Minimum Complexity Density Estimation

Andrew R. Barron, *Member, IEEE* and Thomas M. Cover, *Fellow, IEEE*

Abstract—**The minimum complexity or minimum description-length criterion developed by Kolmogorov, Rissanen, Wallace, Sorkin, and others leads to consistent probability density estimators. These density estimators are defined to achieve the best compromise between likelihood and simplicity. A related issue is the compromise between accuracy of approximations and complexity relative to the sample size. An index of resolvability is studied which is shown to bound the statistical accuracy of the density estimators, as well as the information-theoretic redundancy.**

Index Terms—**Kolmogorov complexity, minimum description-length criterion, universal data compression, bounds on redundancy, resolvability of functions, model selection, density estimation, discovery of probability laws, consistency, statistical convergence rates.**

I. Introduction

THE KOLMOGOROV theory of complexity (Kolmogorov [1]) leads to the notion of a universal minimal sufficient statistic for the optimal compression of data as discussed in V'Yugin [2], Cover [3], [4], and Cover, Gacs, and Gray [5]. The Kolmogorov theory is applicable to arbitrary, possibly nonrandom, data sequences. Related notions of complexity or description length, that are specifically appropriate for making inferences from random data, arise in the work of Rissanen [6]–[11], Wallace *et al.* [12], [13], Sorkin [14], Barron [15]–[17], Cover [3], [4], [18], and V'Yugin [2] and in the context of universal source coding as in Davisson [19]. The goal shared by these complexity-based principles of inference is to obtain accurate and parsimonious estimates of the probability distribution. The idea is to estimate the simplest density that has high likelihood by minimizing the total length of the description of the data. The estimated density should summarize the data in the sense that, given the minimal description of the estimated density, the remaining description length should be close to the length of the best description that could be achieved if the true density were known.

Minimum complexity estimators are treated in a general form that can be specialized to various cases by the choice of a set of candidate probability distributions and by the choice of a description length for each of these distributions, subject to information-theoretic requirements. An idealized form of the minimum complexity criterion is obtained when Kolmogorov's theory of complexity is used to assess the description length of probability laws; however, our results are not restricted to this idealistic framework.

For independent random variables $X_1, X_2, \cdots, X_n$ drawn from an unknown probability density function p, the minimum complexity density estimator $\hat{p}_n$ is defined as a density achieving the following minimization

$$\min_q \left(L(q) + \log \frac{1}{\prod_{i=1}^{n} q(X_i)} \right), \tag{1.1}$$

where the minimization is over a list Γ of candidate probability density functions q, and the logarithm is base 2. As discussed in Section III, this criterion corresponds to the minimization of the total length of a two-stage description of the data. The nonnegative numbers $L(q)$ are assumed to satisfy Kraft's inequality $\sum_q 2^{-L(q)} \leq 1$ and are interpreted to be codelengths for the descriptions of the densities. Although not needed for the information-theoretic interpretation, there is also a Bayesian interpretation of the numbers $2^{-L(q)}$ as prior probabilities. In the Kolmogorov complexity framework, $L(q)$ is equal to the length of the shortest computer code for q as explained in Section IV, and the best data compression and the best bounds on rates of convergence are obtained in this case.

The list Γ of candidate probability densities is often specified from a given sequence of parametric models of dimension $d = 1, 2, \cdots$, with the parameter values restricted to a prescribed number of bits accuracy. The minimum complexity criterion is then used to select the model and to estimate the parameters. Larger lists Γ provide better flexibility to discover accurate yet parsimonious models in the absence of true knowledge of the correct parametric family. In the idealistic case, Γ consists of all computable probability distributions.

The minimum complexity criterion can discover the true distribution. Indeed, it is shown that if the true distribution happens to be on the countable list Γ, then the estimator is exactly correct,

$$\hat{p}_n \equiv p, \tag{1.2}$$

for all sufficiently large sample sizes, with probability one (Theorem 1). Consequently, the probability of error based

Manuscript received February 3, 1989; revised January 25, 1991. A. R. Barron's work was supported by ONR Contracts N00014-86-K-0670 and N00014-89-J-1811. T. M. Cover's work was supported in part by the National Science Foundation under Contract NCR-89-14538.

A. R. Barron is with the Department of Statistics and the Department of Electrical and Computer Engineering, University of Illinois, 101 Illini Hall, 725 S. Wright St., Champaign, IL 61820.

T. M. Cover is with Information Systems Laboratory, 121 Durand, Stanford University, Stanford, CA 94305.

IEEE Log Number 9144768.

on n samples tends to zero as $n \to \infty$. The result is most dramatic in the Kolmogorov complexity framework: if the data are governed by a computable probability law, then, with probability one, this law eventually will be discovered and thereafter never be refuted. Although the law is eventually discovered, one cannot be certain that the estimate is exactly correct for any given n. You know, but you do not know you know.

Consistency of the minimum complexity estimator is shown to hold even if the true density is not on the given countable list, provided the true density is approximated by sequences of densities on the list in the relative entropy sense. Theorems 2 and 3, respectively, establish almost sure consistency of the estimated distribution and (under somewhat stronger assumptions) L^1 consistency of the estimated density. These results, which were announced in [15], [16], are the first general consistency results for the minimum description-length principle in a setting that does not require the true distribution to be a member of a finite-dimensional parametric family.

The main contribution of this paper is the introduction of an index of resolvability,

$$R_n(p) = \min_q \left[\frac{L(q)}{n} + D(p \| q) \right], \qquad (1.3)$$

that is proved to bound the rate of convergence of minimum complexity density estimators as well as the information-theoretic redundancy of the corresponding total description length. Here $D(p\|q)$ denotes the relative entropy. The resolvability of a density function is determined by how accurately it can be approximated in the relative entropy sense by densities of moderate complexity relative to the sample size. Theorem 4 and its corollary state conditions under which the minimum complexity density estimator converges in squared Hellinger distance $d_H^2(p, \hat{p}_n) = \int(\sqrt{p} - \sqrt{\hat{p}_n})^2$ with rate bounded by the index of resolvability, i.e.,

$$d_H^2(p, \hat{p}_n) \le O(R_n(p)) \qquad \text{in probability.} \quad (1.4)$$

Also the complexity of the estimate relative to the sample size, $L_n(\hat{p}_n)/n$ is shown to be not greater than $O(R_n(p))$ in probability.

The results on the index of resolvability demonstrate the statistical effectiveness of the minimum description-length principle as a method of inference. Indeed, with high probability, the estimation error $d_H^2(p, \hat{p}_n)$ plus the complexity per sample size $L_n(\hat{p}_n)/n$, which are achieved by the minimum complexity estimator, are as small as can be expected from an examination of the optimal tradeoff between the approximation error $D(p\|q)$ and the complexity $L(q)/n$, as achieved by the index of resolvability.

It is shown that the index of resolvability $R_n(p)$ is of order $1/n$ if the density is on the list; order $(\log n)/n$ in parametric cases; order $(1/n)^\gamma$ or $((\log n)/n)^\gamma$ in some nonparametric cases, with $0 < \gamma < 1$; and order $o(1)$ in general, provided $\inf_{q \in \Gamma} D(p\|q) = 0$.

It need not be known in advance which class of densities is correct. With minimum complexity estimation, we are free to consider as many models as are plausible and practical. (In contrast, the method of maximum likelihood density estimation fails without constraints on the class of densities.) The minimum complexity estimator converges to the true density nearly as fast as an estimator based on prior knowledge of the true subclass of densities.

The minimum complexity estimator may also be defined for lists of joint densities $q(X_1, X_2, \cdots, X_n)$ that allow for dependent random variables, instead of independence $\prod_{i=1}^n q(X_i)$ as required in (1.1). Indeed, the assumption of stationarity and ergodicity is sufficient for the result on the discovery of the true distribution in the computable case, as shown in [16]. The assumption of independence, however, appears to be critical to our method of obtaining bounds on the rate of convergence of the density estimators in terms of the index of resolvability.

In some regression and classification contexts, a complexity penalty may be added to a squared error or other distortion criterion that does not correspond to the length of an efficient description of the data. Bounds on the statistical risk in those contexts have recently been developed in Barron [17] using inequalities of Bernstein and Hoeffding instead of the Chernoff inequalities used here.

Interpretations and basic properties of minimum complexity estimators are discussed in Sections II–IV. Motivation for the index of resolvability is given in Section V followed by examples of the resolvability for various models in Section VI. The main statistical convergence results are given in Section VII followed by the proofs in Section VIII. Some regression and classification problems that can be examined from the minimum description-length framework are discussed in Section IX.

II. An Informal Example

The minimum description-length criterion for density estimation is illustrated by the following example. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed according to an unknown probability density $p(x)$. Suppose it happens that this density is normal with mean μ and variance $\sigma^2 = \sqrt{2}$. In this example, μ is some fixed uncomputable real number, whereas $\sqrt{2}$ is computable. (Computability means that a fixed-length program exists that can take any integer b as an input and compute the number to accuracy 2^{-b}.)

The minimum description-length idea is to choose a simple density q that yields high likelihood on the data. If $L(q)$ is the number of bits needed to describe q and $\log 1/q(X_1, \cdots, X_n)$ is the number of bits in the Shannon code (relative to q) for the data, then

$$\min_q \left(L(q) + \log \frac{1}{q(X_1, \cdots, X_n)} \right) \qquad (2.1)$$

is the minimum two-stage description length of the data. (The actual Shannon code has length equal to the integer part of the logarithm of the reciprocal of the probability of discretized values of the data; the use of the density is a convenient simplification.)

For the Gaussian example, we may expect the procedure to work as follows. For small sample sizes compared

to the complexity of the normal family (say $n \leq 10$), we would estimate $\hat{p}_n$ to be one of a few very simple densities, such as a uniform density over a simple range that includes the sample. For moderate sample sizes (perhaps $n \approx 100$) we begin to use the normal family. Parameter estimates and associated description lengths that achieve approximately the best tradeoff between complexity and likelihood are derived in [7], [12], [13] and [16, Section 4.2] (see also Section VI where related description lengths are given that optimize the index of resolvability in parametric cases). In particular, we take the maximum likelihood estimates ($\bar{X}_n = (1/n)\Sigma X_i$ and $S_n^2 = (1/n)\Sigma(X_i - \bar{X}_n)^2$) rounded off to the simplest numbers $\hat{\mu}$ and $\hat{\sigma}^2$ in the confidence intervals $\bar{X}_n \pm 1/\sqrt{c_1 n}$ and $S_n^2 \pm 1/\sqrt{c_2 n}$. These numbers are described using roughly $(1/2)\log nc_1$ and $(1/2)\log nc_2$ bits. Here $c_1 = 1/S_n^2$ and $c_2 = 1/(2S_n^4)$ are the empirical Fisher informations, for μ and σ^2 respectively, evaluated at the maximum likelihood.[1] See [7], [12], [13] for relevant discussion on the appropriate constants. In the present example, for which the variance is a relatively simple number, the enumeration of the bits of the estimate is preferred only for sample sizes with $(1/2)\log nc_2$ less than the length of the description of $\sqrt{2}$.

Then when we have enough data to determine the first 10 or so bits of the unknown variance ($n \approx 1\,000\,000$), we begin to believe that the density estimate is normal with variance equal to $\sqrt{2}$. We have guessed correctly that $\sigma^2 = \sqrt{2}$, and this guess results in a shorter description. The estimated mean $\hat{\mu}_n$ is $\bar{X}_n$ rounded to an accuracy of $\sigma/\sqrt{n}$; this requires roughly $(1/2)\log nc_1$ bits where $c_1 = 1/\sigma^2$. For any constant c, no simple number is found in the interval $\bar{X} \pm c/\sqrt{n}$ for large n. We must content ourselves with $\hat{p}_n = \text{normal}(\hat{\mu}_n, \sqrt{2})$.

Note that the complexity of the best density estimate $\hat{p}_n$ grows at first like n, then like $(1/2)\log nc_1 + (1/2)\log nc_2$, and finally like $(1/2)\log nc_1$. For large n, we have discovered that the true density function is Gaussian, that its variance is exactly $\sqrt{2}$, and that its mean is approximately $\bar{X}_n \pm \sigma/\sqrt{n}$. From the data alone, it becomes apparent that the structure of the underlying probability law consists of its Gaussian shape and its special variance. Its mean, however, has no special properties.

Even if the true density is not a member of any of the usual parametric families, the minimum description-length criterion may select a family to provide an adequate approximation for a certain range of sample sizes. Additional samples will then throw doubt on the tentative choice. With the aid of the criterion, we are then free to jump to some other family (and proceed with the estimation of any parameters in this family). The question is whether this disorderly jumping around from procedure to procedure on the basis of some peeking at the data will still allow convergence. We show that indeed convergence does occur for densities estimated by the minimum complexity or minimum description-length criterion.

The formulation of the minimum description-length principle as in [6], [7], [12], [13] leads to a restriction on the parameter estimates in each family to a grid of points spaced at width of order $1/\sqrt{n}$ and the optimization of a criterion for which the dominant terms are

$$\frac{d}{2}\log n + \log 1/p_{\hat{\theta}}(X^n), \tag{2.2}$$

where d is the number of parameters and $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector $\theta \in R^d$, truncated to $(1/2)\log n$ bits per parameter. Rissanen [8], [9] shows that for most parameter points this criterion yields asymptotically the best data compression. Indeed, he shows in [9] that the redundancy of order $(d/2)\log n$ cannot be beaten except for a set of parameter points of measure zero. The theory we develop shows that the minimum description-length criterion for model selection is also justified on the grounds that it produces statistically accurate estimates of the density.

The Gaussian example illustrates an advantage of deviating in some cases from the minimum description-length criterion in the form (2.2), by not necessarily restricting the parameter estimates to a grid of preassigned widths of order $1/\sqrt{n}$. By allowing the search to include simpler parameter values, in particular to include maximum likelihood estimates truncated to fewer than $(1/2)\log n$ bits and nearby numbers of low complexity (such as $\sqrt{2}$ in the previous example), we allow for the possibility of discovery of density functions with special parameter points, which in some cases may govern the distribution of the observed data.

Other departures from the minimum description-length criterion in the form (2.2) are justified when it is not assumed that the density is in a finite-dimensional family. See Case 4 in Section VI for one such example. Nevertheless, it will be seen (Case 3 in Section VI) that criteria of the form (2.2), using sequences of parametric families, continue to be effective for both data compression and inference in an infinite-dimensional context.

III. Some Preliminaries

In this section we set up some notation, define minimum complexity density estimation, and discuss some specializations of the general method.

Let $X_1, X_2, \cdots, X_n, \cdots$ be independent random variables drawn from a (possibly unknown) probability density function $p(x)$. The random variables are assumed to take values in a measurable space $\boldsymbol{X}$ and the density function is taken with respect to a known sigma-finite dominating measure $\nu(dx)$. The joint density function for $X^n =$

[1] It happens in this Gaussian example that the Fisher information matrix is diagonal. For parametric families with nondiagonal information matrices, approximately the best tradeoff is achieved with an estimated parameter vector in an elliptical confidence region centered at the MLE. Such estimates are described in a locally rotated and scaled coordinate system, using about $(1/2)\log nc_1 + \cdots + (1/2)\log nc_d$ bits, which reduces to $(1/2)\log\det(n\hat{I})$, where $c_1, \cdots, c_d$ are the eigenvalues of the empirical Fisher information $\hat{I}$ and d is the dimension of the parameter space, see [16, Sect. 4.2]. Thus $(d/2)\log n$ is the dominant term in the description of the parameters and the $(1/2)\log\det(\hat{I})$ term accounts for the local curvature of likelihood function.

$(X_1, X_2, \cdots, X_n)$ is denoted by $p(X^n) = \prod_{i=1}^n p(X_i)$ for $n = 1, 2, \cdots$; the probability distribution for the process is denoted by P.

For each $n = 1, 2, \cdots$, let Γ_n be a countable collection of probability density functions $q(x)$ (each taken with respect to the same measure $\nu(dx)$). For each q in Γ_n, we let $q(X^n) = \prod_{i=1}^n q(X_i)$ denote the corresponding product density and we let Q denote the corresponding probability distribution (which would make the X_i independent with density q).

