The Minimum Description Length Principle in Coding and Modeling

Andrew Barron, *Member, IEEE*, Jorma Rissanen, *Senior Member, IEEE*, and Bin Yu, *Senior Member, IEEE*

(Invited Paper)

***Abstract*— We review the principles of Minimum Description Length and Stochastic Complexity as used in data compression and statistical modeling. Stochastic complexity is formulated as the solution to optimum universal coding problems extending Shannon's basic source coding theorem. The normalized maximized likelihood, mixture, and predictive codings are each shown to achieve the stochastic complexity to within asymptotically vanishing terms. We assess the performance of the minimum description length criterion both from the vantage point of quality of data compression and accuracy of statistical inference. Context tree modeling, density estimation, and model selection in Gaussian linear regression serve as examples.**

***Index Terms*—Complexity, compression, estimation, inference, universal modeling.**

I. Introduction

In this expository paper we discuss the so-called MDL (Minimum Description Length) principle for model selection and statistical inference in general. In broad terms the central idea of this principle is first to represent an entire class of probability distributions as models by a single "universal" representative model such that it would be able to imitate the behavior of any model in the class. The best model class for a set of observed data, then, is the one whose representative permits the shortest coding of the data.

There are a number of ways to construct representatives of model classes or, what is sufficient, to compute their codelength. The first and the crudest of them, Wallace and Boulton [52], Rissanen [38], is to encode the data with a (parametric) model defined by the maximum-likelihood estimates, quantized optimally to a finite precision, and then encode the estimates by a prefix code. For a reader with any knowledge of information theory there is nothing startling nor objectionable about such a procedure and the principle itself. After all, in order to design a good code for data, the code must capture the statistical characteristics of the data, and in order to be able to decode, the decoder must be given this distribution, which permits the construction of the code, the "codebook," and the particular codeword for the observed data. It is the statisticians for whom the connection between probability distributions and codelengths tends to appear strange and on the surface of it nonexisting. And yet, even a statistician must admit, however grudgingly, that the principle seems to incorporate in a direct way some of the most fundamental, albeit elusive, ideas the founding fathers of statistical inference have been groping for, like the objective of statistics is to reduce data, Fisher [21], and that "we must not overfit data by too complex models." Perhaps, a statistician can take solace in the fact that by the fundamental Kraft inequality, stated below, a codelength is just another way to express a probability distribution, so that the MDL principle becomes the familiar Maximum Likelihood (ML) principle—albeit a global one.

Simple and natural as the MDL principle may be, it nevertheless provides a profound change in the way one should think about statistical problems. About the data themselves, it is not necessary to make the usual assumption that they form a sample from an imagined population, which is something that would be impossible to verify. After all, we are able to design codes for any data that on the whole can be finitely described. However, ever since Shannon's work we know how to design good codes for data generated by sampling a probability distribution, and the same codes will work reasonably well even for data which are not generated that way, provided that they have the kinds of restrictions predicated by the distribution, at least to some degree. Indeed, the greater this degree is the closer the resulting codelength for the data will be to the optimal for the distribution with which the code was designed. This seems to imply that we are just pushing the problem to the selection of the assumed probability distribution, which is exactly what we do. The probability distributions serve us as a means by which to express the regular features in the data; in other words, they serve as *models*. In fact, that ultimately is what all models do, including the deterministic "laws" of nature, which spell out the restrictions to such a high degree that the inevitable deviations between the data and the behavior dictated by the laws give rise to almost singular probability distributions. Prediction is certainly an important motivation for modeling, and one may ask why not use prediction error as a criterion for model selection. Fortunately, almost all the usual prediction

Manuscript received December 2, 1997; revised April 27, 1998. The work of A. Barron was supported in part by NSF under Grants ECE-9410760 and DMS-95-05168. The work of B. Yu was supported in part by ARO under Grant DAAH04-94-G-0232 and by NSF under Grant DMS-9322817.

A. Barron is with the Department of Statistics, Yale University, New Haven, CT 06520-8290 USA.

J. Rissanen is with IBM Research Division, Almaden Research Center, DPE-B2/802, San Jose, CA 95120-6099 USA.

B. Yu is with the Department of Statistics, University of California at Berkeley, Berkeley, CA 94720-3860 USA.

Publisher Item Identifier S 0018-9448(98)05284-5.

error criteria, be they in terms of probability of errors or some distance measure such as the absolute or squared errors, can be expressed in terms of codelength, and there is no conflict between the two [41], [44].

According to this program, the problems of modeling and inference, then, are not to estimate any "true" data generating distribution with which to do inference, but to search for good probability models for the data, where the goodness can be measured in terms of codelength. Such a view of statistics also conforms nicely with the theory of algorithmic complexity, Solomonoff [47], Kolmogorov [34], and can draw on its startling finding about the ultimate limitation on all statistical inference, namely, that there is no "mechanical," i.e., algorithmic, way to find the "best" model of data among all computable models (let alone the metaphysical "true" model).

Although the MDL principle stands on its own and cannot be tampered by findings in analysis, it still leaves a role for probability and coding theories albeit a different one: Analysis can provide support for the principle or pinpoint abnormal behavior, and help provide designs for good codes for data generated by various probability models and classes of them. It happens that such a code design follows a remarkably uniform pattern, which starts with Shannon's basic case of a fixed known data-generating probability distribution, say $P(\underline{x})$, where $\underline{x}$ denotes a data string to be encoded. In this case a meaningful optimization problem is to find a code that has the minimum mean length subject to the restriction that the codeword lengths $\ell(\underline{x})$ satisfy the fundamental Kraft inequality

$$\sum_{\underline{x}} 2^{-\ell(\underline{x})} \leq 1. \tag{1}$$

If we dispose of the somewhat irrelevant restriction that the codeword lengths must be natural numbers, the minimization problem admits the remarkable solution that such *ideal* codeword lengths must coincide with the numbers $-\log P(\underline{x})$, giving the entropy as the minimized mean length. Hence, the optimal codelengths mimic the data generating distribution. Although the MDL principle requires us to find the length of the shortest codeword for the actually observed sequence, rather than a mean length, it is also true that no code exists where the probability of the set of codewords that are shorter than the optimal less c, $-\log P(\underline{x}) - c$, exceeds 2^{-c}. In other words, the codewords of the optimal code are practically the shortest for almost all "typical" strings generated by the distribution.

As stated above, the focus of interest in the MDL principle is in various classes of probability distributions as models, which together with the modeling problems they create are discussed first. For such classes we consider optimization problems that generalize the basic Shannon problem above. If $Q(\underline{x})$ denotes the sought-for universal representative of a model class $\{P(\underline{x}|\theta)\}$ under study, where θ is a parameter vector, the quantity of interest is the difference

$$\log \frac{1}{Q(\underline{x})} - \log \frac{1}{P(\underline{x}|\theta)} = \log \frac{P(\underline{x}|\theta)}{Q(\underline{x})}$$

between the codeword length of the representative and that obtained with Shannon's codes defined by the members $P(\underline{x}|\theta)$ in the class. The first problem calls for a code, defined by Q, which minimizes the difference given maximized over $\underline{x}$ and θ. In the second problem we seek $Q(\underline{x})$ which minimizes the mean difference. For smooth model classes the solutions to these two problems turn out to be virtually the same, and the minimized difference may be interpreted as the *parametric complexity* of the model class involved at the given data sequence of length n. Again generalizing Shannon's basic result the solutions will also be shortest possible for almost all typical strings generated by almost all models in the class. In analogy with the algorithmic or Kolmogorov complexity, the codelength that differs from the ideal by the parametric complexity is called *stochastic complexity*.

Although this paper is tutorial in nature we have decided not to restrict it to an elementary introduction, only, but also to survey some of the more advanced techniques inspired by the theory with the intent to demonstrate how the new ideas contribute to the analysis of the central problems arising in modeling. These include the demonstration of the desirable property of the MDL principle that, if we apply it to data generated by some unknown model in the considered class, then the MDL estimates of both the parameters and their number are consistent; i.e., that the estimates converge and the limit specifies the data generating model. We also discuss the close connection between statistical inference and an important predictive way to do coding, which lends itself to coding data relative to nonparametric models. We conclude the paper with applications of the MDL principle to universal coding, linear regression where the stochastic complexity can be calculated exactly, and density estimation.

II. Modeling Prelude

At two extremes of statistical modeling are issues of parametric inference in a given, perhaps small, finite-dimensional family or model class $\{P(\underline{x}|\theta)\colon\ \theta \in \Theta \subset \Re^d\}$ and issues of nonparametric inference in infinite-dimensional function classes (e.g., of density or regression curves). The MDL principle has implications for each motivated by the aim of providing a good statistical summary of data. The principle is especially useful in the middle ground, where a variety of plausible families $\mathcal{M}_k = \{P_k(\underline{x}|\theta)\colon\ \theta \in \Theta_k \subset \Re^{d_k}\}$ for $k \in \mathcal{K}$ are available, and one seeks to automate the selection of an estimate $\hat{k}$ based on data $\underline{x} = (x_1, \cdots, x_n)$. It is clear that both extremes of fixed small finite-dimensional or infinite-dimensional families have their limitations. Whereas in statistics these limitations and their resolutions via penalized risk criteria are often cast in terms of the bias and variance tradeoff, or the approximation and estimation tradeoff, we will seek here what we regard as a more intrinsic characterization of the quality of the summarization.

In the case that there is a sequence $\mathcal{M}_k, k \in \mathcal{K}$, of classes available for modeling the data, the MDL principle advocates a choice $\hat{k}$ that optimizes a codelength $\log 1/Q_k^*(\underline{x}) + l(k)$. Here $\log 1/Q_k^*(\underline{x})$ is a codelength for description of data using the model class $\mathcal{M}_k$ in accordance with optimal coding criteria discussed below, and $l(k)$ is a codelength for the description

of the class k. For each model class, the codelength criterion involves in an intrinsic way a tradeoff between likelihood $\log 1/P_k(\underline{x}|\theta)$ and a parametric complexity associated with the class $\mathcal{M}_k$ that models data of the given length n. To optimize this tradeoff, we are led to the maximum-likelihood estimator $\hat{\theta} = \hat{\theta}(\underline{x})$ and a parametric complexity that is the minimum additional coding cost $c_{n,k}$ necessary for $\log 1/P_k(\underline{x}|\hat{\theta}) + c_{n,k}$ to be the length of a uniquely decodable code for $\underline{x} \in \underline{\mathcal{X}}$. The MDL estimator $\hat{k}$ achieves minimum total codelength $\log 1/P_{\hat{k}}(\underline{x}|\hat{\theta}) + c_{n,\hat{k}} + l(\hat{k})$.

Rather than being interested in the bits of exact coding of the data, our interest in modeling is to provide quality summarization of the data through the estimated model. An ultimate or idealized summarization is captured by Kolmogorov's minimum sufficient statistic for description. Recall that the Kolmogorov complexity of a data sequence $\underline{x}$ is the length $K(\underline{x})$ of the shortest computer program that outputs $\underline{x}$ on a given universal computer. This complexity is a universal assignment for all strings $\underline{x}$ up to a constant (in the sense that for any given pair of universal computers there is a constant c of translation between computers such that for all sequences, no matter how long, the complexity assigned by the two computers differs by no more than c). Maximally complex sequences are those for which $K(\underline{x})$ equals $\log card(\underline{\mathcal{X}})$ to within a constant. These are sequences which defy interesting summarization or compression.

To get at the idea of optimal summarization, Kolmogorov refined his notion of complexity (see [11, pp. 176, 182]). For each $\underline{x}$, there typically are a number of programs that are minimal in the sense of achieving within a given constant c the complexity $K(\underline{x})$. Among these minimal programs are those which describe the data in two parts. First, some property A (a subset $A \subset \underline{\mathcal{X}}$), satisfied by the sequence $\underline{x} \in A$, is optimally described using $K(A)$ bits, and then $\log card(A)$ bits are used to give the index of $\underline{x} \in A$. When this description of $\underline{x}$ of length $K(A) + \log card(A)$ is minimal, $\log card(A)$ cannot be improved by using the length of any other encoding of $\underline{x}$ in A, and hence the $\log card(A)$ bits are maximally complex (uninteresting bits), conditionally given A. The *interesting part* arises in a property A that does not exactly represent, but rather summarizes the sequence. The best summary is provided by a program for a property A^* satisfied by $\underline{x}$ that has minimum $K(A^*)$ subject to $K(A^*) + \log card(A^*)$ agreeing with $K(\underline{x})$ (to within the specified constant). Such A^* may be called a *Kolmogorov minimal sufficient statistic* for the description of $\underline{x}$.

Our notion of summarization intended for statistical modeling differs from Kolmogorov's in two ways. We do not restrict the first part of the code to be the description of a set containing the sequence, but rather we allow it to be the description of a statistical model (where the counterpart to a set A becomes the uniform distribution on A), and corresponding to a statistical model we replace $\log card(A)$ by the length of Shannon's code for $\underline{x}$ using the model. Secondly, at the expense of ultimate idealism, we do not require that the descriptions of distributions be Kolmogorov optimal (which would be computationally unrealizable), but rather we make our codelength assignments on the basis of principles that capture near optimality for most sequences. Neither do we seek the optimum among all computable distributions but only relative to a given list of models.

