations are sufficiently accurate for the test. Specifically, τ_n is assumed to be a refinement of β_n. If the data partition τ_n is rather course, e.g. if the cell widths w_n satisfy $nw_n \to \infty$, then we are essentially in the discrete case and we may as well set $\beta_n = \tau_n$. However, if the data are observed to high accuracy, e.g. if the cell widths of τ_n satisfy $nw_n \to 0$, then we are essentially in the continuous case. We wonder how refined can be β_n so that we still have consistency.

Given a sequence $\mathbf{X}_n$ in β_n^n, we let $P_{\mathbf{X}_n}$ be the *type* or empirical distribution on β_n, i.e., $P_{\mathbf{X}_n}(a)$ is the relative frequency of occurrence of a in the sequence $(X_1^{(n)},X_2^{(n)},\ldots,X_n^{(n)})$. For a partition with m cells, the number of types is given by $\binom{n+m-1}{m-1}$.

Lemma 4.5: A uniformly consistent test for distributions.

Let β_n be a countable partition of the line into intervals of width w_n such that $nw_n \to \infty$ as $n \to \infty$. Then for any $\epsilon > 0$ there exists a uniformly exponentially consistent test of the hypothesis P^* versus the composite hypothesis N_n^c $= \{P: ||P^*-P||_{\beta_n} \geq \epsilon\}$.

Remarks: A test statistic that works is given by $||P^*-P_{\mathbf{X}_n}||_{\beta_n}$ with the acceptance region $A_n = \{\mathbf{X}_n : ||P^*-P_{\mathbf{X}_n}||_{\beta_n} < \delta\}$ where $0<\delta<\epsilon$. In the proof we will find it convenient to use $||P^*-P_{X_n}||_{\beta_n(C)}$ where $\beta_n(C)$ is a finite partition obtained from β_n by lumping all cells outside a large set C into one cell of low probability.

Proof: We prove a more general result than stated in the Lemma. The space X need not be the real line. The general assumption is that for any $\epsilon>0$ there exists a set C having probability $P^*(C) > 1-\epsilon/4$ such that if $m = m_n$ is the number of cells of β_n which intersect C, then $m/n \to 0$. For the real line we set $C = \{|X| \leq c\}$ with large c, then $m/n = c/(nw_n)$ which tends to zero as $n \to \infty$ provided $nw_n \to \infty$.

Let $\beta_n(C)$ be the partition into $m_n + 1$ cells consisting of the m_n cells which intersect C together with one cell which includes all the rest. The variation distance on β_n satisfies $||P^*-P||_{\beta_n} = 2\sum_{a\in\beta_n}(P^*(a)-P(a))^+$ which is no more than $2\sum_{a\in\beta_n(C)}(P^*(a)-P(a))^+ + 2P^*(C^c)$ which is less than $||P^*-P||_{\beta_n(C)} + \epsilon/2$. Therefore the set of alternatives $N_n^c = \{P: ||P^*-P||_{\beta_n} \geq \epsilon\}$ is contained in the larger set $N_n^c = \{P: ||P^*-P||_{\beta_n(C)} \geq \epsilon/2\}$. Consequently, we may restrict attention to finite partitions with $m_n = o(n)$ cells. We show that for any sequence of partitions β_n of X into $m = m_n$ cells with $m/n \to 0$, and for any $\epsilon>0$ there exists a uniformly exponentially consistent test of P^* versus $N_n^c = \{P: ||P^*-P||_{\beta_n} \geq \epsilon\}$.

For any sequence $\mathbf{X}_n$ in β_n^n and any iid distribution P we have

$$P(\mathbf{X}_n) = 2^{-n(H(P_{X_n}) + D_{\beta_n}(P_{X_n}||P))}$$

where $H(P) = \sum_{a\in\beta_n} P(a) \log 1/P(a)$ is the entropy. This simple fact is readily seen from grouping the product as $P(\mathbf{X}_n) = \prod_{a\in\beta_n} P(a)^{nP_{X_n}(a)}$. A direct consequence is that the number of sequences of a given type P_{X_n} is less than $2^{nH(P_{X_n})}$ (see Csiszár and Korner 1981, Lemma 2.3). Thus we have

$$P(A_n) = \sum_{\mathbf{X}_n \in A_n} 2^{-n(H(P_{X_n}) + D_{\beta_n}(P_{X_n}||P))}$$

and hence

$$P(A_n) \leq \sum_{Q\in A_n'} 2^{-nD_{\beta_n}(Q||P)} \tag{4.20}$$

where A_n' is the set of types P_{X_n} for sequences $\mathbf{X}_n$ in A_n (see Csiszár and Korner 1981, Lemma 2.6). This useful inequality (4.20) holds for any set A_n in the field generated by π_n. In particular for $A_n = \{\mathbf{X}_n : ||P^*-P_{X_n}||_{\beta_n}<\delta\}$, we have $A_n' = \{\text{types } Q : ||P^*-Q||_{\beta_n}<\delta\}$. By the triangle inequality $||Q-P||_{\beta_n} \geq ||P^*-P||_{\beta_n} - ||P^*-Q||_{\beta_n}$ which is greater than $\epsilon-\delta > 0$ for Q in A_n' and P in N_n^c. Therefore for any P in N_n^c we have

$$\begin{aligned} P(A_n) &\leq \sum_{Q\in A_n'} 2^{-nD_{\beta_n}(Q||P)} \\ &\leq \sum_{Q\in A_n'} 2^{-\frac{n}{2}||Q-P||^2_{\beta_n}} \\ &\leq \sum_{Q\in A_n'} 2^{-\frac{n}{2}(\epsilon-\delta)^2} \\ &\leq \binom{n+m}{m} 2^{\frac{-n}{2}(\epsilon-\delta)^2} \end{aligned}$$

Here $m = m_n$ is the number of cells in the partition. The number of types is less than $\binom{n+m}{m} \leq 2^{(n+m)H(m/(n+m))} < 2^{2nH(m/n)}$ where $H(\alpha) = \alpha \log 1/\alpha + (1-\alpha) \log 1/(1-\alpha)$, $0\leq\alpha\leq 1$, is the binary entropy function. We note that $H(\alpha)$ decreases to 0 as $\alpha \rightarrow 0$. Let ϵ' satisfy $0 < \epsilon' < (1/2)(\epsilon-\delta)^2$. Since $m_n/n \rightarrow 0$ we have that $2H(m/n)$ is less than $(\epsilon-\delta)^2 - \epsilon'$ for all large n. Consequently,

$$\sup_{P\in N_n^c} P(A_n) \leq 2^{-n(\frac{1}{2}(\epsilon-\delta)^2-2H(m/n))} \leq 2^{-n\epsilon'}$$

for all large n. Similarly for $0 < \epsilon' < (1/2)\delta^2$, we have

$$P^*(A_n^c) \le \sum_{Q\in A_n^{c\prime}} 2^{-nD_{\beta_n}(Q\|P^*)} \le \binom{n+m}{m} 2^{-n\frac{1}{2}\delta^2} \le 2^{-n\epsilon'}$$

for all large n, since $m_n/n \to 0$. This completes the proof of Lemma 4.5.

