*The Annals of Statistics*
1983, Vol. 11, No. 2, 416–431

# A UNIVERSAL PRIOR FOR INTEGERS AND ESTIMATION BY MINIMUM DESCRIPTION LENGTH

BY JORMA RISSANEN

*IBM Research, San Jose*

An earlier introduced estimation principle, which calls for minimization of the number of bits required to write down the observed data, has been reformulated to extend the classical maximum likelihood principle. The principle permits estimation of the number of the parameters in statistical models in addition to their values and even of the way the parameters appear in the models; i.e., of the model structures. The principle rests on a new way to interpret and construct a universal prior distribution for the integers, which makes sense even when the parameter is an individual object. Truncated real-valued parameters are converted to integers by dividing them by their precision, and their prior is determined from the universal prior for the integers by optimizing the precision.

**1. Introduction.** In this paper we study estimation based upon the principle of minimizing the total number of binary digits required to rewrite the observed data, when each observation is given with some precision. Instead of attempting at an absolutely shortest description, which would be futile, we look for the optimum relative to a class of parametrically given distributions. This Minimum Description Length (MDL) principle, which we introduced in a less comprehensive form in [25], turns out to degenerate to the more familiar Maximum Likelihood (ML) principle in case the number of parameters in the models is fixed, so that the description length of the parameters themselves can be ignored. In another extreme case, where the parameters determine the data, it similarly degenerates to Jaynes's principle of maximum entropy, [14]. But the main power of the new criterion is that it permits estimates of the entire model, its parameters, their number, and even the way the parameters appear in the model; i.e., the model structure. Hence, there will be no need to supplement the estimated parameters with a separate hypothesis test to decide whether a model is adequately parameterized or, perhaps, over parameterized.

We derive the following formula for the description length:

$$L(x, \theta) = -\log P(x \mid \theta) + \log^*[C(k)(\| \theta \|_{M(\theta)})^k]. \tag{1.1}$$

Here, $P(x \mid \theta)$ denotes the likelihood of the data $x$ for the parameter vector $\theta$ with $k$ components, $C(k)$ is the volume of the $k$-dimensional unit ball, and $\| \theta \|_{M(\theta)} = \sqrt{(\theta' M(\theta) \theta)}$ denotes the natural norm induced by the quadratic form associated with the $k \times k$ matrix $M(\theta)$ of the second derivatives of $-\log P(x \mid \theta)$. The function $\log^*$ is defined as $\log^* y = \log y + \log \log y + \cdots$, where only the positive terms are included in the sum. The criterion (1.1) is to be minimized with respect to all the parameters, including their number $k$.

The criterion (1.1), within a constant, is the negative binary logarithm of the joint probability $P(x, \theta)$ of the data and the parameters. It is obtained by optimizing the precision needed to express the parameters, and then using a universal prior distribution for the resulting integers, where the probability of integer $n$ is proportional to $2^{-\log^* n}$. This distribution, which is seen to modify the improper distribution $1/n$ of Jeffreys [18], is derived from coding of the integers in such a manner that certain natural coding theoretic requirements are satisfied. We argue that it correctly expresses one's initial ignorance

Received April 1981; revised September 1982.
*AMS* 1980. *subject classification.* Primary 62A99; secondary 62F10.
*Key words and phrases.* Likelihood, modeling, parameters.

when this notion is made precise. We do this in a novel way. First, the initial knowledge is interpreted to define a class of "test" distributions, and then a best code is found for the worst distribution in the family. The optimum code length defines at once the sought-for prior distribution.

Although the resulting optimum description of the integers is not unique, we show that the code lengths for integer $n$, assigned by all the optimum codes, deviate only slightly from $\log n$, just as the case is seen to be with the $\log^*$ function. This means that the dominant terms in the description length, expressed per observation, are given by the following expression, regardless of which optimum code is selected:

$$-(1/N)\log P(x \mid \theta) + (k/2N)\log(2\pi eN/k) + (k/N)\log \| \theta \|_{I(\theta)}, \tag{1.2}$$

where $I(\theta) = M(\theta)/N$, and $N$ denotes the number of observations. The first two terms define essentially the criteria of Schwarz [29] and Akaike [1] originally derived only for the Koopman-Darmois family of distributions, which now get extended to any smooth family of distributions. In addition, the controversial and restrictive assumption that the parameters be an outcome of a random experiment with a similar distribution, is no longer needed. The third term also improves the earlier criterion of Rissanen [25] in making the criterion invariant with respect to all nonsingular linear transformations.

Our approach, involving the code length of individual strings and their models, is in the spirit of the algorithmic notion of information due to Solomonoff [30] and Kolmogorov [20]. A major difference is that we require the class of models to be given at the outset, which allows us to get directly applicable results. We might say that our approach is more "statistical," although we also need to use simple ideas from information theory. We give at the end of Section 2 a comparative discussion of the earlier approaches that are either similar or otherwise relevant. The main purpose of Section 2 is to develop our view of estimation in an introductory and general way. We take pains to explain the few coding theoretical notions needed, which even a reader without prior exposure to information theory should be able to understand without much difficulty. After all, we never need to actually construct the various codes; it is enough to imagine that they can be constructed and to have an estimate of their length, which we do in the later sections. In Section 3 we discuss the universal priors. In Section 4 we describe how the real-valued ML estimates are truncated to an optimum precision, and we derive the two expressions (1.1) and (1.2) for the final criterion. At the end of the section we list a few of the applications, where this criterion is shown to lead to consistent estimates of models of considerable complexity. In the final Section 5, we show that Jaynes's principle of maximum entropy [14, 15, 16, 17] may be viewed as a particular instance of ours.

**2. Estimation, proper prior, and code length.** We view estimation as the problem of selecting a statistical model out of a parameterized class which best "explains" the observed sequence of data $x = x(1), \cdots, x(N)$ in a very concrete sense. Each observation $x(i)$ is either a real number such that it has no more than a fixed number of fractional digits in its binary representation, or it is a vector of such numbers. We consider a class of statistical *models* as a distributional form $P(\cdot \mid \theta)$, which for each parameter value $\theta$ assigns a distribution to the set of all possible sequences of the same length $N$ and, hence, the probability $P(x \mid \theta)$ to the particular observed sequence. The distribution, in addition, is often defined for every $N$, and it satisfies the compatibility conditions of a random process. Accordingly, we may regard a model to define such a process. Clearly, the process need not be independent nor stationary.

Often in such models the parameter $\theta$ is a vector of $k$ real-valued components, but these may not be independent in that there are either implicitly or explicitly given equations between them. For our purposes, however, we assume that the dependent parameters have been eliminated, and that the remaining $k$ parameters range over the $k$-dimensional euclidean space $\mathbb{R}^k$. We regard $k$ itself as a variable, so that there is such a set of parameters

$\theta$ for each $k$. Typically, we may arrange these spaces to form a nested sequence, each being a subspace of the next, so that $\theta$ may be regarded as a vector of the type $(\theta_1, \cdots, \theta_k, 0, \cdots)$. In the most general application that we are aware of, namely, modeling vector time series with linear systems, the set of the systems of each order is an analytic manifold. But even in that case we may regard the parameters as ranging locally over a $k$-dimensional euclidean space.

Traditionally, the parameter vector $\theta$ has a fixed known number of components. The principle of the maximum likelihood (ML) due to Gauss [7] and Fisher [6] prescribes as the estimate such a parameter value $\theta$ which makes the probability $P(x \mid \theta)$ of the observed sequence maximum. What makes this technique sound is the implicit assumption that the distributions $P(\cdot \mid \theta)$ are simple enough and the number of parameters is small enough not to permit a probability assignment one to the observed sequence. In more general situations, where the number of the parameters also must be estimated, the complexity of the model can no more be taken as fixed, for it clearly increases with the number of the parameters included in it. As a result, the likelihood of the observed sequence increases with this number, and the maximum is reached when there are as many parameters as observations, which clearly is not an interesting nor a useful result. In some cases this defect of the ML principle can be corrected by performing an additional hypothesis test, but in our opinion such a combined technique does not come to grips with the real problem.