We need a notion of the length of a description of q. For each n, let $L_n(q)$ be nonnegative numbers defined for each q in Γ_n. (For convenience, we also define $L_n(q) = \infty$ if q is not in Γ_n.) The following summability requirement,

$$\sum_{q \in \Gamma_n} 2^{-L_n(q)} \le 1, \tag{3.1}$$

is the essential condition assumed of the numbers $L_n(q)$.

The complexity of the data and the minimum complexity density estimate are now defined relative to the lengths $L_n(q)$, $q \in \Gamma_n$. The sample size n is assumed to be given.

Definition: The *complexity* $B(X^n)$ of the data X^n relative to L_n and Γ_n is defined by

$$B(X^n) = \min_{q \in \Gamma_n} \left(L_n(q) + \log \frac{1}{q(X^n)} \right). \tag{3.2}$$

The *minimum complexity estimator* $\hat{p}_n$ of the density relative to L_n and Γ_n is defined by

$$\hat{p}_n = \arg\min_{q \in \Gamma_n} \left(L_n(q) + \log \frac{1}{q(X^n)} \right), \tag{3.3}$$

where, in the case of ties, the density $\hat{p}_n$ is chosen for which $L_n(\hat{p}_n)$ is shortest (and any further ties are broken by selecting the density with least index in Γ_n). It will be seen that a minimizing $\hat{p}_n$ exists with probability one.

There are two fundamental interpretations of the minimum complexity criterion: one from the theory of data compression, the other from Bayesian statistics.

A. Coding Interpretation

The complexity defined in (3.2) is interpreted as a minimal two-stage description length for X^n, for a given sample size n. The terms $L_n(q)$ and $\log 1/q(X^n)$ correspond, respectively, to the length of a description of q and the length of a description of X^n based on q.

To give the precise coding interpretation, assume that X is discrete and that each q is a probability mass function (i.e., q is a density with respect to ν = counting measure). If the numbers $L_n(q)$, $q \in \Gamma_n$ are positive integers satisfying (3.1), then $L_n(q)$ is the length of an instantaneously decodable binary code for $q \in \Gamma_n$. The instantaneous decodability property states that no codeword is the prefix of any other codeword. Since the second-stage description of X^n follows the code for q, the prefix condition is essential for decoding the two stages. The condition (3.1) in this context is Kraft's inequality giving necessary and sufficient conditions for the existence of instantaneous binary codes of the prescribed lengths (see [20, pp. 45–49, 514]).

To explain the second stage of the code, observe that if q is given, then by rounding $\log 1/q(X^n)$ up to the nearest integer, lengths $L_q(X^n) = \lceil \log 1/q(X^n) \rceil$ are obtained that satisfy Kraft's inequality, $\sum_{X^n} 2^{-L_q(X^n)} \le 1$. Hence, as discovered by Shannon, if q is given, then $\lceil \log 1/q(X^n) \rceil$ is the length of an instantaneous code that describes the sequence X^n.

On the other hand, when the density is estimated from the data, then in order for the Shannon code based on an estimate $\hat{p}_n$ to be uniquely decodable, the density $\hat{p}_n$ must first be encoded. The overall length of the code for the data is then (within one bit of)

$$L_n(\hat{p}_n) + \log 1/\hat{p}_n(X^n). \tag{3.4}$$

Thus any density estimator corresponds to a code for the data. The minimum complexity criterion simply chooses the estimator yielding the best compression.

Coding Interpretation in the Continuous Case: If the space X is not discrete, then no finite-length uniquely decodable codes can exist. Nevertheless, quantization of X does lead to outcomes that are finitely describable. In the case of fine quantization, density functions are approximated by ratios of measures. Indeed, if $[x]$ denotes the quantization region that contains x, then $q(x) = \lim Q([x])/\nu([x])$ for almost every x (where the limit is taken for a refining sequence of quantization regions that generates X). Consequently, $\log 1/q(X^n) \cong \log 1/Q([X^n]) + \log \nu([X^n])$ where $[X^n]$ denotes the coordinate-wise quantization of X^n. If this approximation were valid uniformly for $q \in \Gamma_n$, then the minimization as in (3.3) would amount to choosing a density that minimizes the two-stage codelength for the quantized data,

$$L_n(Q) + \log 1/Q([X^n]). \tag{3.5}$$

For simplicity of exposition in this paper, we restrict attention to the minimization involving densities as in (3.3). Discrete random variables are then a special case with ν equal to counting measure. Barron [16] treats the case in which the distribution is estimated by minimizing (3.5); in the theory developed there, the quantization regions are allowed to shrink as the sample size $n \to \infty$. [It is seen that the estimators based on uniformly quantized data on the real line behave in a manner essentially analogous to the continuous case when the width h of the quantization intervals are of smaller order than $1/n$, and in a manner analogous to the discrete case when nh is large. New techniques are also developed there to handle the case when nh is constant.]

B. Bayesian Inference Interpretation

Let $w_n(q)$ be a prior probability mass function on $q \in \Gamma_n$ and set $L_n(q) = \log 1/w_n(q)$ (with the convention that if $w_n(q) = 0$ then $L_n(q) = \infty$). Then the summability

condition (3.1) is satisfied since

$$\sum_q 2^{-L_n(q)} = \sum_q w_n(q) = 1. \tag{3.6}$$

The minimization in (3.3) is seen to be the same as the maximization of

$$2^{-L_n(q)} q(X^n), \tag{3.7}$$

which is proportional (as a function of $q \in \Gamma_n$) to the Bayes posterior probability of q given X^n.

Consequently, the estimator $\hat{p}_n$ defined in (3.3) is the Bayes estimator minimizing the probability of error, $\sum_q w_n(q) Q\{\hat{p}_n \neq q\}$ for a density in Γ_n drawn according to $w_n(q)$.

The connection between the Bayesian and coding interpretations is that if $w_n(q)$ is a prior probability function concentrated on a countable set of densities q, then $\log 1/w_n(q)$ is the length (rounded to an integer) of a Shannon code for q based on w_n. Conversely, if $L_n(q)$ is a codelength for a uniquely decodable code, then $w_n(q) = 2^{-L_n(q)}/c_n$ defines a proper prior probability (where $c_n = \sum 2^{-L_n(q)} \leq 1$ is the normalizing constant).

Thus the minimum description-length principle provides an information-theoretic justification of Bayes' rule. A Bayesian with a discrete prior $w_n(q)$, $q \in \Gamma_n$ chooses the estimate that achieves the minimum total description length $L_n(\hat{p}_n) + \log 1/\hat{p}_n(X^n)$. Of course, Bayesian estimation also has decision-theoretic justification.

We emphasize the necessity of the term $L_n(\hat{p}_n)$ that involves the prior probability. Indeed, in the absence of this term, if $\hat{p}_n$ depends on X^n, then, in general, $\log 1/\hat{p}_n(X^n)$ will not satisfy Kraft's inequality and hence there does not exist a code for X^n with lengths $\log 1/\hat{p}_n(X^n)$. A consequence is that the maximum likelihood rule that selects a density to achieve the minimum value of $\log 1/\hat{p}_n(X^n)$ does not admit a description-length interpretation of this value.

Some basic results for the minimum complexity estimator are a straightforward consequence of the Bayesian interpretation. Define

$$m(X^n) = \sum_q 2^{-L_n(q)} q(X^n). \tag{3.8}$$

In the Bayesian interpretation, $m(X^n)$ is the marginal density function for X^n. It is seen that $m(X^n)$ is finite for almost every X^n (indeed it has integral not greater than one). Thus for almost every X^n, the quantities in (3.7) are summable for $q \in \Gamma_n$ and, consequently, the maximum is achieved. (Indeed, let $\nu > 0$ be the value in (3.7) for some q; by summability there must be a finite set of q such that outside this set the value of $2^{-L_n(q)}\, q(X^n)$ is less than ν, and hence the overall maximum occurs on this finite set.) Thus the following proposition is proved.

Proposition 1: There almost surely exists at least one and no more than finitely many densities achieving the maximum in (3.7). Thus the minimum complexity density estimator $\hat{p}_n$ exists with probability one.

Next we show admissibility of the minimum complexity estimator of a density in the countable set Γ_n, among estimators based on the data $X_1, X_2, \cdots, X_n$. By definition, an estimator $\hat{p}_n$ is *inadmissible* if there is another estimator $\hat{p}_n^{(2)}$ such that $P\{\hat{p}_n^{(2)} \neq p\} \leq P\{\hat{p}_n \neq p\}$ for all $p \in \Gamma_n$, with strict inequality for some $p \in \Gamma_n$. If no such uniformly better estimator exists, then $\hat{p}_n$ is said to be *admissible*. The following proposition is a consequence of the admissibility of Bayes rules.

Proposition 2: The minimum complexity estimator $\hat{p}_n$ is admissible for the estimation of a density in the countable set Γ_n.

IV. Idealized Codelengths and Kolmogorov Complexity

Clearly, a practical requirement on the candidate probability distributions Q is that finite length descriptions exist, i.e., Q must be computable.[2] Subject to this restriction, an idealized form of the minimum complexity criterion is obtained by choosing the descriptions of the probabilities Q to be as short as possible.

Let U be a fixed universal computer with a domain consisting of finite length binary programs ϕ that satisfy the prefix property. Specifically, no acceptable program is a prefix of another, so that the set of binary programs constitutes an instantaneous code and consequently the program lengths satisfy the Kraft inequality. Let Γ^* be the set of all computable probability measures on X. For each $Q \in \Gamma^*$, let $L^*(Q) = L_U^*(Q)$ be the minimum length of programs that recursively enumerate Q,

$$L^*(Q) = \min_{U(\phi) = Q} \text{length}(\phi). \tag{4.1}$$

This $L^*(Q)$ is the Kolmogorov–Solomonoff–Chaitin algorithmic complexity of Q. This measure was independently posed, in different levels of detail, by Kolmogorov [1], Solomonoff [21], and Chaitin [22]. For fundamental properties of L^*, see Chaitin [23] and Levin [24], [25].

We mention that for any two universal computers U and V there exists a finite constant $c = c_{U,V}$ such that

$$|L_U^*(Q) - L_V^*(Q)| \leq c, \qquad \text{for all } Q \in \Gamma^*. \tag{4.2}$$

Moreover, for any computable function $L(Q)$, $Q \in \Gamma$ (on a domain $\Gamma \subset \Gamma^*$) that satisfies the Kraft inequality, there exists a constant $c = c_L$ such that

$$L^*(Q) \leq L(Q) + c, \qquad \text{for all } Q \in \Gamma. \tag{4.3}$$

In the same way, for any computable prior $w(Q)$, $Q \in \Gamma \subset \Gamma^*$ with $\sum_Q w(Q) = 1$, there is an constant $c = c_w$ such that

$$L^*(Q) \leq \log 1/w(Q) + c, \qquad \text{for all } Q \in \Gamma, \tag{4.4}$$

whence

$$2^{-L^*(Q)} \geq w(Q) 2^{-c}, \qquad \text{for all } Q \in \Gamma. \tag{4.5}$$

It is these basic facts about the algorithmic complexity $L^*(Q)$ that provide its appeal as a notion of idealized

[2] A probability measure Q on X is computable, relative to a countable collection of sets $A_1, A_2, \cdots$ that generates the measurable space X, if the set $\{(r_1, r_2, k)\colon r_1 < Q(A_k) < r_2$ for r_1, r_2 rational and $k = 1, 2, \cdots\}$ is recursively enumerable. Thus $Q(A_k)$ can be calculated to any preassigned degree of accuracy.

codelength for the minimum complexity estimation principle. In particular, (4.5) gives a sense in which $2^{-L^*(Q)}$ is a universal prior, giving (essentially) at least as much mass to distributions as would any computable prior.

V. An Index of Resolvability

Minimum complexity density estimation chooses a density that minimizes the quantity

$$\frac{1}{n}L_n(q)+\frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{q(X_i)}. \tag{5.1}$$

This quantity is a random variable depending on X_i, $i=1,2,\cdots,n$, which are assumed to be independent with unknown density p. In order to help explain the behavior of this minimization, we replace (5.1) by its expected value and investigate the corresponding minimization. This expectation is

$$\frac{1}{n}L_n(q)+E_p\left(\log\frac{1}{q(X)}\right)$$

$$=\frac{1}{n}L_n(q)+D(p\|q)+H(p), \tag{5.2}$$

where $H(p)=-\int p(x)\log p(x)\nu(dx)$ is the entropy (of p with respect to ν) and $D(p\|q)=\int p(x)\log(p(x)/q(x))\nu(dx)$ is the relative entropy or Kullback–Leibler distance between p and q.

Since by the law of large numbers, the quantities in (5.1) are close to the expected value for large n, we anticipate that the behavior of the minimization of (5.1) will be largely determined by the minimization of (5.2).

Definition: The *index of resolvability* of p (relative to a list Γ_n, codelengths L_n, and sample size n) is defined by

$$R_n(p)=\min_{q\in\Gamma_n}\left(\frac{1}{n}L_n(q)+D(p\|q)\right). \tag{5.3}$$

An interpretation of the index of resolvability is the following. If we know p, then $nH(p)$ bits are required to describe X^n on the average. If we do not know p, then $n(H(p)+R_n(p))$ bits suffice to describe X^n on the average. This is proved shortly in Proposition 4. The index of resolvability may be interpreted as the minimum description-length principle applied on the average. The index of resolvability is used (in the proof of Theorem 4) to bound the rate of convergence of the density estimator $\hat{p}_n$ that minimizes (5.1) in terms of a density $\tilde{p}_n$ that achieves the minimum in (5.3).

The density $\tilde{p}_n$ minimizing $L_n(q)$ among those that achieve the minimum in (5.3) is regarded as the density that best resolves p for sample size n. A compromise is achieved between densities that closely approximate p and densities with logical simplicity.

For example, suppose the true density p is a standard normal perturbed by having zero density in a small segment accounting for about 0.001 of the mass of the normal curve and having density scaled up by a factor of 1.001 on the rest of the line (so that the total area remains one). Then, for sample sizes n much less than 1000, it is unlikely for a normal density to have observations in the perturbed segment. The true density and the normal density are indistinguishable in this case. Indeed, the relative entropy distance between the true density and the standard normal is $\log 1.001$, which is approximately equal to $0.001\log e$. If $L(p)$ and $L(\phi)$ are the description lengths of the true density and the normal density, respectively, then from definition (5.3), the normal density has better resolvability for $n<1000(L(p)-L(\phi))/\log e$.

The density $\tilde{p}_n$ is a theoretical analog of the sample-based minimum complexity estimator $\hat{p}_n$. In our analysis of $\hat{p}_n$, we regard it as being more directly an estimator of $\tilde{p}_n$ than an estimator of p. The total error between $\hat{p}_n$ and p involves contributions from the estimation of $\tilde{p}_n$ by $\hat{p}_n$ and from the approximation of p by $\tilde{p}_n$.

Observe that in general the resolvability can be improved by increasing n, enlarging Γ_n, or decreasing the lengths $L_n(q)$.

In the limit as $n\to\infty$, the index of resolvability $R_n(p)$ converges to zero if and only if there is a sequence of densities q_n in Γ_n such that $D(p\|q_n)\to 0$ and $L_n(q_n)/n\to 0$.

Definition: The *information closure* of Γ, denoted by $\bar{\Gamma}$, is the set of all probability densities p for which $\inf_{q\in\Gamma}D(p\|q)=0$.

In the case of all computable probability measures on the real line, it is shown in Barron [16] that the information closure $\bar{\Gamma}^*$ consists of all densities p for which $D(p\|q)$ is finite for some computable measure Q. Moreover, $\bar{\Gamma}^*$ includes all bounded densities with finite support and all densities with tails or peaks bounded by a computable integrable function.

Here we show that the information closure is the set of all distributions for which the resolvability tends to zero as $n\to\infty$. A condition is required to force regular behavior of the numbers $L_n(q)$ as a function of n. Let $\Gamma=\cup_n\Gamma_n$ be the union of the lists of densities Γ_n.

Growth restriction:

$$L_n(q)=o(n), \qquad \text{for each } q\in\Gamma. \tag{5.4}$$

Note that this condition requires that each $q\in\Gamma$ is in Γ_n for all large n. The growth restriction is automatically satisfied for a constant ($\Gamma_n=\Gamma$) or increasing ($\Gamma_n\uparrow\Gamma$) sequence of sets of densities with a constant ($L_n(q)=L(q)$) or convergent ($\lim_n L_n(q)=L(q)$) sequence of codelengths.