Lemma 4.6: A uniformly consistent test for densities.

Let β_n be a countable partition into intervals of width w_n such that w_n is proportional to $1/n$ (or $0< c_1 < nw_n < c_2$ for all large n). Suppose P^* has a bounded density function $p^*(x)$ with finite mean $\int xp^*(x)dx$. Then there exists a uniformly exponentially consistent test of the hypothesis P^* versus $N_n^c = \{P : \|P^*-P\|_{\beta_n} \ge \epsilon\}$.

Roughly the test statistic that works is of the form $\sum_{a\in\beta_n}(e^{nP^*(a)}I_{\{a \text{ empty}\}} - 1)$, except that we truncate to a finite partition and randomize the number of samples that are used in the test. Note that data does not accumulate in cells of width $1/n$, so the ordinary tests involving the sample distribution do not work here.

Proof: Let β_n' be the set of all cells in β_n which intersect $\{|X|\le c\}$. We let R denote the union of the remaining cells. The constant c will be specified later. The number of cells in β_n' is no more than a multiple of n, indeed $m_n \le (2c/c_1)n$. Let c^* be the bound on the density function p^*. Then we have that $P^*(a)$ is less than $c^*w_n < c^*c_2/n$ for all cells a in β_n' .

Let S_n be a Poisson$(n\lambda)$ random variable with $0<\lambda<1$ and let the random variable $N(a)$ be the number of occurrences of the symbol a in the sequence of observations indexed from 1 to S_n' where $S_n' = \min\{S_n,n\}$. The test statistic is

$$T_n' = \sum_{a\in\beta_n'} \left(e^{n\lambda P^*(a)}I_{\{N(a)=0\}} - 1 \right).$$

and the acceptance regions are of the form $A_n(S_n) = \{\mathbf{X}_n : T_n' < n\delta\}$.

First we argue that the Poisson($n\lambda$) variable S_n is less than n except in a set with exponentially small probability. Indeed $P\{S_n \geq n\} = P\{\exp(tS_n) \geq \exp(tn)\}$ which by Markov's inequality is less than $\exp(-tn)\, E\exp(tS_n)$. But the moment generating function of the Poisson is known to be $E\exp(tS_n) = \exp(n\lambda(e^t-1))$. Thus $P\{S_n \geq n\} \leq \exp(-n(t-(e^t-1)\lambda))$. The best t is seen to be $t = \ln 1/\lambda$, in which case the bound becomes $e^{-n(\lambda-1-\ln\lambda)} = e^{-nr}$ where $r = \lambda-1-\ln\lambda$ is strictly positive for $0<\lambda<1$.

Let T_n' and T_n denote the test statistic with and without the truncation of S_n at n. We have

$$P^*\{T_n' \geq n\delta\} \leq P^*\{T_n \geq n\delta\} + e^{-nr}$$

and

$$P\{T_n' < n\delta\} \leq P\{T_n < n\delta\} + e^{-nr}$$

So it suffices to examine T_n with no truncation of S_n.

With respect to P^*, the cell counts $N(a)$ are independent Poisson random variables with parameter $n\lambda P^*(a)$. Thus $P^*\{N(a) = 0\} = e^{-n\lambda P^*(a)}$. Let E^* and E denote expectation with respect to P^* and P respectively. Then

$$E^*\, T_n = E^* \sum_{a\in\beta_n'} \left(e^{n\lambda P^*(a)} I_{\{N(a)=0\}} - 1 \right)$$

$$= \sum \left(e^{n\lambda P^*(a)} e^{-n\lambda P^*(a)} - 1 \right)$$

$$= 0.$$

Similarly, with respect to P the cell counts $N(a)$ are independent Poisson($n\lambda P(a)$) random variables. Thus

$$E\, T_n = \sum_{a\in\beta_n'} \left(e^{n\lambda(P^*(a)-P(a))} - 1 \right). \tag{4.21}$$

We now show that $E\,T_n$ exceeds a small multiple of n. Consider the sets $G_n = \{a \in \beta_n' \ : P^*(a) > P(a)\}$ and $F_n = \{a \in \beta_n' \ : P^*(a) \leq P(a)\}$. Note that $||P^*\text{-}P||_{\beta_n}$ may be expressed as $2\sum_{a\in\beta_n}(P^*(a) - P(a))^+$ which is no more than $2(P^*(G_n)\text{-}P(G_n)) + P^*(R)$. Writing (4.21) as a sum over G_n plus a sum over F_n and applying Jensen's inequality, we obtain the following

$$E\,T_n = \sum_{a\in G_n}\Big(\exp(n\lambda(P^*(a)\text{-}P(a))) - 1\Big) + \sum_{a\in F_n}\Big(\exp(n\lambda(P^*(a)\text{-}P(a))) - 1\Big)$$

$$\geq |G_n|\left(\exp(\frac{n\lambda}{|G_n|}\sum_{a\in G_n}(P^*(a)\text{-}P(a))) - 1\right) + |F_n|\left(\exp(\frac{n\lambda}{|F_n|}\sum_{a\in F_n}(P^*(a)\text{-}P(a))) - 1\right)$$

$$\geq |G_n|\left(\exp\frac{n\lambda}{|G_n|}(\frac{1}{2}||P^*\text{-}P||_{\beta_n}\text{-}P^*(R)) - 1\right) + |F_n|\left(\exp(\frac{-n\lambda}{|F_n|}\frac{1}{2}||P^*\text{-}P||_{\beta_n}) - 1\right)$$