We then see that the source of the problem is an issue of complexity of the model, which the likelihood function as such is not equipped to deal with. However, by a shift of view the maximum likelihood principle can be reinterpreted to measure the amount of complexity in the data, from which it can be extended to include even the complexity of the model. Indeed, we may regard the binary logarithm, $-\log P(x \mid \theta)$, otherwise known in information theory as the self-information and in statistics just as the negative log likelihood, to be the number of bits it takes to redescribe or encode $x$ with an ideal code relative to the assumed statistical model of the data, see Rissanen and Langdon [28]. In fact, one can show that only a vanishing fraction of long strings can be encoded with appreciably fewer bits than $-\log P(x \mid \theta)$, and that the ideal is reachable within a small fixed number of excess bits by use of a suitably designed coding. For a pertinent discussion of these matters in the context of algorithmic notion of information, see Levin [23].

With this interpretation the ML principle is equivalent with selecting that probabilistic model $P(x \mid \theta)$ which permits the shortest ideal code length for the observed sequence, provided that the model used in the encoding, i.e., the parameter $\theta$ is given, too. Unlike the ML principle, the new interpretation does not make appeal to one's subjective value judgement, and hence it seems more objective. But more importantly, now it makes perfect sense to talk about the storage requirement in describing the model itself, which after all must be done if the decoder wants to recreate the data $x$ from its code. We wish to emphasize that our information theoretic view of estimation differs in an essential manner from the earlier approaches that depend on the mean of the self information, the entropy, Jaynes [14], or on the so-called mutual information, Kullback [21]. In traditional information theory, the entropy is regarded as the ideal mean code length, while the expression $-\log P(x \mid \theta)$ has not been of direct interest in terms of a code length. With the introduction of the powerful "arithmetic codes," see e.g. Rissanen and Langdon [28], the measure $-\log P(x \mid \theta)$ for the ideal code length is more appropriate. For instance, it is applicable even when the source is non-stationary. In the currently studied case of estimation, its use is, of course, indispensable.

If we work with a fixed family of models, such that the description of the family does not depend on $x$ nor the parameters, the cost of the complexity of a model may be taken as the number of bits it takes to describe its parameters. Clearly now, when adding new parameters to the model, we must balance their own cost against the reduction they permit in the ideal code length, $-\log P(x \mid \theta)$, and we get the desired effect in a most natural manner. If we denote the total number of bits required to encode the parameters $\theta$ by

$L(\theta)$, then we can write the total code length as

$$L(x, \theta) = -\log P(x \mid \theta) + L(\theta), \tag{2.1}$$

which we seek to minimize over $\theta$. (Throughout this paper we write the binary logarithm just as "log"). The formula (2.1) reminds us strongly of Bayes's formula, written in a logarithmic form. In fact, if we write $Q(\theta)$ for $2^{-L(\theta)}$, then $P(x, \theta) = 2^{-L(x,\theta)}$ becomes a joint probability of $x$ and $\theta$, provided that the prior $Q(\cdot)$ indeed is a distribution, which in general it need not be. There are fortunately natural coding theoretic requirements, to be discussed in a moment, that do make $Q(\theta)$ a probability, and (2.1) then becomes a form of Bayes's law. We remind the reader that the parameters are truncated numbers, rather than elements of a continuum, and we do not need to deal with density functions.

There are two important points worth emphasizing in this view of estimation. First, the minimized total code length (2.1) provides a self-contained measure of the goodness of the fitted models. Different classes of models can perfectly well be tried and their performance compared against the same set of data. Ordinarily this is not possible because of the danger of "tailoring" the models to the data, and a fresh batch is needed to evaluate the results. What prevents "cheating" here is the requirement that the parameters, too, must be described, and it does not pay to let the parameters incorporate the coding of $x$ and hence claim a zero length code for it. The second point is that the meaning of the description length $L(\theta)$ and the associated probability $2^{-L(\theta)}$ of the parameter $\theta$ remain conceptually the same, whether we assume the existence of a frequency distribution or whether we view the parameters as individual objects for which no prior frequency distribution makes sense. In both cases, the parameters must be described one way or another, which offers a neat solution to the qualitative part of the bitter Bayesian—non-Bayesian dispute about priors; the quantitative part, namely, which prior distribution to assign to the parameters, remains, and it perhaps will never be settled to everyone's satisfaction. However, in the next section, we do our utmost to narrow the available rational choices to what we feel is quite irrelevant variations from something that may be regarded as a fundamental property of the natural numbers.

Our plan is to reduce the coding of the parameters to that of the natural numbers. In order to get an idea of this process, suppose the parameter is a single binary number such as 1011.11, written in two fractional places. By dividing this number by the precision ¼ we get the integer $n(\theta) = 101111$, which is seen to have the length $1 + [\log n(\theta)] = 6$, where $[y]$ is the greatest integer not greater than $y$. Hence we conclude that to describe the scalar parameter $\theta$, it takes about $\log n(\theta)$ bits, where $n(\theta)$ is obtained by dividing $\theta$ by its precision. The case with a vector parameter is similar, albeit more involved, and we leave its dicussion to Section 4.

The description length of the parameters may thus be reduced to the calculation of the length required to describe integers (provided that the precision is also communicated to the decoder), and it may seem that this induces a natural prior distribution for them, namely, $2^{-\log n} = 1/n$. These inverses, however, are not summable, and they therefore do not define a proper prior distribution. Curiously enough, Jeffreys [18] argued that for a real valued parameter $z$ the correct prior density, expressing complete ignorance, should be $dz/z$, which for integer valued parameters gives the distribution $1/z$. The fact that the integral of $dz/z$ diverges was not regarded as too disturbing; such priors are simply accepted as "improper" densities. No other than a pragmatic ad hoc meaning has been offered, however, for such "improper" densities. For an illuminating discussion of the various approaches to capture the elusive notion of a prior that adequately reflects initial ignorance or near ignorance about the parameters, we refer to Cox and Hinkley [3, pages 377–379].

There is a quite natural interpretation of the properness requirement in terms of code lengths, which will allow us to modify the preliminary choice $1/n$ so as to give a universal prior distribution for the integers. Since the codes of both $n(\theta)$ and $x$ are binary strings, we

cannot just attach them next to each other, for the decoder must be able to tell which portion of the total code string includes the code of $n(\theta)$. After all, he must be able to decode the parameters before he can start decoding the observed sequence from the rest of the code. We cannot use a special third symbol as a separating comma, because then the code string is no longer binary. It is well known, see eg. Elias [5] and Chaitin [2], and with a little thought easily appreciated, that if the decoder always reads the code string from left to right, then a both necessary and sufficient condition for the decoder to be able to separate the codeword from whatever binary string follows it, is that the codewords for the integers form a *prefix* set. This means that no codeword is allowed to be a prefix of another. Indeed, otherwise, how possibly could the decoder know, having reached the end of the shorter codeword, that this is the codeword meant to stand for the encoded integer rather than the longer one.