III. Optimal Coding Methods

A. Shannon Coding

Let $\underline{x} \in \underline{\mathcal{X}}$ refer to data to be described and modeled, where $\underline{\mathcal{X}}$ is a given countable set. Typically we have the set $\underline{\mathcal{X}} = \mathcal{S}^n$ of length n sequences $\underline{x} = (x_1, \cdots, x_n)$ for $x_i \in \mathcal{S}$ from some discrete alphabet $\mathcal{S}$ such as English or ASCII characters or a discretization of real-valued variables.

Description of $\underline{x} \in \underline{\mathcal{X}}$ is accomplished by means of a mapping into finite-length binary sequences, called *codewords*, where the map, called a (binary) *code*, is required to be one-to-one, and concatenations of codewords are also required to be in one-to-one correspondence with sequences of symbols $\underline{x}$, themselves sequences, from $\underline{\mathcal{X}}$. That is the requirement of unique decodability. It is accomplished in particular by arranging the codewords to satisfy the property that no codeword is a prefix for a codeword of another $\underline{x} \in \underline{\mathcal{X}}$. This yields a correspondence between codes and labeled binary trees, where the codeword for $\underline{x}$ is the sequence of zeros and ones that gives the path from the root to the leaf labeled $\underline{x}$. Given a code tree let $\ell(\underline{x})$ denote the length of the codeword (or path) that describes $\underline{x}$. According to the theory of Shannon, Kraft, and McMillan, see, e.g., [11], there exists a uniquely decodable code with lengths $\ell(\underline{x})$ for $\underline{x} \in \underline{\mathcal{X}}$ if and only if the Kraft inequality (1) holds. Indeed, to each code there corresponds a subprobability mass function $Q(\underline{x}) = 2^{-\ell(\underline{x})}$. For a complete tree, in which every internal node has both descendants and all leaves are codewords, an interpretation is that a random walk starting at the root ends up at $\underline{x}$ with probability $Q(\underline{x}) = 2^{-\ell(\underline{x})}$, and hence $\Sigma_{\underline{x}} 2^{-\ell(\underline{x})} = 1$. Shannon gave an explicit construction of a code with length equal to $\log 1/Q(\underline{x})$, rounded up to an integer, as follows: Order the strings $\underline{x}$ by decreasing value of $Q(\underline{x})$ and define the codeword of $\underline{x}$ as the first $\ell(\underline{x})$ bits of the cumulative probability $\Sigma_{\underline{x}' < \underline{x}} Q(\underline{x}')$.

Shannon also posed the following optimization problem. If we are given a probability mass function $P(\underline{x})$ on $\underline{\mathcal{X}}$, then what codelengths achieve the minimum expected value $E_P \ell(\underline{x})$? From the correspondence between codes and subprobability mass functions Q it is seen that the solution is to take $\ell(\underline{x}) = \log 1/P(\underline{x})$ if we ignore the integer codelength constraint. Indeed, with any other choice the excess codelength

$$\ell(\underline{x}) - \ell_P(\underline{x}) = \log 1/Q(\underline{x}) - \log 1/P(\underline{x}) = \log \frac{P(\underline{x})}{Q(\underline{x})}$$

has positive expected value, given by the relative entropy or Kullback–Leibler distance

$$E_P(\ell(\underline{x}) - \ell_P(\underline{x})) = E_P \log \frac{P(\underline{x})}{Q(\underline{x})} = D(P\|Q)$$

which equals zero only if $Q = P$. Thus given P, the Shannon codelength $\ell_P(\underline{x}) = \log 1/P(\underline{x})$ is optimum, and $D(P\|Q)$ is the expected codelength difference (redundancy) when Q is used in the absense of knowledge of P. This property,

together with the simple probability inequality that $\ell(\underline{x}) - \ell_P(\underline{x})$ exceeds $-c$ except in a set of probability not greater than 2^{-c} for all $c>0$, leads us to call $\ell_P(\underline{x}) = \log 1/P(\underline{x})$ the optimal or *ideal* codelength.

B. Coding with a Model Class

The subject of universal data compression deals with describing data when the source distribution P is unknown. A most useful coding theory, as an extension of Shannon's theory with P given, can be developed if the distributions P are, instead of being completely unknown, restricted to a class of parametric distributions $\mathcal{M}_k = \{P_k(\underline{x}|\theta)\colon\ \theta \in \Theta_k \subset \Re^{d_k}\}$, refered to above as a model class. The results also turn out to provide the codelengths required for the MDL criterion.

Suppose we are given a parametric family of probability mass functions $\mathcal{M} = \{P(\underline{x}|\theta)\colon\ \theta \in \Theta\}$, which have corresponding Shannon codelengths $\log 1/P(\underline{x}|\theta)$. There is a collection of data compressors, indexed by θ. With hindsight, after observation of $\underline{x}$, the shortest of these is $\log 1/P(\underline{x}|\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(\underline{x})$ is the maximum-likelihood estimate (MLE) achieving $P(\underline{x}|\hat{\theta}) = \max_\theta P(\underline{x}|\theta)$. This is our target level of performance. Though its value can be computed, it is not available to us as a valid codelength, for without advance knowledge of $\hat{\theta}(\underline{x})$ we do not know which Shannon tree to decode. If we code data using a distribution $Q(\underline{x})$, the excess codelength, sometimes called regret, over the target value is

$$\log 1/Q(\underline{x}) - \log 1/P(\underline{x}|\hat{\theta}(\underline{x})) = \log \frac{P(\underline{x}|\hat{\theta})}{Q(\underline{x})} \tag{2}$$

which has the worst case value $\max_{\underline{x}} \log P(\underline{x}|\hat{\theta}(\underline{x}))/Q(\underline{x})$. Shtarkov [46] posed the problem of choosing $Q(\underline{x})$ to minimize the worst case regret, and he found the unique solution to be given by the maximized likelihood, normalized thus

$$Q^*(\underline{x}) = \frac{P(\underline{x}|\hat{\theta})}{C}$$
$$C = \sum_{\underline{x}\in\underline{X}} P(\underline{x}|\hat{\theta}(\underline{x})). \tag{3}$$

This distribution plays an important role in the MDL theory, and we refer to it as the normalized maximum-likelihood (NML) distribution. Notice that $C = C(\mathcal{M}, n)$ depends on the model class $\mathcal{M}$ and the size n of the sample space $\underline{X} = \mathcal{S}^n$. The corresponding codelength is

$$\log 1/Q^*(\underline{x}) = \log 1/P(\underline{x}|\hat{\theta}) + \log C(\mathcal{M}, n) \tag{4}$$

which gives the minimax regret

$$\min_Q \max_{\underline{x}} \log \frac{P(\underline{x}|\hat{\theta}(\underline{x}))}{Q(\underline{x})} = \log C(\mathcal{M}, n). \tag{5}$$

The proof of the optimality of Q^* is simply to note that $\log P(\underline{x}|\hat{\theta})/Q^*(\underline{x}) = \log C(\mathcal{M}, n)$ for all $\underline{x}$, and for any other subprobability mass function $Q(\underline{x})$ we have $Q(\underline{x}) < Q^*(\underline{x})$ for at least one $\underline{x}$, where $\log P(\underline{x}|\hat{\theta}(\underline{x}))/Q(\underline{x})$ is strictly worse.

This optimal codelength $\log 1/P(\underline{x}|\hat{\theta}) + c_n$ associated with the NML distribution is what we call the *stochastic complexity* of data relative to the model class $\mathcal{M}$. It exceeds the $\log 1/$*maximized likelihood* term by the additional coding cost

$$c_n = \log C(\mathcal{M}, n) = \log \Sigma_{\underline{x}} p(\underline{x}|\hat{\theta}(\underline{x})).$$

Because this additional cost rises due to the unknown parameter, we call it the *parametric complexity*. Also in support of this terminology we note that other coding schemes, such as two-part codes as in [43] (which first describe parameter estimates to an optimal precision and then the data conditional on the parameter estimates), achieve a similar complexity term expressed in terms of the length of the description of optimally discretized parameter estimates. We emphasize that in the case of the code with respect to the NML distribution, the normalization insures Kraft's inequality, and hence encoding of $\underline{x}$, can be done directly without the need for separate encoding of $\hat{\theta}$.

C. Codes Optimal for Average Regret

While we are interested in the regret defined in (2), we do not presume to be interested only in its worst case value. Thus we consider expected regrets with respect to distributions in $\mathcal{M}$ and with respect to mixtures of these distributions, and we discuss the behavior of the corresponding minimax and maximin values. A mixture that achieves an (asymptotic) minimax and maximin solution forms an alternative MDL coding procedure that will be related to the NML code. With respect to any distribution $P(\underline{x}|\theta) = P_\theta(\underline{x})$ in $\mathcal{M}$, the expected value of the regret of Q is

$$R_n(P_\theta, Q) = E_{P_\theta} \log \frac{P(\underline{x}|\hat{\theta}(\underline{x}))}{Q(\underline{x})}$$

where in the right side we left the dummy variable $\underline{x}$ over which the expectation is taken. Averaging the expected regret further with respect to any probability distribution w on Θ is the same as averaging with respect to the mixture (marginal)

$$Q^w(\underline{x}) = \int P(\underline{x}|\theta)\, dw(\theta)$$

and the resulting average regret is $E_{Q^w} \log P(\underline{x}|\hat{\theta})/Q(\underline{x})$, which equals

$$E_{Q^w} \log \frac{P(\underline{x}|\hat{\theta})}{Q^w(\underline{x})} + D(Q^w||Q).$$

Thus $Q = Q^w$ is the unique choice to minimize the average regret with respect to the distribution w. In decision-theoretic terminology, the Bayes optimal code is of length

$$\log 1/Q^w(\underline{x}) = \log 1/\int P(\underline{x}|\theta)\, dw(\theta).$$

The expected regret has a minimax value

$$\overline{\mathcal{R}}_n(\mathcal{M}) = \min_Q \max_\theta E_{P_\theta} \log P(\underline{x}|\hat{\theta}(\underline{x}))/Q(\underline{x})$$

which agrees with the maximin value

$$\underline{\mathcal{R}}_n(\mathcal{M}) = \max_w \min_Q E_{Q^w} \log \frac{P(\underline{x}|\hat{\theta}(\underline{x}))}{Q(\underline{x})}$$
$$= \max_w E_{Q^w} \log \frac{P(\underline{x}|\hat{\theta}(\underline{x}))}{Q^w(\underline{x})} \tag{6}$$

where the maximization is enlarged from Θ to distributions on Θ, as is standard to allow equality of the minimax and maximin values. The maximization over w yields least favorable priors $\overline{w}$ for which the corresponding procedure, coding based on $Q^{\overline{w}}(\underline{x})$, is both maximin and minimax.

Related quantities studied in universal data compression are based on the expected codelength difference (redundancy)

$$D(P_\theta||Q) = E_\theta \log \frac{P(\underline{x}|\theta)}{Q(\underline{x})}$$

which uses the unknown θ in the target value $\log 1/P(\underline{x}|\theta)$ rather than the MLE. The average redundancy with respect to a distribution on θ is equal to Shannon's mutual information $I(\theta;\underline{x})$ when the Bayes optimal code is used. Consequently, the maximin average redundancy is $\max_w I(\theta;\underline{x})$, which is recognized as the Shannon information capacity of the class $\{P(\underline{x}|\theta)\colon\ \theta \in \Theta\}$ (Davisson [12]). The minimax value $\min_Q \max_\theta E_\theta \log P(\underline{x}|\theta)/Q(\underline{x})$ and the maximin value of the redundancy (i.e., the capacity) are equal (see Davisson *et al*. [14] and Haussler [28]). In subsequent sections we will have more to say about the Kullback–Leibler divergence $E_\theta \log P(\underline{x}|\theta)/Q(\underline{x})$, including interpretations in coding and prediction, its asymptotics, and useful finite sample bounds.

Both of the target values $\log 1/P(\underline{x}|\theta)$ and $\log 1/P(\underline{x}|\hat{\theta})$ are unrealizable as codelengths (because of lack of knowledge of θ in one case and because of failure of Kraft's inequality in the other) and an extra descriptional price is to be paid to encode $\underline{x}$. In this section we retain $\log 1/P(\underline{x}|\hat{\theta})$ as the idealized target value for several reasons, not the least of which is that (unlike the other choice) it can be evaluated from the data alone and so can be a basis for the MDL criterion. By use of the same quantity in pointwise, worst case, and average value analyses we achieve a better understanding of its properties. We identify that the parametric complexity $\log C(\mathcal{M}, n)$ (and its asymptotic expression) arises in characterization of the minimax regret, the minimax expected regret, and in pointwise bounds that hold for most sequences and most distributions in the model class.

Note that the minimax value of the expected regret, through its maximin characterization in (6), may be expressed as

$$\begin{aligned}\underline{\mathcal{R}}_n(\mathcal{M}) &= \max_w E_{Q^w} \log \frac{P(\underline{x}|\hat{\theta}(\underline{x}))}{Q^w(\underline{x})} \\ &= \log C(\mathcal{M}, n) - \min_w D(Q^w(\underline{x})||Q^*(\underline{x})). \quad (7)\end{aligned}$$

Thus optimization over w to yield a minimax and maximin procedure is equivalent to choosing a mixture $Q^{\overline{w}}(\underline{x})$ closest to the normalized maximum likelihood $Q^*(\underline{x})$ in the sense of Kullback–Leibler divergence (see also [56]). Moreover, this divergence $D(Q^{\overline{w}}||Q^*)$ represents the gap between the minimax value of the regret and the minimax value of the expected regret. When the gap is small, optimization of the worst case value of the regret is not too different from optimization of the worst expected value over distributions in the class. In particular, if for some w the average regret $E_{Q^w} \log (P(\underline{x}|\hat{\theta}(\underline{x}))/Q^w(\underline{x}))$ and the NML regret $\log C(\mathcal{M}, n)$ agree asymptotically, then $D(Q^w(\underline{x})||Q^*(\underline{x})) \to 0$ and, consequently, w is asymptotically least favorable and the asymptotic minimax regret and minimax expected regret coincide. Such asymptotic agreement of (average) regret for the NML and mixture distributions is addressed next.