Proposition 3: If the numbers $L_n(q)$ satisfy the growth restriction (5.4), then

$$\lim_{n\to\infty}R_n(p)=0, \tag{5.5}$$

if and only if p is in $\bar{\Gamma}$, the information closure of Γ.

Proof of Proposition 3: Clearly $R_n(p)\to 0$ implies $D(p\|\tilde{p}_n)\to 0$ and hence p is in $\bar{\Gamma}$. Suppose conversely that $\inf_{q\in\Gamma}D(p\|q)=0$. Given any $\epsilon>0$, choose q in Γ

such that $D(p\|q) < \epsilon$. Then by the growth restriction (5.4),

$$\lim_{n\to\infty} R_n(p) \le \lim_{n\to\infty} \frac{1}{n} L_n(q) + D(p\|q) < \epsilon. \quad (5.6)$$

Now $\epsilon > 0$ is arbitrary, so $\lim R_n(p) = 0$ as desired. □

The *redundancy* $\Delta_n(p)$ of a code is defined to be the expected value of the difference between the actual and ideal codelengths divided by the sample size. For the minimum two-stage codelengths $B(X^n)$ defined as in (3.2) we have

$$\Delta_n(p) = \frac{1}{n} E(B(X^n) - \log 1/p(X^n)). \quad (5.7)$$

Here $\log 1/p(X^n)$ is interpreted as the ideal codelength: it can only be achieved with true knowledge of the distribution p. Its expected length is the entropy of p. When the entropy is finite, the redundancy measures the excess average description length beyond the entropy. The redundancy, which plays a role similar to that of a risk function in statistical decision theory, is the basis for information-theoretic notions of the efficiency of a code, as developed in Davisson [19].[3]

Proposition 4: The redundancy of the minimum two-stage code is less than or equal to the index of resolvability, i.e.,

$$\Delta_n(p) \le R_n(p). \quad (5.8)$$

Proof: We have

$$\frac{1}{n}(B(X^n) - \log 1/p(X^n)) = \min_{q\in\Gamma_n} \left(\frac{1}{n} L_n(q) + \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \right). \quad (5.9)$$

Taking the expected value with respect to P, we have

$$\Delta_n(p) = E \min_q (\cdot) \le \min_q E(\cdot) = R_n(p),$$

as desired. □

Remarks: In nondiscrete cases, $B(X^n)$ and $\log 1/p(X^n)$ are not actually codelengths. Nevertheless, the log density ratio $\log p(X^n)/q(X^n)$ in (5.9) does represent the limit, as the quantization regions become vanishingly small, of the log probability ratio $\log P([X^n])/Q([X^n])$. Ignoring the necessary rounding to integer lengths, this log probability ratio is the difference between the codelength $\log 1/Q([X^n])$ and the ideal codelength $\log 1/P([X^n])$.

For quantized data, the redundancy of the minimum two-stage code is the expected value of $(B([X^n]) - \log 1/P([X^n]))/n$ where $B([X^n])$ is the minimum of the codelengths from expression (3.5). In this case, the redundancy is bounded by $R_n^{[\cdot]}(p) = \min_q (L_n(q)/n + D^{[\cdot]}(p\|q))$. Here $D^{[\cdot]}(p\|q) = \sum_A P(A) \log P(A)/Q(A)$ is the discrete relative entropy obtained by summing over sets in the partition formed by the quantization regions. As a consequence of familiar inequality $D^{[\cdot]}(p\|q) \le D(p\|q)$, we have $R_n^{[\cdot]}(p) \le R_n(p)$ uniformly for all quantizations. Consequently, the index of resolvability $R_n(p) = \min(L_n(q)/n + D(p\|q))$ provides a bound on the redundancy that holds uniformly over all quantizations.

The key role of the resolvability for estimation by the minimum description length criterion will be given in Section VII. There it will be shown that $R_n(p)$ bounds the rate of convergence of the density estimator.

VI. Examples of Resolvability

In this section we present bounds on the index of resolvability for various classes of densities. In each case the list Γ is chosen to have information closure which includes the desired class of densities. The bounds on resolvability are obtained with specific choices of $L_n(q)$. Nevertheless, in each case these bounds lead to bounds on the resolvability using $L^*(q)$ (the algorithmic complexity of q). With L^*, the best rates of convergence of the resolvability hold without prior knowledge of the class of densities.

We show that the resolvability $R_n(p)$ is $O(1/n)$ in computable cases, $O((\log n)/n)$ in smooth parametric cases, and $O(1/n)^\gamma$ or $O(((\log n)/n)^\gamma)$ in some nonparametric cases, where $0 < \gamma < 1$.

The bounds on resolvability in these examples are derived in anticipation of the consequences for the rates of convergence of the density estimator (Section VII). We intersperse the examples and resolvability calculations with remarks on the implications for parametric model selection. There is also opportunity to compare some choices of two stage codes in the parametric case using average and minimax criteria involving the index of resolvability.

Case 1) P is computable: $R_n(p) = L(p)/n$ for all large n.

Suppose $L_n(q) = L(q)$ does not depend on n. Let $\tilde{p}_n$ be the density that achieves the best resolution in (5.3). If the density p is on the list Γ, then, for all sufficiently large n,

$$\tilde{p}_n = p \quad (6.1)$$

and

$$R_n(p) = \frac{L(p)}{n}. \quad (6.2)$$

If there is more than one density on the list that is a.e. equal to p, then in (6.1) and (6.2) we take the one for which $L(p)$ is shortest.

To verify (6.1) and (6.2) we first note that for all n, $0 < L(\tilde{p}_n) \le L(p)$ (because any q with $L(q) > L(p)$ results in a higher value of $L(q)/n + D(p\|q)$ than the value $L(p)/n$ that is achieved at $q = p$). Now for small n compared to $L(p)$, densities q that are simpler than p may be preferred. However, for all $n \ge L(p)/D_{\min}$, it

[3] A referee has suggested another relevant notion of redundancy, namely, $E_m(B(X^n) - \log 1/m(X^n))$, where $m(X^n) = \sum_q 2^{-L(q)} q(X^n)$ and the expectation E_m is taken with respect to $m(x^n)$. This measures the average deficiency of the minimal two-stage description compared to the code that is optimal for minimizing the Bayes average description length with prior $w(q) = 2^{-L(q)}$.

must be that $\tilde{p}_n = p$ and $R_n(p) = L(p)/n$ where

$$D_{\min} = \min_q \{D(p\|q): L(q) < L(p)\}. \quad (6.3)$$

Indeed for such n, we observe that for each q with $L(q) < L(p)$, the value of $L(q)/n + D(p\|q)$ is greater than $D_{\min}$ and hence greater than $L(p)/n$, which is the value at p, whence $\tilde{p}_n = p$.

Case 2) $\mathbf{P}$ is in a d-dimensional parametric family

$$R_n(p) \sim (d/2)(\log n)/n.$$

For sufficiently regular parametric families $\{p_\theta: \theta \in \Theta\}$, $\Theta \subset R^d$, there exists Γ_n, L_n and constants c_θ such that for every θ

$$R_n(p_\theta) \le \frac{(d/2)\log n + c_\theta + o(1)}{n}. \quad (6.4)$$

Moreover, for every Γ_n and L_n and for all θ except in a set of Lebesgue measure zero,

$$R_n(p_\theta) \ge (1 - o(1))\frac{(d/2)\log n}{n}. \quad (6.5)$$

The lower bound (6.5) is a consequence of a bound on redundancy proved in Rissanen [9, Theorem 1], and the regularity conditions stated there are required. The upper bound (6.4) is closely related to a result in Rissanen [8, Theorem 1b)] for the redundancy of two-stage codes. Here we derive (6.4) requiring only that Θ be an open set and that for each θ the relative entropy $D(p_\theta\|p_{\tilde{\theta}})$ is twice continuously differentiable as a function of $\tilde{\theta}$ (so that the second order Taylor expansion (6.7) holds). For compact subsets of the parameter space, bounds on the minimax resolvability are also obtained.

First to establish (6.4), we let Γ_n be the set of densities p_θ for which the binary expansions of the parameters terminate in $(1/2)\log n$ bits to the right of the decimal point and we set the corresponding description length to be

$$L_n(p_\theta) = l_{[\theta]} + \frac{d}{2}\log n, \quad (6.6)$$

for $p_\theta \in \Gamma_n$ where $l_{[\theta]}$ denotes the length of a code for the vector of integer parts of the components of θ. Thus Γ_n corresponds to a rectangular grid of parameter values with cells of equal width $\delta = 1/\sqrt{n}$. The choice of δ of order $1/\sqrt{n}$ is seen to optimize the resolvability, which is of order $(-\log\delta)/n + \delta^2$. For $\theta \in \Theta$, the truncation of the binary expansion of the coordinates to $(1/2)\log n$ bits yields an approximation to the density with relative entropy distance of order $1/n$ and a codelength of $l_{[\theta]} + (d/2)\log n$. Consequently, the redundancy satisfies $R_n(p_\theta) \le ((d/2)\log n + O(1))/n$. This verifies (6.4).

This derivation uses the fact that the relative entropy satisfies $D(p_\theta\|p_{\tilde{\theta}}) = O(\|\theta - \tilde{\theta}\|^2)$ as $\tilde{\theta} \to \theta$ for any given θ. Indeed, since $D(p_\theta\|p_{\tilde{\theta}})$ achieves a minimum at $\tilde{\theta} = \theta$, it follows that the gradient with respect to $\tilde{\theta}$ is zero at θ and the second order Taylor expansion is

$$D(p_\theta\|p_{\tilde{\theta}}) = \frac{1}{2}(\theta - \tilde{\theta})^T J_\theta(\theta - \tilde{\theta})\log e + o(\|\theta - \tilde{\theta}\|^2), \quad (6.7)$$

where J_θ is the nonnegative definite matrix of second partial derivatives (with respect to $\tilde{\theta}$) of $E\ln(p_\theta(X)/p_{\tilde{\theta}}(X))$ evaluated at $\tilde{\theta} = \theta$. Although we do not need a further characterization of J_θ here, it is known that under additional regularity conditions J_θ is the Fisher information matrix with entries $-E(\partial^2 \ln p_\theta(X)/\partial\theta_j\partial\theta_k)$.

To optimize the constant c_θ in (6.4) according to average or minimax resolvability criteria, Γ_n should correspond to a nonuniform grid of points to account for the curvature and scaling reflected in J_θ. Assume that J_θ is positive definite. In the Appendix, it is shown that, given $\epsilon > 0$, the best covering of the parameter space (such that for every θ there is a $\tilde{\theta}$ in the net with $(\theta - \tilde{\theta})^T J_\theta(\theta - \tilde{\theta}) \le \epsilon^2$) is achieved by a net having an asymptotic density of $\lambda_d(1/\epsilon)^d \det(J_\theta)^{1/2}$ points per unit volume in neighborhoods of θ, where λ_d is a constant (equal to the optimum density for the coverage of R^d by balls of unit radius). We set $\epsilon = \sqrt{d/n}$, which optimizes the bound on the resolvability. We need a code for the points in the net. It is shown in the Appendix, that if $w(\theta)$ is a continuous and strictly positive prior density on Θ, then the points in the net can be described using lengths $L_n(p_{\tilde{\theta}})$, $p_{\tilde{\theta}} \in \Gamma_n$, such that for any given θ,

$$L_n(p_{\tilde{\theta}}) = \frac{d}{2}\log n + \frac{1}{2}\log\det(J_\theta) + \log\frac{1}{w(\theta)} - \frac{d}{2}\log c_d + o(1), \quad (6.8)$$

where $\tilde{\theta}$ is the point in the net that best approximates θ and $o(1) \to 0$ as $n \to \infty$. Here $c_d = d/(\lambda_d)^{2/d}$ is a constant which is close to $2\pi e$ for large d. In (6.8) the term $\log 1/w(\theta)$ may be regarded as the description length per unit volume, for a small set that contains θ, and the remaining terms account for the log of the number of points per unit volume in this set. Sets Γ_n and codelengths L_n with properties similar to (6.8) are derived in Barron [16] and Wallace and Freeman [13]. The principle difference is that here the codelengths are designed to optimize the resolvability, which involves the expected value of the log-likelihood, whereas in [16] the codelengths are designed to optimize the total description-length based on the sample value. (This accounts for the use of the Fisher information J_θ in (6.8) instead of the empirical Fisher information $\hat{J}$.)

With the given choice of L_n and Γ_n and using $R_n(p_\theta) \le L_n(p_{\tilde{\theta}})/n + D(p_\theta\|p_{\tilde{\theta}})$, we obtain the following bound on the resolvability,

$$R_n(p_\theta) \le \frac{1}{n}\left(\frac{d}{2}\log n + \log\frac{\det(J_\theta)^{1/2}}{w(\theta)} - \frac{d}{2}\log c_d/e + o(1)\right). \quad (6.9)$$

Moreover, it is seen that this bound holds uniformly on compact subsets of the parameter space. For any compact set $B \subset \Theta$, the asymptotic minimax value of the right side of (6.9) is obtained by choosing the prior $w(\theta)$ such that

the bound is asymptotically independent of θ, i.e., we set

$$w(\theta) = \frac{\det(J_\theta)^{1/2}}{c_{J,B}}, \tag{6.10}$$

where $c_{J,B} = \int_B \det(J_\theta)^{1/2} d\theta$. With this choice, it is seen that the minimax resolvability is bounded by

$$R_n \le \frac{1}{n}\left(\frac{d}{2}\log n + \log \int_B \det(J_\theta)^{1/2} d\theta - \frac{d}{2}\log\frac{c_d}{e} + o(1)\right). \tag{6.11}$$

Corresponding to this bound is the choice of a constant codelength

$$L_n(p_{\hat\theta}) = \frac{d}{2}\log n + \log c_{J,B} + \log c_d + \delta_n, \tag{6.12}$$

which is equal to the log of the minimum cardinality of nets that cover B in such a way that for every $\theta \in B$ there is a $\tilde\theta$ in the net with $(\theta-\tilde\theta)^T J_\theta(\theta-\tilde\theta) \le (d/n)$. Here $\lim \delta_n = 0$.

Similar lower bounds on minimax resolvability can be obtained from known lower bounds on minimax redundancy. Indeed, it is shown in Barron and Clarke [26] (with uniformity on compact sets B shown in Clark [27]) that, under suitable regularity conditions, the code that optimizes the average redundancy, i.e., the code based on the density $m(X^n) = \int p_\theta(X^n) w(\theta)\, d\theta$, has asymptotic redundancy given by

$$\Delta_n(p_\theta) = \frac{1}{n}\left(\frac{d}{2}\log n + \log\left((\det J_\theta)^{1/2}/w(\theta)\right) - \frac{d}{2}\log 2\pi e + o(1)\right). \tag{6.13}$$

Consequently the prior in (6.10) yields the asymptotically minimax redundancy as well as bounds on the minimax resolvability. (A similar role for this prior is given in Krichevsky and Trofimov [28] for the special case of the redundancy of codes for the multinomial family.) Note that the expression (6.13) for the redundancy and the bound (6.9) for the resolvability differ in the constant term, but otherwise they are the same. The prior in (6.10), which is defined to be proportional to the square-root of the determinant of the Fisher information matrix, was introduced by Jeffreys [29, pp. 180–181] in another statistical context.

Remarks: Consider the index of resolvability in a model selection context. We are given a list of parametric families from which one is to be selected from the data by the minimum description-length criterion. The previous analysis applies (with slight modification) to bound the index of resolvability in this case. Indeed, let $\{p_\theta^{(k)}\}$, $k = 1, 2, \cdots$ be a list of families, with corresponding sets $\Gamma_n^{(k)}$ and codelengths $L_n^{(k)}(q)$, each of which is designed to satisfy (6.4). In this case $\Gamma_n = \cup_k \Gamma^{(k)}$ is taken to be the union of the sets of candidate densities and $L_n'(q) = L_n^{(k)}(q) + L(k)$ for q in Γ', where k is the index of the family that contains the density q. Here $L(k)$ is chosen to satisfy $\sum_k 2^{-L(k)} \le 1$ so that it is interpretable as a codelength for k. If k^* is the index of the family that contains the true density, then without prior knowledge that this is the right family, we obtain an index of resolvability that differs by only $L(k^*)/n$ when compared to the resolvability attained with true knowledge of the family. Consequently, the index of resolvability remains of order $(\log n)/n$, when the true density is in one of the parametric families, even though the true family is unknown to us.