The prefix property has the important implication that necessarily the length $L(n)$ of the codeword for integer $n$ satisfies the Kraft inequality,

$$\sum_{n\geq 1} 2^{-L(n)} \leq 1. \tag{2.2}$$

To see why this is true, consider the infinite binary tree with the root node up, the nodes marked 0 and 1 in the first level below the root, the nodes marked 00, 01, 10, and 11 in the second level, and so on. Observe, that the sum of the terms $2^{-L(i)}$ equals one, when $i$ runs through all the nodes at each level, and $L(i)$ denotes the number of symbols in the marking of node $i$. Now, assign any so-marked node as the codeword for the integer 1, say in the level $L(1)$, and prune the tree by dropping all its descendant nodes so that this node becomes a terminal node; i.e., a "leaf." Form the sum $S = 2^{-L(1)}$. Next, assign another node to the next integer, 2, such that it is not in the path leading to any previously assigned codeword nodes, and that when the node's descendants are pruned, the remaining tree still has infinitely many nodes left. That is to say, we do not "close up" the tree, which in this case would happen only if the two first level nodes were assigned as codewords. Add the term $2^{-L(2)}$ to $S$. Continue by assigning codewords in this manner to all the integers, and it is clear that $S$ remains less than one, which implies (2.2).

In order to satisfy the requirement of a prefix code, we then must assign a longer codeword to the integer $n$ than $[\log n]$, the additional bits required to describe where the codeword ends. And by (2.2) the minimum requirement is to make the numbers $2^{-L(n)}$ add up to one, which is precisely what we need in order to get a proper prior. Exactly how prefix codes are designed for the set of integers is not an entirely trivial matter, because clearly we cannot just list them in a table but, instead, an algorithm is needed to define them. Although we do not actually need to construct them either, we illustrate a basic idea which also serves another purpose. Suppose that we were to store the binary integer $n =$ 101101000010110 in a computer, where other binary data follows it (after all, we never have the computer for ourselves alone!). We attach in front the binary number 1111 to tell the decoder that the next 15, or about $\log n$ symbols, define our integer. However, this is not enough, because how do we know that the given $4 \simeq \log 15 \simeq \text{loglog } n$ symbols, rather than 3 or 5 or any other number of initial symbols, have the required length information? So, we tell this by assigning $3 \simeq \log 4$ symbols, namely 100, to form another preamble. We end up assigning $22 \simeq \log n + \text{loglog } n + \text{logloglog } n$ symbols to define the integer $n$ in a self containing manner. (The description as given is not quite complete; for a complete description, see Elias [5].)

We conclude this section with a review of the relevant works that may be regarded as the predecessors of our approach. The outlined minimum description length principle to do estimation has its roots in the algorithmic notion of information due to Solomonoff [30] and Kolmogorov [20]; see also Chaitin [2] and Levin [24]. In that notion the information content of a (binary) string is defined to be the length of its shortest description; i.e., program that generates it, in a universal computer. Hence, in a sense, any of its generating programs may be viewed as a model for the string, and the shortest ones are the best. A theorem of considerable conceptual importance due to Kolmogorov states that there is no

program that would accept strings as the input and would produce as the output their shortest generating programs. Accordingly, it would be futile to construct an optimization problem, which seeks to isolate the best model among the class of all computable ones. The reason we do not get into a fundamental problem of this kind is that in our set-up we do not attempt to find the absolutely best model; rather, we seek a model within a restricted class, whose exact specification is left to the designer to suit the application at hand. For instance, for time series we may wish to consider the class of ARMA (Autoregressive Moving Average) models, without claiming that the best description of the observed data is to be found within this class.

While the models involved in the algorithmic theory of information are non-probabilistic, in the work of Solomonoff [30, 31], a very general apparatus is proposed for the generation of probabilities for all finite strings. In essence, the idea is to assign a probability to a given string proportional to $\sum 2^{-L(i)}$, where $L(i)$ denotes the length in the $i$th generating program for the string, and $i$ runs through all such programs. Hence, a string with many short programs in a universal computer gets assigned a high probability. This way, in effect, each universal computer generates a random process with its probabilities, which then can be used as a basis for inference. Because of the halting problem, however, the resulting probability function is not computable, and certain computable approximations are required. Much as in the algorithmic theory of information, Solomonoff seeks a globally best probabilistic model for observed strings. While this approach has undeniable, and should we say, philosophical appeal, it is futile to expect that one in any way would ever be able to find all the possible generating programs for long strings, and therefore the evaluation of the so defined universal probability assignment cannot be done, except, perhaps, for some very special strings. And again we are led to the more modest approach, where we consider a restricted class of models, already given in terms of probability distributions (except for the prior for the integers, to be discussed!), whose selection requires human ingenuity and intuition, but where the results are directly applicable.

Further, we mention the works of Good [8, 9], where information-like measures for "explicativity" and weight of evidence are studied. Although the proposed measures do not generally admit a description length interpretation, no more than the penalty term added to the likelihood in Good and Gaskins [10], the deep and philosophical observations in [8] and [9] do have some relevance to our basic notions. For example, Good in [9] maintains that the number of dimensionless parameters in a "law" is not an adequate measure of its complexity, because a parameter such as 5.4603 is more complex than another such as 2. Our Section 4 deals exclusively with this issue. Good also discusses the celebrated "Ockham's razor," which in its statement "plurality is not to be assumed without necessity" aptly captures the guiding principle in this paper.

After the first writing of this paper we became aware of a beautiful study of simplicity in induction by J. Kemeny [19], which in spirit is perhaps the closest prior publication to ours. Criticising the naive use of "Ockham's razor" to eliminate excess complexity in hypotheses, Kemeny suggests the following rule for model selection: Pick the simplest model *compatible* with the observed values. This he argues to be the way scientists actually construct their theories (= models). He then sets out to explicate the two vague terms in the rule, "simplest" and "compatible." The latter is done by declaring a hypothesis compatible, if the observations fall within its 99 percent confidence interval, while the notion of simplicity is entirely in line with the modern notions of complexity of description. Moreover, at the end of the paper Kemeny suggests (crediting N. Goodman for it) that one might try to find a criterion which combines simplicity and compatibility in an optimum manner, which would eliminate the need for the arbitrarily selected confidence level. And this is exactly what our principle does!

Having done a fair amount of literature search, we are led to the conclusion that while the basic elements in the principle of minimum description length can be found in several earlier papers, and they have been applied to various related ends, a clear formal declaration of the principle itself as a means to do estimation does not seem to have been

made prior to our paper in [25]. In addition, as will be evident, quite a bit is needed to convert this commonsensical principle to an explicit formula, applicable directly to a variety of estimation problems.

**3. A universal prior for integers.** Although the MDL principle appears to be free from subjective value judgments, there still remains something to be selected, namely, the coding of the integers. It is true that we could adopt Solomonoff's probability assignment to the integers, but what would be the point? We might just as well take the step all the way and adopt his probability assignment to all strings, and there would be no need for our MDL principle at all. It seems clear that it is at this point that we must make a compromise and try to construct a probability assignment to the integers, or, equivalently, a coding system, which is as free from subjective preferences as we can make it, while giving at the same time a formula, which permits us to evaluate the probability of any integer. Although we will have to settle for something less than a unique undisputedly superior coding system for the integers, that would satisfy the philosophically minded reader, we believe to have isolated the essence of such a system. This, fortunately, is all we need, because it allows us to work out an approximation to the total description length (2.1), which is sufficiently accurate to allow us to estimate parameters in models of unprecedented complexity.

In this section we draw on the work of Elias [5], where several coding systems for integers are described. Since we really do not need the codes themselves, we abstract from them a mathematical length function, which we prove to have the desired properties. But first we approach the problem from a different direction, mostly in order to get perspective.

Traditionally, statisticians have been concerned with the problem of how to express one's initial ignorance about parameters, which range over a subset $Z$ of integers. The crux of the problem is to define precisely the notion of "initial knowledge." We take a direct approach and define *initial knowledge* to be any set $\Omega$ of distributions $P = (P(1), P(2), \cdots)$ on the set $Z$. We call $P$ a "test" distribution. The intent is to regard the initial knowledge about the integers to act as constraints, so that little knowledge corresponds to a large class, and vice versa. For example, let $\Omega$ consist of all non-singular distributions on the interval $[1, M]$, where a distribution is non-singular if $P(i) < 1$ for all $i$. This set reflects the initial knowledge when only the upper bound $M$ is known about the integer.