Here $|G_n|$ and $|F_n|$ denotes the cardinality of the sets G_n and F_n. Now using the familiar inequalities $e^x - 1 \geq x + x^2/2$ and $e^{-x} - 1 \geq -x$, we obtain

$$E\,T_n \geq \left(n\lambda(\frac{1}{2}||P^*\text{-}P||_{\beta_n} - P^*(R)) + \frac{1}{2}\frac{(n\lambda)^2}{|G_n|}(\frac{1}{2}||P^*\text{-}P||_{\beta_n} - P^*(R))^2\right) - \left(n\lambda\frac{1}{2}||P^*\text{-}P||_{\beta_n}\right)$$

$$= \frac{1}{2}\frac{(n\lambda)^2}{|G_n|}(\frac{1}{2}||P^*\text{-}P||_{\beta_n} - P^*(R))^2 \ - n\lambda P^*(R)$$

$$\geq \frac{1}{2}\frac{(n\lambda)^2}{m_n}(\frac{1}{2}||P^*\text{-}P||_{\beta_n} - P^*(R))^2 \ - n\lambda P^*(R)$$

$$\geq n\lambda\left(\frac{c_1\lambda}{4c}(\frac{\epsilon}{2})^2 - 2P^*(R)\right)$$

for $||P^* - P||_{\beta_n} \geqslant \epsilon$. Recall that $E^*|X|$ is finite by assumption. By Markov's inequality $cP^*(R) \leqslant cP^*\{|X|>c\} \leqslant E^*|X|I_{\{|X|>c\}}$ which is made less than $\epsilon' = c_1\lambda\epsilon^2/64$ by choosing c sufficiently large. Consequently for all P in N_n^c we have

$$E\,T_n \geq 2n\lambda\epsilon'$$

Finally we invoke Hoeffding's inequality (Hoeffding 1963, Theorem 2) for sums of bounded independent random variables. If V_i are independent random variables with bounds $\alpha_i \leq V_i \leq \gamma_i$, then for $S = \sum_i V_i$ and $t > 0$

$$P\{S - ES \geq t\} \leq \exp\left(\frac{-2t^2}{\sum_i (\gamma_i - \alpha_i)^2} \right)$$

where i ranges over a finite index set. Applying Hoeffding's inequality to the sum $T_n = \hat{\sum}_{a \in \beta_n'} (e^{nP^*(a)} I_{\{N(a) = 0\}} - 1)$ yields for $0 < \delta < 2\lambda\epsilon'$,

$$P\{T_n < n\delta\} = P\{-(T_n - E\,T_n) > E\,T_n - n\,\delta\}$$

$$\leq P\{-(T_n - E\,T_n) \geq n\delta'\}$$

$$\leq \exp\left(\frac{-2(n\delta')^2}{\sum_a e^{2nP^*(a)}} \right)$$

$$\leq e^{-n\alpha}$$

uniformly for all P in N_n^c, where $\delta' = 2\lambda\epsilon' - \delta$ and $\alpha \leq (c_1/c)e^{-2c^*c_2}(\delta')^2$. Similarly for the P^* probability of error,

$$P^*\{T_n \geq n\delta\} \leq \exp\left(\frac{-2(n\delta)^2}{\sum_a e^{2nP^*(a)}} \right)$$

$$\leq e^{-n\alpha}$$

where $\alpha \leq (c_1/c)e^{-2c^*c_2}\delta^2$. We conclude that T_n' is a uniformly exponentially consistent test as desired. This completes the proof of Lemma 4.6.

The consequences of Lemmas 4.4, 4.5, and 4.6 for consistency of estimates are summarized as follows. We assume that $X_1, X_2, \ldots$ are independent with distribution P^*. The data $\mathbf{X}_n$ is $(X_1^{(n)}, \ldots, X_n^{(n)})$, where $X_i^{(n)}$ is the cell in a partition τ_n

which contains X_i. For the next two theorems the partition β_n has cell widths w_n which satisfy $nw_n \to \infty$. Or if P^* has a bounded density with finite mean, the cell widths may satisfy $c_1/n < w_n < c_2/n$ for some constants $c_2 > c_1 > 0$. We assume that τ_n refines β_n. The cell widths of τ_n may be arbitrarily small. Recall that if v is a prior on a measurable space of distributions P then the posterior distribution given data $\mathbf{X}_n$ is $v(B|\mathbf{X}_n) = \int_B P(\mathbf{X}_n) v(dP) / \int P(\mathbf{X}_n) v(dP)$.

Theorem 4.5: Bayes consistency in n^{-1}–variation distance.

Let v be a prior on iid distributions which assigns strictly positive mass to the relative entropy neighborhoods $\{P : D(P^*||P) \leq \epsilon\}$. Then the posterior distribution given data $\mathbf{X}_n$ is asymptotically concentrated on β_n–variation neighborhoods $N_n = \{P : ||P^*-P||_{\beta_n} < \epsilon\}$ in the sense that

$$v(N_n|\mathbf{X}_n) \to 1 \quad P^* \text{ almost surely.}$$

Moreover, if v is a countable prior and if $\hat{P}_n$ maximizes the posterior probability $v(P|\mathbf{X}_n)$ then

$$||P^*-\hat{P}_n||_{\beta_n} \to 0 \quad P^* \text{ almost surely.}$$

Theorem 4.6: Non-parametric consistency of logical smoothing.

If the density p^* has finite relative entropy from some computable density (in particular if $p^*(x)$ is less than some computable integrable function) and if $\hat{P}_n$ is any distribution for which the two-stage description is less than $\min\{L(P) + \lceil \log 1/P(\mathbf{X}_n) \rceil\} + o(n)$, *then*

$$||P^*-\hat{P}_n||_{\beta_n} \to 0 \quad P^* \text{ almost surely.}$$

Moreover, if $\hat{P}_n$ is modified to be flat (conditionally uniform) in each cell of β_n while preserving the probabilities of these cells, then the corresponding density $\hat{p}_n$ is consistent in L_1 distance

$$\int |p^*(x)-\hat{p}_n(x)|\,dx \to 0 \quad P^* \text{ almost surely.}$$

Proof of Theorems 4.5 and 4.6: Uniformly exponentially consistent tests for P^* versus $\{P : \|P^*-P\|_{\beta_n} \geqslant \epsilon\}$ are demonstrated in Lemmas 4.5 and 4.6. Recalling the remarks preceding Lemma 4.4, it follows that $\nu(N_n \mid X_n) \to 1$ and $\hat{P}_n$ is consistent in β_n-variation. Moreover, by the triangle inequality

$$\int |p^*(x)-\hat{p}_n(x)|\,dx \leqslant \|P^*-\hat{P}_n\|_{\beta_n} + \sum_{A \in \beta_n} \int_A |p^*(x)-(P^*(A)/\mu(A))|\,dx$$

The first term in the bound we have shown to converge. The second term is a "bias" associated with probability histograms and it tends to zero as shown in Abou-Jaoude (1976). (Note that here we may have less bias since nw_n may be bounded, whereas for ordinary histograms consistency requires $nw_n \to \infty$). Therefore, the logically smooth density estimate $\hat{p}_n$ is consistent in L_1 distance. This completes the proof.