A related criterion for the selection of parametric models was introduced by Schwarz [30], with a Bayesian interpretation, and by Barron [16] and Rissanen [10], with a minimum two-stage description-length interpretation. In this method the index $\hat k_n$ of the family is chosen to minimize $L(k) + \log 1/m_k(X^n)$, where $m_k(X^n) = \int p_\theta^{(k)}(X^n) w_k(\theta)\, d\theta$ is the marginal density of X^n obtained by integrating with respect to a given prior density $w_k(\theta)$ for the kth family. Schwarz [30] and Rissanen [10] have obtained approximations to the criterion showing that it amounts to the minimization of $(d_k/2)\log n + \log 1/p_\theta^{(k)}(X^n)$ as in the minimum description-length criterion. Here d_k is the dimension of the kth family. A more detailed analysis as in [16], applying Laplace's method to approximate the integral defining $m_k(X^n)$, yields exact asymptotics, including terms involving the prior density and the determinant of the empirical Fisher information matrix. This analysis is the basis for (6.13) as derived in [26]. Moreover, examination of the approximation to the criterion shows that it is very similar to minimum complexity estimation with codelengths $L_n(p_\theta^{(k)})$ approximated as in (6.8).

Case 3) Sequences of parametric families:

$$R_n(p) \le O\left(\frac{(\log n)}{n}\right)^{2r/(2r+1)}.$$

What if the true density is not in any of the finite-dimensional families? We show that for a large nonparametric class of densities, a sequences of parametric families continues to yield a resolvability of order $(d_n/2)(\log n)/n$, except that now the best dimension d_n grows with the sample size.

Consider the class of all densities $p(x)$ with $0 < x < 1$ for which the smoothness condition

$$\int_0^1 \left(\tfrac{d^r}{dx^r}\log p(x)\right)^2 dx < \infty$$

is satisfied for some $r \ge 1$. We find sequences of parametric families with the property that for every such density, the resolvability satisfies

$$R_n(p) \le O\left(\frac{\log n}{n}\right)^{2r/(2r+1)}. \tag{6.14}$$

Moreover, this rate is achieved by minimum complexity density estimation without prior knowledge of the degree of smoothness r.

Consider sequences of exponential families of the form

$$p_\theta^{(d)}(x) = \exp\left(\sum_{j=1}^{d} \theta_j \phi_j(x) - \psi_d(\theta) \right), \quad (6.15)$$

where $\psi_d(\theta) = \log \int_0^1 \exp(\sum_{j=1}^d \theta_j \phi_j(x))\, dx$, $\theta \in R^d$, and $1, \phi_1(x), \cdots, \phi_d(x)$ are orthonormal functions on $L^2[0,1]$ that are chosen to form a basis for polynomials (of degree d), splines (of order $s \geq 1$ with m equally spaced knots and $d = m + s - 1$), or trigonometric series (with a maximal frequency of $d/2$). We focus on the polynomial and spline cases, since the trigonometric case requires that the periodic extension of $\log p(x)$ must also be r-times differentiable for (6.17) below to hold.

In Barron and Sheu, bounds are determined for the relative entropy distances $D(p \| p_{\theta^*}^{(d)})$ and $D(p_{\theta^*}^{(d)} \| p_\theta^{(d)})$, where θ^* in R^d is chosen to minimize $D(p \| p_\theta^{(d)})$. There the bounds are used to determine the rate at which $D(p \| p_{\hat{\theta}}^{(d_n)})$ converges to zero in probability, where $\hat{\theta}$ is the maximum likelihood estimator of the parameter and d_n is a prescribed sequence of dimensions. Here we use the bounds on the relative entropy from Barron and Sheu [31] to derive bounds on the index of resolvability. This bound on the resolvability will lead to the conclusion that, with a sequence of dimensions $\hat{d}_n$ estimated by the minimum description-length criterion, the density estimator converges at rate bounded by $((\log n)/n)^{2r/(2r+1)}$.

Let Γ_n consist of the union for all $d \geq 1$ of the sets of densities $p_\theta^{(d)}$ for which the binary expansion of the coordinates of θ terminate in $(1/2)\log n$ bits to the right of the binary point. Also let $w^d(k_1, \cdots, k_d) = \prod_{j=1}^d w(k_j)$ be a prior for vectors of integers that makes the coordinates independent with a probability mass function $w(k)$, $k = 0, \pm 1, \pm 2, \cdots$. Assume, for convenience, that $w(k)$ is symmetric and decreasing in $|k|$. (Assume other choices for the prior distribution can also be shown to lead to bounds of the desired form.) Then set the codelengths for $p_\theta^{(d)}$ in Γ_n to equal

$$L_n\left(p_\theta^{(d)}\right) = \frac{d}{2}\log n + \log 1/w^d([\theta]) + 2\log d + c. \quad (6.16)$$

Here $\log 1/w^d([\theta])$ is the codelength for the integer part of the parameter vector and $2\log d + c$ is a codelength for the dimension d where $c = \sum_{d=1}^{\infty} d^{-2}$. (In the spline case, if the order s is not fixed, then we add an additional $\log d$ bits for the description of $s \leq d$.)

Note that in this set up, the minimum complexity criterion is used to automatically select a sequence of dimensions $\hat{d}_n$ that provide parsimonious yet accurate density estimates.

In order to verify (6.14) we proceed as follows. Let θ^* in R^d be chosen to minimize $D(p \| p_\theta^{(d)})$, i.e., $p_{\theta^*}^{(d)}$ is that member of the family that provides the best approximation to p in the relative entropy sense. Set $\gamma = \max_x |\log p(x)|$ (which is finite as a consequence of the integrability of the derivative). It is shown in [31] that in the polynomial case and in the spline case (with $r \leq s \leq d$), there exists a constant c (that depends on r, but does not depend on the density p or the dimension d) such that

$$D\left(p \| p_{\theta^*}^{(d)}\right) \leq \frac{ce^\gamma}{d^{2r}} \int (D^r \log p)^2. \quad (6.17)$$

Moreover, there exists a constant γ^* depending on γ such that $\max_x |\log p_{\theta^*}^{(d)}(x)| \leq \gamma^*$ for all large d. For simplicity we assume that γ^* is an integer. By [31, (5.3)] we have for any parameter vector θ that

$$D\left(p_{\theta^*}^{(d)} \| p_\theta^{(d)}\right) \leq \frac{1}{2} e^{\gamma^*} e^{a_d \|\theta^* - \theta\|} \|\theta^* - \theta\|^2 \log e, \quad (6.18)$$

where a_d is a sequence of order $O(d)$ in the polynomial case and $O(\sqrt{d})$ in the spline and trigonometric cases.

Now let θ be chosen to equal θ^* with each coordinate truncated to $(1/2)\log n$ bits accuracy (to the right of the binary point). As a consequence of the inequality $(\theta_1^*)^2 + \cdots + (\theta_d^*)^2 \leq \int (\log p_{\theta^*}^{(d)})^2 \leq (\gamma^*)^2$, it is seen that the integers $[\theta_j]$ are bounded by γ^*. Consequently, from (6.16) the description length for this density is bounded by

$$L_n\left(p_\theta^{(d)}\right) \leq (d/2)\log n + d \log 1/w(\gamma^*) + 2\log d + c. \quad (6.19)$$

With the given choice of θ we have $\|\theta^* - \theta\|^2 \leq d/n$.

Now we combine the bounds from (6.17) and (6.18). It is seen that for any constant c_0, there exist constants c_1 and c_2, such that for all d satisfying $a_d^2 d/n \leq c_0$, the relative entropy distance satisfies

$$\begin{aligned} D\left(p \| p_\theta^{(d)}\right) &= D\left(p \| p_{\theta^*}^{(d)}\right) + D\left(p_{\theta^*}^{(d)} \| p_\theta^{(d)}\right) \\ &\leq c_1 \left(\frac{1}{d}\right)^{2r} + c_2 \frac{d}{n}. \end{aligned} \quad (6.20)$$

The first identity in (6.20) is a Pythagorean-like identity from [31, Lemma 3] that is valid when the family is of the exponential form. As a consequence of this bound, if a sequence of dimensions $d = d_n$ is chosen such that $a_d^2 d/n$ is bounded, then the index of resolvability satisfies

$$\begin{aligned} R_n(p) &\leq \frac{1}{n} L_n\left(p_\theta^{(d)}\right) + D\left(p \| p_\theta^{(d)}\right) \\ &\leq O\left(\frac{d}{n}\log n\right) + O\left(\frac{1}{d}\right)^{2r} + O\left(\frac{d}{n}\right). \end{aligned} \quad (6.21)$$

This bound is optimized with $d = O(n/\log n)^{1/(2r+1)}$ (for which the condition $a_d^2 d/n \leq O(1)$ will be satisfied in the polynomial, spline and trigonometric cases for all $r \geq 1$), which yields

$$R_n(p) \leq O\left(\frac{\log n}{n}\right)^{2r/(2r+1)}. \quad (6.22)$$

Remarks: As a consequence of this bound, using the results of Section VII, it is seen that the minimum complexity density estimator converges in squared Hellinger distance at rate $((\log n)/n)^{2r/(2r+1)}$. Moreover, as previously noted, the minimum complexity criterion automatically chooses an appropriate sequence of dimensions d from the data without knowledge of the degree of smoothness r. In contrast, the rates of convergence of order $n^{-2r/(2r+1)}$ obtained in [31] are for density estima-

tors in families with a sequence of dimensions d of order $n^{1/(2r+1)}$, is preselected with knowledge of the degree of smoothness r.

Therefore, with minimum complexity estimation, we converge at a rate within a logarithmic factor of the rate obtainable with knowledge of the smoothness class of the density. This remains true whether the true density is in a finite- or infinite-dimensional class.

In related contexts of model selection (in particular in the context of selecting the order of a polynomial regression), Shibata [32] and Li [33] have shown that criteria closely related to criteria proposed by Akaike [34] are asymptotically optimal (in the sense that the risk of the estimated model is asymptotically equivalent to the risk achievable by knowledge of the sequence of model dimensions that minimize the risk), provided the true distribution is not in any of the finite-dimensional families; whereas this asymptotic optimality fails for other criteria including the minimum description-length criterion. However, to achieve this optimality property in infinite-dimensional cases, the criteria used by Shibata and Li sacrifices strong consistency in finite-dimensional cases. It is reasonable to conjecture that results similar to those obtained by Shibata carry over to the case of density estimation with sequences of exponential families. Unfortunately, the methodology used by Shibata and Li relies heavily on linearity properties of the models that limit the validity of the criteria.

In contrast, the minimum complexity criterion does not require the candidate parametric models to be approximately linear. We are free to add to the list densities having arbitrary and possibly irregular form, in hopes of obtaining better estimates in some cases, without hurting the bounds on the rates of convergence in the best understood cases.

Concerning splines, we remark that ideally the minimum description-length criterion should be used to select the order s. If instead we fix s, then the above analysis holds only for $r \le s$. With splines of a fixed order, it is not possible to take advantage of smoothness of order $r > s$ to get the faster rates of convergence that are possible with polynomials or variable-order splines.

Histograms, which are piecewise constant density estimators, are a special case of spline models in which the order of the spline is fixed at $s = 1$. Therefore, the results of this section apply to histograms in the case that the minimum description-length criterion is used to select the number of cells. The index of resolvability converges to zero at rate $((\log n)/n)^{2/3}$ for log-densities with at least one square-integrable derivative. Other results that involve the stochastic complexity and the relative entropy in the histogram setting may be found in Hall and Hannon [35], Yu and Speed [36], and Barron, Gyorfi, and van der Meulen [42]. In particular, Yu and Speed [36] demonstrate that the redundancy is $c((\log n)/n)^{2/3}(1 + o(1))$ and explicitly identify the constant c, for a class of universal codes that (as they point out) are closely related to the two-part codes we consider here. A slightly faster convergence rate of order $n^{-2/3}$ is possible for the relative entropy and the redundancy, as shown in [31], [36], and [54] using other histogram-based methods with a predetermined sequence of number of bins. Yu and Speed [36, Theorem 3.1] demonstrate that $n^{-2/3}$ is the optimal redundancy in a minimax setting involving first derivative assumptions on the density function.

Minimum complexity criteria may also be used to select the boundaries of the cells (or more generally to select the locations of the knots for the spline models), leading to improved resolvability in some cases. Nevertheless, equal-spaced boundaries are sufficient to obtain the indicated bounds on the index of resolvability.

Case 4) Fully nonparametric:

$$R_n(p) = O(n^{-2r/(2r+1)}).$$

We show that by a special selection of the set Γ_n, that does not involve the use of a sequence of smooth parametric families, a resolvability of $O((1/n)^{2r/(2r+1)})$ instead of $O((\log n)/n)^{2r/(2r+1)}$ can be attained using assumptions on derivatives of the density up to order r. Moreover, it is shown that $O(n^{-2r/(2r+1)})$ is asymptotically the minimax resolvability as well as being the minimax rate of convergence of density estimators.

First consider the class of density functions p on the unit interval for which the log-density $f(x) = \log p(x)$ is in the Sobolev ball,

$$W_2^r = \Big\{ f \text{ on } [0,1] : -\gamma \le f(x) \le \gamma, \int_0^1 \left(D^k f(x)^2\right) dx \le \gamma^2,\ k = 1,2,\cdots,r \Big\},$$

where γ is an arbitrary positive constant.

The Kolmogorov ϵ-entropy H_ϵ of a set of functions W is the log of the cardinality of the smallest net of functions $\tilde{f}$ such that for every function f in W there is an $\tilde{f}$ with $|f(x) - \tilde{f}(x)| < \epsilon$ for all x (Kolmogorov and Tihomirov [37]). In Birman and Solomjak [38], it is shown that for all sufficiently small ϵ, the ϵ-entropy of the Sobolev ball is bounded by $c(1/\epsilon)^{1/r}$ where c is a constant depending only on γ and r.

Fix an ϵ-net with log-cardinality satisfying $H_\epsilon \le c(1/\epsilon)^{1/r}$. We let Γ_n consist of the densities proportional to $e^{\tilde{f}(x)}$ for $\tilde{f}$ in the net. (Here ϵ will be chosen as a function of n.) Thus each q in Γ_n is of the form $q(x) = e^{\tilde{f}(x) - c_{\tilde{f}}}$ where $c_{\tilde{f}} = \log \int_0^1 e^{\tilde{f}(x)} dx$. Now by [31, Lemma 1], if $\|\cdot\|$ denotes the supremum norm, we have

$$D(p\|q) \le \frac{1}{2} e^{\|f - \tilde{f}\|} \int p(x) \left(f(x) - \tilde{f}(x)\right)^2 dx$$

$$\le \frac{1}{2} e^{\|f - \tilde{f}\|} \|f - \tilde{f}\|^2, \tag{6.23}$$

which is less than $(1/2)e^{\epsilon}\epsilon^2$ by the choice of $\tilde{f}$. Setting

$L_n(q) = H_\epsilon$, we obtain the following bound on the resolvability

$$R_n(p) \le \frac{1}{n} H_\epsilon + \frac{1}{2} e^\epsilon \epsilon^2$$

$$\le \frac{c}{n} \left(\frac{1}{\epsilon}\right)^{1/r} + \frac{1}{2} e^\epsilon \epsilon^2, \qquad (6.24)$$

which holds uniformly for all log-densities in the Sobolev ball. Noting that e^ϵ tends to one for small ϵ, it is readily seen that choosing $\epsilon_n = O(n^{-r/(2r+1)})$ gives the best rate in (6.24). With such a choice we have resolvability bounded by

$$R_n(p) = O(n^{-2r/(2r+1)}), \qquad (6.25)$$

uniformly for all log-densities in the Sobolev ball, for all large n.

By adding description-length terms for r and for γ, we may use the minimum complexity criterion to automatically select a suitable Sobolev ball from the data. The indicated rate on the index of resolvability will hold without prior knowledge of the best smoothness class.

