Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity*

Paul Vitányi†
CWI and University of Amsterdam

Ming Li†
City University of Hong Kong

November 27, 1996

Abstract

Ignoring practicality, we investigate the ideal form of minimum description length induction where each individual hypothesis and each individual data sample is maximally compressed. For various reasons this is not done or considered desirable in the praxis of the ‘minimum description length (MDL)’ method of J. Rissanen and the ‘minimum message length (MML)’ method of C. Wallace. The shortest descriptions which are conceivably effective are those of least Kolmogorov complexity. Such shortest effective descriptions cannot be constructively found. Theoretically, however, they can be used to rigorously analyse the relation between shortest description length reasoning and Bayesianism. The analysis gives evidence why in practice the latter is prone to overfitting and the former isn’t. Technically, we use the novel notion of individual randomness of objects as expressed by P. Martin-Löf’s randomness tests. Minimum description length reasoning using shortest effective descriptions coincides with Bayesian reasoning using the universal prior distribution, provided the minimum description length is reached for those hypotheses with respect to which the data sample is individually random in the sense of Martin-Löf. The analysis holds both for hypothesis identification and for prediction methods in the style of R. Solomonoff.

AMS Subject Classification: Primary 68S05, 68T05; Secondary 62C10, 62A99.

Keywords: MDL, MML, Bayes’s rule, Kolmogorov complexity, universal distribution, randomness test

1 Introduction

Given a body of data concerning some phenomenon under investigation, we want to select the most plausible hypothesis from among all appropriate hypotheses, or predict future data. The

---

*Parts of the ideas in this paper appeared in preliminary form in: M. Li and P.M.B. Vitanyi, Computational Machine Learning in Theory and Praxis. In: ‘Computer Science Today’, J. van Leeuwen, Ed., Lecture Notes in Computer Science, Vol. 1000, Springer-Verlag, Heidelberg, 1995, 518-533; P.M.B. Vitanyi and M. Li, Ideal MDL and Its Relation To Bayesianism, Proc. ISIS: Information, Statistics and Induction in Science’, World Scientific, Singapore, 1996, 282-291.

†Partially supported by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, by NWO through NFI Project ALADDIN under Contract number NF 62-376, and NSERC under International Scientific Exchange Award ISE0125663. Address: CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl

‡Supported in part by the NSERC Operating Grant OGP0046506, ITRC, a CGAT grant, and the Steacie Fellowship. On sabbatical leave from: Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: mli@cs.cityu.edu.hk
maediaeval dictum known as ‘Occam’s razor’ tells us that, all other things being equal, the simplest explanation is the most likely one. Interpreting ‘simplest’ as ‘having shortest description’, the most likely hypothesis is the most compressed one. The length of the shortest effective description of some object is its Kolmogorov complexity. If the data are non-probabilistic and noiseless (error-free), the argument says that among all hypotheses consistent with the data the one with least Kolmogorov complexity is the most likely one. This idea of hypothesis identification is closely related to Solomonoff’s original ‘universal prior’ which is used for prediction, [15].

The problem becomes more difficult if we have to deal with probabilistic data, and background knowledge concerning the a priori probability of hypotheses before any data are known. The classical method is Bayes’s rule. The problem with applying Bayes’s rule is that one requires the prior probabilities of the hypotheses first. Unfortunately, it is often impossible to obtain these. In the unlikely case that we possess the true prior distribution, in practice the data tend to be noisy due to the measuring process or other causes. The latter confuses Bayes’s rule into overfitting the hypothesis by adding random features while trying to fit the data.

One way out of the conundrum of a priori probabilities is to require inference of hypotheses to be completely or primarily data driven. The minimum description length (MDL) approach, [11, 12] embodies this idea. It is usually presented as justified in and of itself by philosophical persuasion. The minimum message length (MML) approach, [20, 21], while relying on priors, in practice is a related approach. We are only interested in the following common idea shared between all such methods.

“Select the hypothesis which minimizes the sum of the length of the description of the hypothesis (also called ‘model’) and the length of the description of the data relative to the hypothesis.”

This approach balances the complexity of the model (and its tendency for overfitting) against the preciseness of fitting the data (the error of the hypothesis).

Ideally, the description lengths involved should be the shortest effective description lengths. (We use ‘effective’ in the sense of ‘Turing computable’, [17].) Shortest effective description length is asymptotically unique and objective and known as the Kolmogorov complexity of the object being described. Such shortest effective descriptions are ‘effective’ in the sense that we can compute the described objects from them. Unfortunately, it can be shown, see [8], that one cannot compute the length of a shortest description from the object being described. This obviously impedes actual use. Instead, one needs to consider computable approximations to shortest descriptions, for example by restricting the allowable approximation time. This course is followed in one sense or another in the practical incarnations such as MML and MDL. There one often uses simply the Shannon-Fano code, which assigns prefix code length \(-\log P(x)\) to \(x\) irrespective of the regularities in \(x\). If \(P(x) = 2^{-l(x)}\) for every \(x \in \{0,1\}^n\), then the code word length of an all-zero \(x\) equals the code word length of a truly irregular \(x\). While the Shannon-Fano code gives an expected code word length close to the entropy, it does not distinguish the regular elements of a probability ensemble from the random ones.