Remarks: The almost-sure convergence of $\nu(N \mid X_n)$ to one for any fixed partition β establishes weak consistency (the sequence of posterior distributions converges weakly to point mass at P^*). This generalizes the result of Freedman (1963) who showed weak consistency under the additional assumptions that the data partition $\tau_n = \tau$ is fixed (i.e., the discrete case) and that the entropy $H = \sum_{A \in \tau} P^*(A) \log 1/P^*(A)$ is finite. These assumptions are now seen to be extraneous. Moreover, Theorem 4.5 shows that stronger forms of consistency also hold. L. Schwartz (1965) is often credited with extending Freedman's result to the non-discrete case. However, she assumed but did not demonstrate the existence of a uniformly consistent test.

Why does our method obtain convergence in β_n—variation using cells of width w_n proportional to $1/n$, but not using narrower cells? Surprisingly, if $\liminf nw_n = 0$, there does not exist a uniformly exponentially consistent test for

P^* versus $\{P : \| P^* - P \|_{\beta_n} > \epsilon\}$. This is shown by constructing a Bayes estimate which is inconsistent in β_n–variation (even though the relative entropy neighborhoods are assigned positive mass). Indeed, the prior is constructed to assigns mass to countably many distributions P_n, $Q_{k,n}$, $k=1,2,...,2^n$, $n=1,2,...$. The distributions labeled P_n tend to P^* in the relative entropy sense whereas the others labeled $Q_{k,n}$ are highly oscillatory distributions which are far from P^*. Indeed the foil distributions $Q_{k,n}$ are randomly constructed to assign either $Q_{k,n}(a)=0$ or $Q_{k,n}(a)=2P^*(a)$ for each $a \in \beta_n$, according to the outcome of independent Bernoulli(1/2) coin-flips. The prior assign mass $2^{-n}/n^2$ to each $Q_{k,n}$ for $k=1,2,...,2^n$. But the prior on the good distributions P_n is set to drop off exponentially fast. The proof of inconsistency involves random coding techniques borrowed from Shannon's channel coding theory. The details will appear in a later paper.

There is also a description length justification for quantization using cells of width equal to $1/n$. For smooth densities the extra bits of accuracy beyond $\log n$ are shown to be asymptotically incompressible Bernoulli(1/2) random variables. Therefore, we expect to have nearly minimal two-stage description length by using a distribution $\hat{P}_n$ which is conditionally uniform in each cell of width $1/n$.

Let $X_1, X_2, ...$ be independent random variables with density function $p(x)$. Let $\mathbf{X}_n = (X_1, ..., X_n)$ and let $\mathbf{X}_n^b = (X_1^b, ..., X_n^b)$ where X_i^b is the dyadic interval of width $w = w_n = 2^{-b_n}$ which contains X.

Lemma 4.7: Approximating probabilities with densities.

Suppose that the density function $p(x)$ is continuously differentiable and that for any $\delta > 0$, there exists $g(x)$ with finite $Eg(X)$ such that $|p'(\tilde{x})/p(\tilde{x})|$ and $|p'(\tilde{x})/p(x)|$ are less than $g(x)$ for all $|\tilde{x}-x| < \delta$. Let $c = \int |p'(x)| dx$. If $w_n \to 0$, then we have

$$\limsup_{n \to \infty} \frac{1}{n w_n} \left| \ln \frac{P(\mathbf{X}_n^b)}{w_n^n p(\mathbf{X}_n)} \right| \leq c/2. \tag{4.22}$$

with probability one.

Remarks: Thus if nw_n is bounded by ϵ then $P(X_n^b)$ and $w^n p(X_n)$ agree to within a multiplicative constant $e^{\pm\epsilon c/2}$. Moreover, the Shannon codelength based on P given by $\lceil \log 1/P(X_n^b) \rceil$ is within $1 + c\,\epsilon$ bits of the density approximation $\log 1/p(X_n) + nb_n$.

Consider the conditional distribution $P(X_n^b \mid X_n^l) = P(X_n^b)/P(X_n^l)$ where $b_n > l_n$. This is the distribution for the $n(b_n - l_n)$ remaining bits past X_n^l. Note that the same $p(X_n)$ appears in the density approximation to both $P(X_n^b)$ and $P(X_n^l)$, so the density term vanishes when approximating the ratio. From (4.22) we obtain

$$\limsup \frac{1}{nh_n} \mid \log 1/P(X_n^b \mid X_n^l) - n(b_n - l_n) \mid \leqslant c \log e \tag{4.23}$$

where $h_n = 2^{-l_n}$. If nh_n is less than ϵ, then (4.23) shows that the brute force description using $n(b_n - l_n)$ bits to describe X_n^b given X_n^l is within $1 + c\,\epsilon \log e$ bits of the optimum conditional codelength $\lceil \log 1/P(X_n^b \mid X_n^l) \rceil$.

Proof of Lemma 4.7:

Given any $\epsilon > 0$, let δ be sufficiently small that there is a function $g(x)$ which dominates both $\mid p'(\tilde{x})/p(\tilde{x}) \mid$ and $\mid p'(\tilde{x})/p(x) \mid$ and which has expectation $Eg(X) \leqslant E \mid p'(X)/p(X) \mid + \epsilon = c + \epsilon$. Let X^b be the dyadic interval of width 2^{-b} containing X. Choose $w = 2^{-b}$ to be less than δ. Then by Jensen's inequality and the mean value theorem we obtain

$$\ln \frac{P(X^b)}{wp(X)} = \ln \frac{1}{w} \int_{X^b} \frac{p(x)}{p(X)} dx$$

$$\geqslant \frac{1}{w} \int_{X^b} \ln \frac{p(x)}{p(X)}\, dx$$

$$\geqslant -\frac{1}{w} \int_{X^b} \mid x - X \mid dx \sup_{\mid \tilde{x} - X \mid \leqslant \delta} \mid \frac{\partial}{\partial x} \ln p(\tilde{x}) \mid$$

$$\geqslant -\frac{1}{w}\int_0^w x dx \ g(X)$$

$$= -\frac{w}{2} g(X).$$

On the otherhand, from $\ln x \leqslant x-1$ and the mean value theorem we obtain