D. Asymptotic Equivalence of Optimal Solutions in Average and Worst Cases

The solutions to the two related minimax problems in the preceding subsection, namely, the NML distribution $Q^*(\underline{x})$ and the mixture $Q^w(\underline{x})$ with respect to a distribution $w(\theta)$

$$Q^w(\underline{x}) = \int P(\underline{x}|\theta)\, dw(\theta)$$

both have merits as defining the codelength for the MDL principle, and deserve to be studied more closely. The mixtures, in particular, for a fixed-weight distribution $w(\theta)$ have the advantage, in addition to average regret optimality, that they extend to a distribution on infinite sequences when $P(\underline{x}|\theta)$ is defined consistently for all n, and hence they define a random process. To do the same for the NML distribution, a construct of the type

$$Q^*(x_{n+1}|x^n) = Q^*(x^{n+1})/\Sigma_u\, Q^*(x^n, u)$$

may be used.

We can study the asymptotic codelength of these distributions for smooth parametric families $\{P(\underline{x}|\theta)\colon \theta \in \Theta \subset R^d\}$ on $\underline{x} = \mathcal{S}^n$ possessing an empirical Fisher information matrix $\hat{I}(\theta)$ of second derivatives of $(1/n)\log 1/P(\underline{x}|\theta)$. Let $I(\theta)$ be the corresponding Fisher information. We are interested both in the mean codelength and the pointwise codelength. We begin with the mixtures, for which the main technique is Laplace's approximation. Let $w(\theta)$ the prior density assumed to be continuous and positive. For smooth independent and identically distributed (i.i.d.) models, the expected regret is given by

$$\begin{aligned}&E_{P(\underline{x}|\theta)}[\log 1/Q^w(\underline{x}) - \log 1/P(\underline{x}|\hat{\theta})] \\ &\qquad = (d/2)\log \frac{n}{2\pi} + \log |I(\theta)|^{1/2}/w(\theta) + o(1)\end{aligned}$$

where the remainder tends to zero uniformly in compact sets in the interior of Θ (see Clarke and Barron, [9, p. 454], where references are given for suitable conditions on the family to ensure the regularity of the MLE). This expected regret expression leads naturally to the choice of $w(\theta)$ equal to the Jeffreys prior proportional to $|I(\theta)|^{1/2}$ to achieve an approximately constant expected regret when $|I(\theta)|^{1/2}$ is integrable on Θ. The Jeffreys prior is

$$w_J(\theta) = |I(\theta)|^{1/2}/c_J$$

where $c_J = \int |I(\theta)|^{1/2}\, d\theta$. This gives, uniformly in sets interior to the parameter space, a value for the expected regret of

$$(d/2)\log \frac{n}{2\pi} + \log \int |I(\theta)|^{1/2}\, d\theta + o(1) \qquad (8)$$

and, consequently, this is also the asymptotic value of the average regret $E_{Q^w} \log (P(\underline{x}|\hat{\theta}(\underline{x}))/Q^w(\underline{x}))$ with $w = w_J$. As discussed above, if the minimax regret has the same

asymptotics as this average regret, then the minimax regret and minimax expected regret agree asymptotically and $D(Q^{w_J}(\underline{x})||Q^*(\underline{x}))$ tends to zero. This asymptotic equivalence has been identified in the special case of the class of all discrete memoryless (i.i.d.) sources on a given finite alphabet in Xie and Barron [56]. Here we show it holds more generally.

The key property of Jeffreys prior [33] for statistics and information theory is that it is the locally invariant measure that makes small Kullback–Leibler balls have equal prior probability (see Hartigan [24, pp. 48–49]).

We next study the NML distribution and its asymptotic pointwise codelength. In Rissanen [43] conditions are given (without i.i.d. requirement) such that the code based on the NML distribution

$$Q^*(\underline{x}) = P(\underline{x}|\hat{\theta}(\underline{x}))/C_{\mathcal{M},n}$$

achieves regret that satisfies asymptotically

$$\log C(\mathcal{M}, n) = (d/2)\log\frac{n}{2\pi} + \log\int |I(\theta)|^{1/2}\, d\theta + o(1).$$

That gives the asymptotics of what we have called the parametric complexity. The stochastic complexity is the associated codelength based on the NML distribution, which satisfies

$$\log 1/Q^*(\underline{x}) = \log 1/P(\underline{x}|\hat{\theta}) + (d/2)\log\frac{n}{2\pi} + \log\int |I(\theta)|^{1/2}\, d\theta + o(1) \tag{9}$$

where the remainder does not depend on $\underline{x}$ and tends to zero as $n \to \infty$. The derivation in [43] directly examines the normalization factor in the NML code using a uniform central limit theorem assumption for the parameter estimates and does not involve Laplace's approximation.

The regret of this NML code is seen to agree with the average regret (8) of the mixture with Jeffreys' prior, in the sense that the difference tends to zero, which means that $D(Q^{w_J}(\underline{x})||Q^*(\underline{x}))$ tends to zero as $n \to \infty$, providing the desired asymptotic equivalence of the Jeffreys mixture and normalized maximum likelihood.

Though $Q^{w_J}(\underline{x})$ and $Q^*(\underline{x}))$ merge in the Kullback–Leibler sense, the ratio need not converge to one for every data sequence. Indeed, Laplace's approximation can be applied to obtain the pointwise codelength for the mixtures

$$\log 1/Q^w(\underline{x}) = \log 1/P(\underline{x}|\hat{\theta}) + \frac{d}{2}\log\frac{n}{2\pi} + \log\frac{|\hat{I}(\hat{\theta})|^{1/2}}{w(\hat{\theta})} + \epsilon_n(\underline{x}) \tag{10}$$

where for the remainder to be small it is necessary that $\underline{x}$ be such that $\hat{\theta}(\underline{x})$ is in the interior of Θ. Here again a choice for $w(\theta)$ as Jeffreys' prior yields the same parametric cost as in (9), except for the remainder terms $\frac{1}{2}\log|\hat{I}(\hat{\theta})|/|I(\hat{\theta})| + \epsilon_n(\underline{x})$. These should have the desired stochastic behavior of converging to zero in probability for each P_θ with θ interior to Θ. By arranging for modifications to the mixture to better encode sequences with $\hat{\theta}$ near the boundary, or with $\hat{I}$ not close to I, it is possible to obtain codelengths under suitable conditions such that uniformly in $\underline{x}$ they do not exceed the minimax regret asymptotically. See Xie and Barron [56] and Takeuchi and Barron [50].

E. Strong Optimality of Stochastic Complexity

We have seen that the solutions to the two minimax optimization problems behave in a similar manner, and the expressions (9) and (10) for the asymptotic codelength have the built-in terms we would like to see. First, there is the target value $\log 1/P(\underline{x}|\hat{\theta})$ that would be achievable only with advance knowledge of the maximum-likelihood estimate $\hat{\theta} = \hat{\theta}(\underline{x})$. Secondly, the remaining terms, dominated by the ubiquitous $(d/2)\log n$ penalty on the number of parameters, express the codelength price of our lack of advance knowledge of the best $\hat{\theta}$. Still, since the solutions are based on minimax criteria (for the regret or expected regret), a nagging suspicion remains that there might be another codelength which cannot be beaten except for some very rare sequences or a very small subset of the models. Reassuringly enough, it was shown in Rissanen [41] that the common behavior described above, in fact, is optimal for most models in the sense of the following theorem, which generalizes Shannon's noiseless coding theorem, by showing a positive lower bound on the redundancy of order $(k/2)\log n$ for most θ.

Assume that there exist estimates $\hat{\theta}(x^n)$ which satisfy the central limit theorem at each interior point of Θ, such that $\sqrt{n}(\hat{\theta}(x^n) - \theta)$ converges in distribution (or, more generally, such that $\sqrt{n}(\hat{\theta}(x^n) - \theta)$ is $O(1)$ in probability). Assume that the boundary of Θ has zero volume. If $Q(x^n)$ is any probability distribution for x^n, then (Rissanen [56]) for each positive number ϵ and for all $\theta \in \Theta$, except in a set whose volume goes to zero as $n \to \infty$

$$E_{P_\theta}\log\frac{P(x^n||\theta)}{Q(x^n)} \geq \frac{k-\epsilon}{2}\log n.$$

Later Merhav and Feder [37] gave similar conclusions bounding the measure of the set of models for which the redundancy is a specified amount less than a target value. They use the minimax redundancy for the target value without recourse to parametric regularity assumptions, and they use any asymptotically least favorable (asymptotically capacity achieving) prior as the measure. Under the parametric assumptions, Rissanen [42], with later refinements by Barron and Hengartner [6], shows that the set of parameter values with $\overline{\lim_n} E_{P_\theta}\log P(x^n||\theta)/Q(x^n)/\log n < (k/2)$ has volume equal to zero, and [6] shows how this conclusion may be used to strengthen classical statistical results on the negligibility of superefficient parameter estimation. Barron [2] obtained strong pointwise lower bounds that hold for almost every $\underline{x}$ sequence and almost every θ. Related almost sure results appear in Dawid [17].

Here we show that the technique in [2] yields asymptotic pointwise regret lower bounds for general codes that coincide (to within a small amount) with the asymptotic minimax regret including the constant terms.

Assume for the moment that $\underline{x}$ is distributed according to P and coded using Q. We recall the basic Markov-type inequality

$$P\{Q(\underline{x})/P(\underline{x}) \geq 2^c\} \leq 2^{-c}$$

and the implication that the codelength difference $\log 1/Q(\underline{x}) - \log 1/P(\underline{x})$ is greater than $-c$ except in a set of probability less than 2^{-c} for all $c \geq 0$. We apply the inequality with P replaced by the mixture distribution

$$P^w(\underline{x}) = \int P_\theta(\underline{x})\, dw(\theta)$$

to yield a conclusion for the codelength difference

$$d_n(\underline{x}) = \log 1/Q(\underline{x}) - \log 1/P^w(\underline{x}).$$

We find that for any choice of code distribution $Q(\underline{x})$, the P^w probability that $d_n \leq -2c$ is less than 2^{-2c}, so applying Markov's inequality to the prior probability

$$W\{\theta\colon\ P_\theta\{d_n \leq -2c\} > 2^{-c}\}$$

we find that it is not larger than 2^{-c}. The conclusion we will use is that the P_θ probability that $\log 1/Q(\underline{x})$ is less than $\log 1/P^w(\underline{x}) - 2c$ is less than 2^{-c}, except for a set B of θ with $w(B) \leq 2^{-c}$.

The Laplace approximation reveals under suitable conditions that

$$\log 1/P^w(\underline{x}) = \log 1/P(\underline{x}|\hat{\theta}) + \frac{d}{2}\log\frac{n}{2\pi} + \log |I(\theta)|^{1/2}/w(\theta) + o(1)$$

where the remainder $o(1)$ tends to zero in P_θ probability for each θ in the interior of the parameter space. Take w to be the Jeffreys prior density, which, because of its local invariance property for small information-theoretic balls, is natural to quantify the measure of exceptional sets of θ. The conclusion in this case becomes for any competing code distribution Q the code regret is lower-bounded by

$$\log 1/Q(\underline{x}) - \log 1/P(\underline{x}|\hat{\theta}) \geq \frac{d}{2}\log\frac{n}{2\pi} + \log\int |I(\theta)|^{1/2}\, d\theta - 2c + o(1)$$

where $o(1)$ tends to zero in P_θ-probability, for all θ in the interior of the parameter set, except for θ in a set B of Jeffreys probability less than 2^{-c}. This shows that asymptotically, the minimax regret cannot be beaten by much for most $\underline{x}$ with distribution P_θ for most θ.

Serendipitously, the basic inequality remains true with a uniformity over all n inside the probability. That is, $P\{\sup_n Q(\underline{x})/P(\underline{x}) \geq 2^c\}$ remains not greater than 2^{-c}, provided that the the sequences of distributions for $\underline{x} = x_1, x_2, \cdots, x_n$ remain compatible as n is increased (Barron [2, p. 28], Tulcea [51]). Consequently, setting

$$T = \sup_n \log^+ Q(\underline{x})/P(\underline{x})$$

we see that uniformly in n the excess codelength $\log 1/Q(\underline{x}) - \log 1/P(\underline{x})$ remains bigger than $-T$ where $T \geq 0$ has mean $E_P T$ not larger than $\log e$ and it is stochastically dominated by an exponential random variable $P\{T > c\} \leq 2^{-c}$. Using the Jeffreys mixture as the standard, it follows that for any competing compatible sequences of code distributions $Q(\underline{x})$ we have that for all n the codelength $\log 1/Q(\underline{x})$ is at least $\log 1/P^{w_J}(\underline{x}) - T$, which shows that the following strong pointwise lower bound holds P^{w_J}-almost surely, and hence also P_θ-almost surely for w_J-almost every θ

$$\log 1/Q(\underline{x}) > \log 1/P(\underline{x}|\hat{\theta}) + \frac{d}{2}\log\frac{n}{2\pi} + \log\int |I(\theta)|^{1/2}\, d\theta - T + o(1)$$

provided the remainder in the Laplace approximation tends to zero P_θ-almost surely, for almost every θ. To quantify the behavior of T we note that $P^w\{T > 2c\} \leq 2^{-2c}$ and hence $P_\theta\{T > 2c\} \leq 2^{-c}$ except in a set B of θ with Jeffreys probability less than 2^{-c}.