We regard the test distributions as adversaries for the task of constructing a prefix code for the integers, i.e., a code where the lengths of the codewords satisfy the Kraft inequality (2.2). Instead of the entire code we really need only the sequence of the codeword lengths $L = (L(1), L(2), \cdots)$, defined on the same set of the integers as the test distributions. Being forced to use just one code for all the test distributions, we find it sensible to look for one which is no worse than it has to be. In other words, we seek to optimize the worst case code performance, where the natural measure of a code's performance is the code efficiency, defined to be the ratio of the entropy to the mean code length, or, equivalently, its inverse

$$\min_L \sup_{P\in\Omega} \sum_{i\in Z} [P(i)L(i)]/H(P). \tag{3.1}$$

In addition, we require $L$ to satisfy (2.2), which the optimal code is seen to do with equality. This causes the distribution $2^{-L(i)}$, defined by the code, to be necessarily proper.

To test this approach we work out two examples.

EXAMPLE 1. Let $\Omega$ consist of the single member $P$, as in a Bayesian case. The solution to (3.1) is by the fundamental inequality in information theory, Gibbs' theorem, seen to be $L(i) = -\log P(i)$. (This result, provable by direct optimization, states that $-\sum P(i) \log Q(i) \geq -\sum P(i) \log P(i)$ whenever the $Q(i)$'s add up to one, and the equality holds if and only if $Q(i) = P(i)$, for all $i$.) The prior defined by $L(i)$ is therefore $P$, which, of course, is an entirely noncontroversial result.

EXAMPLE 2. Let $\Omega$ consist of all non-singular distributions on $[1, M]$. Because for a

length law there is no finite supremum to the ratio in (3.1), we restrict the test distributions further such that $H(P) \geq \varepsilon$. This does not, however, weaken the result, because we have the solution: For every $\varepsilon$ and every length sequence $L$,

$$\max_P(\log M)/H(P) \leq \max_P\{\sum_{i=1}^{M} P(i)L(i)\}/H(P), \tag{3.2}$$

where $P$ is restricted such that $H(P) \geq \varepsilon$. The equality holds if and only if $L(i) = \log M$. Hence, again, the resulting prior $Q(i) = 1/M$ agrees with the generally accepted uniform distribution, which, however, is usually obtained on the dubious grounds of "insufficient reason." This, incidentally, appears to be only the second formulation of an optimum problem with the uniform distribution as the solution. The first was done by Jaynes; it follows at once from the results in the last section in the absence of the constraints (5.2).

To prove (3.2), let $L$ be a minimizing sequence, and suppose $L(i) < L(k)$ for some indices $i$ and $k$. Then for any maximizing $P$ it must be the case that $P(i) \leq P(k)$, because a permutation of the indices does not change the entropy. Add $\Delta(i)$ to $L(i)$ and subtract $\Delta(k)$ from $L(k)$, where these increments are chosen so that the equality in (2.2) holds. This means that within terms of order $\{\Delta(i)\}^2$, $\Delta(k) = -2^{L(k)-L(i)}\Delta(i)$. Hence, for these increments small enough

$$\Delta(i)P(i) + \Delta(k)P(k) = \Delta(i)\{P(i) - P(k)2^{L(k)-L(i)}\} < 0,$$

which reduces the right hand side ratio in (3.2) for every such maximizing $P$ and hence contradicts the optimality of $L$. Therefore, $L(i) = L(k)$ for all $i$ and $k$, and by the equality in (2.2), $L(i) = \log M$.

These two examples, where generally accepted priors exist, suggest that our formal approach appears to work in the desired manner, and we are ready for the main case: the set $Z$ consists of all positive integers. The first task is to select the set of test distributions to reflect the prior knowledge, which in this case amounts to ignorance or near ignorance. In fact, we do have some idea of the integer that we expect to turn out in, say, estimation: We expect its size to depend on the number of parameters, and hence very large integers requiring hundreds of digits to write down may occur. Although we cannot claim that the bigger the integer the less likely its occurrence, we still feel that this is so eventually. Such considerations cause us to put the constraints as follows

$$\begin{aligned} &\text{(i) } 1 > P(i) \text{ for all } i, \text{ and } P(i) \geq P(i+1),\ i > M, \text{ for some } M, \\ &\text{(ii) } H(P) = -\sum P(i)\log P(i) = \infty. \end{aligned} \tag{3.3}$$

Clearly, we have no idea of the number $M$. We may view the second condition, namely, that the distributions have infinite entropy, as a technical condition, needed to get a solution. It plays a role similar to the non-singularity condition in Example 2. One might also justify it by saying that a finite entropy on an infinite set is a nongeneric exceptional case, because it requires the bulk of the probability mass to lie in a finite interval. If we knew that to be the case, we would use the solution in Example 2. The fact that the mean ideal code length in this case is inifinite, does not, of course, lead to absurdities; every integer still has a finite code length.

Consider a code which maps the set of all positive integers $Z$ to the set of all finite binary strings in a one-to-one fashion, where, moreover, the codeword lengths $L(i)$ satisfy the Kraft inequality (2.2) with equality. We also put $0 < L(i) \leq L(i+1)$ for all $i$. This turns out to be a harmless restriction, because we will not find a better code even if we allow the wider class of codes where this constraint is relaxed. Furthermore, we feel that the universal prior $2^{-L(i)}$, defined by such a code, ought to be monotone nonincreasing, for otherwise one might wonder what the remarkable integers are where the successors have a higher probability. Such codes were called *representations* of integers by Elias [5].

We are then led to considering the following problem

$$\min_L \sup_P \lim_{N\to\infty}\{\sum_{i=1}^{N} P(i)L(i)\}/\{-\sum_{i=1}^{N} P(i)\log P(i)\}, \tag{3.4}$$

where $P = (P(1), P(2), \cdots)$ is a sequence of probabilities satisfying (3.3), and the monotone nondecreasing sequence $L = (L(1), L(2), \cdots)$ is to satisfy (2.2). We call any solution to (3.4) an *optimum* length sequence.

Define for any positive number $x$, $x \geq 1$, see Leung-Yan-Cheong and Cover [22], who used the construct to obtain lower bounds for the mean code length, $\log^*(x) = \log x + \log\log x + \cdots$, where the sum involves only the non-negative terms, whose number is clearly finite. These authors also proved that the sum $\sum 2^{-\log^*(n)} = c$ is finite; in fact, we show in Appendix A that $c \simeq 2.865064$. If we put $L^0(n) = \log^*(n) + \log(c)$, the Kraft-inequality holds with equality, and the sequence $L^0$ defines at once a universal prior for the natural numbers:

$$Q(n) = 2^{-L^0(n)}, \quad \text{for} \quad n > 0; \qquad L^0(n) = \log^*(n) + \log c. \tag{3.5}$$

THEOREM 1. *The sequence $L^0(n)$ is optimum.*

The proof is given in Appendix B.

We see that $Q(n)$ is given by

$$Q(n) = (1/n) \times (1/\log n) \times \cdots \times (1/\log \cdots \log n) \times (1/c), \tag{3.6}$$

where the first factor corresponds to Jeffreys's improper prior. The other terms reduce this dominant factor just enough to make $Q(n)$ a proper distribution. We can extend this prior to all non-negative integers by putting $Q(0) = \frac{1}{2}$ and replacing $c$ by $2c$. To extend this distribution to the set of all integers, add one to $L^0(n)$, and put $Q(-n) = Q(n)$.