$$\ln \frac{P(X^b)}{w\ p(X)} \leqslant \frac{P(X^b)-w\ p(X)}{w\ p(X)}$$

$$= \frac{1}{w}\int_{X^b} \frac{p(x)-p(X)}{p(X)} dx$$

$$\leqslant \frac{1}{w}\int_{X^b} |x-X|\,dx \quad \sup_{|\tilde{x}-X|\leqslant\delta} \left|\frac{p'(\tilde{x})}{p(X)}\right|$$

$$\leqslant \frac{w}{2} g(X).$$

Therefore, assuming that $w_n = 2^{-b_n}$ tends to zero, we have

$$\limsup_{n\to\infty} \frac{1}{nw_n} \left| \ln \frac{P(\mathrm{X}_n^b)}{w_n^n p(\mathrm{X}_n)} \right| = \limsup \frac{1}{n}\sum_{i=1}^n \frac{1}{w_n} \left| \ln \frac{P(X_i^b)}{w_n^n p(X_i)} \right|$$

$$\leqslant \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \frac{1}{2} g(X_i)$$

$$= \frac{1}{2}E(g(X)) \leqslant \frac{1}{2}(c+\epsilon).$$

Letting $\epsilon \rightarrow 0$ completes the proof of Lemma 4.7.

Finally we examine some density estimation schemes which are related to logical smoothing. Let $\Gamma = \{P_k : k=1,2,...\}$ be a countable list of distributions on the measurable space X. Each distribution is assumed to have a density function p_k with respect to some sigma-finite measure μ. The joint density functions are $p_k(\mathrm{x}_n) = \prod_{i=1}^n p_k(x_i)$ for $\mathrm{x}_n = (x_1,x_2,...,x_n)$. Let the data $X_1,X_2,...$ be independent random variables with density p^*. This true density p^* need not be on the list Γ. However, we do assume that p^* may be approximated in the relative entropy sense by densities in Γ (recall section 3.4).

Theorem 4.7: The consistency of Cover's density estimator

Let $\hat{p}$ *be a density estimate which achieves* $\min\{\log 1/p_k(X_n) : k \leqslant 2^{c_n}\}$ *where* $c_n \to \infty$ *and* $c_n/n \to 0$ *as* $n \to \infty$. *Then the* L_1 *distance* $\int |p^*(x)-\hat{p}(x)|\mu(dx)$ *tends to zero with* P^* *probability one.*

Remarks: Thus if we restrict attention to densities p with complexity $L(p)$ less than c_n (as suggested in Cover 1972), then the sequence of densities which minimize $\log 1/p(X_n)$ is consistent in L_1 distance (since the number programs with length less than c_n is less than 2^{c_n}). Moreover, the proof of Theorem 4.7 is easily modified to show that any sequence of densities $\hat{p}_n$ for which $\log 1/\hat{p}_n(X_n)$ is within $o(n)$ of $\min\{\log 1/p(X_n) : L(p) < c_n\}$ is also consistent in L_1 distance. In particular, the logically smooth density estimate which minimizes $L(p) + \log 1/p(X_n)$ subject to $L(p)<c_n$ is consistent in L_1 distance.

Proof of Theorem 4.7: The L_1 distance $\int |p-q|$ is equivalent to the Hellinger distance defined by $d^2(p,q) = \int(\sqrt{p}-\sqrt{q})^2$. Indeed, $d^2(p,q) \leqslant \int |p-q| \leqslant 2d(p,q)$ (see Kraft 1955 or Pittman 1979, p.7). Thus it is equivalent to prove convergence in Hellinger distance.

Let $\delta>0$ be given and set $0<\epsilon<\delta$. The assumption that p^* is approximated in the relative entropy sense by densities on the list, ensures that $\hat{p}(X_n)$ exceeds $p^*(X_n)e^{-n\epsilon}$ for all large n with probability one. We show that $p^*(X_n)e^{-n\epsilon}$ exceeds $\max\{p_k(X_n) : k<2^{c_n} \text{ and } d^2(p^*,p_k)\geqslant\delta\}$ except in a set A_n with exponentially small probability. It then follows by the Borel-Cantelli Lemma that $d^2(p^*,\hat{p})<\delta$ for all large n with probability one. Indeed, by the union of events bound we have

$$P^*(A_n) \leqslant \sum P^*\{\, p^*(\mathrm{X}_n) \leqslant e^{n\epsilon} p_k(\mathrm{X}_n)\} \qquad (4.24)$$

where the sum is over all $k < 2^{c_n}$ for which $d^2(p^*, p_k) \geqslant \delta$. Applying standard bounds (as in Chernoff 1952) to the terms in the above sum yield

$$P^*\{\, p^*(\mathrm{X}_n) \leqslant e^{n\epsilon} p_k(\mathrm{X}_n)\} \leqslant e^{n\epsilon/2} \int (p_k(\mathrm{x}_n) p^*(\mathrm{x}_n))^{1/2}\, d\,\mathrm{x}_n$$

$$= e^{n\epsilon/2} \left[\int (p_k(x) p^*(x))^{1/2} dx \right]^n$$

$$\leqslant e^{-n(\delta-\epsilon)/2}. \qquad (4.25)$$

The last inequality follows from $\int (p_k p^*)^{1/2} = 1 - (1/2) d^2(p^*, p_k) \leqslant e^{-(1/2)d^2}$ $\leqslant e^{-\delta/2}$. Summing over k as in (4.24) yields

$$P^*(A_n) \leqslant 2^{c_n} e^{-n(\delta-\epsilon)/2}$$

which tends to zero exponentially fast since $c_n / n \rightarrow 0$. Consequentially, $d^2(p^*, \hat{p}) < \delta$ for all large n with probability one. Thus Cover's density estimate is consistent in L_1 distance. This completes the proof of Theorem 4.7.

The same proof technique also yields a result on the consistency of Bayes estimates with a countable prior. Again we assume that the true density p^* is approximated in the relative entropy sense by densities in the countable set $\{p_k\}$.