In summary, these results provide a grand generalization of Shannon's noiseless coding theorem in setting the limit to the available codelength and also demonstrating coding techniques which achieve the limit. For such reasons and due to the accurate evaluation of the codelength in (9) it was called in [43] the *stochastic complexity* of the data string, given the model class involved.

F. Simplification via Sufficiency

Both the NML and mixture codes have a decomposition, based on likelihood factorization for sufficient statistics, that permits insightful simplification of the computations in some cases. In this section we change the notation for the members of the parametric family to $f_\theta(\underline{x})$ or $p_\theta(\underline{x})$ rather than $p(\underline{x}|\theta)$ so as to maintain a clearer distinction from conditional distributions, given estimators, or other functions of the data. In particular, in this section, $p(\underline{x}|\hat{\theta})$ refers to the conditional distribution given the maximum-likelihood estimator rather than the likelihood evaluated at the MLE.

For a sufficient statistic $S = S(\underline{x})$ the probability of sequences $\underline{x}$ factors as $f_\theta(\underline{x}) = p(\underline{x}|s)p_\theta(s)$ where $p(\underline{x}|s)$ is the conditional probability function for $\underline{x}$ given $S(\underline{x}) = s$ (independent of θ by sufficiency) and

$$p_\theta(s) = \sum_{\underline{x}\colon S(\underline{x})=s} f_\theta(\underline{x})$$

is the probability function for the statistic S. As a consequence of the factorization, the maximum-likelihood estimate $\hat{\theta} = \hat{\theta}(s)$ may be regarded as a function of the sufficient statistic. Consequently, the maximized likelihood is $f_{\hat{\theta}}(\underline{x}) = p(\underline{x}|s)p_{\hat{\theta}}(s)$ at $s = S(\underline{x})$, and the normalizing constant simplifies to

$$C = C(\mathcal{M}, n) = \sum_{\underline{x}} f_{\hat{\theta}(\underline{x})}(\underline{x}) = \sum_s p_{\hat{\theta}(s)}(s)$$

since

$$\sum_{\underline{x}\colon s(\underline{x})=s} p(\underline{x}|s) = 1.$$

Thus there is a close connection between the NML distribution for s, namely, $p^*(s) = p_{\hat{\theta}(s)}(s)/C$, and the NML distribution for $\underline{x}$

$$f^*(\underline{x}) = p(\underline{x}|s)p^*(s)$$

at $s = S(\underline{x})$. The stochastic complexity, then, splits as

$$\log 1/f^*(\underline{x}) = \log 1/p(\underline{x}|s) + \log 1/p^*(s)$$

into the complexity of s plus the Shannon codelength for $\underline{x}$ given s. In much the same manner, the Bayes mixtures factor as

$$f_w(\underline{x}) = p(\underline{x}|s)p_w(s)$$

where $p_w(s) = \int p_\theta(s)\, dw(\theta)$ and Bayes optimal codelengths split as

$$\log 1/f_w(\underline{x}) = \log 1/p(\underline{x}|s) + \log 1/p_w(s).$$

Of particular interest is the case (which holds true in exponential families) that the maximum-likelihood estimator is itself a sufficient statistic $s = \hat{\theta}(\underline{x})$. In this case, the NML distribution for $\underline{x}$ is

$$f^*(\underline{x}) = p(\underline{x}|\hat{\theta})p^*(\hat{\theta})$$

and the NML distribution for $S = \hat{\theta}$ becomes

$$p^*(s) = p_{\hat{\theta}}(\hat{\theta})/C = g(\hat{\theta})/C,$$

where $g(\theta) = p_\theta(\theta)$ which is obtained simply as a density on the range of the parameter estimator by plugging into the distribution of the estimator the same value for the estimate as for the parameter, with normalization constant $C = \Sigma_{\hat{\theta}}\, p_{\hat{\theta}}(\hat{\theta})$. For example, in the Bernoulli(p) model the NML distribution of the relative frequency of ones $\hat{p} = y/n$ is

$$\binom{n}{y}(y/n)^y(1-y/n)^{n-y}/C_n$$

with

$$C_n = \sum_{y=0}^{n} \binom{n}{y} (y/n)^y(1-y/n)^{n-y}$$

which by Stirling's formula can be shown to be close to the Jeffreys Beta $(1/2, 1/2)$ distribution for p internal to $(0, 1)$. In the Gaussian model studied in Section V, the NML distribution for the sufficient statistic subject to certain constraints on the parameters is shown to be exactly Jeffreys' distribution.

IV. Inference

A. Predictive Coding and Estimation

A given joint distribution $Q(\underline{x}) = Q(x^n)$ on n-tuples can be written in the predictive or sequential form

$$Q(x^n) = \prod_t^n Q(x_t|x^{t-1}). \tag{11}$$

The converse is also true; that is, a joint distribution $Q(x^n)$ can be constructed by specifying the predictive distributions $\{Q(\cdot|x^{t-1})\}$.

For a given joint distribution Q, the factorization in (11) implies a predictive implementation of coding based on Q that encodes $x_1, x_2, \cdots, x_n$ one by one in that order. The codelength of x_1 is $-\log_2 Q(x_1)$. After the transmission of x_1, it will be known both to the sender and the receiver, and x_2 can be transmitted using the predictive distribution $Q(\cdot|x_1)$, which results in a codelength $-\log_2 Q(x_2|x_1)$. At time t the first $t-1$ data points $x^{t-1} = (x_1, \cdots, x_{t-1})$ are known to the sender and the receiver, and $Q(\cdot|x^{t-1})$ can be used to transmit x_t, which results in the codelength $-\log_2 Q(x_t|x^{t-1})$. In other words,

$$-\log_2 Q(x^n) = \sum_{t=1}^{n} -\log_2 Q(x_t|x^{t-1}) \tag{12}$$

which means that the total codelength for encoding x^n using $Q(\cdot)$ is the same as encoding the symbols one by ony using the predictive or conditional distribution $Q(\cdot|x^{t-1})$.

If we now postulate that the underlying source distribution is $P(x^n)$, the expected redundancy of Q with respect to P, which is different from the expected regret considered in Section III, is the Kullback–Leibler divergence between P and Q

$$\begin{aligned} D(P(x^n)||Q(x^n)) &= E_P \log_2 \frac{P(x^n)}{Q(x^n)} \\ &= E_P \log_2 \frac{\prod_t P(x_t|x^{t-1})}{\prod_t Q(x_t|x^{t-1})} \\ &= \sum_{t=1}^{n} E_{P(x^n)} \log_2 \frac{P(x_t|x^{t-1})}{Q(x_t|x^{t-1})} \\ &= \sum_{t=1}^{n} E_{P(x^{t-1})} E_{P(x_t|x^{t-1})} \\ &\quad \cdot \log_2 \frac{P(x_t|x^{t-1})}{Q(x_t|x^{t-1})} \\ &= \sum_{t=1}^{n} E_{P(x^{t-1})} D(P(\cdot|x^{t-1})\, Q(\cdot|x^{t-1})). \end{aligned}$$

This identity links the fundamental quantity, *expected redundancy*, from coding theory with statistical estimation, because the right-hand side is precisely the accumulated prediction error of the Kullback–Leibler risk of the sequence $\{Q(x_t|x^{t-1})\colon\ t = 1, \cdots, n\}$. This risk is equivalent to the mean squared error (MSE) when both P and Q are Gaussian distributions with the same covariance structure. In general, when P and Q are bounded away from zero, the Kullback–Leibler risk has a close connection with more traditional statistical estimation measures such as the square of the L^2 norm (MSE) and the Hellinger norm.

When Q is the mixture $Q(x^n) = \int P_\theta(x^n)\, dw(\theta)$ over a regular parametric family

$$\{P_\theta(x^n)\colon\ \theta \in \Theta \subset R^d\}$$

of d parameters with the mixing distribution or prior w, the tth summand in the accumulated risk is the risk of the Bayesian predictive distribution

$$Q(x_t|x^{t-1}) = \int P_\theta(x_t|x^{t-1}) P_w(\theta|x^{t-1})\, d\theta$$

where P_w is the posterior distribution of θ given x^{t-1}. In coding again, in order to build a code for x^n predictively, the predictive distribution $Q(\cdot|x^{t-1})$ allows us to revise the code in light of what we have learned from data prior to time t. For example, frequently appearing symbols should be assigned short codewords and less frequent ones long codewords. This predictive form lends itself naturally to the on-line adaptation

of coding or estimation to the underlying source. Moreover, it has an intimate connection with the prequential approach to statistical inference as advocated by Dawid, [15], [16].

Let Q be built on the plug-in predictive distribution based on an estimator $\hat{\theta}(x^n)$, which often is a suitably modified maximum-likelihood estimator to avoid singular probabilities

$$Q(x_t|x^{t-1}) = P_{\hat{\theta}(x^{t-1})}(x_t|x^{t-1}).$$

Then the tth summand in the accumulated risk is

$$E_{P_\theta(x^{t-1})} D(P_\theta(x_t|x^{t-1})||P_{\hat{\theta}(x^{t-1})}(x_t|x^{t-1}))$$

which is approximately $d/(2t)$ if $\hat{\theta}(x^t)$ is an efficient sequence of estimators. Summing up gives

$$\sum_{t=1}^{n} d/(2t) \approx \frac{d}{2}\log n$$

and this is exactly the leading term in the parametric complexity at sample size n. Hence, whether we consider estimation or predictive coding or, for that matter, any form of coding, we meet this same optimal leading term $(d/2)\log n$ in the regular parametric case, and it plays a fundamental role in both.

To bolster the connection given here between the individual risk of efficient estimators of order $d/2n$ and the optimal cumulative risk or redundancy of order $(d/2)\log n$, we mention here that classic results on negligibility of the set of parameter values for which an estimator is superefficient (LeCam [35] assuming bounded loss) are extended in Barron and Hengartner [6] to the Kullback–Leibler loss using results of Rissanen [42] on the negligibility of the set of parameter values with coding redundancy asymptotically less than $(1-\epsilon)(d/2)\log n$.

Frequently, we wish to fit models where the number of parameters is not fixed, such as the class of all histograms. For such the tth term in the accumulated risk

$$E_{P(x^{t-1})} D(P(x_t|x^{t-1})||P_{\hat{\theta}\hat{d}}(x_t|x^{t-1}))$$

where $\hat{d} = \hat{d}(x^{t-1})$ denotes the maximium-likelihood estimate of the number of parameters in $\hat{\theta} = \hat{\theta}(x^{t-1})$, behaves under suitable smoothness conditions as $t^{-\alpha}$ for some $0<\alpha<1$. Then the accumulated risk itself behaves as

$$\frac{1}{n}\sum_{t}^{n} t^{-\alpha} \approx n^{-\alpha}$$

and $n^{-\alpha}$ may be called the *nonparametric complexity per sample* at sample size n.

B. Consistency of the MDL Order Estimates

A test for any model selection and estimation procedure is to apply it to the selection of a model class and then analyze the result under the presumption that the data are generated by a model in one of the classes. It is to the credit of the MDL principle that the model-selection criteria derived from it are consistent although there are obviously other ways to devise directly consistent model-selection criteria, see, for example, Hannan [22] and Merhav and Ziv [36].

Consider first a family of parametric model classes, one for each k in a countable set K

$$\mathcal{M}_k = \{P(\underline{x}|\theta):\ \theta \in \Theta_k \subset R^{d_k}\}.$$

If we use the mixture model for each $\mathcal{M}_k$ to represent the class, we need to minimize

$$\log 1/Q_k(\underline{x}) + l(k)$$

where

$$Q_k(\underline{x}) = \int_{\Theta_k} w_k(\theta) P_k(\underline{x}|\theta)\, d\theta$$

and

$$\sum_{k\in K} 2^{-l(k)} \le 1.$$

Denote the data-generating class by $\mathcal{M}_{k_0}$. The MDL principle identifies $\mathcal{M}_{k_0}$ with Q_{k_0} probability tending to 1. That is, the MDL prinple leads to consistent-order selection criteria, on average, provided that Q_k are singular relative to Q_{k_0} on the space of infinite sequences. This is true, for example, if the Q_k are distinct stationary and ergodic distributions, or they are mixtures of such distributions, provided that the priors induced on the space of distributions are mutually singular. For instance, we may have parametric families of i.i.d., or Markov, distributions, where the parameter spaces are of different dimensions and absolutely continuous prior densities are assigned to each dimension.