Similar results for the index of resolvability can be obtained in the case of Sobolev conditions imposed on the density itself (instead of the log-density), assuming that the density function is bounded away from zero. Indeed, let $W_{2,+}^r = \{f \in W_2^r: f(x) \ge 1/\gamma\}$, $\gamma > 1$, for which the ϵ-entropy must have the same bound $H_\epsilon \le c(1/\epsilon)^{1/r}$, let Γ_n be the set of probability density functions proportional to $\tilde{f}(x)$ for $\tilde{f}$ in the ϵ-net of $W_{2,+}^r$, and let $L_n(q)$ be the log of the cardinality of this net. Each q in Γ_n is of the form $q(x) = \tilde{f}(x)/c_{\tilde{f}}$ where now $c_{\tilde{f}} = \int_0^1 \tilde{f}(x)\,dx \le \gamma$ and $q(x) \ge 1/\gamma^2$. Using inequalities between the relative entropy and Chi-square distance ($D(p\|q) \le \int (p-q)^2/q \le \int (p-cq)^2/q$ for $c > 0$), which may be deduced as in [31, Section 3], it follows that for probability density functions $p(x) = f(x)$ in $W_{2,+}^r$, we have $D(p\|q) \le \gamma^2 \int_0^1 (f(x) - \tilde{f}(x))^2\,dx$, which is less than $\gamma^2\epsilon^2$ by suitable choice of $\tilde{f}$ in the ϵ-net. As in the previous case it follows that the index of resolvability satisfies

$$R_n(p) \le \frac{1}{n} H_\epsilon + \gamma^2 \epsilon^2,$$

and optimizing the choice of ϵ yields $R_n(p) = O(n^{-2r/(2r+1)})$ as before.

As a consequence of this bound on the index of resolvability (and by application of Theorem 4, Section VII), we see that the minimum complexity density estimator, specialized to the current case, converges to the true density in squared Hellinger distance at rate $n^{-2r/(2r+1)}$, uniformly for all densities in the Sobolev class $W_{2,+}^r$. Now when both p and q are bounded and bounded away from zero (here $1/\gamma \le p(x) \le \gamma$ and $1/\gamma^2 \le q(x) \le \gamma^2$) the squared Hellinger distance, the relative entropy and the integrated squared error are equivalent to within a constant factor: indeed, $\int (\sqrt{p} - \sqrt{q})^2 \le D(p\|q) \le \int (p-q)^2/q \le \gamma^2 \int (p-q)^2 \le 4\gamma^4 \int (\sqrt{p} - \sqrt{q})^2$. It follows that the density estimator also converges in relative entropy and integrated squared error at rate $n^{-2r/(2r+1)}$ uniformly for densities in the Sobolev class. Now this rate is known to be asymptotically minimax for the integrated squared error for densities in $W_{2,+}^r$ (see Bretagnolle and Huber [40], Efroimovich and Pinsker [41]); also, it is the minimax rate for the redundancy (formulated as a cumulative relative entropy) as recently shown in Yu and Speed [36]. It follows therefore that $n^{-2r/(2r+1)}$ is also the minimax rate for the index of resolvability of densities in this space. (Indeed, any faster uniform convergence of the resolvability would yield a faster convergence of the density estimator resulting in a contradiction.)

Other classes of functions may be considered for which if the density functions in the class are bounded away from zero by an amount γ, then the metric entropy H_ϵ is known. By the same argument, the resolvability of densities in the class by densities in the ϵ-net automatically satisfies

$$R_n(p) \le \frac{H_\epsilon}{n} + \frac{1}{2} e^\epsilon \epsilon^2. \qquad (6.26)$$

For each such class of functions, optimization of the choice ϵ leads to a rate of convergence for the index of resolvability.

Minimum complexity estimation with the ϵ-net of functions is analogous to Grenander's method of sieve estimation [39]. The important difference is that with minimum complexity estimation we can automatically estimate the sieve of the best granularity. Moreover, with the index of resolvability we have bounds on the rate of convergence of the sieve estimator.

The Kolmogorov metric entropy has also been used by Yatracos [42] to obtain rates of convergence in L^1 for a different class of density estimators. However, it is not known to us whether the metric entropy has previously been used to give bounds on redundancy for universal codes. The new ideas here are the relationships between redundancy, resolvability, and rates of convergence of minimum complexity estimators.

Remarks: In the Examples 2, 3, and 4, we permitted the lengths $L_n(q)$ to depend on the given sample size. Nevertheless, by paying a price of order $(\log\log n)/n$, comparable resolvability can be achieved using lengths $L'(q)$ which do not depend on n. The advantage is that the growth and domination conditions (7.3), (7.4), and (7.6) which are used in Theorems 1, 2, and 3 will then be satisfied. To construct such an assignment of description lengths $L'(q)$, we first note that positive integers k can be encoded using $2\log k + c$ bits where c is a constant. Given Γ_n and $L_n(q)$ for $n = 1, 2, \cdots$, define a new list $\Gamma' = \bigcup_k \Gamma_{2^k}$ to be the union of the sets for indices equal to powers of two and define, for $q \in \Gamma'$,

$$L'(q) = L_{n_k}(q) + 2\log\log n_k + c, \qquad (6.27)$$

with $n_k = 2^k$, where k is the first index such that $q \in \Gamma_{2^k}$. It is seen that L' satisfies Kraft's inequality on Γ'. With L' in place of L_n we achieve resolvability satisfying $R_n'(p) \le ((d/2)\log n + 2\log\log n + O(1))/n$ in the parametric case. In general, since between the powers of two the

resolvability $R_n'(p) = \min(L'(q)/n + D(p\|q))$ is never more than twice the resolvability at the next power of two, we conclude that $R_n'(p) \le O(R_{n'}(p) + (\log\log n')/n')$, where $n' = 2^{\lceil \log n \rceil}$. In particular, when $R_n(p)$ is of larger order than $(\log\log n)/n$, the overall rate is unaffected by the addition of the $(\log\log n)/n$ term, so it follows that $R_n'(p) = O(r_{n'}(p))$.

For practical estimation of a density function, we are more inclined to use sequences of parametric families as in Case 3, instead of using the "fully nonparametric" estimators as in Case 4, despite the fact that for a large class of functions the index resolvability tends to zero at a slightly faster rate in Case 3. There are two reasons for this. Firstly, the metric entropy theory does not provide an explicit choice for the net of density functions with which we can compute. Secondly, with sequences of parametric families, while converging at a nearly optimal rate even in the infinite-dimensional case, we retain the possibility of delight in the discovery of the correct family in the finite-dimensional case.

VII. The Convergence Results

In this section we present our main theorems establishing convergence of the sequence of minimum complexity density estimators. The first three theorems concern the statistical consistency of the estimators. Bounds on rates of convergence are given in Theorem 4 and its corollary.

Conditions: For each of the results, one or more of the following conditions are assumed. Given a sequence of lists Γ_n and numbers $L_n(q)$ for densities q in Γ_n, let $\Gamma = \bigcup_n \Gamma_n$. Set $L_n(q) = \infty$ for g not in Γ_n.

Summability: There exists a constant $b > 0$ such that

$$\sum_{g \in \Gamma_n} 2^{-L_n(q)} \le b, \qquad \text{for all } n. \tag{7.1}$$

Light tails: There exist constants $0 < \alpha < 1$ and b' such that

$$\sum_{q \in \Gamma_n} 2^{-\alpha L_n(q)} \le b', \qquad \text{for all } n. \tag{7.2}$$

Growth restriction:

$$\limsup_n \frac{L_n(q)}{n} = 0, \qquad \text{for every } q \in \Gamma. \tag{7.3}$$

Nondivergence:

$$\limsup_n L_n(q) < \infty, \qquad \text{for every } q \in \Gamma. \tag{7.4}$$

Nondegeneracy:

$$L_n(q) \ge l, \quad \text{for all } q \in \Gamma_n \text{ and all } n, \text{ for some constant } l > 0. \tag{7.5}$$

Domination: There exists $L(q)$, $q \in \Gamma$ and a constant c such that

$$L(q) \le L_n(q) + c, \quad \text{for all } q \text{ and all } n \text{ and } \sum_q 2^{-L(q)} \le 1. \tag{7.6}$$

Remarks Concerning the Conditions: The main condition for all of our results is the summability condition (7.1). It is implied by Kraft's inequality in the data compression framework or it is implied by the requirement of a proper prior in the Bayesian framework. This condition (or the closely related condition (7.6)) is used to obtain the results of Theorems 1 and 2 on the consistency of the estimator of the distribution. The somewhat more stringent assumption (7.2) is used to get the rate of convergence results for the estimator of the density. The Corollary to Theorem 4 shows how this more stringent condition can be circumvented by restricting the minimization to densities that are not excessively complex.

Either the growth restriction (7.3) or the boundedness (7.4) is used with the almost sure results (Theorems 1, 2, 3), but they are not needed for the main result (Theorem 4) on the rate of convergence in probability. For condition (7.5), the constant l can be taken to equal 1 when the lengths $L_n(q)$ are positive integers.

For given $L(q)$ and Γ that do not depend on n, if $\sum_q 2^{-L(q)} \le 1$ and if Γ contains more than one point, then all of these conditions are satisfied except perhaps for the tail condition (7.2). A modified criterion with $\lambda L(q)$ used in place of $L(q)$, where $\lambda > 1$ is a constant, is seen to satisfy all of the conditions, provided $\sum_q 2^{-L(q)} \le 1$. In particular (7.2) will hold with $\alpha = 1/\lambda$. Note that this modification will not increase the index of resolvability by more than the factor λ. In particular $R_n(p)$ will have the same rates of convergence.

For the case of complexity constrained maximum likelihood estimators in Cover [18], the density estimate $\hat{p}_n$ is selected by maximizing the likelihood in Γ_n, where $\Gamma_1, \Gamma_2, \cdots$ is an increasing sequence of collections of densities. This is a special case of minimum complexity density estimation with $L_n(q)$ set to a constant on Γ_n. We impose the cardinality restriction $\log\|\Gamma_n\| = o(n)$. In this case we set $L_n(q) = 2\log\|\Gamma_n\|$ for $q \in \Gamma_n$ and ∞ otherwise. Then conditions (7.1), (7.2), (7.3), and (7.5) are satisfied, so all of the convergence results except Theorem 1 hold in this case. Even if the collections Γ_n are not increasing, the conditions are still satisfied for Theorem 4. The proofs of the theorems are in Section VIII.

Let $X_1, X_2, \cdots$ be independent and identically distributed with probability density function $p(x)$. Let $\hat{p}_n$ be the minimum complexity density estimate defined by (3.3). Thus $\hat{p}_n$ achieves

$$\min_{q \in \Gamma_n} \left(L_n(q) + \log 1/q(X^n)\right). \tag{7.7}$$

Theorem 1 (Discovery of the true density): Assume L_n satisfies the nondivergence condition (7.4) and the domination condition (7.6). If

$$p \in \Gamma, \tag{7.8}$$

then

$$\hat{p}_n \equiv p, \tag{7.9}$$

for all sufficiently large n, with probability one.

Thus, in the important case that $\Gamma = \Gamma^*$, if the data are governed by a computable law then this law eventually

will be discovered and thereafter never be refuted. However, although the estimator eventually will be precisely correct, it is never known for any given sample size whether the true density has been discovered.

Next we present convergence properties that do not require that the true density be in Γ. It is assumed to be an information limit of such densities. The next result establishes convergence of the estimated distributions.

Theorem 2 (Consistency of the minimum complexity estimator of the distribution): Assume L_n satisfies the summability condition (7.1) and the growth restriction (7.3). If $p \in \Gamma$, then for each measurable set S,

$$\lim_{n\to\infty} \hat{P}_n(S) = P(S) \qquad \text{with probability one.} \quad (7.10)$$

Assuming that X is a separable Borel space, it follows that, with probability one,

$$\hat{P}_n \Rightarrow P \quad (7.11)$$

in the sense of weak convergence.

In Barron [43] a technique is developed that shows convergence of a sequence of distance functions stronger than distances corresponding to weak convergence but not as strong as convergence in total variation. See the remark following the proof in Section VIII.

The next two results show convergence of the density estimates in L^1 and hence convergence of the distributions in total variation. However, the stronger summability condition (7.2) is required.

Theorem 3 (Consistency of the minimum complexity estimator of the density): Assume L_n satisfies the tail condition (7.2) and the growth restriction (7.3). If $p \in \bar{\Gamma}$, then with probability one,

$$\lim_{n\to\infty} \int |p - \hat{p}_n| = 0 \quad (7.12)$$

and

$$\lim_{n\to\infty} \frac{L_n(\hat{p}_n)}{n} = 0. \quad (7.13)$$

Let $d_H^2(p,q) = \int(\sqrt{p} - \sqrt{q})^2$ denote the Hellinger distance. Convergence of densities in L^1 distance and convergence in the Hellinger distance are equivalent as is evident from the following equalities (Pitman [44, p. 7])

$$d_H^2(p,q) \le \int |p-q| \le 2 d_H(p,q). \quad (7.14)$$

The Hellinger distance is also related to the entropy distance. Indeed $d_H^2(p,q) \le \int p \ln p/q$ and if $p(x)/q(x) \to 1$ in sup norm, then

$$d_H^2(p,q) \sim \frac{1}{2}\int p \ln \frac{p}{q} \quad (7.15)$$

in the sense that the ratio of the two sides converges to one.

For sequences of positive random variables Y_n, the notation $Y_n \lesssim R_n$ *in probability* is used to denote convergence in probability at the indicated rate. This means that the ratio Y_n/R_n is bounded in probability, i.e., for every $\epsilon > 0$, there is a $c > 0$, such that $P\{Y_n/R_n > c\} \le \epsilon$ for all large n.

The following result relates the accuracy of the density estimator to the information-theoretic resolvability. It is this result that demonstrates the importance of the index of resolvability for statistical estimation by the minimum-description length principle.

Theorem 4 (Convergence rates bounded by the index of resolvability): Assume L_n satisfies the tail condition (7.2) and the nondegeneracy condition (7.5). If $\lim R_n(p) = 0$, then $\hat{p}_n$ converges to p in Hellinger distance with rate bounded by the resolvability $R_n(p)$, i.e.,

$$d_H^2(p, \hat{p}_n) \lesssim R_n(p) \qquad \text{in probability.} \quad (7.16)$$

Moreover,

$$\frac{L_n(\hat{p}_n)}{n} \lesssim R_n(p) \qquad \text{in probability.} \quad (7.17)$$

Remarks: The conclusion (7.16) has recently been strengthened in [17] (building on the proof technique developed here in Section VIII), to yield that for all $n \ge 1$,

$$E\big(d_H^2(p, \hat{p}_n)\big) \le c R_n(p),$$

where c is a constant. A bound on the constant obtained in [17] is $(2 + 4(1 + (b + e^{-1})/l)/(1-\alpha))/\log e$.

A consequence of Theorem 4 for the classes of densities considered in Section VI, is that the minimum complexity density estimators converge at rate $1/n$, $(\log n)/n$, $((\log n)/n)^{2r/(2r+1)}$ or $n^{-2r/(2r+1)}$, respectively. To obtain these rates, the lengths $L_n(q)$ used in Section VI are replaced by $\lambda L_n(q)$ where $\lambda > 1$, so that the tail condition (7.2) is satisfied, or we use the modification indicated below.

If weights $2^{-L_n(q)}$ are summable but do not satisfy the tail condition, we show how a slight modification results in a convergent density estimator. Fix $\lambda > 1$ (in particular, we suggest $\lambda = 2$), and let $\hat{L}_n^{(\lambda)}$ be the value of $L_n(q)$ for a density that achieves $\min_q(\lambda L_n(q) + \log 1/q(X^n))$. Now define $\hat{\hat{p}}_n$ to be the density that achieves the minimum of $L_n(q) + \log 1/q(X^n)$ subject to the constraint that $L_n(q) \le 2\hat{L}_n^{(\lambda)}$. Thus

$$\hat{\hat{p}}_n = \arg \min_{q:\ L_n(q) \le 2\hat{L}_n^{(\lambda)}} \big(L_n(q) + \log 1/q(X^n)\big), \quad (7.18)$$

where ties are broken in the same way as for $\hat{p}_n$ (by choosing a minimizing density with least $L_n(q)$). Here the constant 2 could be replaced by any constant $c > 1$.