Theorem 4.8: On Bayes consistency for root summable priors

Let $\hat{p}$ *be a density estimate which achieves* $\max_k v_k p_k(\mathrm{X}_n)$ *or equivalently* $\min\{\log 1/v_k + \log 1/p_k(\mathrm{X}_n)\}$ *where* $v_k > 0$ *and* $\sum_k (v_k)^\alpha < \infty$ *for some* $0 < \alpha < 1$. *Then*

$$\lim_{n \rightarrow \infty} \int | p^*(x) - \hat{p}(x) | \, \mu(dx) = 0 \quad P^* \text{ almost surely.}$$

In particular, suppose the density estimate $\hat{p}_n$ is defined to achieve $\min\{cL(P) + \log 1/p(\mathrm{X}_n)\}$ where $c > 1$ and the minimum is over computable iid distributions with densities. Note that $2^{-cL(P)}$ is root summable with $\alpha=1/c$. Thus the sequence of modified logically smooth density estimators $\hat{p}_n$ is consistent in L_1 distance.

Proof of Theorem 4.8: As before, we show convergence in the Hellinger distance which is equivalent to convergence in L_1. Without loss of generality assume that $1/2\leqslant\alpha<1$. Given any $\delta>0$ set $0<\epsilon<\delta(1-\alpha)/\alpha$. Again, the assumption that p^* is approximated in the relative entropy sense ensures that $\max_k\, v_k p_k(X_n)$ exceeds $p^*(X_n)e^{-n\epsilon}$ for all large n with probability one. So it remains to show that $p^*(X_n)e^{-n\epsilon}$ exceeds $\max\{v_k p_k(X_n) : d^2(p^*,p_k)\geqslant\delta\}$ except in a set with exponentially small probability. Using the union of events bound as in (4.24) it is enough to show that the following sum is exponentially small

$$\sum_k P^*\{\, p^*(X_n) \leqslant e^{n\epsilon} v_k p_k(X_n)\} \tag{4.26}$$

where the sum is over all k such that $d^2(p^*,p_k)\geqslant\delta$. As in (4.25) these terms are not greater than $e^a = v_k^{1/2}e^{-n(\delta-\epsilon)/2}$. By Markov's inequality they are also not greater than $e^b = v_k e^{n\epsilon}$. Now $\min\{a,b\} \leqslant \beta a+(1-\beta)b$ for $0<\beta<1$. In particular, choosing $\beta=2\alpha-1$ yields

$$\begin{aligned}
P^*\{p^*(X_n) \leqslant e^{n\epsilon} v_k p_k(X_n)\} &\leqslant \min\{\, v_k^{1/2}\, e^{-n(\delta-\epsilon)/2},\; v_k e^{n\epsilon}\} \\
&= \min\{e^a, e^b\} \\
&\leqslant e^{\beta a+(1-\beta)b} \\
&= v_k^{(1+\beta)/2}\, e^{\beta n\epsilon}\, e^{-n(1-\beta)(\delta-\epsilon)/2} \\
&= v_k^{\alpha}\, e^{-n[(1-\alpha)\delta-\alpha\epsilon]}.
\end{aligned}$$

Summing over k we have that expression (4.26) is less than $ce^{-n[(1-\alpha)\delta-\alpha\epsilon]}$ which is exponentially small (here $c = \sum_k v_k^{\alpha}$). Consequently, the Bayes density estimate $\hat{p}$ is consistent in L_1 distance. This completes the proof of Theorem 4.8.

Conclusions

Inference and data compression are intimately related. The common goal is to find a sufficient summary of the data. Exact descriptions of data have informative and uninformative parts. For stochastic data, we have shown that there is no loss in assuming that the informative part is the description of a probability distribution. The uninformative part is then the best code for the data given the distribution (namely, the Shannon code with length equal to minus the logarithm of the likelihood). Thus the goal of data summarization leads to logical statistical inference. The estimates of probability distributions and densities which provide the best summaries of data are those for which the total description length is minimized. We have shown that these estimates consistently infer the true laws.

Some suggestions for additional investigation are to determine whether the mode of convergence can be refined in the non-parametric case, to determine rates of convergence, to investigate logically smooth regression, and to continue the search for feasible estimators achieving nearly minimal description length.

References

Abou-Jaoude, S. (1976) Conditions necessaires et suffisantes de convergence L^1 en probabilité de l'histogramme pour une densite. *Annals de l'Institut Henri Poincaré*, Vol.12, pp.213-231.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *Proceedings of the 2nd International Symposium on Information Theory*, P.N. Petrov and F. Csaki, editors, Akademiai Kiado, Budapest, pp.267-281.

Ash, R.B. (1972) *Real Analysis and Probability*, Academic Press, New York.

Barron, A.R. (1982) *Polynomial Network Training Program: User's Guide and Technical Description*, Adaptronics, General Research Corporation, McLean, VA.

Barron, A.R. (1984) Predicted sqaured error: a criterion for automatic model selection, in *Self Organizing Methods in Modeling*, J. Farlow editor, Marcell Dekker, New York.

Barron, A.R. (1985) The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem, *Annals of Probability*, Vol.13, No.4.

Barron, A.R. and Cover, T.M. (1983) *Convergence of Logically Simple Estimates of Unknown Probability Densities*. Presented at the 1983 International Symposium on Information Theory, St. Jovite, Canada.

Barron, A.R., Van Straten, F.W. and Barron, R.L. (1977) Adaptive learning network approach to weather forecasting, *Proceedings IEEE International Conference on Cybernetics and Society*, pp. 724-727.

Barron, R.L. (1975) Learning networks improve computer-aided prediction and control, *Comput. Des.*, pp.65-70.

Barron, R.L., Mucciardi, A.N., Cook, F.J, Craig, J.N, and Barron, A.R. (1984) Adaptive learning networks: Development and application in the United States of algorithms related to GMDH, in *Self Organizing Methods in Modeling*, Marcell Dekker, New York.

Billingsley, P. (1961) *Statistical Inference for Markov Processes*, Unversity of Chicago Press.

Boehner, P. (1957) *Ockham: Philosophical Writings*, Nelson, New York.

De Bruijn, N.G. (1958) *Asymptotic Methods in Analysis*, Dover, New York.

Chaitin, G.J. (1966) On the length of programs for computing finite binary sequences, *Journal of the ACM*, Vol.13, pp.547-569.

References

Chaitin, G.J. (1969) On the length of programs for computing finite binary sequences: statistical considerations, *Journal of the ACM*, Vol.16, pp.145-159.

Chaitin, G.J. (1975) A theory of program size formally identical to information theory, *Journal of the ACM*, Vol.22, pp.329-340.

Chaitin, G.J. (1976) Algorithmic Entropy of Sets, *Computers and Mathematics with Applications*, Vol.2, pp.233-245.

Chomsky, N. (1959) On certain formal properties of grammers, *Information and Control*, Vol.2, pp.137-161.