A coding theoretic interpretation of the length sequence $L^0$ was already given at the end of Section 2, which, perhaps, together with Theorem 1 makes it a natural choice as the representative of the optimum solutions. We point out that in the log* function we cannot increase the base of the logarithm so as to generate a more efficient code without losing the summability of $2^{-\log^*(n)}$. However, other optimum solutions exist, such as $L^1(n) = \log n + 2\log\log n + c'$, where $c'$ is a positive constant, but we can prove that they all have $\log n$ as the dominant term, so that at least to a first approximation all universal priors for the integers behave alike. One may object to the choice of $L^0$ as the representative of the optimum solutions on the grounds that it is more complex than, say, $L^1$. A counter argument is that complexity here refers to the amount of computation needed to evaluate the function, which we are not particularly concerned with. An additional supporting argument is that $L^0(n)$ is shorter than $L^1(n)$ when the integers get larger. For example, for $n = 2^{32} \simeq 10^9$, $L^0(n) \simeq 40$ while $L^1(n) > 42$. In fact, we do not know of any other explicitly given optimum solution which would have a shorter length than $L^0(n)$ for large integers. For this reason, our view in the issue is this: Until someone comes up with a more efficient representation of the integers, we might as well use the best known to us; if again the emphasis is in simplicity of computation, we may pick as few of the leading terms as desired, often just the first.

THEOREM 2. *For every optimum length sequence $L$,*

$$\log n < L(n) < \log n + r(n),$$

*where $r(n)/\log n \to 0$ and $r(n) \to \infty$ as $n \to \infty$.*

PROOF. The first inequality follows from the fact that $L(n)$ is monotone nondecreasing and satisfies the Kraft-inequality. Indeed (Wyner's inequality [33]),

$$1 > 2^{-L(1)} + \cdots + 2^{-L(n)} \geq n2^{-L(n)}.$$

To show the rest, suppose $r(n)/\log n > \varepsilon$ for all $n$. With the notations $L_N = Q(1)L(1) + \cdots + Q(N)L(N)$ and $H_N = -Q(1)\log Q(1) - \cdots - Q(N)\log Q(N)$, where $Q(i)$ is from

(3.6), we would then have

$$L_N/H_N > (1+\varepsilon)(\sum_{i=1}^{N} Q(i)\log i)/H_N = (1+\varepsilon)(\sum_{i=1}^{N} q_{Ni}\log i)/(-\sum_{i=1}^{N} q_{Ni}\log q_{Ni} - \log Q_N),$$

where $Q_N = Q(1) + \cdots + Q(N)$ and $q_{Ni} = Q(i)/Q_N$. We show in Appendix A that $H_N$ goes to infinity with $N$, which with the fact that $Q_N \to 1$ implies that also the entropy in the denominator goes to infinity. Now, by Theorem 4 in Elias [5], the right hand side ratio tends to 1 as $N$ grows. Hence, the sequence $L(i)$, when applied to the distribution $Q(i)$, which satisfies the conditions (3.3), is seen to contradict the assumption of its optimality. Hence, $r(n)/\log n \to 0$. Clearly, $L(i)$ does not satisfy the Kraft-inequality unless $r(n) \to \infty$. □

It is worth summarizing in conclusion the meaning of the so-found prior distribution. First, it is based upon a description of integers, and hence it is independent of the assumptions about the role the parameters play. They may have an unknown frequency distribution, or, more importantly, they are allowed to be unique individual objects, such as a design parameter in a new industrial pilot plant which has never before been in existence. In contrast to the traditional and controversial interpretation of the prior as a subjectively selected "degree of rational belief" or "odds" in an imaginary betting game, our view is quite concrete and objective: The parameter value, being just an integer, must be described among all its possible values; i.e., all the integers. That the coding ought to be done the best we know how appears only to be a prudent requirement.

Ideally, we would have liked the set of the constraints (3.3) to have ensured a unique solution, but that was not to be. A consolation is, of course, provided by Theorem 2, which does not permit much leeway for the universal priors. With this theorem we not only get added confidence in using Jeffreys' improper prior $1/n$, but our interpretation also provides a natural explanation of what is being sacrificed in this approximation: the comma between the code of the integer and its preceding length information is being omitted. This is of some importance, because this writer, at least, has always felt uncomfortable when a distribution is being approximated by a nondistribution, as the case was with Jeffreys's interpretation.

**4. Optimal precision.** Having a universal distribution on the integers, we now return to the calculation of the length $L(\theta)$, required to describe vector parameters of components that are truncated real numbers. In many cases of interest the first term in (2.1) is dominant, and for each number of parameters the minimizing parameter values are close to the ML estimates. Hence, the problem is to decide on the precision to be used for the ML estimates of the vector $\theta$. Clearly, if we use a coarse precision and describe each parameter with only a few bits, the second term $L(\theta)$ will be small, but the first term will grow from its minimum, since we in general are no longer using the correct ML estimates. In [25], following a suggestion by Wallace and Boulton [32], we solved the optimum precision problem in a manner which initially appeared to be adequate but which now can be seen to have the defect that the resulting estimation criterion lacks the necessary invariance properties (This is no reflection on the cited authors, who discuss a different case where no such defects occur.)

Recall in Section 2 that we regard the space of the parameters $\theta$ for each value of $k$ to be the space $R^k$. The relevant point here is that the parameters are not constrained by equations so as to make their image in $R^k$ a hyper surface. Originally, the parameters may well be so constrained, but the parameters that we are considering here are thought to form an independent generating set. In the following discussion let $\theta$ denote the ML estimate.

Fundamentally, a truncation process calls for a definition of an equivalence relation for the pairs of vectors in $R^k$ and a selection of a representative in each equivalence class as the truncated value for each of its elements. We pick the equivalence classes to be of some

easily described type, such as translates of a $k$-dimensional parallelepiped $A$, and the truncated value $\theta'$ of the parameter $\theta$ may be taken as the center of the parallelepiped in which $\theta$ falls. Further, these representatives $\theta'$ are converted efficiently to integers by ordering them, and the code of $\theta'$ is the position index of the first rectangle in the ordered list that includes $\theta$. Clearly, the ordering is to be done independently of any particular element of $R^k$, except the origin, and all orderings are equivalent unless we add other desired requirements, as we shall do in a moment when we also sketch in a bit more detail a particular enumeration of the rectangles. We also comment on the selection of the origin later.

The knowledge of the class of the models $\{P(x/\theta)\}$ permits the selection of the basic parallelepiped $A$ to be done in a manner which in some cases is optimum and in other cases locally optimum. For this, we assume the usual smoothness of the term $-\log P(x/\theta)$, so that we can expand it around its minimizing point $\theta$, the maximum likelihood estimate, to within second order terms as

$$-\log P(x/\theta') \simeq -\log P(x/\theta) + \tfrac{1}{2}(\delta, M(\theta)\delta). \tag{4.1}$$

Here $\delta = \theta' - \theta$, and $(x, y)$ denotes the inner product. $M(\theta)$ is the symmetric operator defined by the second partials of $-\log P(x/\theta)$. Since we picked the truncated value $\theta'$ as the center of the parallelepiped of the same shape and size as $A$, we conclude that the maximum value for the quadratic term results when $\theta$ falls at one corner. Instead of the maximum value, we could calculate a mean increase by assuming some distribution for the deviation of $\theta$ from the center of its enclosing rectangle. Using the maximum value has the advantage, however, that it is independent of such distributions. We further see that for a fixed permitted maximum deviation $d = (\delta, M(\theta)\delta)$, we achieve the coarsest precision if we pick the parallelepiped $A$ a rectangle within this ellipsoid so that its volume $V$ is maximized. Such a rectangle turns out to have the edges parallel to the principal axes of the ellipsoid, and the maximum volume is given by

$$V(d) = (4d/k)^{(1/2)k}/\sqrt{\det M(\theta)}. \tag{4.2}$$

This rectangle $A$ may well be regarded to provide the precision on the level $V(d)$, depending on the parameter $d$, for the space. In the ideal case, where $M(\theta)$ does not depend on $\theta$, the decoder can calculate $A$ from the description of the model class alone. But even in other cases in which $M(\theta)$ changes smoothly with $\theta$, we may use a fixed approximation of the eigenvectors and eigenvalues of $M(\theta)$ to provide a near optimum truncation of $\theta$ in a neighborhood, and hence to arrive at an estimation criterion to evaluate the parameters within that neighborhood.