Observe that if the minimum complexity density estimate $\hat{p}_n$ has length $L_n(\hat{p}_n)$ less than $2\hat{L}_n^{(\lambda)}$, then the resulting estimate is unchanged, i.e., $\hat{\hat{p}}_n = \hat{p}_n$. The intention of the modification is to change the estimate only when unconstrained use of the criterion would result in a density with complexity $L_n(\hat{p}_n)$ much larger than the complexity of densities that optimize the resolvability.

Corollary to Theorem 4: Suppose L_n satisfies the summability condition (7.1) and the nondegeneracy condition (7.5). Let $\hat{\hat{p}}_n$ be defined by (7.18). Then

$$d_H^2\big(p, \hat{\hat{p}}_n\big) \lesssim R_n(p) \qquad \text{in probability.} \quad (7.19)$$

Moreover,

$$\frac{L_n(\hat{p}_n)}{n} \le R_n(p) \qquad \text{in probability.} \tag{7.20}$$

VIII. Proofs

The minimization of $L_n(q)+\log 1/q(X^n)$ is the same as the maximization of $q(X^n)2^{-L_n(q)}$. We shall find it mathematically convenient to treat the problems from the perspective of maximizing $q(X^n)2^{-L_n(q)}$.

A tool we will use repeatedly in the proofs is Markov's inequality applied as in Chernoff [45] to yield the following inequalities:

$$P\{p(X^n) \le cq(X^n) \text{ and } X^n \in B\} \le cQ\{p(X^n) \le cq(X^n) \text{ and } X^n \in B\}, \tag{8.1}$$

for any measurable set B in X^n and any constant $c>0$, in particular

$$P\{p(X^n) \le cq(X^n)\} \le c, \tag{8.2}$$

and, in the same manner,

$$\begin{aligned} P\{p(X^n) &\le cq(X^n)\} \\ &= P\{(p(X^n))^{1/2} \le c^{1/2}(q(X^n))^{1/2}\} \\ &\le \rho^n c^{1/2} \\ &\le 2^{-nd^2(p,q)/2}c^{1/2}, \end{aligned} \tag{8.3}$$

where $\rho = \int (pq)^{1/2}$ and $d(p,q)$ is defined to be a multiple of the Hellinger distance

$$d^2(p,q) = \int \left(\sqrt{p}-\sqrt{q}\right)^2 \log e. \tag{8.4}$$

The inequality $\log \rho \le -d^2/2$ follows from $(1/2)\int(\sqrt{p}-\sqrt{q})^2 = 1-\rho$ and $\log \rho \le (\rho-1)\log e$. The factor of $\log e$ in (8.4) is chosen for convenience so that all exponents in (8.3) are base 2.

We note here that these inequalities are applied in each case with c proportional to $2^{-L_n(q)}$. The summability of the resulting bounds in (8.1) and (8.2), summing over q in Γ_n, is key to the proof of consistency of the minimum complexity estimator. The presence of the fractional power of c in the bound (8.3) forces more stringent summability hypotheses to be imposed to get the rate of convergence results.

Proof of Theorem 1: We are to show that

$$P\{\hat{p}_n \ne p \text{ infinitely often}\} = 0. \tag{8.5}$$

For a decreasing sequence of sets, the probability of the limit is the limit of the probabilities. Thus

$$P\{\hat{p}_n \ne p \text{ infinitely often}\} = \lim_{k\to\infty} P\{\hat{p}_n \ne p \text{ for some } n \ge k\}. \tag{8.6}$$

For $\hat{p}_n$ to not equal p, it is necessary that $p(X^n)2^{-L_n(p)} \le q(X^n)2^{-L_n(q)}$ for some $q \ne p$. Consequently, by the union of events bound,

$$\begin{aligned} P\{\hat{p}_n \ne p &\text{ for some } n \ge k\} \\ &\le \sum_q P\{p(X^n)2^{-L_n(p)} \le q(X^n)2^{-L_n(q)} \text{ for some } n \ge k\} \\ &= \sum_q P(A_k^{(q)}), \end{aligned} \tag{8.7}$$

where the sum is for q in Γ with $q \ne p$. Here $A_k^{(q)}$ is the event that $p(X^n)2^{-L_n(p)} \le q(X^n)2^{-L_n(q)}$ for some $n \ge k$. We will show that the probabilities $P(A_k^{(q)})$ are dominated by a summable bound $2^{-L(q)+c+c_0}$ and that they converge to zero as $k\to\infty$ for each q.

First we show the domination. To exclude small n for which $L_n(p)$ may be infinite, we use condition (7.4) to assert that given p there exists c_0 and k_0 such that $L_n(p) \le c_0$ for all $n \ge k_0$. Consider $k \ge k_0$. Momentarily fix q. The event $A_{k_0}^{(q)}$ is a disjoint union of the events A_{n,k_0} that $p(X^n)2^{-L_n(p)} \le q(X^n)2^{-L_n(q)}$ occurs for the first time at n (i.e., the opposite inequality obtains for $k_0 \le n' < n$). Then, by inequality (8.1) and condition (7.6),

$$\begin{aligned} P(A_k^{(q)}) &\le P(A_{k_0}^{(q)}) \\ &= \sum_{n=k_0}^{\infty} P(A_{n,k_0}) \\ &\le \sum_{n=k_0}^{\infty} Q(A_{n,k_0})2^{-L_n(q)+L_n(p)} \\ &\le \sum_{n=k_0}^{\infty} Q(A_{n,k_0})2^{-L(q)+c+c_0} \\ &\le 2^{-L(q)}2^{c+c_0}. \end{aligned} \tag{8.8}$$

This bound is summable for q in Γ, so it gives the desired domination.

Now we show convergence of the probabilities $P(A_k^{(q)})$ to zero as $k\to\infty$. By inequality (8.3), the event $\{p(X^n)2^{-L_n(p)} \le q(X^n)2^{-L_n(q)}\}$ has probability bounded by $2^{-nd^2(p,q)}2^{c_0}/2$ that is exponentially small. Whence by the Borel–Cantelli Lemma, $P(A_k^{(q)})$ tends to zero for each $q \ne p$.

By the dominated convergence theorem, as $k\to\infty$, the limit of the sum in (8.7) is the same as the sum of the limits. Consequently,

$$P\{\hat{p}_n \ne p \text{ infinitely often}\} = 0.$$

This completes the proof of Theorem 1. □

Remark: Two other proofs of this theorem can be found in Barron [16], based on martingale convergence theory. The present proof shares the greatest commonality with the developments forthcoming.

For the proofs of Theorems 2 and 3 we will use the following.

Lemma 1: Suppose L_n satisfies the growth restriction (7.3). If $p \in \overline{\Gamma}$ then for any $\epsilon > 0$, if $\tilde{p} \in \Gamma$ satisfies $D(p\|\tilde{p}) < \epsilon$, then

$$2^{-L_n(\tilde{p})}\tilde{p}(X^n) \ge p(X^n)2^{-n\epsilon}, \qquad \text{for all large } n, \tag{8.9}$$

with probability one. Moreover, for any positive sequence

c_n for which $\lim c_n/n = 0$, the left side of (8.9) exceeds the right side by at least the factor 2^{c_n}, for all large n, with probability one.

Proof of Lemma 1: Taking the logarithm and dividing by n, the desired inequality (8.9) is seen to be the same as

$$\frac{L_n(\tilde{p})}{n} + \frac{1}{n}\log\frac{p(X^n)}{\tilde{p}(X^n)} < \epsilon, \qquad \text{for all large } n, \quad (8.10)$$

with probability one. This is true by application of (7.3) and the strong law of large numbers. The second claim follows in the same way. □

Remark: We note that the left side of (8.10) is an upper bound to the pointwise redundancy per sample defined by $(B(X^n) - \log 1/p(X^n))/n$ (compare with 5.7). Thus a consequence of the Lemma is the following.

Corollary to Lemma 1: If (7.3) is satisfied and if $p \in \bar{\Gamma}$, then the pointwise redundancy per sample $(B(X^n) - \log 1/p(X^n))/n$ converges to zero with probability one.

Proof of Theorem 2: We are to show that if $p \in \bar{\Gamma}$, then

$$\lim_{n\to\infty} \hat{P}_n(S) = P(S) \qquad \text{with probability one,}$$

for arbitrary measurable subsets S in X. Toward this end, given any $\delta > 0$, choose $0 < \epsilon < \delta/2$ and choose $\tilde{p} \in \Gamma$ such that $D(p\|\tilde{p}) < (1/2)\epsilon^2 \log e$. Then $|P(S) - \tilde{P}(S)| < (1/2)\int |p - \tilde{p}| < (1/2)\epsilon$. From Lemma 1 we have

$$2^{-L_n(\tilde{p})}\tilde{p}(X^n) > p(X^n)e^{-n\epsilon^2/2}, \qquad \text{for all large } n, \quad (8.11)$$

with probability one.

Let $N(S, X^n) = \sum_{i=1}^n 1_{\{X_i \in S\}}$ be the number of observations in S. Then $N(S, X^n)$ has a binomial $(n, P(S))$ distribution when the X_i are independent with distribution P, whereas it would have a binomial $(n, Q(S))$ distribution if the X_i were independent with distribution Q. Define the set

$$B_n = \left\{ \left| \frac{N(S, X^n)}{n} - P(S) \right| < \epsilon \right\}. \quad (8.12)$$

Then the Hoeffding [46] or by standard type-counting arguments in information theory, $P(B_n^c) \le 2e^{-n2\epsilon^2}$ and $Q(B_n) \le e^{-n(\delta-\epsilon)^2/2}$ uniformly for all Q with $|Q(S) - P(S)| \ge \delta$, where B_n^c denotes the complement of the event B_n. (Thus (8.12) defines the acceptance region of a test for P versus $\{Q: |Q(S) - P(S)| \ge \delta\}$ that has uniformly exponentially small probabilities of error, [43].)

We want to show that with high probability

$$p(X^n)e^{-n\epsilon^2/2} > \max_q q(X^n)2^{-L_n(q)}, \quad (8.13)$$

where the maximum is for all q in Γ_n with $|Q(S) - P(S)| \ge \delta$. Let A_n be the event that (8.13) does not occur: this is a union of the events $A_n^{(q)}$ defined by

$$A_n^{(q)} = \left\{ p(X^n)e^{-n\epsilon^2/2} \le q(X^n)2^{-L_n(q)} \right\}. \quad (8.14)$$

To bound the probability of A_n we use the union of events bound and (8.1) to obtain

$$\begin{aligned} P(A_n) &\le P(A_n \cap B_n) + P(B_n^c) \\ &\le \sum_q P\left(A_n^{(q)} \cap B_n\right) + P(B_n^c) \\ &\le \sum_q 2^{-L_n(q)} e^{n\epsilon^2/2} Q\left(A_n^{(q)} \cap B_n\right) + P(B_n^c) \\ &\le \sum_q 2^{-L_n(q)} e^{n\epsilon^2/2} e^{-n(\delta-\epsilon)^2/2} + P(B_n^c) \\ &\le be^{-nr} + e^{-n\epsilon^2/2}, \end{aligned} \quad (8.15)$$

where the sum is for all q in Γ_n with $|Q(S) - P(S)| \ge \delta$. Here $r = ((\delta - \epsilon)^2 - \epsilon^2)/2$, which is strictly positive by the choice of ϵ. Thus $P(A_n)$ is exponentially small. Using the Borel–Cantelli lemma and combining (8.11) with (8.13) we have

$$2^{-L_n(\tilde{p})}\tilde{p}(X^n) > \max_q q(X^n)2^{-L_n(q)}, \qquad \text{for all large } n, \quad (8.16)$$

with probability one, where the maximum is for all q in Γ_n with $|Q(S) - P(S)| \ge \delta$. Thus there exists densities in Γ_n with $|Q(S) - P(S)| < \delta$ that have a larger value for $q(X^n)2^{-L_n(q)}$ than all q with $|Q(S) - P(S)| \ge \delta$. Consequently, the minimum complexity estimator, which is defined to achieve the overall maximum, must satisfy

$$\left|\hat{P}_n(S) - P(S)\right| < \delta, \qquad \text{for all large } n, \quad (8.17)$$

with probability one. Since $\delta > 0$ is arbitrary, it follows that, with probability one,

$$\lim_{n\to\infty} \hat{P}_n(S) = P(S),$$

for any measurable set S in X. Consequently, for any countable collection G of sets, we have

$$P\left\{ \lim_{n\to\infty} \hat{P}_n(S) = P(S), \text{ for all } S \in G \right\} = 1. \quad (8.18)$$

Assuming that X is a separable Borel space (e.g., the real line), it follows that there exists a countable collection of sets that generates the Borel sigma-field. Applying (8.18) to this countable collection, it follows that

$$P\left\{\hat{P}_n \Rightarrow P\right\} = 1, \quad (8.19)$$

where $\Rightarrow$ denotes weak convergence. □

Remark: A similar proof using more elaborate hypothesis tests, as in Barron [43], shows that

$$\lim_{n\to\infty} \sum_{S \in \pi_n} \left|\hat{P}_n(S) - P(S)\right| = 0 \qquad \text{with probability one,} \quad (8.20)$$

for any sequence of partitions π_n of X for which the effective cardinality is of order $O(n)$.

Proof of Theorem 3: Here we show almost sure convergence of the minimum complexity density estimate, in Hellinger distance, and almost sure convergence of $L_n(\hat{p}_n)/n$, for weights $2^{-L_n(q)}$ that satisfy the tail condition (7.2).

Note that since $\sum 2^{-\alpha L_n(q)}$ is decreasing in α, condition (7.2) is unchanged if it is assumed that $1/2 \le \alpha < 1$. Given $\delta > 0$ and $1/2 \le \alpha < 1$, set $0 < \epsilon < \delta(1-\alpha)$. For $p \in \bar{\Gamma}$ there exists a density $\tilde{p} \in \Gamma$ with $D(p\|\tilde{p}) < \epsilon$ so that $d^2(p,\tilde{p}) < \epsilon < \delta$. Then by Lemma 1,

$$2^{-L_n(\tilde{p})}\tilde{p}(X^n) > p(X^n)2^{-n\epsilon}, \qquad \text{for all large } n, \quad (8.21)$$

with probability one. Consequently, to show that $d^2(p,\hat{p}_n) < \delta$ and $L_n(\hat{p}_n) < n\delta$, it is enough to show that

$$p(X^n)2^{-n\epsilon} > \max_q q(X^n)2^{-L_n(q)}, \qquad \text{for all large } n, \quad (8.22)$$

with probability one, where the maximum is for all q with $d^2(q,p) \ge \delta$ or $L_n(q) \ge n\delta$. Using the Borel–Cantelli lemma and the union of events bound, it is enough to show that the following sum is exponentially small:

$$\sum_q P\{p(X^n)2^{-n\epsilon} \le q(X^n)2^{-nL_n(q)}\}, \quad (8.23)$$

where the sum is for all q with $d^2(q,p) \ge \delta$ or $L_n(q) \ge n\delta$. For the terms in the sum with $L_n(q) \ge n\delta$ we use the upper bound from (8.2):

$$2^{-L_n(q)}2^{n\epsilon} \le 2^{-\alpha L_n(q)}2^{-n(\delta(1-\alpha)-\epsilon)}. \quad (8.24)$$

These terms have a sum less than $b2^{-n(\delta(1-\alpha)-\epsilon)}$, which is exponentially small by the choice of ϵ. For the terms in the sum with $d^2(p,q) \ge \delta$ we use (8.2) and (8.3) to obtain the upper bound:

$$\begin{aligned}&\min\{2^{-L_n(q)}2^{n\epsilon}, 2^{-L_n(q)/2}2^{-n(\delta-\epsilon)/2}\}\\ &\quad\le (2^{-L_n(q)}2^{n\epsilon})^{2\alpha-1}(2^{-L_n(q)/2}2^{-n(\delta-\epsilon)/2})^{2(1-\alpha)}\\ &\quad= 2^{-\alpha L_n(q)}2^{-n(\delta(1-\alpha)-\epsilon\alpha)}, \end{aligned} \quad (8.25)$$

where we have used the fact that $\min\{c_1,c_2\} \le c_1^\beta c_2^{1-\beta}$ for $0 \le \beta \le 1$ and any positive c_1, c_2. This bound also has a sum less than $b'2^{-n(\delta(1-\alpha)-\epsilon)}$ that is exponentially small.