The proof is simple, [2], [5]. Let Q^* be the mixture of Q_k except for k_0

$$Q^*(\underline{x}) = c_l^{-1} \sum_{k\neq k_0} 2^{-l(k)} Q_k(\underline{x})$$

where

$$c_l = \sum_{k\neq k_0} 2^{-l(k)} \le 1.$$

Because all the models in the summation are singular relative to Q_{k_0}, Q^* must be mutually singular with Q_{k_0}. It follows that the log-likelihood ratio or redundancy

$$\log Q_{k_0}(X^n)/Q^*(X^n) = \log 1/Q^*(X^n) - \log 1/Q_{k_0}(X^n)$$

tends almost surely to infinity, Doob [20]. We find that with Q_{k_0} probability one, for n large

$$\begin{aligned} l(k_0) + \log 1/Q_{k_0}(X^n) &< \log 1/Q^*(X^n) \\ &\le \min_{k\neq k_0}\{1/Q_k(X^n) + l(k)\}. \end{aligned}$$

The second inequality holds, because the sum

$$Q^*(X^n) = \sum_{k\neq k_0} 2^{-l(k)} Q_k(X^n)$$

is larger than the maximum of the summands. Thus the minimizing distribution is the distribution Q_{k_0} from the correct model class $\mathcal{M}_{k_0}$ as n tends to infinity and under probability Q_{k_0}, provided that Q_k are singular relative to Q_{k_0} on the infinite-sequence space, that is,

$$Q_{k_0}(\hat{k} = k_0) = \int w_{k_0}(\theta) P(\hat{k} = k_0|\theta)\, d\theta \to 1.$$

Moreover, $Q_{k_0}(\hat{k} = k_0 \text{ for all large } n) = 1$. This implies that $P_\theta(\hat{k} = k_0 \text{ for all large } n) = 1$ and hence that as $n \to \infty$

$$P(\hat{k} = k_0 | \theta) \to 1$$

for w_{k_0}-almost all $\theta \in \mathcal{M}_{k_0}$.

In many situations, such as nested exponential families, the above result holds for all $\theta \in \mathcal{M}_{k_0}$. The proof is more involved, but gives more insight. Roughly speaking, the mixture version of the MDL is an approximate penalized likelihood criterion just as the two-stage MDL, which asymptotically behaves as the Bayesian Information Criterion (BIC) of Schwartz [49].

For n large, in probability or almost surely

$$\log 1/Q_k(X^n) = \log 1/P(X^n|\hat{\theta}_k) + \frac{d_k}{2} \log n + O(1).$$

From classical parametric estimation theory for regular families, such as nested exponential families, we have the following asymptotic expansion:

$$\log 1/P(X^n|\hat{\theta}_k) - \log 1/P(X^n|\hat{\theta}_{k_0}) \begin{cases} \Rightarrow \chi^2_{k-k_0}, & \text{if } k \geq k_0 \\ n||\theta_k - \theta_{k_0}||^2(1 + o(1)), & \text{if } k < k_0. \end{cases}$$

This gives the consistency of the mixture MDL for all θ_{k_0}.

Since other forms of the MDL share the same asymptotic expression with the mixture, they also identify the correct model with probability tending to 1 as the sample size gets large. Consistency results for the predictive MDL principle can be found in [15], [17], and [32] for regression models, [23], and [31] for time-series models, and [53] for stochastic regression models. For exponential families, [27] gives a consistency result for BIC. Predictive, two-stage, and mixture forms of the MDL principle are studied and compared in [48] in terms of misfit probabilities and in two prediction frameworks for the regression model. It is worth noting that searching through all the subsets to find codelengths $l(\cdot)$ can be a nontrivial task on its own.

We note that for consistency any $\ell(k)$ satisfying Kraft's inequality is acceptable. However, for good finite sample behavior, as well as asymptotic behavior of risk and redundancy, one should pay closer attention to the issue of choice of description length of the models. The index of resolvability provides a means to gauge, in advance of observing the data, what sort of accuracy of estimation and data compression is to be expected for various hypothetical distributions, and thereby yields guidance in the choice of the model descriptions.

C. Resolvability

Perhaps more relevant than consistency of a selected model, which as formulated above would presume that the data are actually generated by a model in one of the candidate classes, is the demonstration that the MDL criterion is expected to give a suitable tradeoff between accuracy and complexity relative to the sample size, whether or not the models considered provide an exact representation of a data generating distribution. The index of resolvability from Barron and Cover [5] provides a tool for this analysis.

Consider first the case that the description length entails multiple stages, yielding a minimum description length of the form

$$\log 1/P_{\hat{k}}(\underline{x}|\hat{\theta}) + L_{\hat{k}}(\hat{\theta}) + \ell(\hat{k}) = \min_{k \in K} \min_{\theta \in \Theta_{k,\delta}} \{\log 1/P_k(\underline{x}|\theta) + L_k(\theta) + \ell(k)\}$$

where $\ell(k)$ is the codelength for the class index k in K, the term $L_k(\theta)$ is the codelength for parameter values of precision δ in a quantized parameter space $\Theta_{k,\delta}$, and, finally, $\log 1/P_k(\underline{x}|\theta)$ is the codelength for the data given the described class index and parameter values. (Typically, the precision δ is taken to be of order $1/\sqrt{n}$ so as to optimize the tradeoff between the terms in the description length, yielding $(d_k/2) \log n$ as a key component of $L_k(\theta)$.) Minimizing the description length in such a multistage code yields both a model selection $\hat{k}$ by MDL and a parameter estimate $\hat{\theta}$ (close to the maximum-likelihood estimate) in the selected family.

As in [5], it can be conceptually simpler to think of the pair k and θ as together specifying a model index, say m. Selection and estimation of $\hat{k}, \hat{\theta}$ provides an estimate $\hat{m}$. Then the above minimization is a special case of the following minimum description length formulation, where $\Sigma_m 2^{-L(m)} \leq 1$:

$$B(\underline{x}) = \log 1/P_{\hat{m}}(\underline{x}) + L(\hat{m}) = \min_m \{\log 1/P_m(\underline{x}) + L(m)\}.$$

The corresponding *index of resolvability* of a distribution P by the list of models P_m with sample size n is defined by

$$R_n(P) = \min_m \{D(P(x^n)||P_m(x^n))/n + L(m)/n\}$$

which expresses, in the form of the minimum expected description length per sample, the intrinsic tradeoff between Kullback–Leibler approximation error and the complexity relative to the sample size.

It is easy to see that $R_n(P)$ upper-bounds the expected redundancy per sample, which is

$$(1/n)E_P(B(X^n) - \log 1/P(X^n)).$$

It is also shown in Barron and Cover [5] that if the models P_m are i.i.d. and the data are indeed i.i.d. with respect to P, then the cumulative distribution corresponding to $P_{\hat{m}}$ converges (weakly) to P in probability, provided $\inf_m D(P||P_m) = 0$. Moreover, if $L(m)$ are modified to satisfy a somewhat more stringent summability requirement $\Sigma_m 2^{-\alpha L(m)}$ for some positive $\alpha < 1$, then the rate of convergence of $P_{\hat{m}}$ to P is bounded by the index of resolvability, in the sense that

$$H^2(P, P_{\hat{m}}) \leq O(R_n(P)) \tag{13}$$

in probability, where

$$H^2(P, Q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx$$

is the squared Hellinger norm between distributions with densities p and q. These bounds are used in [5] and [7] to derive convergence rates in nonparametric settings with the use of sequences of parametric models of size selected by MDL.

For description length based on a mixture model $Q^w(\underline{x}) = \int P(\underline{x}|\theta)\, dw(\theta)$ analogous performance bounds are available from a related quantity. In particular, the index of resolvability of a distribution $P(\underline{x})$ using the mixture of models $P(\underline{x}|\theta)$ with prior w and any parameter set A (usually chosen as a neighborhood around P) and sample size n is defined by

$$R_{n,A}(P) = \max_{\theta \in A} (1/n) D(P(\underline{x})||P(\underline{x}|\theta)) + (1/n) \log 1/w(A)$$

which when optimized over parameter sets A yields Kullback–Leibler balls

$$A_\epsilon = \{\theta\colon (1/n) D(P(\underline{x})||P(\underline{x}|\theta)) \le \epsilon^2\}$$

and index of resolvability

$$R_n(P) = \min_\epsilon \{\epsilon^2 + (1/n) \log 1/w(A_\epsilon)\}.$$

As shown in [3] (see also [4], [6], [29], and [57]), this quantity provides for the mixture code an upper bound to the expected redundancy per sample and thereby it also provides and upper bound to the Cesaro average of the Kullback–Leibler risk of the Bayes predictive estimators already discussed in Section IV-A.

Various parametric and nonparametric examples of determination of risk bounds of MDL estimators are possible as demonstrated in the cited literature; here we shall be content to give, in the next subsection, a general determination of the minimax rates in nonparametric settings.

D. Optimal Rate Minimax Estimation and Mixture Coding

In Section III we used Laplace's approximation method to obtain the behavior of the mixture distributions as solutions to the minimax mean redundancy problem for parametric models. In this subsection, we base our approach on a mixture code (or distribution) over a finite net to provide a unified approach to the upper and lower bounds on the optimal estimation rate in the minimax density estimation paradigm. However, the corresponding NML results in this nonparametric density estimation problem are yet to be developed, and NML's connection to the mixture distributions in this context is yet to be explored.

Fano's inequality from Information Theory has always been used to derive the lower bounds [8], [18], [25], [26], [58]. MDL-based density estimators now provide refinement to the lower bound and a matching upper bound as shown in Yang and Barron [57], revealing a Kolmogorov capacity characterization of the minimax values of risk and redundancy.

Consider a class $\mathcal{M}$ of i.i.d. densities, for which the distances between pairs of densities for x satisfy $D(f(\cdot)||g(\cdot)) \asymp H^2(f,g)$ for f and g in $\mathcal{M}$. This equivalence is satisfied by many, if not all, smooth function classes. The advantage of H^2 is that it satisfies the triangle inequality while D does not. However, D brings in clean information-theoretic identities and inequalities. Taking advantage of the equivalence of H^2 and D, we can switch between D and H^2 when appropriate to obtain a clear picture on optimal rate minimax estimation. Metric-entropy nets into which the estimation problem will be transferred turn out to be useful. Because such nets are finite, information-theoretic results such as Fano's inequality are easy to apply.

We are interested in the minimax estimation rates

$$\min_{\hat{f}} \max_{f \in \mathcal{M}} E_f H^2(f, \hat{f})$$

and

$$\min_{\hat{f}} \max_{f \in \mathcal{M}} E_f D(f(x)||\hat{f}(x))$$

where the minimum is over estimators $\hat{f}$ based on an i.i.d. sample of size n drawn from f, the divergence D is evaluated by averaging x with respect to f, independent of the sample, and E_f is taking the expected value of D as a function of the sample from f. Morever, we are interested in the minimax nonparametric complexity (redundancy)

$$\min_{Q} \max_{f \in \mathcal{M}} E_f D(f(x^n)||Q(x^n))$$

which, in accordance with Section IV-A, is the same as the minimax cumulative Kullback–Leibler risk. Here

$$f(x^n) = \prod_{i=1}^n f(x_i)$$

and the minimum is taken over all joint probability densities Q on $\mathcal{X}^n$ (which provide codes for x^n in $\mathcal{X}^n$). For a recent treatment of asymptotics and metric entropy characterization of the latter quantity see Haussler and Opper [30]. Here, following Yang and Barron [57], we focus on the relationship of the minimax risk to the nonparametric complexity and the Kolmogorov metric entropy as revealed through the resolvability and an improved application of Fano's inequality.

Let $V(\epsilon_n)$ be the Kolmogorov metric entropy of the ϵ_n-net $N(\epsilon_n)$ of $\mathcal{M}$ in terms of H or $\sqrt{D}$. That is, we need $N(\epsilon_n) = 2^{V(\epsilon_n)}$ number of ϵ_n balls to cover the class and no fewer. We use the code corresponding to the uniform mixture distribution $f_{\epsilon_n}(x^n)$ of the centers in the ϵ_n-cover. We examine the redundancy of the mixture with respect to the ϵ_n-net

$$R_n^{\epsilon_n}(f) = D(f(x^n)||f_{\epsilon_n}(x^n))$$

which from (13) is also the accumulated Kullback–Leibler prediction error of $f_{\epsilon_n}(x^n)$.

This is a crucial quantity in both upper and lower bounds on the minimax estimation error. It also bounds from above the risk of the mixture-induced density estimator $\hat{f}^{\epsilon_n}(x)$

$$E_f H^2(f, \hat{f}^{\epsilon_n}) \asymp E D(f||\hat{f}^{\epsilon_n}) \le \frac{1}{n} R_n^{\epsilon_n}(f),$$

where

$$\hat{f}^{\epsilon_n}(x) = (1/n) \sum_{i=0}^{n-1} f_{\epsilon_n}(x_{i+1} = x|x^i)$$

is the Cesaro average of the predictive density estimator induced by the mixture density $f_{\epsilon_n}(x^n)$ on the ϵ_n-net.

Moreover, there is a bound on $R_n^{\epsilon_n}(f)$ in terms of an index of resolvability. Let $g_f(x^n)$ be the closest member to

$f(x^n)$ in $N(\epsilon_n)$. Then $D(f(x^n)||g_f(x^n)) \le n\epsilon_n^2$, and

$$\begin{aligned} R_n^{\epsilon_n}(f) &= D(f(x^n)||f_{\epsilon_n}(x^n)) \\ &= E_f \log \left\{ f(x^n)/(2^{-V(\epsilon_n)} \sum_{g \in N(\epsilon_n)} g(x^n)) \right\} \\ &\le E_f \log \{f(x^n)/(2^{-V(\epsilon_n)} g_f(x^n))\} \\ &= V(\epsilon_n^2) + D(f(x^n)||g_f(x^n)) \\ &\le V(\epsilon_n) + n\epsilon_n^2. \end{aligned} \tag{14}$$

It follows that

$$E_f H^2(f, \hat{f}^{\epsilon_n}) \le V(\epsilon_n)/n + \epsilon_n^2. \tag{15}$$

The same order upper bound holds for a minimum complexity estimator as shown in Barron and Cover [5], in which one minimizes a two-stage codelength over f in $N(\epsilon_n)$.