As stated above, the shifted rectangle in which $\theta$ falls, and therefore its index $n(\theta)$, too, as the code of its truncation, depends on the particular way the space $R^k$ is covered by these rectangles. Just as the universal code length for the integers is a monotone nondecreasing sequence, we clearly want an enumeration of the rectangles to be such that the index grows with the distance from the origin. The question is which norm to use to measure the distance. By and large, we will end up in a similar code length for $\theta$ no matter which norm we use, but we will get a particularly satisfying invariance property if we pick the natural norm, $\|\theta\|_{M(\theta)} = \sqrt{(\theta, M(\theta)\theta)}$, which also has to do with statistical "curvature," Efron [4]. We sketch how the rectangles are made to cover the space so as to correspond to this norm. First, the maximal rectangle $A(1) = A$ within the ellipsoid $(y, M(\theta)y) = d$ is centered at the origin. It cuts the principal axes at points, defined by the vectors $a_i$ of length $\sqrt{(d/(k\lambda_i))}$, where $\lambda_i$ denotes the $i$th eigenvalue of $M(\theta)$. The other rectangles $A(2)$, $A(3)$, $\cdots$, are translations of $A$ to the centers $z = n_1a_1 + \cdots + n_ka_k$, where $n_i$ are integers. Moreover, we place these centers so that the set of the rectangles within each ellipsoid of the form $(y, M(\theta)y) =$ constant forms a consecutive sequence $A(1), \cdots, A(n)$. In other words, we cover the space in a growing spiral fashion following the elliptic surfaces.

As a consequence of this enumeration, the index $n(\theta)$ is given approximately by the ratio of the volume enclosed by the ellipsoid $(y, M(\theta)y) = D = (\theta, M(\theta)\theta)$ to the volume

$V(d)$, or $n(\theta) = C(k)(kD/4d)^{k/2} + O(|A|)$, where $O(|A|)$ represents a portion of the very last layer of the rectangles intersecting the surface of the ellipsoid. Further, $|A| = 2\sqrt{(d/k\lambda_{\min})}$ denotes the maximum edge of the rectangle $A$, where $\lambda_{\min}$ is the minimum eigenvalue of $M(\theta)$. Because $\lambda_{\min}$ grows asymptotically like $N$, $|A|$ shrinks asymptotically like $1/\sqrt{N}$. The coefficient $C(k)$ is defined by the volume of the $k$-dimensional unit ball, and it is given by

$$C(k) = \begin{cases} (2\pi)^{(1/2)k}/(\tfrac{1}{2}k)!2^{(1/2)k}, & k \text{ even}, \\ \pi^{(1/2)(k-1)}2^{k+1}(\tfrac{1}{2}(k+1))!/(k+1)!, & k \text{ odd}. \end{cases}$$

Using Stirling's formula we get

$$\log C(k) = -\tfrac{1}{2}(k+1)\log k + \tfrac{1}{2}k\log(2\pi e) + r(k), \tag{4.3}$$

where $r(k)$ satisfies $c < r(k) < c + 1$ for a constant $c$.

From (4.2) and the just derived formula for $n(\theta)$ we see that the larger the permitted deviation $d/2$ is from the minimum $-\log P(x/\theta)$, the larger volume $V(d)$ we get, and the smaller the index $n(\theta)$ will be. We may ask for the value of $d$ which minimizes the sum

$$\log^* n(\theta) + d/2. \tag{4.4}$$

The solution is easy if we replace log* by log. Omitting the term $O(|A|)$, the optimum value for $d$ is by a straightforward differentiation seen to be $ka$, where $a = 1/\log e$, or about 0.693.

When the resulting optimum value is added to $-\log P(x/\theta)$, we get the approximation to the total code length as

$$L(x, \theta) = -\log P(x/\theta) + k \log \|\theta\|_{M(\theta)} + \log C(k).$$

A normalization with (4.3) gives

$$\ell(x, \theta) = -(1/N)\log P(x/\theta) + (k/2N)\log(2\pi eN/k) + (k/N)\log \|\theta\|_{I(\theta)} + O((\log k)/N), \tag{4.5}$$

where $I(\theta) = M(\theta)/N$. If the sum of the second and the third terms in (4.5) is negative we drop them both, because it means that the index $n(\theta)$ is one, which, in turn, means that $\theta$ is truncated to the 0-vector. The intended use of this criterion is to successively minimize it for each $k = 0, 1, \cdots$ over the parameters $\theta$ until a minimizing $k$ and the associated parameters are found.

The minimization problem (4.4), where log* is retained, can also be solved, numerically at least, but a new difficulty arises. To illustrate this, let us use the approximation $\log n + \log\log n$ for log*. Then the minimizing value for $d$ is given by $ak(1 + a/\log n(\theta))$. The fact that $d$ now depends on $\theta$ implies that each truncated $\theta$ requires its own precision. This leads to a complex coding system, although still an implementable one. To see the peculiarity of the resulting coding, consider the $j$th rectangle, $j < n(\theta)$, in the sequence $1, \cdots, j, \cdots, n(\theta)$ of the rectangles that are closer to the origin than $n(\theta)$. This rectangle is not the optimum equivalence class of its center, say. If we wish to encode this center optimally, we shall have to construct a new sequence of rectangles $1, \cdots, j(\theta)$ of a different size than those in the sequence for $n(\theta)$. We might say that each point in the space carries along its own yardstick to measure its path length to the origin. In order to avoid such a cumbersome coding system, we use the same value for $d$ as above, namely, $d = ka$, which gives an approximation to the ideal code length as

$$l^*(x, \theta) = -(1/N)\log P(x/\theta) + (1/N)\log^*\{C(k)[N(\theta, I(\theta)\theta)]^{(1/2)k}\}. \tag{4.6}$$

This has the advantage over (4.5) that it admits a Bayesian interpretation.

REMARKS. If the criteria (4.5) and (4.6) purport to express an intrinsic property of the data and the given family of models, then, clearly, the result ought to be independent of

scaling of the parameters and, more generally, of selection of coordinate systems. The dominant first term, the maximum likelihood term, causes no problems, for any invertible transformation $g$: $R^k \to R^k$, $\theta \to \theta'' = g(\theta)$, only relabels the same set of the parameters, and the value of the minimum $-\log P(x/\theta)$ certainly does not change with such relabelings. The second term in (4.5), clearly, remains the same if we for the moment keep the number of the parameters fixed. The third term in (4.5) and the second in (4.6) are the ones of possible concern, because they involve a specific functional form, namely, the quadratic.

Let $g$: $R^k \to R^k$ be a twice differentiable transformation such that the Jacobian $\partial\theta/\partial\theta''$ is nonsingular. If we differentiate $-(1/N) \log P(x/\theta'')$ with respect to the variables $\theta''$, we then have

$$I''(\theta'') = (\partial\theta/\partial\theta'')'I(\theta)(\partial\theta/\partial\theta''), \tag{4.7}$$

where we used the fact that the first partials of $-\log P(x/\theta)$ vanish at the ML estimate $\theta$. The prime $'$ denotes the adjoint, or the transpose of the Jacobian matrix, if we regard the parameters as expressed relative to a coordinate system. We see immediately that the quadratic forms $(\theta, I(\theta)\theta)$ and $(\theta'', I(\theta'')\theta'')$ are equal if $g$ is a non-singular *linear* transformation, for example, one resulting from a coordinate change. This undoubtedly is reassuring.