Church (1936) An Unsolved problem of elementary number theory, *Am. J. Math.*, Vol.55, pp.345-363.

Cover, T.M. (1972) A hierarchy of probability density function estimates, in *Frontiers in Pattern Recognition*, Academic Press, New York, pp. 83-98.

Cover, T.M. (1973) Enumerative source encoding *IEEE Transactions on Information Theory*, Vol.29, pp.73-77.

Cover, T.M. (1973) On determining the irrationality of the mean of a random variable, *Annals of Statistics*, Vol.1, pp.862-871.

Cover, T.M. (1973) Generalizations on patterns using Kolmogorov complexity, *Proceedings First International Joint Conference on Pattern Recognition*, Washington, D.C.

Cover, T.M. (1983) Kolmogorov complexity, data compression, and inference. In *The Impact of Processing Techniques on Communications*, edited by J.K. Skwirzynski, Martinus Nijhoff Publ., Boston, MA, pp.23-34.

Cover, T.M. and King, R.C. (1978) A convergent gambling estimate of the entropy of English, *IEEE Transactions on Information Theory*, Vol.24, pp.413-421.

Cover, T.M. and Leung-Yan-Cheong, S.K. (1978) Some equivalences between Shannon entropy and Kolmogorov complexity, *IEEE Transactions on Information Theory*, Vol.24, pp.331-338.

Crossley, J.N. et.al. (1972) *What is Mathematical Logic?* Oxford Unversity Press.

Csiszár, I. (1967) Information-type measures of difference of probability distributions and indirect observers. *Studia Sci. Math. Hungar.* Vol.2, pp.299-318.

Csiszár, I. and Korner, J. (1981) *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York.

Davisson, L.D. (1973) Universal noiseless coding, *IEEE Transactions on Information Theory*, Vol.19, pp.783-795.

References

Davisson, L.D. (1983) Minimax Noiseless Universal Coding for Markov Sources *IEEE Transactions on Information Theory*, Vol.29, pp.211-215.

Davisson, L.D. and Leon-Garcia, A. (1980) A source matching approach to finding minimax codes, *IEEE Transactions on Information Theory*, Vol.26, pp.166-174.

Davisson, L.D., McEliece, R.J, Pursley, M.B., and Wallace, M.S. (1981) Efficient universal noiseless source codes, *IEEE Transactions on Information Theory*, Vol.27, pp.269-279.

Devroye, L. (1983) The equivalence of weak, strong, and complete convergence in L^1 for kernel density estimates, *Annals of Statistics*, Vol.11, pp.896-904.

Devroye, L. and Gyorfi, L. (1985) *Nonparametric Density Estimation: The L^1 View*, Wiley, New York.

Doob, J.L. (1949) Application of the theory of martingales. *Le Calcul de Probabilities et ses Applications. Colloques Internationaux du Centre National de la Resherche Scientifique*, Paris, pp.23-27.

Doob, J.L. (1953) *Stochastic Processes*, Wiley, New York.

Duhem, P. (1909) *Etudes sur Leonard de Vinci* II, Paris.

Elias, P. (1975) Universal codeword sets and representation of the integers, *IEEE Transactions on Information Theory*, Vol.21, pp.194-203.

Farlow, J. (1984) The GMDH algorithm, in *Self Organizing Methods in Modeling*, Marcell Dekker, New York.

Freedman, D.A. (1963) On the asymptotic behavior of Bayes' estimates in the discrete case, *Annals of Mathematical Statistics*, Vol.34, pp.1386-1403.

Gallager, R.G. (1968) *Information Theory and Reliable Communication*, Wiley, New York.

Good, I.J. and Gaskins, R.A. (1971) Nonparametric roughness penalties for probability densities, *Biometrika*, Vol.58, pp.255-277.

Geman, S. and Hwang, C.R. (1982) Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics*, Vol.10, pp.401-414.

Gilbert, E.W. and Moore, E.F. (1959) Variable-length binary encodings, *Bell System Technical Journal*, pp.933-967.

Goddu, A. (1984) *The Physics of William of Ockham*, E.J. Brill Publ., Leiden-Koln, Netherlands.

Grenander, U. (1981) *Abstract Inference*, Wiley, New York.

References

Hannan, E.J. (1980) The estimation of the order of an ARMA process, *Annals of Statistics*, Vol.8, pp.1071-1081.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, Vol.58, pp.13-30.

Hopcroft, J.E. and Ullman, J.D. (1969) *Formal Languages and their Relations to Automata*, Addison-Wesley, Reading, MA.

Ivakhnenko, A.G. (1971) Polynomial Theory of Complex Systems, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.4, pp.364-378.

Karush, J. (1961) A simple proof of an inequality of McMillan, *IEEE Transactions on Information Theory*, Vol.7, p.118.

Kieffer, J.C. (1978) A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, Vol.26, pp.674-682.

Kfoury, A.J., Moll, R.N., and Arbib M.A. (1982) *A Programming Approach to Computability*, Springer-Verlag, New York.

Kleene, S.C. (1936) General recursive functions of natural numbers, *Math. Ann.* Vol.112, pp.727-742.

Kolmogorov, A.N. (1965) Three approaches to the quantitative definition of information, *Problemy Peredachi Informatsii*, Vol.1, pp.3-11.

Kraft, C. (1955) Some conditions for consistency and uniform consistency of statistical procedures, *University of California Publications in Statistics*, Vol.2, pp.125-141.

Kraft, L.G. (1949) *A Device for Quantizing, Grouping, and Coding Amplitude Modulated Pulses*, M.S. Thesis, Department of Electrical Engineering, MIT, Cambridge, MA.

Krichevsky, R.E. and Trofimov, V.K. (1981) The performance of universal encoding, *IEEE Transactions on Information Theory*, Vol.27, pp.199-207.

Kullback, S. (1967) A lower bound for discrimination in terms of variation, *IEEE Transactions on Information Theory*, Vol.13, pp.126-127.

Longo, G. and Sgarro, A. (1979) The source coding theorem revisited: a combinatorial approach. *IEEE Transactions on Information Theory*, Vol.25, pp.544-548.

Loux, M.J. (1974) *Ockham's Theory of Terms* (translation of part I of the *Summa Logicae*), University of Notre Dame Press.

Markov, A.A. (1954) Theory of Algorithms, *Trudy Mathematicheskego Instituta imeni V.A. Steklova*, Vol.42. (Translation by Israel Program for Scientific

Translations, Jerusalem, 1961.)

Mallows, C.L. (1973) Some comments on C_p, *Technometrics*, Vol.15, pp.661-675.