By adjusting the choice of ϵ_n these upper bounds yield the minimax rate for the redundancy and consequently for the cumulative Kullback–Leibler risk of predictive estimation.

If the function class $\mathcal{M}$ is sufficiently large (in the sense that $\tilde{\epsilon}$, achieving $V(\tilde{\epsilon}) = V(\epsilon)/4$, is of the same order as ϵ, as ϵ tends to zero), then the bounds here also yield the minimax rate for the estimation of f in the traditional noncumulative formulation.

Indeed, by a now standard Information Theory technique introduced to the statistics community by Hasminskii [25] (see also [8], [18], [26], [57], and [58]) the estimation error in terms of a metric can be bounded from below by Fano's inequality via the probability of testing error on the finite ϵ_n-net. Here we choose the net $N(\epsilon_n)$ as a maximal packing set in which we have the largest number of densities in $\mathcal{M}$ that can be separated by ϵ_n in the Hellinger metric (consequently, it is also a cover in the sense that every f in $\mathcal{M}$ is within ϵ_n of a density in the net). For any given estimator $\hat{f}$ of f, by consideration of the estimator $\tilde{f}$ which replaces $\hat{f}$ by the closest density in the ϵ_n net and use of the triangle inequality for Hellinger distance, one has that

$$\max_{f \in \mathcal{M}} E_f H^2(f, \hat{f}) \ge (\epsilon_n/2)^2 \max_{f \in N(\epsilon_n} P_f(\tilde{f} \ne f).$$

Then by application of Fano's inequality there is a positive constant c such that for any estimator

$$\max_{f \in \mathcal{M}} E_f H^2(f, \hat{f}) \ge c\epsilon_n^2 (1 - (I_{\epsilon_n} + \log 2)/V(\epsilon_n))$$

where I_{ϵ_n} is the mutual information between Θ and X^n when Θ takes a uniform distribution on the ϵ_n-net and the conditional distribution of X^n, given a particular value f in the net, is $f(x^n) = \Pi_i f(x_i)$. Now as we recall this mutual information has been well studied, and ever since the development of Fano's inequality in the 1950's the precise nature of the capacity (the maximum value of the information over choices of input distribution) has played a central role in applications of Fano's inequality in Information Theory [11]. However, prior to reference [57], the mutual information in these statistical bounds had been bounded from above by the Kullback–Leibler diameter

$$n \times \max_{f,f'} D(f||f').$$

To yield satisfactory rate bounds, from what would otherwise be a crude bound on mutual information, required first restricting f, f' to a subset of $\mathcal{M}$ of special structure in which the diameter is of the same order as the separation $n\epsilon_n^2$ and the same order as its log-cardinality (typically, via a hypercube construction), plus a hope that the minimax rate on the subset would be as large as on the original family $\mathcal{M}$, and the existence of such a special structure was a condition of the theory, so that application of that theory requires the invention of a hypercube-like construction in each case. However, the requirement of such construction can be easily bypassed.

Indeed, since I_{ϵ_n} is the mimimum Bayes average redundancy with respect to a prior, it is not larger than the maximum redundancy of any given procedure. That is,

$$\begin{aligned} I_{\epsilon_n} &= \min_Q (1/2^{V(\epsilon_n)}) \sum_{f \in N(\epsilon_n)} D(f(x^n)||Q(x^n)) \\ &\le \min_Q \max_{f \in N(\epsilon_n)} D(f(x^n)||Q(x^n)) \\ &\le \min_Q \max_{f \in \mathcal{M}} D(f(x^n)||Q(x^n)). \end{aligned}$$

Hence for any joint distribution Q on x^n

$$I_{\epsilon_n} \le \max_{f \in \mathcal{M}} D(f(x^n)||Q(x^n)).$$

For $\tilde{\epsilon}_n$ to be chosen later, take $Q(x^n) = f_{\tilde{\epsilon}_n}(x^n)$ as the uniform mixture over the net $N(\tilde{\epsilon}_n)$ to get

$$I_{\epsilon_n} \le \max_{f \in \mathcal{M}} D(f(x^n)||f_{\tilde{\epsilon}_n}(x^n)) = R_n^{\tilde{\epsilon}_n}(f).$$

It follows from the resolvability bound (14) that

$$I_{\epsilon_n} \le V(\tilde{\epsilon}_n) + n\tilde{\epsilon}_n^2.$$

Hence

$$\max_{f \in \mathcal{M}} E_f H^2(f, \hat{f}) \ge c\epsilon_n^2 (1 - (V(\tilde{\epsilon}_n) + n\tilde{\epsilon}_n^2 + \log 2)/V(\epsilon_n)).$$

It is clear that the $V(\epsilon) + n\epsilon^2$ acts as the critical index of resolvability since it appears in both upper and lower bounds on the H^2 (or D) error in density estimation. It determines the minimax rate when $H^2 \asymp D$ as follows. Set $\tilde{\epsilon}_n$ to achieve $V(\tilde{\epsilon}_n) = n\tilde{\epsilon}_n^2$, thereby achieving the minimum order for $V(\epsilon) + n\epsilon^2$, and then choose ϵ_n somewhat smaller, but of the same order, such that

$$V(\tilde{\epsilon}_n) + n\tilde{\epsilon}_n^2 + \log 2 = V(\epsilon_n)/2.$$

Then we have

$$\max_{f \in \mathcal{M}} E_f H^2(f, \hat{f}) \ge c\epsilon_n^2/2.$$

Since the upper and lower bounds are of the same order we conclude that we have characterized the asymptotic rate of the minimax value.

Indeed, we find there is asymptotic agreement among several fundamental quantities: the nonparametric complexity (redundancy) per symbol

$$\min_g \max_{f \in \mathcal{M}} D(f(x^n)||g(x^n))/n,$$

the Shannon capacity

$$\max_{w(\theta) \text{ on } \mathcal{M}} I(\Theta; X^n)/n,$$

the Kolmogorov capacity $V(\epsilon_n)/n$, the critical radius ϵ_n^2, the minimax Cesaro average prediction risk

$$\min_{\hat{f}_0, \hat{f}_1, \cdots, \hat{f}_{n-1}} \max_{f \in \mathcal{M}} (1/n) \sum_{t=0}^{n-1} E_f D(f || \hat{f}_t),$$

the minimax Kullback–Leibler risk, and the minimax squared Hellinger risk based on a sample of size n.

These metric entropy characterizations of minimax rate in a nonparametric class $\mathcal{M}$ determine not only the minimax rate but also the rate achievable for most functions in the class, in the sense that for any sequence of estimators (or for any code distribution) the subclass of functions estimated at a better rate have a cover of asymptotically negligible size in comparison to $\mathcal{M}$. This is shown in Barron and Hengartner [6], extending the arguments of Rissanen [42] and in [45], and can also be shown by the methods of Merhav and Feder [37].

In the case of a Lipschitz or Sobolev class of functions on a bounded set, with s the order of smoothness, and several other function classes discussed in [57], the metric entropy is of order $V(\epsilon) \asymp \epsilon^{-1/s}$ for the L_2 metric and this remains the order of the metric entropy of the subclass of densities that are bounded and are bounded away from zero using $H, \sqrt{D}$, or L_2 for the distance. This leads to the optimal density estimation rate in terms of H^2 or D of $n^{-2/(2s+1)}$, which remains the optimal rate also in mean integrated squared error even if the densities are not bounded away from zero.

V. Applications

We discuss three applications of the MDL principle, the first on coding, the second on linear Gaussian regression, and the third on density estimation. As often is the case in nontrivial applications of the principle the model classes suggested by the nature of the applications turn out to be too large giving an infinite parametric or nonparametric complexity. A problem then arises regarding how to carve out a relevant subclass and how to construct a representative for it, the ideal being the stochastic complexity by the formula (9). However, computational issues often force us to use suitable mixtures or even combinations of the two perhaps together with the predictive method.

A. Universal Coding

Despite the close connection between the MDL principle and coding, the theory of universal coding and the code designs were developed without cognizance of the principle. This is perhaps because most universal codes, such as the widely used codes based on Lempel–Ziv incremental parsing, are predictive by nature, which means that there is no codebook that needs to be encoded, and hence the connection between the code redundancy and the number of bits needed to transmit the codebook was not made explicit until the emergence of a universal code based on context models, Rissanen [40]. We discuss briefly this latter type of universal codes.

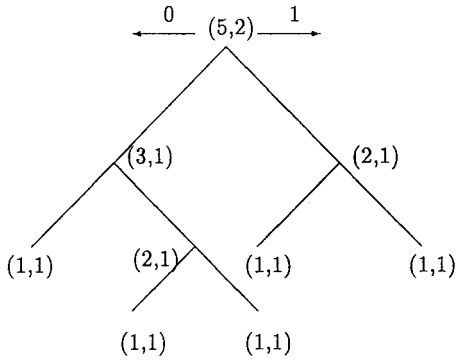

Fig. 1. Context tree for string 00100.

An obvious way to design a universal code for data modeled by a finite-state machine is to estimate the parameters from the data, including their number and the associated structure, and then use the result to encode the data. A particularly convenient way to do it is by an Arithmetic Code, see, e.g., Rissanen and Langdon [39], which is capable of encoding the individual symbols, even if they are binary, without the need to block them as required in the conventional Huffman codes. However, a direct execution of this program would require several passes through the data, which would result in an awkward code. In [40], an algorithm, called Context, was described which collects in a growing tree recursively, symbol for symbol, all the symbol occurrence counts in virtually all possible configurations of the immediately preceding symbols, called *contexts*, that the data string has. Hence, for instance, in the string 00100 the symbol value 0 of the fifth symbol x_5 occurs in the empty context λ four times. Out of these the preceding symbol is 0 twice, or, as we say, it occurs in the context 0 two times, and further out of these occurrences the symbol preceding the 0-context is 1 once. In other words, the substring 100 occurs once. Since extending the context to the left reduces the set of symbol occurrences it will be convenient to read the contexts from right to left. And this same phenomenon allows us to organize the nested sets of contexts in a binary tree, which can be grown recursively while also collecting the symbol occurrence counts. In such a representation, each node corresponds to a context, the root, in particular, to the empty context. We first spell out the relatively simple tree-growing algorithm, and show the tree obtained from the string 00100 in Fig. 1. We then describe how the special "encoding" nodes are chosen by use of the predictive version of the MDL principle described in Section IV, and, finally, we discuss to what extent the so-obtained universal model and data compression system achieve the ideal as defined in (9).

For the binary alphabet the tree-growing algorithm constructs a tree T_n for data string $x^n = x_1, \cdots, x_n$ with two counts c_0, c_1 at each node indicating the numbers of occurrences of the two symbols 0 and 1 at the context identified with the node, as follows.

1) Initialize T_0 as the 1-node tree with counts $(1, 1)$.
2) Having the tree T_{t-1}, read the next symbol $x_t = i$. "Climb" the tree along the path into the past x_{t-1}, $x_{t-2}, \cdots$ starting at the root and taking the branch

specified by x_{t-1}, and so on. At each node visited update the count c_i by 1. Climb until a node is reached whose count $c_i = 2$ after the update.

3) If the node is an internal node return to step 2). But if the node is a leaf, add two son nodes and initialize their counts to $(1, 1)$ and return to step 2).

Because of the initialization the counts exceed the real occurrence counts by unity, and they satisfy the important condition

$$c_i(s) = c_i(s0) + c_i(s1), \qquad i = 0, 1 \tag{16}$$

where $s0$ and $s1$ are the son nodes of s, whenever the sons' counts $c_i(s0)$ and $c_i(s1)$ are greater than 1.

Suppose we have constructed the tree T_t and intend to encode the symbol x_{t+1}. The values of the past symbols $\cdots, x_{t-2}, x_{t-1}, x_t$, when read in reverse, define a path from the root through consecutive nodes, each having the two counts with which the symbol could be encoded. Which node along this path should we choose? A quite convenient way is to apply the MDL principle and to search for the earliest node s^* along this path where the sum of the sons' stochastic complexity of the substrings of x^t, defined by their symbol occurrences, is larger than that of s^*. Indeed, the stochastic complexity of the symbol occurrences at each node defines an ideal code length for the same symbols, and the node comparison is fair, because by the condition (16) each symbol occurring at the father node also occurs at one of the son nodes. The symbol occurrences at each node or context s may be viewed as having being generated by a Bernoulli source, and we can apply (9) to compute the stochastic complexity, written here as $L(s)$, to a good approximation as follows:

$$L(s) = c(s)\log c(s) - \sum_{i=0,1} c_i(s)\log c_i(s) + \frac{1}{2}\log\frac{c(s)\pi}{2} + O(1/c(s)). \tag{17}$$

Instead of computing the stochastic complexities for every symbol occurrence by this formula it is much simpler to do it recursively as follows:

$$P(i|s) = \frac{c_i(s) + 1/2}{c(s) + 1} \tag{18}$$

and the counts are the ones when the symbol i occurred at the node s. This recursive implementation, when cumulated over all the past symbol occurrences at this node, gives to within the last term the stochastic complexity in (17). To get a universal code we encode the symbol x_{t+1} with the ideal codelength $-\log P(i|s^*)$ at the selected node s^*, which can be approximated with an arithmetic code as well as desired.