Therefore, (8.22) is established. From (8.21) and (8.22) we deduce that all maximizers $\hat{p}_n$ of $q(X^n)2^{-L_n(q)}$ must satisfy

$$d^2(\hat{p}_n,p) < \delta \quad \text{and} \quad \frac{L_n(\hat{p}_n)}{n} < \delta, \qquad \text{for all large } n, \quad (8.26)$$

with probability one. Here $\delta > 0$ is arbitrary. Consequently,

$$\lim_{n\to\infty} d^2(\hat{p}_n,p) = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{L_n(\hat{p}_n)}{n} = 0,$$

with probability one. □ (8.27)

Remark: If $c_n > 0$ is any sequence with $\lim c_n/n = 0$, then by the same reasoning, with the second claim of Lemma 1 used in place of (8.21), it is seen that the value of $q(X^n)2^{-L_n(q)}$ at $\tilde{p}$ will exceed the maximum value for all q with $d^2(q,p) \ge \delta$ or $L_n(q) \ge n\delta$ by at least the factor 2^{c_n} for all large n, with probability one. Consequently, every density that achieves within c_n of the minimum two-stage description, will simultaneously satisfy (8.26) for all large n, with probability one. That is, they are all close to the true density p, and none of them has complexity larger than $n\delta$.

The following result will be useful in the proof of Theorem 4.

Lemma 2: Let p and q be any two probability density functions on X and let $X_1, \cdots, X_n$ be independent random variables with density p or q. Then

$$P\{X^n \in B\} \le Q\{X^n \in B\}2^{nr} + \frac{D(p\|q)}{r} + \frac{1}{nr}\frac{\log e}{e}, \quad (8.28)$$

for all measurable subsets B of X^n, all $r > 0$ and all n.

Proof: The inequality is trivial if $D(p\|q)$ is infinite. Now suppose $D(p\|q)$ is finite. Let $A_n = \{p(X^n) \le q(X^n)2^{nr}\}$ and $B_n = \{X^n \in B\}$, then as in (8.1),

$$\begin{aligned}P(B_n) &\le P(A_n \cap B_n) + P(A_n^c)\\ &\le Q(B_n)2^{nr} + P(A_n^c). \end{aligned} \quad (8.29)$$

Now by Markov's inequality,

$$\begin{aligned}P(A_n^c) &= P\{\log p(X^n)/q(X^n) > nr\}\\ &\le \frac{E_P(\log P(X^n)/q(X^n))^+}{nr}\\ &\le \frac{nD(p\|q) + (\log e)/e}{nr}, \end{aligned} \quad (8.30)$$

where we have used the fact that the expectation with respect to P of the negative part of $\log p(X^n)/q(X^n)$ is the expectation with respect to Q of $(p(X^n)/q(X^n)) \cdot (\log p(X^n)/q(X^n))^-$ that is bounded by $(1/e)\log e$. Together (8.29) and (8.30) prove the lemma. □

Proof of Theorem 4: We show that if the weights $2^{-L_n(q)}$ satisfy the tail condition (7.2) and if the resolvability $R_n(p)$ tends to zero, then the minimum complexity density estimate $\hat{p}_n$ converges in squared Hellinger distance with rate bounded by $R_n(p)$ in probability. Also $L_n(\hat{p}_n)/n$ converges with rate bounded by $R_n(p)$.

Choose $\tilde{p}_n$ to achieve the best resolution $R_n(p) = L_n(\tilde{p}_n)/n + D(p\|\tilde{p}_n)$. Let $1/2 \le \alpha < 1$ be such that condition (7.2) is satisfied. For $c > 1$, let

$$\begin{aligned}B_n = \{&d^2(p,\hat{p}_n) > 4cR_n(p)/(1-\alpha) \text{ or}\\ &L_n(\hat{p}_n)/n > cR_n(p)/(1-\alpha)\}. \end{aligned} \quad (8.31)$$

The factor of $1-\alpha$ in the denominators is for convenience in the proof. Given $\epsilon > 0$, we show that $P(B_n)$ has limit less than ϵ for c sufficiently large. Applying Lemma 2 with $r = (c-1)R_n(p)/2$ and $Q = \tilde{P}_n$ and using $R_n(p) \ge D(p\|\tilde{p}_n)$, we have

$$\begin{aligned}P(B_n) \le \tilde{P}_n(B_n)2^{(c-1)nR_n(p)/2} &+ \frac{2}{c-1}\\ &+ \frac{2}{(c-1)nR_n(p)}\frac{\log e}{e}. \end{aligned} \quad (8.32)$$

Next we bound the $\tilde{P}_n$ probability of the event B_n. Using the triangle inequality $d(p,\hat{p}_n) \le d(\tilde{p}_n,\hat{p}_n) + d(p,\tilde{p}_n)$ and $d(p,\tilde{p}_n) \le \sqrt{D(p\|\tilde{p}_n)} \le \sqrt{R_n(p)}$, it is seen that B_n is a subset of the event

$$\begin{aligned}\tilde{B}_n = \{&d^2(\tilde{p}_n,\hat{p}_n) > cR_n(p)/(1-\alpha) \text{ or}\\ &L_n(\hat{p}_n)/n > cR_n(p)/(1-\alpha)\}. \end{aligned} \quad (8.33)$$

For the event $\bar{B}_n$ to occur there must be some q with $d^2(\bar{p}_n, q) > cR_n(p)/(1-\alpha)$ or $L_n(q)/n > cR_n(p)/(1-\alpha)$ for which the value of $2^{-L_n(q)}q(X^n)$ is at least as large as the value achieved at $\bar{p}_n$. Thus by the union of events bound

$$\bar{P}_n(\bar{B}_n) \le \sum_q \bar{P}_n\{2^{-L_n(\bar{p}_n)}\bar{p}_n(X^n) \le 2^{-L_n(q)}q(X^n)\}, \tag{8.34}$$

where the sum is for q with $d^2(\bar{p}_n, q) > cR_n(p)/(1-\alpha)$ or $L_n(q)/n > cR_n(p)/(1-\alpha)$. As in the proof of Theorem 3, the terms in this sum are not greater than

$$\min\{2^{-L_n(q)+L_n(\bar{p}_n)}, 2^{-(L_n(q)+L_n(\bar{p}_n)-nd^2(\bar{p}_n,q))/2}\} \le 2^{-\alpha L_n(q)}2^{-cnR_n(p)}2^{L_n(\bar{p}_n)}. \tag{8.35}$$

Summing this bound, using (7.2) and $L_n(\bar{p}_n) \le nR_n(p)$, yields

$$\bar{P}_n(\bar{B}_n) \le b'2^{-(c-1)nR_n(p)}. \tag{8.36}$$

Plugging this result into (8.32) and using $L_n(\bar{p}_n) \ge l$ by condition (7.5), we obtain

$$P(B_n) \le b'2^{-(c-1)nR_n(p)/2} + \frac{2}{c-1} + \frac{2}{(c-1)nR_n(p)}\frac{\log e}{e} \le b'2^{-(c-1)l/2} + \frac{2}{c-1} + \frac{2}{(c-1)l}\frac{\log e}{e}. \tag{8.37}$$

Taking c sufficiently large yields $P(B_n) \le \epsilon$. This completes the proof of Theorem 4. □

Remarks:

a) From (8.37) we have a bound on the probability of interest that holds uniformly for all densities p, for all sample sizes n, for all L_n, and for all $1/2 < \alpha < 1$ satisfying $\sum_q 2^{-\alpha L_n(q)} \le b'$ and $L_n(q) \ge l$, namely,

$$P\{d^2(p, \hat{p}_n) > 4cR_n(p)/(1-\alpha) \text{ or } L_n(\hat{p}_n)/n > cR_n(p)/(1-\alpha)\} \le b'2^{-(c-1)l/2} + \frac{2}{c-1} + \frac{2}{(c-1)l}\frac{\log e}{e}. \tag{8.38}$$

b) If the tail condition $\sum 2^{-\alpha_n L_n(q)} \le b'$ holds for some sequence $\alpha_n = 1 - 1/c_n$, where $c_n \to \infty$, and $c_n R_n(p) \to 0$, then $d^2(p, \hat{p}_n)$ and $L_n(\hat{p}_n)/n$ converge in probability at rate bounded by $c_n R_n(p)$.

c) A consequence of the previous remark is that if $2^{-L_n(q)}$ are weights that satisfy the summability condition (7.1) but not the tail condition (7.2), then by replacing $L_n(q)$ with $(1+1/c_n)L_n(q)$, new weights are obtained for which the minimum complexity estimator will converge at rate bounded by $c_n R_n(p)$.

d) With a slight modification of the proof of Theorem 4, it is seen that the value of $2^{-L_n(q)}q(X^n)$ at $\bar{p}_n$ will exceed the maximum value for all densities with $d^2(\bar{p}_n, q) > cR_n(p)/(1-\alpha)$ or $L_n(q)/n > cR_n(p)/(1-\alpha)$ by at least the factor $2^{c_0 nR_n(p)}$, except in an event of probability less than $b'2^{-(c-1-c_0)l/2} + (2/(c-1-c_0))(1+(\log e)/el)$, for $c-1 > c_0 \ge 0$. Consequently, except in this event of small probability, all densities that achieve values of $L_n(q) + \log 1/q(X^n)$ that are within $c_0 nR_n(p)$ of the minimum will satisfy $d^2(p,q) \le 4cR_n(p)/(1-\alpha)$ and $L_n(q)/n \le cR_n(p)/(1-\alpha)$.

Proof of the Corollary to Theorem 4: Assuming only that $\sum 2^{-L_n(q)} \le 1$, we are to show that the density $\hat{\hat{p}}_n$ that minimizes $L_n(q) + \log 1/q(X^n)$ subject to $L_n(q) \le 2\hat{L}_n$ will converge in squared Hellinger distance at rate bounded by $R_n(p)$ in probability. Here $\hat{L}_n$ is the length $L_n(\hat{p}_n)$ for a density $\hat{p}_n$ that achieves the minimum of $\lambda L_n(q) + \log 1/q(X^n)$ where $\lambda > 1$.

First we verify that $\hat{\hat{p}}_n$ achieves a value of $\lambda L_n(q) + \log 1/q(X^n)$ that is within $(\lambda-1)\hat{L}_n$ of the minimum. Indeed, $\lambda L_n(\hat{\hat{p}}_n) + \log 1/\hat{\hat{p}}_n(X^n)$ is equal to $(\lambda-1)L_n(\hat{\hat{p}}_n) + \min\{L_n(q) + \log 1/q(X^n)\colon L_n(q) \le 2\hat{L}_n\}$, which is less than or equal to $(\lambda-1)2\hat{L}_n + L_n(\hat{p}_n) + \log 1/\hat{p}_n(X^n)$. This last expression reduces to $(\lambda-1)\hat{L}_n + \lambda L_n(\hat{p}_n) + \log 1/\hat{p}_n(X^n)$ as desired.

Now, setting $c_0 = (c-1)/2$, we have

$$P\{d^2(p, \hat{\hat{p}}_n) > cR_n(p)\} \le P\{d^2(p, \hat{\hat{p}}_n) > cR_n(p) \text{ and } (\lambda-1)\hat{L}_n \le c_0 nR_n(p)\} + P\{\hat{L}_n/n > c_0 R_n(p)/(\lambda-1)\}. \tag{8.39}$$

The first event on the right is included in the event that $d^2(p,q) > cR_n(p)$ for some density that achieves within $c_0 nR_n(p)$ of the minimum of $\lambda L_n(q) + \log 1/q(X^n)$; by Remark d), this event has a probability that is made arbitrarily small by the choice of c sufficiently large. Also, the second event on the right has small probability for c large, by direct application of Theorem 4. This completes the proof of the corollary. □

IX. Remarks on Regression and Classification

The results in this paper have been developed in the context of density estimation. Nevertheless, it is possible to apply the convergence results to problems in nonparametric regression and classification. For instance, in regression it might be assumed that the data is of the form $X_i = (U_i, Y_i)$, $i = 1, \cdots, n$, where the input random variables U_i are drawn from a design density $p(u)$ and the output random variables Y_i are conditionally distributed as Normal$(f(u), \sigma^2)$ given that $U_i = u$. Suppose the error variance σ^2 is known. The conditional mean $f(u)$ is the unknown function that we wish to estimate. Assigning complexities $L(g)$ to a countable set of candidate functions g, we select $\hat{f}_n$ to minimize

$$L(g) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - g(U_i))^2 \log e. \tag{9.1}$$

This $\hat{f}_n$ is the minimum complexity regression estimator.

The index of resolvability in this context equals

$$R_n(f) = \min_g \left(\frac{L(g)}{n} + \frac{1}{2\sigma^2} \|f - g\|^2 \log e \right), \quad (9.2)$$

where $\|f-g\|^2 = \int (f(u)-g(u))^2 p(u)\,du$ (here the relative entropy reduces to a multiple of the L^2 distance). Using results for the L^2 approximation rates for smooth functions, as in Cox [47], bounds on the index of resolvability can be obtained that yield the same rates of convergence as we have given for density estimation. For instance, consider least squares polynomial regression with the degree of the polynomial automatically determined by the minimum complexity criterion. If $p(u)$ is bounded and has bounded support on the real line and if the rth derivative of f is square integrable then $R_n(f) \le O(((\log n)/n)^{2r/(2r+1)})$.

By Theorem 4, the squared Hellinger distance between the densities that have conditional mean functions $\hat{f}_n$ and f converges to zero in probability with rate bounded by $R_n(f)$ (provided $L(q)$ is chosen such that $\Sigma 2^{-\alpha L(g)}$ is finite for some $0 < \alpha < 1$). The squared Hellinger distance in this context is seen to equal $\int (1 - e^{(\hat{f}_n(u)-f(u))^2/8\sigma^2}) p(u)$ from which it is straightforward to obtain the lower bound $c \int \min((\hat{f}_n(u)-f(u))^2, 8\sigma^2) p(u)\,du$, where $c = (1-e^{-1})/8\sigma^2$. Consequently, the squared distance $\int \min((\hat{f}_n - f)^2, 8\sigma^2)$ converges to zero in probability with rate bounded by the index of resolvability.

Similar results hold for classification problems. Consider for instance the two-class case with class labels $Y \in \{0,1\}$. Here (U_i, Y_i), $i = 1,2,\cdots,n$ are independent copies of the random pair (U,Y). The conditional probability $f(u) = P\{Y=1 | U=u\}$ denotes the optimal discriminant function that we wish to estimate. Suppose complexities $L(g)$ are assigned to a countable set of functions $g(u)$ each with range restricted to $0 \le g \le 1$. (For instance, these functions may be obtained by logistic transformations of linear models, $g(u) = 1/(1 + \exp(-\Sigma \theta_{j=1}^{d} \phi_j(u))$, where the ϕ_j are polynomial or spline basis functions and the θ_j are restricted to $(1/2)\log n$ bits accuracy. The dimension d is automatically selected by the minimum complexity criterion.) It is seen that the minimum complexity estimator selects $\hat{f}_n$ to minimize

$$L(g) + \sum_{i=1}^{n} Y_i \log \frac{1}{g(U_i)} + \sum_{i=1}^{n} (1-Y_i) \log \frac{1}{1-g(U_i)}. \quad (9.3)$$

The index of resolvability in this classification context is

$$R_n(f) = \min_g \frac{L(g)}{n} + \int p(u) \left(f(u) \log \frac{f(u)}{g(u)} + (1-f(u)) \log \frac{(1-f(u))}{(1-g(u))} \right).$$

Rates of convergence for $R_n(f)$ can be obtained in the same manner as for density estimation. For instance, in the case of the logistic models with polynomial basis functions, if $p(u)$ is bounded and has bounded support on the real line and if the rth derivatives of $\log f(u)$ and $\log(1-f(u))$ are square integrable, then $R_n(f) \le O(((\log n)/n)^{2r/(2r+1)})$.