Martin-Lof, P. (1966) The definition of random sequences, *Information and Control*, Vol.9, pp.602-619.

McMillan, B. (1956) Two inequalities implied by unique decipherability, *IRE Transactions on Information Theory*, Vol.2, pp.115-116.

Minsky, M.L. (1967) *Computation: Finite and Infinite Machines*, Prentice-Hall, Englewood Cliffs, N.J.

De Montricher, G.M., Tapia, R.A., and Thompson, J.R. (1975) Nonparametric maximum likelihood estimation of probability densities by penalty function methods, *Annals of Statistics*, Vol.3, pp.1329-1348.

Moody, E.A. (1935) *The Logic of William of Ockham*, Sheed and Ward, Inc., New York.

Oxtoby, J.C. (1952) Ergodic Sets, *Bull. Amer. Math. Soc.* Vol.58, pp.116-136.

Parzen, E. (1962) On the estimation of a probability density function and the mode, *Annals of Mathematical Statistics*, Vol.33, pp.1065-1076.

Pitman, E.J.G. (1979) *Some Basic Theory for Statistical Inference*, Chapman and Hall, London.

Pólya, G. and Szego, G. (1972) *Problems and Theorems in Analysis* **I**, Springer-Verlag, New York.

Post, E.L. (1936) Finite combinatory processes - formulation 1, *Jounral of Symbolic Logic*, Vol.1, pp.103-105.

Post, E.L. (1943) Formal reductions of the general combinatorial decision problem, *American Journal of Mathematics*, Vol.65, pp.197-215.

Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation*, Academic Press, Orlando, FL.

Rissanen, J. (1978) Modeling by shortest data description, *Automatica*, Vol.14, pp.465-471.

Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length, *Annals of Statistics*, Vol.11, pp.416-431.

Rissanen, J. (1984a) Universal coding, information, prediction, and estimation, *IEEE Transactions on Information Theory*, Vol.30, pp.629-636.

Rissanen, J. (1984b) *A Predictive Inference Principle for Estimation*, IBM Research

Lab, San Jose, CA.

Robbins, H.E. (1955) *American Math Monthly*, Vol.62, pp.26-29.

Robinson, J. (1950) General Recursive Functions *Proceedings of the American Mathematical Society*, Vol.1, pp.703-718.

Rogers, H. (1967) *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, New York.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function, *Annals of Math. Statistics*, Vol.127, pp.832-837.

Schnorr, C.P. (1977) A survey of the theory of random sequences. In *Basic Problems in Methodology and Linguistics*, Butts and Hintikka, editors, Reidel Publ., Dordrechi, Holland, pp.193-211.

Schnorr, C.P. and Fuchs, P. (1977) General random sequences and learnable sequences, *Journal of Symbolic Logic*, Vol.42, pp.329-340.

Schwartz, L. (1965) On Bayes Procedures, *Z. Warhrsch. Verw. Gebiete*, Vol.4, pp.10-26.

Schwartz, S.C. (1967) Estimation of a probability density by an orthogonal series, *Annals of Mathematical Statistics*, Vol.38, pp.1262-1265.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, Vol.6, pp.461-464.

Shannon, C.E. (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.

Solomonoff, R.J. (1964) A formal theory of inductive inference, *Information and Control*, pp.1-22, 224-254.

Sorkin, R. (1983) A quantitative Occam's Razor, *International Journal of Theoretical Physics*, Vol.22, 1091-1103.

Tapia, R.A. and Thompson, J.R. (1978) *Nonparametric Probability Density Estimation*, Johns Hopkins University Press, Baltimore, MD.

Tierney, L. and Kadane, J.B. (1972) *Accurate Approximations for Posterior Moments and Marginals*, Technical report no.431, School of Statistics, University of Minnesota.

Tornay, S.C. (1938) *Ockham: Studies and Selections*, Open Court Publ., La Salle, IL.

Trofimov, V.K. (1974) Redundancy of universal coding of arbitrary Markov Sources, *Problemy Peredachi Informatsii*, Vol.10, pp.16-24.

Lab, San Jose, CA.

Robbins, H.E. (1955) *American Math Monthly*, Vol.62, pp.26-29.

Robinson, J. (1950) General Recursive Functions *Proceedings of the American Mathematical Society*, Vol.1, pp.703-718.

Rogers, H. (1967) *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, New York.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function, *Annals of Math. Statistics*, Vol.127, pp.832-837.

Schnorr, C.P. (1977) A survey of the theory of random sequences. In *Basic Problems in Methodology and Linguistics*, Butts and Hintikka, editors, Reidel Publ., Dordrechi, Holland, pp.193-211.

Schnorr, C.P. and Fuchs, P. (1977) General random sequences and learnable sequences, *Journal of Symbolic Logic*, Vol.42, pp.329-340.

Schwartz, L. (1965) On Bayes Procedures, *Z. Warhrsch. Verw. Gebiete*, Vol.4, pp.10-26.

Schwartz, S.C. (1967) Estimation of a probability density by an orthogonal series, *Annals of Mathematical Statistics*, Vol.38, pp.1262-1265.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, Vol.6, pp.461-464.

Shannon, C.E. (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.

Solomonoff, R.J. (1964) A formal theory of inductive inference, *Information and Control*, pp.1-22, 224-254.

Sorkin, R. (1983) A quantitative Occam's Razor, *International Journal of Theoretical Physics*, Vol.22, 1091-1103.

Tapia, R.A. and Thompson, J.R. (1978) *Nonparametric Probability Density Estimation*, Johns Hopkins University Press, Baltimore, MD.

Tierney, L. and Kadane, J.B. (1972) *Accurate Approximations for Posterior Moments and Marginals*, Technical report no.431, School of Statistics, University of Minnesota.

Tornay, S.C. (1938) *Ockham: Studies and Selections*, Open Court Publ., La Salle, IL.

Trofimov, V.K. (1974) Redundancy of universal coding of arbitrary Markov Sources, *Problemy Peredachi Informatsii*, Vol.10, pp.16-24.

References

Turing, A.M. (1936) On Computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Vol.42.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate, *Annals of Mathematical Statistics*, Vol.20, pp.595-601.

Wolfowitz, J. (1949) On Wald's proof of the consistency of the maximum likelihood estimate, *Annals of Mathematical Statistics*, Vol.20, pp.601-602.

Wozencraft, J.W. and Reiffen, B. (1961) *Sequential Decoding*, MIT Technology Press, Wiley, New York.