The penalty term, the second term in (4.6), can be brought to 0 by a shift of the origin to $\theta$. Does this invalidate the entire process? We think not. When we selected the enumeration of the rectangles $A$, we assumed that the origin of the space $R_k$ is determined by the entire family of the models $\{P(x/\theta)\}$ and by the functions that map the permitted set of the independent parameters into $R^k$. Hence the origin gets determined without knowledge of any particular value of $\theta$, and the decoder can calculate it from the description of the background information, whose length adds just a constant to the criterion. If we wish to change the origin to the minimizing parameter $\theta$, so as to make its index 1, we clearly must describe this special choice to the decoder, in which case its cost is the same as that of $\theta$, and nothing is saved. The situation is analogous to the choice of the universal computer in the algorithmic theory of complexity with its representation of the integers. It is meant to be selected impartially without having any particular integer value in mind, whose representation we would like to and which we could make very short.

Neither does the third term in (4.5) remain invariant under invertible non-linear transformations. Without having a clear understanding of the implications of this, we can report a significant class of applications, where it gives the criterion (4.5) a singular power to discriminate between "good" and "bad" models with the same number of parameters; Rissanen [27]. These are the so-called ARMA models for vector time series, which are described by points on an analytic manifold. To give a glimpse of the relevant point, consider the unit circle $\{(x, y) \mid x^2 + y^2 = 1\}$ as a manifold. (We are indebted to one of the referees for suggesting this example.) We cover the circle with two open sets defined by one parameter each: $\{(x, y) = ((1 - t^2)/(1 + t^2), 2t/(1 + t^2)) \mid -\infty < t < \infty\}$ and $\{(x, y) = (2s/(1 + s^2), (1 - s^2)/(1 + s^2)) \mid -\infty < s < \infty\}$. If we were to use the first parameterization, then the point $(x, y) = (-1, 0)$ in the manifold could be approximated well with only very large values of $t$. The vector time series case is more complicated, because the relevant matrix $I(\theta)$ in a bad coordinate system tends to singular in such a way that its determinant stays constant.

The behavior of the criterion (4.5), when only the first two terms are retained, has been analyzed for scalar time series. The criterion has been shown to lead to strongly consistent estimates of the parameters, including their number, in AR-processes, Hannan and Quinn [12] and Rissanen [26], and in ARMA-processes, Hannan [11] and Hannan and Rissanen [13]. In contingency tables, the criterion, measuring the total amount of information, offers a perhaps speculative but nonetheless intriguing possibility to discover automatically inherent links as "laws of nature" in experimentally collected data. In the usual analyses such links had to be first proposed by humans for a statistical verification or rejection.

**5. Connection with maximum entropy principle.** In this final section we show that Jaynes's maximum entropy principle is a special case of ours. This is of some significance because there are a number of important applications where the ML principle fails but where the maximum entropy formalism has been highly successful [16, 17]. The maximum entropy principle in its general formulation is as follows. Let a random experiment have $q$ possible values at each trial, and hence $q^m$ values in $m$ trials. If the value $i$ for $i \leq q$ occurs $m_i$ times in the sequence $s$ of length $m$, then $\theta_i = m_i/m$ defines a probability for the outcome $i$, and the associated entropy is given by

$$H = -\sum_{i \leq q} \theta_i \log \theta_i. \tag{5.1}$$

Consider the subset of all sequences $s$ of length $m$ which satisfy $n$ linearly independent constraints ($n < q$)

$$\sum_{i \leq q} A_{ji}\theta_i = x_j, \quad 1 \leq j \leq n. \tag{5.2}$$

The set $x = \{x_1, \cdots, x_n\}$ constitutes the observed data, measuring $n$ "physical quantities" defined by a known matrix with elements $A_{ji}$. The probabilities $\theta = (\theta_1, \cdots, \theta_q)$ define the parameters, which are to be determined so that the entropy $H$ is maximized subject to (5.2) and the equation that the sum of the parameters is one.

Because of the constraints (5.2) we define our model class as follows: $P(x \mid \theta) = 1$, for all distributions $\theta$ in the set determined by (5.2). Indeed, since for any such $\theta$ there is a unique sequence of data $x$ (even if we let the equations (5.2) be non-linear in $\theta$), the ideal code length $-\log P(x \mid \theta)$ ought to be zero. Further, in this case we clearly should not use the universal prior for the integers, which presupposes no prior knowledge, because here we know so much more about the parameters. In fact, by the way probabilities are originally defined, the counts $m_i$ are associated with equivalence classes of sequences, and it seems natural to define the probability $Q(\theta)$ as the ratio of the number of sequences with counts $m_i$ to the total number of sequences of length $m$, or $Q(\theta) = C(m_1, \cdots, m_q)/q^m$, where $C(m_1, \cdots, m_q) = m!/m_1! \cdots m_q!$. Minimizing the total description length $L(x, \theta)$, (2.1), or in this case minimizing $-\log Q(\theta)$, is now seen to be equivalent to maximizing $\log C(m_1, \cdots, m_q)$. This with Stirling's formula amounts essentially to maximizing the entropy (5.1). The latter will, of course, hold exactly if we divide by $m$ and let $m$ go to infinity, as we should in conformity with the customary frequency interpretation of probabilities.

## APPENDIX A.

We calculate the sum $c = \sum 2^{-\log^* n}$ by sharpening the arguments given in [22]. Let $\log^{(k)}$ denote the $k$-fold composition of the log-function, and let $\exp^{(k)}$ denote its inverse. For example, $\exp^{(0)}(x) = x$, $\exp^{(1)}(x) = 2^x$, $\exp^{(2)}(2) = 2^4 = 16$, and $\exp^{(3)}(2) = 2^{16}$. For brevity, we write $e(k) = \exp^{(k)}(2)$. The derivative of the $k$-fold logarithm function, written as $D\log^{(k)}(x)$, is seen to be given by

$$D \log^{(k)}(x) = 1/[a^k x(\log x) \cdots (\log^{(k-1)}(x))] = (a^{-k})2^{-\log^*(x)},$$

for $e(k-2) \leq x \leq e(k-1)$, where $a = 1/\log e = .69 \cdots$, and $e(i) = 1$, for $i < 0$. The first equality holds for any $x$ for which the $k$-fold logarithm is defined, while the second equality holds only for the stated numbers $x$. Because of this we can evaluate the integral

$$\int_{e(k-1)}^{e(k)} 2^{-\log^* x} dx = a^{k+1},$$

which in turn gives

$$\int_{e(k)}^{\infty} 2^{-\log^* x} dx = a^{k+2}/(1-a) = S(k).$$

Next observe that

$$2^{-\log^* n} < \int_{n-1}^{n} 2^{-\log^* x} < 2^{-\log^*(n-1)}, \tag{A.1}$$

which implies that we can bracket the desired infinite sum $c$ by the inequalities

$$\sum_1^{n-1} 2^{-\log^* i} + S(n) < c < \sum_1^{n-1} 2^{-\log^* i} + S(n-1). \tag{A.2}$$

The difference between the two bounds is by (A.1) less than $2^{-\log^*(n-1)}$. By putting $n - 1 = 16 = e(2)$ and calculating the upper bound in (A.2), we get $c$ as 2.86 with error less than $2^{-7}$. By putting $n - 1 = 2^{16} = e(3) = 65536$, we get $c$ as 2.865064 with 6 decimals.

We observe in passing that $f(x) = 2^{-\log S - \log^* x}$, where $S = a/(1-a) \simeq 2.257$, defines a density function for the reals in the interval $[1, \infty)$. This, moreover, can be extended to all nonnegative reals by defining for $0 < x < 1$, $\log^*(x) = \log x + \log\log 1/x + \cdots$, where the $k$-fold logarithms are continued until the last positive one, and by replacing $S$ by $2S$ in the above expression for $f(x)$. This density function is seen to modify Jeffreys' improper density $1/x$ to make its integral to unity just as the case was with the integers. Because the probabilities of the truncated reals, resulting from this density, do not admit the same code theoretic interpretations as those derived from our universal prior for the integers, we do not base our estimation on this density as the prior.