Collectively, all the special "encoding" nodes s^* carve out from the tree T_t a complete subtree $\mathcal{T}_t^*$, which defines a *Tree Machine* (Weinberger *et al*. [54]). If the data are generated by some *Tree Machine* in a class large enough to include the set of Markov chains as special *Tree Machines*, then with a somewhat more elaborated rule for selecting the encoding nodes the algorithm was shown to find the data-generating machine almost surely. (The algorithm given above differs slightly from the one in the cited reference but the results proven still hold.) Moreover, the ideal codelength for long strings defined by the resulting universal model, given as $L(x^n) = \Sigma_{s^*} L(s^*)$, differs from the stochastic complexity in (9) for the considered class of models by $o(\log n)$ or less. It cannot, however, agree with it completely, because the algorithm models the data as being generated by a collection of Bernoulli sources. In reality, the various Bernoulli processes at the states of a, say, Markov chain, are linked by the state transitions, which means that the stochastic complexity of a string, relative to Markov chains is smaller than the one defined by the ideal codelength of the universal Context algorithm. The same of course is true of the class of *Tree Machines*.

We conclude this subsection by mentioning another universal code (Willems *et al*. [55]), where no *Tree Machine* needs to be found. Instead, by an algorithm one can compute the weighted sum over all complete subtrees of T_t of the probabilities assigned to x^t by the leaves of the subtrees. When the weights are taken as "prior" probabilities we get a mixture of all *Tree Machine* models, each corresponding to a complete subtree. Again, since the codelengths defined by the complete subtrees differ from their stochastic complexities, the codelength of the mixture, which is comparable to that obtained with algorithm Context, will be larger than the stochastic complexity of data-generating *Tree Machines*.

B. Linear Regression

We consider the basic linear regression problem, where we have data of type $(y_t, x_{1t}, x_{2t}, \cdots)$ for $t = 1, 2, \cdots, n$, and we wish to learn how the values y_t of the *regression* variable y depend on the values $x_{it}, i = 1, 2, \cdots, K$, of the *regressor* variables $\{x_i\}$. There may be a large number of the regressor variables, and the problem of interest is to find out which subset of them may be regarded to be the most important. This is clearly a very difficult problem, because not only is it necessary to search through 2^K subsets but we must also be able to compare the performance of subsets of different sizes. Traditionally, the selection is done by hypothesis testing or by a variety of criteria such as AIC, BIC, [1], [49], and cross validation. They are approximations to prediction errors or to Bayes model selection criterion but they are not derived from any principle outside the model selection problem itself. We shall apply the MDL principle as the criterion, and the problem remains to find that subset of, say k, regressor variables which permit the shortest encoding of the observed values of the regression variables y^n, given the values of the subset of the regressor variables.

For small values of K, a complete search of the $\binom{K}{k}$ subsets for $k = 1, \cdots, K$ is possible, but for a large value we have to settle for a locally optimal subset. One rather convenient way is to sort the variables by the so-called "greedy" algorithm, which finds first the best single regressor variable, then the best partner, and so on, one at a time. In order to simplify the notations we label the regressor variables so that the most important is x_1, the next most important x_2, and so on, so that we need to find the value k such that the subset $\{x_1, x_2, \cdots, x_k\}$ is the best as determined by the MDL criterion.

We fit a linear model of type

$$y_t = \beta' \underline{x}_t + \epsilon_t = \sum_{i=1}^{k} \beta_i x_{it} + \epsilon_t \tag{19}$$

where the prime denotes transposition, and for the computation of the required codelengths the deviations ϵ_t are modeled as samples from an i.i.d. Gaussian process of zero mean and variance $\tau = \sigma^2$, also as a parameter. In such a model, the response data y^n, regarded as a column vector $y_1, \cdots, y_n$, is also normally distributed with the density function

$$f(y^n|X, \beta, \tau) = \frac{1}{(2\pi\tau)^{n/2}} e^{-(1/2\tau)\sum_t (y_t - \beta' \underline{x}_t)^2} \tag{20}$$

where $X_k' = \{x_{it}\}$ is the $k \times n$ matrix defined by the values of the regressor variables. Write $Z_k = X_k' X_k = nS_k$, which is taken to be positive definite. The development until the very end will be for a fixed value of k, and we drop the subindex k in the matrices above as well as in the parameters. The maximum-likelihood solution of the parameters is given by

$$\hat{\beta}(y^n) = Z^{-1} X' y^n \tag{21}$$

$$\hat{\tau}(y^n) = \frac{1}{n} \sum_t (y_t - \hat{\beta}'(y^n) \underline{x}_t)^2. \tag{22}$$

We next consider the NML density function (3)

$$\hat{f}(y^n|X) = \frac{f(y^n|X, \hat{\beta}(y^n), \hat{\tau}(y^n))}{\displaystyle\int_{Y(\tau_0, R)} f(z^n|\hat{\beta}(z^n), \hat{\tau}(z^n))\, dz^n} \tag{23}$$

where y^n is restricted to the set

$$Y(\tau_{0,R}) = \{z^n | \hat{\tau}(z^n) \geq \tau_0\}. \tag{24}$$

In this the lower bound τ_0 is determined by the precision with which the data are written. This is because we use the normal density function (20) to model the data and approximate the induced probability of a data point y, written to a precision δ, by δ times the density at y. For an adequate approximation, δ should be a fraction of the smallest value of σ of interest, namely, $\sqrt{\tau_0}$, which, in turn, has to be no larger than $\hat{\sigma}$. Put $R \geq \hat{\beta}'(y^n) S \hat{\beta}(y^n)$.

The numerator in (23) has a very simple form

$$f(y^n|X, \hat{\beta}(y^n), \hat{\tau}(y^n)) = 1/(2\pi e \hat{\tau}(y^n))^{n/2} \tag{25}$$

and the problem is to evaluate the integral in the denominator. In [19], Dom evaluated such an integral in a domain that also restricts the range of the estimates $\hat{\beta}$ to a hypercube. He did the evaluation in a direct manner using a coordinate transformation with its Jacobian. As discussed in Subsection III-D we can do it more simply and, more importantly, for the given simpler domain by using the facts that $\hat{\beta}$ and $\hat{\tau}$ are sufficient statistics for the family of normal models given, and that they are independent by Fisher's lemma. Hence if we with $\theta = (\beta, \tau)$ rewrite $f(y^n|X, \beta, \tau) = f(y^n; \theta)$ in order to avoid confusion, then we have first the factorization of the joint density function for y^n and $\hat{\theta}(y^n)$, which, of course, is still $f(y^n; \theta)$, as the product of the marginal density of $\hat{\theta}$ and the conditional density of y^n given $\hat{\theta}$

$$f(y^n; \theta) = p(y^n|\hat{\theta}(y^n); \theta) p(\hat{\theta}(y^n); \theta). \tag{26}$$

By the sufficiency of the statistic $\hat{\theta}$ we also have

$$f(y^n; \theta) = h(y^n) p(\hat{\theta}(y^n); \theta) \tag{27}$$

which shows that $p(y^n|\hat{\theta}(y^n); \theta) = h(y^n)$ is actually independent of θ. Moreover,

$$p(\hat{\theta}; \theta) = p_1(\hat{\beta}; \theta) p_2(\hat{\tau}; \tau) \tag{28}$$

where p_1 is normal with mean β and covariance $(\tau/n)S^{-1}$ while p_2 is obtained from the χ^2 distribution for $n\hat{\tau}/\tau$ with $n - k$ degrees of freedom.

Integrating the conditional $p(y^n|\hat{\theta}(y^n); \theta) = h(y^n)$ over y^n such that $\hat{\theta}(y^n)$ equals any fixed value $\hat{\theta}$ yields unity. Therefore, with

$$p(\hat{\theta}; \hat{\theta}) \equiv g(\hat{\tau}) = p_1(\hat{\beta}; \hat{\theta}) p_2(\hat{\tau}; \hat{\theta})$$

we get from the expression for the χ^2 density function in (28)

$$C = \int f(y^n|X, \hat{\theta}(y^n))\, dy^n \tag{29}$$

$$= A_{n,k} \int_{\tau_0, B}^{\infty} \tau^{-(k+2/2)}\, d\tau\, d\beta \tag{30}$$

$$= \frac{A_{n,k}}{k/2} \tau_0^{-k/2} \cdot V \tag{31}$$

where V is the volume of $B = \{\beta' S \beta \leq R\}$ and

$$A_{n,k} = \frac{|S|^{1/2}}{(2\pi)^{k/2}} \frac{\left(\dfrac{n-k}{2e}\right)^{n-k/2}}{\Gamma\left(\dfrac{n-k}{2}\right)}. \tag{32}$$

We then have the NML density function itself

$$\hat{f}(y^n|X) = \frac{1}{(2\pi e\hat{\tau})^{n/2} C} = \frac{1}{(2\pi e\hat{\tau})^{n/2} g(\hat{\tau})} w(\hat{\tau}) \tag{33}$$

$$w(\hat{\tau}) = g(\hat{\tau})/C. \tag{34}$$

Equations (34) and (32) give the stochastic complexity in exact form. However, the evaluation of the gamma function has to be done from an approximation formula, such as Stirling's formula. When this is done the stochastic complexity reduces to the general formula (9) with a sharper estimate for the remainder term $o(1)$. For this the Fisher information is needed, which is given by $|I(\beta, \tau)| = |S|/(2\tau^{k+2})$ and the integral of its square root by

$$\int_{\tau_0}^{\infty} |I(\beta, \tau)|^{1/2}\, d\tau = \frac{\sqrt{2}}{k} |S|^{1/2} \tau_0^{-k/2}. \tag{35}$$

We see in passing that the density function $w(\tau)$ agrees with

$$w(\tau) = |I(\beta, \tau)|^{1/2} \Big/ \int_{\tau_0, B}^{\infty} |I(\theta, \tau)|^{1/2}\, d\tau\, d\beta. \tag{36}$$

If we then apply Stirling's formula to the gamma function in (32) we get

$$-\ln \hat{f}(y^n|X) = \frac{n}{2}\ln(2\pi e\hat{\tau}) + \frac{k+1}{2}\ln\frac{n}{2\pi} + \ln V + \ln\int_{\tau_0}^{\infty} |I(\theta,\tau)|^{1/2}\,d\tau + R(k,n) \quad (37)$$

where

$$R(k,n) = -1/(12(n-k)) - k/(2n) + O(k^2/n^2) + O(1/(n-k)^3)$$

in agreement with the general formula (9) except that the term $o(1)$ gets sharpened.

This formula can be used as a criterion for selecting k provided the regressor variables are already sorted so that we only want to find the first k most important ones. This is because we may safely encode each k with the fixed codelength $\log K$, or $\log n$ if no other upper bound exists. If by contrast the variables are not sorted by importance we have to add to the criterion the codelength $\log\binom{K}{k}$ needed to encode each subset considered.

It is of some interest to compare the stochastic complexity derived with the mixture density with respect to Jeffreys' prior $|I(\beta,\tau)|^{1/2}$ divided by its integral, which, however, cannot be taken as in (35) but it must be computed over a range of both β and τ. The latter can be taken as above, but the former will have to be a set such that it includes $\hat{\beta}$ in its interior. A natural choice is a k-dimensional hyperball B or an ellipsoid defined by the matrix S of volume V_k. Jeffreys' prior, then, is given by $\pi(\beta,\tau)$ in (34). We need to calculate the mixture

$$f_\pi(y^n|X) = \frac{k}{2V_k}\tau_0^{k/2}\int_B d\beta\int_{\tau_0}^{\infty}\frac{1}{(2\pi\tau)^{n/2}} \cdot e^{(-n/2\tau)(\hat{\tau}+(\beta-\hat{\beta})'S(\beta-\hat{\beta}))}\tau^{-(k+2/2)}\,d\tau \quad (38)$$

$$< \frac{k}{2V_k(2\pi)^{n/2}}\tau_0^{k/2}|S|^{-1/2}\left(\frac{2\pi}{n}\right)^{-k/2} \cdot \int_{\tau_0}^{\infty} e^{(-n/2\tau)\hat{\tau}}\tau^{-(n+2/2)}\,d\tau \quad (39)$$

$$< \frac{k}{2V_k(2\pi)^{n/2}}\tau_0^{k/2}|S|^{-1/2}\left(\frac{2\pi}{n}\right)^{-k/2} \cdot \left(\frac{n\hat{\tau}}{2}\right)^{-n/2}\Gamma\left(\frac{n}{2}\right). \quad (40)$$

The first inequality comes from the fact that B does not capture all of the probability mass of the normal density. The second approximation is better; only the small probability mass falling in the initial interval $(0,\tau_0)$ of the inverse gamma distribution for τ is excluded. If we apply Stirling's formula to the gamma function we get

$$\ln\frac{\hat{f}(y^n|X)}{f_\pi(y^n|X)} > \ln V_k + \mathit{Rem}\,(k,n) \quad (41)$$

where $\mathit{Rem}\,(k,n)$ is a term similar to $R(k,n)$ in (37). For fixed k and large n the two criteria are essentially equivalent. This is because then the fixed set B will include virtually all of the probability mass of the normal density function in the mixure centered at $\hat{\beta}$, and the left-hand side of (41) will exceed the right-hand side only slightly. However, for small n, the set B will have to be taken relatively large to capture most of the said probability mass, which means that $f_{\pi(y^n|X)}$ will be a lot smaller than $\hat{f}(y^n|X)$, and the mixture criterion will not be as sharp as the one provided by the NML density.