By Theorem 4, the square of the Hellinger distance, and hence also the square of the L^1 distance $\int p(u) |f(u) - \hat{f}_n(u)|$, converges at rate bounded by the index of resolvability $R_n(f)$. From accurate estimates of the discriminate function, good classification rules are obtained. Indeed, let P_e be the Bayes optimal probability of error, which corresponds to the classification rule that decides class 1 if and only if $f(U) \ge 1/2$, and let $P_e^{(n)}$ be the probability of error for the rule that decides class 1 if and only if $\hat{f}_n(U) \ge 1/2$. It can be shown that $|P_e^{(n)} - P_e| \le 2 \int p(u) |\hat{f}_n(u) - f(u)|$. Consequently, $P_e^{(n)}$ converges to the optimal probability of error at rate bounded by $\sqrt{R_n(f)}$.

The convergence results for minimum complexity regression and classification estimators are particularly useful for problems involving complicated multidimensional models, such as multilayered artificial neural networks, see [17], [48], [49]. The minimum complexity criterion is used to automatically select a network structure of appropriate complexity.

X. Conclusion

The minimum complexity or minimum description-length principle, which is motivated by information-theoretic considerations, provides a versatile criterion for statistical estimation and model selection. If the true density is finitely complex, then it is exactly discovered for all sufficiently large sample sizes. For large classes of infinitely complex densities, the sequence of minimum complexity estimators is strongly consistent. An index of resolvability has been introduced and characterized in parametric and nonparametric settings. It has been shown that the rate of convergence of minimum complexity density estimators is bounded by the index of resolvability.

Appendix
Details on Resolvability in the Parametric Case

Here we verify bounds on the optimum resolvability in parametric cases that are stated in Section VI.

Let $w(\theta)$ be a continuous and positive prior density function on the parameter space and suppose that the matrix J_θ (obtained from second-order derivatives of the relative entropy) is continuous and positive definite. We are to establish the existence of Γ_n and L_n satisfying the properties indicated in Section VI (Case 2). In particular Γ_n is to correspond to a net of points, such that for every θ there is a $\tilde{\theta}$ in the net satisfying

$$(\theta - \tilde{\theta})^T J_\theta (\theta - \tilde{\theta}) \le \frac{d + o(1)}{n} \quad (A.1)$$

and

$$L_n(p_{\tilde{\theta}}) = \log \left(\lambda_d (n/d)^{d/2} \det(J_\theta)^{1/2} \right) + \log 1/w(\theta) + o(1). \quad (A.2)$$

The set Γ_n is obtained in the following way. First, the parameter space is partitioned into disjoint rectangles A within which the prior density $w(\theta)$ and the matrix J_θ are nearly constant. Then in each set A, an ϵ-net of points $\tilde{\theta}$ is chosen such that for every θ in A, there is a $\tilde{\theta}$ with $(\theta-\tilde{\theta})^T J_A(\theta-\tilde{\theta}) \le \epsilon^2$. The minimal such net requires $N_\epsilon(A)$ points, where for small ϵ,

$$N_\epsilon(A) \sim \lambda_d (1/\epsilon)^d \operatorname{vol}(A)(\det J_A)^{1/2} \qquad \text{(A.3)}$$

and λ_d is a constant (see Lorentz [50, p. 153]). This amounts to taking the rotated and scaled parameter vectors $\xi = J_A^{1/2}\theta$ and finding economical coverings of the parallelograms $\{J_A^{1/2}\theta\colon \theta \in A\}$, using Euclidean balls of radius ϵ. The constant λ_d is the optimal density (in points per unit volume) for the coverage of R^d using balls of unit radius. Now for large d, it is seen that $\lambda_d^{2/d}/d \sim 1/2\pi e$. [This asymptotic density is found by combining the bounds of Rogers [51] and Coxeter, Few, and Rogers [52] for the thickness of the optimal covering with the Stirling approximation to the volume of the unit ball, see Conway and Sloane [53, ch. 1, (18) and ch. 2, (2) and (19)]. Consequently, the constants $c_d = d/\lambda_d^{2/d}$ are bounded independently of d.

We let Γ_n consist of the densities $p_{\tilde{\theta}}$ for $\tilde{\theta}$ in the ϵ-nets of the rectangles. The bound on resolvability will depend on ϵ through the terms $-(d/n)\log\epsilon + (1/2)\epsilon^2 \log e$ for which the optimum ϵ is seen to equal $\sqrt{d/n}$, so we now set $\epsilon = \sqrt{d/n}$ accordingly.

Now we define the codelengths $L_n(p_\theta)$. Let $W(A)$ denote the prior probability of the rectangles A. For $p_{\tilde{\theta}} \in \Gamma_n$ set

$$L_n(p_{\tilde{\theta}}) = \log 1/W(A) + \log N_\epsilon(A), \qquad \text{(A.4)}$$

for $\tilde{\theta}$ in A, for each A in the partition. Clearly,

$$\sum 2^{-L_n(p_{\tilde{\theta}})} = 1.$$

The matrices J_A are chosen such that J_A is positive definite and $\det J_A / \det J_\theta$ is arbitrarily close to one for all θ in A. This can be done by a choice of sufficiently small rectangles A because of the assumed continuity and positive definiteness of J_θ. In the same way $w(\theta)\operatorname{vol}(A)/W(A)$ is arbitrarily close to one, uniformly for θ in A. Moreover, by uniform continuity, these approximations are valid uniformly for all rectangles in a compact subset of the parameter space. Then from (A.3), (A.4), and the Taylor expansion of D, we have that for any given $\delta > 0$ and any compact set $B \subset \Theta$, there exists set $\Gamma_n^{(\delta,B)}$, codelengths $L_n^{(\delta,B)}(q)$ and $n_{\delta,B}$ such that for all $n \ge n_{\delta,B}$,

$$\left| L_n(p_{\tilde{\theta}}) - \log\left(\lambda_d (n/d)^{d/2} (\det J_\theta)^{1/2}/w(\theta)\right)\right| < \delta, \qquad \text{(A.5)}$$

$$(\theta-\tilde{\theta})^T J_\theta (\theta-\tilde{\theta}) \le \frac{d}{n}(1+\delta), \qquad \text{(A.6)}$$

and

$$D(p_\theta \| p_{\tilde{\theta}}) \le \frac{(d/2)\log e}{n}(1+2\delta), \qquad \text{(A.7)}$$

uniformly for all $\theta \in B$, where $\tilde{\theta}$ is the point in the net that minimizes the left side of (A.6).

Now let δ_k be a sequence decreasing to zero and let B_k be a sequence of compact sets increasing to Θ. Without loss of generality n_{δ_k, B_k} is an increasing sequence diverging to infinity as $k \to \infty$. For each $n \ge 1$, let k_n be the last index such that $n_{\delta_k, B_k} \le n$. Then $\lim k_n = \infty$. Setting $\Gamma_n = \Gamma_n^{(\delta_{k_n}, B_{k_n})}$ and $L_n(q) = L_n^{(\delta_{k_n}, B_{k_n})}(q)$ we have that for all n, (A.5)–(A.7) are satisfied with δ_{k_n} in place of δ, uniformly on B_{k_n}. Since any compact subset of Θ is eventually contained in B_{k_n}, this establishes the existence of a single set Γ_n and length function L_n for which (A.1) and (A.2) are satisfied uniformly on compacts.

Finally, from (A.5) and (A.7) it follows that the index of resolvability satisfies

$$\begin{aligned} R_n(p_\theta) &\le \frac{1}{n} L_n(p_{\tilde{\theta}}) + D(p_\theta \| p_{\tilde{\theta}}) \\ &\le \frac{1}{n}\Big((d/2)\log n/c_d + \log\left(\det(J_\theta)^{1/2}/w(\theta)\right) \\ &\qquad + (d/2)\log e + o(1)\Big), \end{aligned} \qquad \text{(A.8)}$$

where $o(1)$ tends to zero uniformly on compacts.

The minimax bound on resolvability now follows as in Section VI, upon taking $w(\theta)$ to be proportional to $\det(J_\theta)^{1/2}$. In particular, for each compact set $B \subset \Theta$,

$$\inf_{\Gamma_n} \sup_{L_n \theta \in B} R_n(p_\theta) \le \frac{1}{n}\left(\frac{d}{2}\log n + \log \int_B \det(J_\theta)^{1/2}\, d\theta - \frac{d}{2}\log\frac{c_d}{e} + o(1)\right). \qquad \text{(A.9)}$$

In this analysis, we used a minimal net of points for covering the parameter space to a prescribed covering radius for a locally specified metric, so as to bound the minimax resolvability. If nets based on other coverings are used (such as cubes in the locally transformed parameter ξ), similar terms still appear involving the Fisher information and the prior density, but somewhat worse constants are obtained in the minimax bound.

As pointed out by a referee, a different net can yield improved bounds for the average resolvability,

$$\int w(\theta) R_n(p_\theta)\, d\theta.$$

To bound the average resolvability, it is suggested that optimal quantization regions (with centroids $\tilde{\theta}$) be selected subject to a constraint on the average value for $(\theta-\tilde{\theta})^T J_A(\theta-\tilde{\theta})$ (instead of a constraint on the maximum value). Indeed, suppose we constrain the average value to equal ϵ^2. Using optimum quantization results as in [53], it is seen by an analysis similar to that previously given that the minimum number of quantization points in each set A is the same as in (A.3) but with $(dG_d)^{d/2}$ in place of λ_d, where G_d is the coefficient of optimum mean-square quantization as characterized in [53, pp. 58–59]. In particular, from a result of Zador, $G_d \sim 1/2\pi e$ for large d. This yields codelengths $L_n(p_{\tilde{\theta}})$ that are the same as

before, but with $c'_d = 1/G_d$ in place of c_d. Both c_d and c'_d are close to $2\pi e$, for large d. Thus for large dimensions, there is not much difference in the codelengths designed from optimum covering and optimal quantization considerations.

REFERENCES

[1] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Peredach. Inform.*, vol. 1, pp. 3–11, 1965.

[2] V. V. V'Yugin, "On the defect of randomness of a finite object with respect to measures with given complexity bounds," *Theory Probab. Appl.*, vol. 32, pp. 508–512, 1987.

[3] T. M. Cover, "Generalization on patterns using Kolmogorov complexity," in *Proc. First Int. Joint Conf. Pattern Recog.*, Washington, DC, Oct. 1973.

[4] ____, "Kolmogorov complexity, data compression, and inference," in *The Impact of Processing Techniques on Communications*, J. K. Skwirzynski, Ed. Boston, MA: Martinus Nijhoff Publ., 1985, pp. 23–34.

[5] T. M. Cover, P. Gacs, and R. M. Gray, "Kolmogorov's contributions to information theory and algorithmic complexity," *Ann. Probab.*, vol. 17, pp. 840–865, July 1989.

[6] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[7] ____, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, pp. 416–431, June 1983.

[8] ____, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. 30, pp. 629–636, July 1984.

[9] ____, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, Sept. 1986.

[10] ____, "Stochastic complexity and sufficient statistics," *J. Roy. Statist. Soc. B*, vol. 49, pp. 223–239, 1987.

[11] ____, *Stochastic Complexity in Statistical Inquiry*. Teaneck, NJ: World Scientific Publ., 1989.

[12] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Comput. J.*, vol. 11, pp. 185–194, 1968.

[13] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Roy. Statist. Soc. B*, vol. 49, pp. 240–265, 1987.

[14] R. Sorkin, "A quantitative Occam's razor," *Int. J. Theoretic Phys.*, vol. 22, pp. 1091–1103, 1983.

[15] A. R. Barron, "Convergence of logically simple estimates of unknown probability densities," presented at *IEEE Int. Symp. Inform. Theory*, St. Jovite, Canada, Sept. 26–30, 1983.

[16] ____, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, Aug. 1985.

[17] ____, "Complexity regularization," in *Proceedings NATO Advanced Study Institute on Nonparametric Functional Estimation*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer Academic Publ., 1991.

[18] T. M. Cover, "A hierarchy of probability density function estimates," in *Frontiers in Pattern Recognition*. New York: Academic Press, 1972, pp. 83–98.

[19] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. 19, pp. 783–795, Nov. 1973.

[20] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[21] R. J. Solomonoff, "A formal theory of inductive inference," *Inform. Contr.*, vol. 7, pp. 224–254, 1964.

[22] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *J. Assoc. Comput. Machine*, vol. 13, pp. 547–569, 1966.

[23] ____, "A theory of program size formally identical to information theory, *J. Assoc. Comput. Machine*, vol. 22, pp. 329–340, 1975.

[24] L. A. Levin, "On the notion of a random sequence," *Soviet Math. Dokl.*, vol. 14, pp. 1413–1416, 1973.

[25] ____, "Laws of information conservation and aspects of the foundations of probability theories," *Probl. Inform. Transm.*, vol. 10, pp. 206–210, 1974.

[26] B. S. Clarke and A. R. Barron, "Information theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[27] B. S. Clarke, "Asymptotic cumulative risk and Bayes risk under entropy loss, with applications." Ph.D. dissertation, Dept. Statist., Univ. of Illinois, Urbana, IL, July 1989.

[28] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encodings," *IEEE Trans. Inform. Theory*, vol. 27, pp. 199–207, Mar. 1981.

[29] H. Jeffreys, *Theory of Probability*. Oxford: Oxford Univ. Press, 1967.

[30] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[31] A. R. Barron and C. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 19, no. 3, Sept. 1991.

[32] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, pp. 45–54, 1981.

[33] K. C. Li, "Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set," *Ann. Statist.*, vol. 15, pp. 958–975, Sept. 1987.

[34] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Inform. Theory*, P. N. Petrov and F. Csaki, Eds. Budapest: Akademia Kiado, 1983, pp. 267–281.

[35] P. Hall and E. J. Hannan, "On stochastic complexity and nonparametric density estimation," *Biometrika*, vol. 75, pp. 705–714, 1988.

[36] B. Yu and T. P. Speed, "Stochastic complexity and model selection II: Histograms," Tech. rep. no. 241, Dept. of Statist. Univ. of California, Berkeley, Mar. 1990.

[37] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Uspehi*, vol. 3, pp. 3–86, 1959.

[38] M. S. Birman and M. Z. Solomjak, "Piecewise-polynomial approximations of functions of the classes W_p^α," *Mat. USSR-Sbornik*, vol. 2, pp. 295–317, 1967.

[39] U. Granander, *Abstract Inference*. New York: Wiley, 1981.

[40] J. Bretagnolle and C. Huber, "Estimation des densités: Risque minimax", *Z. Wahrscheninlichkeitstheorie verw. Gebiete*, vol. 47, pp. 119–137, 1979.

[41] S. Y. Efroimovich and M. S. Pinsker, "Estimation of square-integrable probability density of a random variable," *Probl. Inform. Transm.*, vol. 18, pp. 175–189, 1983.

[42] Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy," *Ann. Statist.*, vol. 13, pp. 768–774, June 1985.

[43] A. R. Barron, "Uniformly powerful goodness of fit tests," *Ann. Statist.*, vol. 17, no. 1, Mar. 1989.

[44] E. J. G. Pitman, *Some Basic Theory for Statistical Inference*. London: Chapman and Hall, 1979.

[45] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–507, 1952.

[46] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, Mar. 1963.

[47] D. D. Cox, "Approximation of least squares regression on nested subspaces," *Ann. Statist.*, vol. 16, pp. 713–732, June 1988.

[48] A. R. Barron and R. L. Barron, "Statistical learning networks: A unifying view," *Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface*. Fairfax, Virginia, April 21–23, 1988, E. Wegman, D. T. Gantz, and J. J. Miller, Eds. Alexandria, VA: Amer. Statist. Assoc., 1988.

[49] A. R. Barron, "Statistical properties of artificial neural networks," presented at *Proc. 28th IEEE Conf. Decision Contr.*, Tampa, Florida, Dec. 1989.

[50] G. G. Lorentz, *Approximation of Functions*. New York: Holt, Rinehart, and Winston, 1966.

[51] C. A. Rogers, "Lattice coverings of space," *Mathematika*, vol. 6, pp. 33–39, 1959.

[52] H. S. M. Coxeter, L. Few, and C. A. Rogers, "Covering space with equal spheres," *Mathematika*, vol. 6, pp. 147–157, 1959.

[53] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*. New York: Springer-Verlag, 1988.

[54] A. R. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimation convergent in total variation and in informational divergence," to appear in *IEEE Trans. Inform. Theory*.