We conclude this appendix by showing that the distribution $Q(i)$ has infinite entropy, written as $H(Q)$, which is needed in Theorem 2. By retaining from $L^0$ only the first term $\log i$, we have

$$H(Q) \geq \sum_1^4 1/n + \sum_5^{e(2)} 1/n \log^{(2)}(n) + \sum_{e(2)+1}^{e(3)} 1/n(\log^{(2)}(n)) \log^{(3)}(n) + \cdots .$$

The $k$th term is strictly greater than the quantity

$$(1/e(0) \cdots e(k-2)) \sum_{e(k-1)}^{e(k)} 1/n > [(1/e(0) \cdots e(k-2))e(k-1) - 1]\ln(2) > 1.$$

Here we used the inequality $H_n > \ln(n)$ for the sum of the first $n$ terms in the harmonic series, and ln denotes the natural logarithm.

## APPENDIX B

PROOF OF THEROEM 1. Let $P(n)$ satisfy (3.3) and let $L(n)$ denote any sequence satisfying the Kraft inequality. Denote $P_N = P(1) + \cdots + P(N)$, $L_N = P(1)L(1) + \cdots + P(N)L(N)$, and $H_N = -P(1)\log P(1) - \cdots - P(N)\log P(N)$. We show first that

$$\lim_{N\to\infty} L_N/H_N \geq 1. \tag{B.1}$$

Let $Q_N = Q(1) + \cdots + Q(N)$, where $Q(i) = 2^{-L(i)}$. Then, by Gibbs's inequality,

$$\sum_{i=1}^N P(i)\log\{(Q_N/P_N)P(i)/Q(i)\} = L_N - H_N + P_N\log(Q_N/P_N) \geq 0.$$

The third term goes to 0 as $N$ goes to infinity, and (B.1) follows.

We show next that with $L(i) = L^0(i)$ the equality in (B.1) holds for all $P$ satisfying (3.3). We have immediately (Wyner's inequality [33]), $(j - M) \leq 1/P(j)$, from which,

$$H_N \geq \sum_{j=M+1}^N P(j)\log(j(j-M)/j) \geq \sum_{j=M+1}^N P(j)\log(j) - \log(M+1). \tag{B.2}$$

Next,

$$L_N \leq \sum_{j=M+1}^N P(j)\log(j) + \sum_{j=M+1}^N P(j)r(j) + C,$$

where $r(j) = L(j) - \log j$, and $C$ is the maximum of $L(i)$ for $i \leq M$. Let $f(i) = r(i)/\log(i)$, and let $K(\varepsilon)$ be an index such that $K(\varepsilon) > M$ and $f(i) \leq \varepsilon$ for all $i \geq K(\varepsilon)$. Clearly, such indices exist for every positive $\varepsilon$. Then, for $N > K(\varepsilon)$,

$$\sum_{i=M+1}^N P(i)r(i) \leq \varepsilon \sum_{i=K(\varepsilon)+1}^N P(i)\log(i) + R(\varepsilon),$$

where $R(\varepsilon)$ denotes the maximum value of $r(i)$ for $i \leq K(\varepsilon)$. Using this and (B.2) we get

$$L_N/H_N \leq 1 + \varepsilon + \{R(\varepsilon) + C + (1 + \varepsilon) + \log(M + 1)\}/H_N.$$

When we let $N$ go to infinity and observe that the resulting inequality holds for every $\varepsilon$, the theorem follows.

## REFERENCES


[1] AKAIKE, H. (1977). On entropy maximization principle. *Applications of Statistics*, Ed. P. R. Krishnaiah, pages 27–41. North Holland, Amsterdam.

[2] CHAITIN, G. J. (1975). A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.* **22** 329–340.

[3] COX, D. R., and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

[4] EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.

[5] ELIAS, P. (1975). Universal codeword sets and representations of the integers. *IEEE Trans. Inform. Theory* **IT-21** 194–203.

[6] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

[7] GAUSS, K. F. (1809). Teoria motus corporum coelestium in sectionibus conicis solem ambientium. Reprinted translation: Theory of motion of the heavenly bodies moving about the sun in conic sections. Dover, New York, 1963.

[8] GOOD, I. J. (1968). Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *British J. Philos. Sci.* **19** 123–143.

[9] GOOD, I. J. (1977). Explicativity: a mathematical theory of explanation with statistical applications. *Proc. Roy. Soc. London Ser. A* **354** 303–330.

[10] GOOD, I. J., and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 2, 255–277.

[11] HANNAN, E.J. (1978). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 762–777.

[12] HANNAN, E. J., and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B.* **41** No. 2, 190–195.

[13] HANNAN, E. J., and RISSANEN, J. (1982). Recursive estimation of ARMA order. *Biometrika* **69** 81–94.

[14] JAYNES, E. T. (1961). *Probability Theory in Science and Engineering*. McGraw-Hill, New York.

[15] JAYNES, E. T. (1968). Prior probabilities. *IEEE Trans. Systems Sci. and Cybernet.* SSC-4 227–241.

[16] JAYNES, E. T. (1982). On the rationale of maximum entropy methods. September 1982, *Proc. of IEEE, Special Issue on Spectral Estimation*, S. Haykin, editor, **70** 939–952.

[17] JAYNES, E. T. (1982). Prior information in inference. *J. Amer. Statist. Assoc.*, to appear

[18] JEFFREYS, H. (1961). *Theory of Probability*, Third Edition. Oxford at the Clarendon Press.

[19] KEMENY, JOHN G. (1953). The use of simplicity in induction. *Philos. Rev.* **62** 391–315.

[20] KOLMOGOROV, A. N. (1965). Three approaches to the quantitative definition of information. *Problems Inform. Transmission* **1** 4–7.

[21] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[22] LEUNG-YAN-CHEONG, S. K., and COVER, T. (1978). Some equivalences between Shannon entropy and Kolmogorov complexity. *IEEE Trans. Inform. Theory* **IT-24** 331–338.

[23] LEVIN, L. (1973). On the notion of a random sequence. *Soviet Math. Dokl.* **14/5** 1413–1416.

[24] LEVIN, L. (1974). Laws of information conservation (nongrowth) and aspects of the foundations of probability theory. *Problems Inform. Transmission* **10/3** 206–210.

[25] RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

[26] RISSANEN, J. (1980). Consistent order estimates of autoregressive processes by shortest description of data. *Analysis and Optimisation of Stochastic Systems*. (ed. O. Jacobs, M. Davis, M. Dempster, C. Harris, P. Parks). Académic, New York.

[27] RISSANEN, J. (1982). Estimation of structure by minimum description length. *Circuits, Systems, and Signal Processing*, special issue on Rational Approximations **1** No. 3–4, 395–406.

[28] RISSANEN, J. and LANGDON, JR., G. G. (1981). Universal modeling and coding. *IEEE Trans. on Inform. Theory* **IT-27** 12–23.

[29] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

[30] SOLOMONOFF, R. J. (1964). A formal theory of inductive inference. Part I, *Inform. and Control* **7** 1–22; Part II, *Inform. and Control* **7** 224–254.

[31] SOLOMONOFF, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory* **IT-24/4** 422–432.

[32] WALLACE, C. S., and BOULTON, D. M. (1968). An information measure for classification. *Comput. J.* **11** No. 2 185–194.

[33] WYNER, A. D. (1972). An upper bound on entropy series. *Inform. Control.* **20** 176–181.


IBM CORPORATION-K54/282
5600 COTTLE ROAD
SAN JOSE, CALIFORNIA 95193