C. Density Estimation

In this section we discuss a simple density estimator based on histograms. Consider a histogram density function on the unit interval with m equal-length bins, defined by the m bin probabilities $p = p_1,\cdots,p_m$ satisfying $p_1+\cdots+p_m = 1$

$$f(y|p,m) = mp_{i(y)} \quad (42)$$

where $i(y)$ denotes the index of the bin where y falls.

This extends to sequences x^n by independence. Write the resulting joint density function as $f(x^n|p,m)$. We are interested in calculating the NML density by use of (9). The Fisher information is given by $|I(p)| = \Pi_i p_i^{-1}$, and the integral of its square root, which is of Dirichlet's type, is given by $\pi^{m/2}/\Gamma(m/2)$. Equation (9) then gives

$$-\log\hat{f}(x^n|m) = -\log f(x^n|\hat{p},m) + \frac{m-1}{2}\log\frac{n}{2\pi} + \log\frac{\pi^{m/2}}{\Gamma(m/2)} + o(1) \quad (43)$$

where the components of $\hat{p}\equiv\hat{p}(x^n)$ are $c_j(x^n)/n$ and $c_j(x^n)$ denoting the number of data points from x^n that fall into the jth bin. Just as in the previous subsection one can obtain sharper estimates for the remainder than $o(1)$, but we will not need them.

Next, consider the mixture

$$\hat{f}(y|x^n) = \sum_{m=1}^{m(n)} w_n(m)\hat{f}(y|x^n,m) \quad (44)$$

where

$$w_n(m) = \frac{\hat{f}(x^n|m)}{\sum_{k=1}^{m(n)}\hat{f}(x^n|k)} \quad (45)$$

and $m(n) = \lceil n^{1/3}\rceil$ for large values of n. This number comes from analysis done in [45], where such a value for the number of bins was shown to be optimal asymptotically, when the ideal codelength for a predictive histogram estimator, equivalent to $\hat{f}(x^n|m)$, is minimized. For small values of n the choice of $m(n)$ could be made by the desired smoothness.

This estimator has rather remarkable properties. If the data are samples from some histogram with the number of bins less than $m(n)$, then the corresponding weight gets greatly emphasized, and the mixture behaves like the data-generating histogram. If again the data are generated by a smooth density function, then the mixture will also produce a surprisingly smooth estimate. To illustrate we took a test case and generated 400 data points by sampling a two-bin histogram on the unit

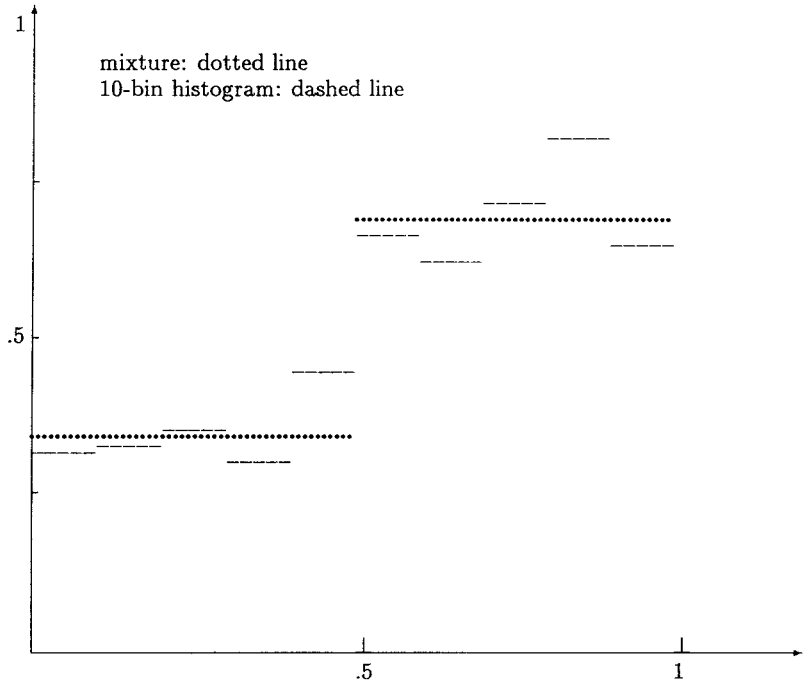

Fig. 2. Mixture and ten-bin histograms.

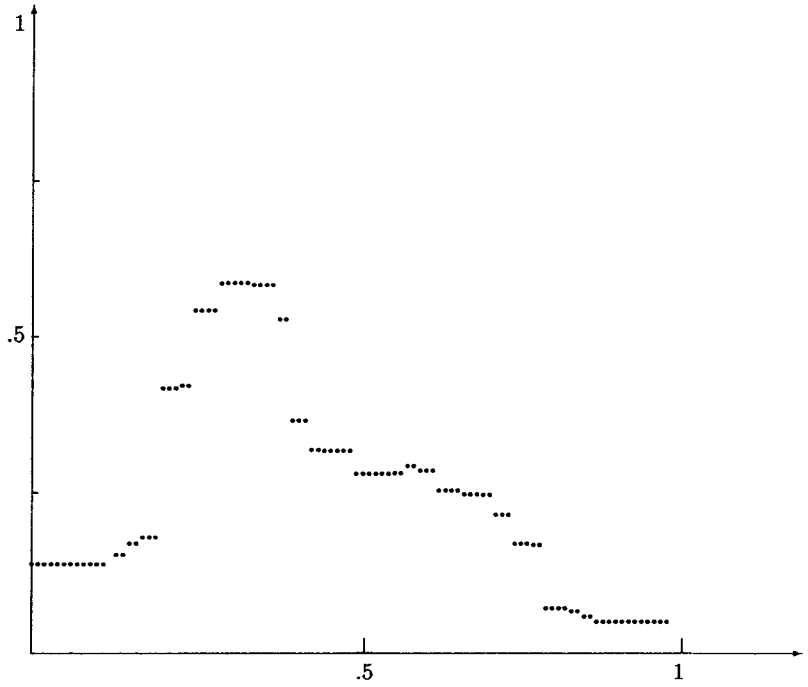

Fig. 3. A mixture histogram.

interval, where the first half has the probability mass 0.3 and the second 0.7, so that at the middle there is an abrupt change. We took $m(400) = 10$. Fig. 2 shows the result, dotted line, together with a ten-bin histogram, dashed line. The mixture nails down the two-bin data generating density just about perfectly, while the ten-bin histogram shows rather severe swings.

In Fig. 3 we have depicted the mixture density function with $m(n) = 10$ for another data set of size 76, not generated by any density function. The length of the steps is seen to be short in the rapidly changing regions of data density, which gives the illusion of smoothness and flexibility. Generating a continuous density function by connecting the dots with a curve would be easy, but to do so would require prior knowledge not present in the discrete data.

REFERENCES

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory*, Petrov

and Csaki, Eds. Budapest: Akademia Kiado, 1973, pp. 267–281.
[2] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, Sept. 1985.
[3] ———, "Are Bayes rules consistent in information?" in *Problems in Communications and Computation*, T. M. Cover and B. Gopinath, Eds. New York: Springer-Verlag, 1987, pp. 85–91.
[4] ———, "The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions," Tech. Rep. 7, Dept. Statist., Univ. Illinois, Apr. 1988.
[5] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, July 1991.
[6] A. R. Barron and N. Hengartner, "Information theory and superefficiency," *Ann. Statist.*, 1998, to be published.
[7] A. R. Barron, Y. Yang, and B. Yu, "Asymptotically optimal function estimation by minimum complexity criteria," in *Proc. 1994 IEEE Int. Symp. Information Theory*.
[8] L. Birgé, "Approximation dans les espaces metriques et theorie de l'estimation," *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, vol. 65, pp. 181–237, 1983.
[9] B. S. Clarke and A. R. Barron "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.
[10] ———, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Infer.*, vol. 41, pp. 37–60, 1994.
[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[12] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[13] ———, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, Mar. 1983.
[14] L. D. Davison and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, 1980.
[15] A. P. Dawid, "Present position and potential developments: Some personal views, statistical theory, the prequential approach," *J. Roy. Statist. Soc. A*, vol. 147, pt. 2, pp. 278–292, 1984.
[16] ———, "Prequential analysis, stochastic complexity and Bayesian inference," presented at the Fourth Valencia International Meeting on Bayesian Statistics, Peniscola, Spain, Apr. 15–20, 1991.
[17] ———, "Prequential data analysis," *Current Issues in Statistical Inference: Essays in Honor of D. Basu, IMS Monograph 17*, M. Ghosh and P. K. Pathak, Eds., 1992.
[18] L. Devroye, *A Course in Density Estimation*. Basel, Switzerland: Birkhäuser-Verlag, 1987.
[19] B. Dom, "MDL estimation for small sample sizes and its application to linear regression," IBM Res. Rep. RJ 10030, June 13, 1996, also submitted for publication.
[20] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1953.
[21] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Phil. Trans. Roy. Soc. London, Ser. A*, vol. 222, pp. 309–368, 1922.
[22] E. J. Hannan, "The estimation of the order of an ARMA process" *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[23] E. J. Hannan, A. J. McDougall, and D. S. Poskitt, "The determination of the order of an autoregression," *J. Roy. Statist. Soc. Ser. B*, vol. 51, pp. 217–233, 1989.
[24] J. A. Hartigan, *Bayes Theory*. New York: Springer-Verlag, 1983.
[25] R. Z. Hasminskii, "A lower bound on the risks of nonparametric estimates of densities in the uniform metric." *Theory Probab. Its Applic.*, vol. 23, pp. 794–796, 1978.
[26] R. Z. Hasminsikii and I. A. Ibragimov, "On density estimation in the view of Kolmogorov's ideas in approximation theory," *Ann. Statist.*, vol. 18, pp. 999–1010, 1990.
[27] D. Haughton, "Size of the error in the choice of a model fit data from an exponential family," *Sankhya Ser. A*, vol. 51, pp. 45–58, 1989.
[28] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, July 1997.
[29] D. Haussler and A. Barron, "How well do Bayes methods work for online prediction of $\{\pm 1\}$ values?" in *Proc. NEC Symp. Computation and Cognition*, 1992.
[30] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, 1997.
[31] E. M. Hemerly and M. H. A. Davis, "Strong constistency of the predictive least squares criterion for order determination of autoregressive processes," *Ann. Statist.*, vol. 17, pp. 941–946, 1989.
[32] U. Hjorth, "Model selection and forward validation," *Scand. J. Statist.*, vol. 9, pp. 95–105, 1982.
[33] J. Jeffreys, *Theory of Probability*, 3rd ed. Oxford, U.K.: Oxford Univ. Press, 1961.
[34] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Transm.*, vol. 1, pp. 4–7, 1965.
[35] L. LeCam, "On some asymptotic properties of maximum likelihood and related Bayes' estimates," in *University of California Publications in Statistics*, J. Neyman, M. Loeve, and O. Struve, Eds. London, U.K.: Cambridge Univ. Press, 1953, pp. 277–329.
[36] N. Merhav and J. Ziv, "Estimating the number of states of a finite-state source," *IEEE Trans. Inform. Theory*, vol. 41, pp. 61–65, Jan. 1992.
[37] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theory of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 714–722, May 1995.
[38] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[39] J. Rissanen and G. G. Langdon, Jr., "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, Jan. 1981.
[40] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656–664, Sept. 1983.
[41] ———, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.
[42] ———, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.
[43] ———, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.
[44] ———, "Hypothesis selection and testing by the MDL principle," invited paper for the special issue of *Comp. J.* devoted to Kolmogorov complexity, 1997.
[45] J. Rissanen, T. Speed, and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 38, pp. 315–323, Mar. 1992.
[46] Y. M. Shtarkov, "Universal sequential coding of single messages," *Probl. Inform. Transm.*, vol. 23, no. 3, pp. 3–17, July–Sept. 1987.
[47] R. J. Solomonoff, "A formal theory of inductive inference," Part I, *Inform. Contr.*, vol. 7, pp. 1–22; Part II, *Inform. Contr.*, vol. 7, pp. 224–254. 1964.
[48] T. P. Speed and B. Yu, "Model selection and prediction: Normal regression," *Ann. Inst. Statist. Math.*, vol. 45, pp. 35–54, 1994.
[49] G. Schwarz, "Estimation the dimension of a model" *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[50] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret for exponential and curved exponential families," in *Proc. IEEE 1998 Int. Symp. Information Theory* (Boston, MA, Aug. 1998).
[51] A. I. Tulcea, "Contributions to information theory for abstract alphabets," *Arkiv für Matematik*, vol. 4, no. 18, pp. 235–247, 1960.
[52] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Comp. J.*, vol. 11, no. 2, pp. 185–195, 1968.
[53] C. Z. Wei, "On the predictive least squares principle," *Ann. Statist.*, vol. 20, pp. 1–42, 1992.
[54] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, May 1995.
[55] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.
[56] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction,"*IEEE Trans. Inform. Theory*, submitted for publication, 1996.
[57] Y. H. Yang and A. R. Barron, "Information theoretic determination of minimax rates of convergence," *Annals of Statistics*, submitted for publication Oct. 1995, revised Mar. 1997.
[58] B. Yu, "Assouad, Fano, and Le Cam," *Festschrift in Honor of L. Le Cam on His 70th Birthday*, D. Pollard and G. Yang, Eds. Berlin, Germany: Springer-Verlag, 1997.
[59] B. Yu and T. P. Speed, "Data compression and histograms," *Probab. Theory Related Fields*, vol. 92, pp. 195–229, 1992.