The code of the shortest effective descriptions, with the Kolmogorov complexities as the code word length set, also gives an expected code word length close to the entropy yet compresses the regular objects until all regularity is squeezed out. All shortest effective descriptions are completely random themselves, without any regularity whatsoever. Kolmogorov complexity can be used to develop a theory of (idealized) minimum description length reasoning. In particular, shortest effective descriptions enable us to rigorously analyse the relation between shortest description length
reasoning and Bayesianism. This provides a theoretical basis for, and gives confidence in, practical uses of the various forms of minimum description length reasoning mentioned.

We rigorously derive and justify this Kolmogorov complexity based form of minimum description length, ‘Ideal MDL’, via the Bayesian approach using a particular prior distribution over the hypotheses (the so-called ‘universal distribution’). This leads to a mathematical explanation of correspondences and differences between Ideal MDL and Bayesian reasoning, and in particular it gives some evidence under what conditions the latter is prone to overfitting while the former isn’t. Namely, for hypothesis identification Ideal MDL using Kolmogorov complexity can be reduced to the Bayesian approach using the universal prior distribution, provided the minimum description length is reached for those hypotheses with respect to which the data sample is *individually random* in the sense of Martin-Löf, [10]. Under those conditions Ideal MDL, Bayesianism, MDL, and MML, select pretty much the same hypothesis. A similar analysis is provided for the prediction problem.

The notion of randomness of individual objects has a long history which goes back to the initial attempts by von Mises, [18], to formulate the principles of application of the calculus of probabilities to real-world phenomena. Classical probability theory cannot even express the notion of ‘randomness of individual objects’. Following almost half a century of unsuccessful attempts, the theory of Kolmogorov complexity, [6], and Martin-Löf tests for randomness, [10], finally succeeded in formally expressing the novel notion of individual randomness in a correct manner, see [8]. Objects which are random in this sense will satisfy all effective tests for randomness properties—those which are known and those which are yet unknown alike.

## 2 Hypothesis Identification by Minimum Description Length

The minimum description length principle is an algorithmic paradigm that is widely applied. That is, it is widely applied at least in spirit; to apply it literally may run in computation difficulties since it involves finding an optimum in a exponentially large set of candidates as noted for example in [8]. Yet in some cases one can approximate this optimum, [19, 22]. For the theoretical case where the minimum description lengths involved are the Kolmogorov complexities, we mathematically derive the minimum description length paradigm from first principles, that is, Bayes’s rule. To do so we need auxiliary notions of universal distribution and Martin-Löf tests for randomness of individual objects.

Before proceeding it is useful to point out that the idea of a two-part code for a body of data $D$ is natural from the perspective of Kolmogorov complexity. If $D$ does not contain any regularities at all, then it consists of purely random data and there is no hypothesis to identify. Assume that the body of data $D$ contains regularities. With help of a description of those regularities (a model) we can describe the data compactly. Assuming that the regularities can be represented in an effective manner (that is, by a Turing machine), we encode the data as a program for that machine. Squeezing all effective regularity out of the data, we end up with a Turing machine representing the meaningful regular information in the data together with a program for that Turing machine representing the remaining meaningless randomness of the data. This is the intuition, which finds its basis in the Definitions 2 and 3. However, it is difficult to find a valid mathematical way to force a sensible division of the information at hand in a meaningful part and a meaningless part. One way to proceed is suggested by the analysis below.
2.1 Bayesian Reasoning

Consider a situation in which one has a set of observations of some phenomenon, and also a finite or countably infinite set of hypotheses which are candidates to explain the phenomenon. For example, we are given a coin and we flip it 1000 times. We want to identify the probability that the coin has outcome ‘head’ in a single coin flip. That is, we want to find the bias of the coin. The set of possible hypotheses is uncountably infinite if we allow each real bias in [0, 1], and countably infinite if we allow each rational bias in [0, 1).

For each hypothesis $H$ we would like to assess the probability that $H$ is the ‘true’ hypothesis, given the observation of $D$. This quantity, $\Pr(H|D)$, can be described and manipulated formally in the following way.

Consider a sample space $\Omega$. Let $D$ denote a sample of outcomes, say experimental data concerning a phenomenon under investigation. Let $H_1, H_2, \ldots$ be an enumeration of countably many hypotheses concerning this phenomenon, say each $H_i$ is a probability distribution over $\Omega$. The list $\mathcal{H} = \{H_1, H_2, \ldots\}$ is called the hypothesis space. The hypotheses $H_i$ are exhaustive and mutually exclusive.

For example, say the hypotheses enumerate the possible rational (or computable) biases of the coin. As another possibility there may be only two possible hypotheses: hypothesis $H_1$ which says the coin has bias 0.2, and hypothesis $H_2$ which puts the bias at 0.8.

Suppose we have a priori a distribution of the probabilities $P(H)$ of the various possible hypotheses in $\mathcal{H}$ which means that $\sum_{H \in \mathcal{H}} P(H) = 1$. Assume furthermore that for all $H \in \mathcal{H}$ we can compute the probability $\Pr(D|H)$ that sample $D$ arises if $H$ is the case. Then we can also compute (or approximate in case the number of hypotheses with nonzero probability is infinite) the probability $\Pr(D)$ that sample $D$ arises at all

$$\Pr(D) = \sum_{H \in \mathcal{H}} \Pr(D|H)P(H).$$

From the definition of conditional probability it is easy to derive Bayes’s formula

$$\Pr(H|D) = \frac{\Pr(D|H)P(H)}{\Pr(D)}.$$  \hspace{1cm} (1)

The prior probability $P(H)$ is often considered as the learner’s initial degree of belief in hypothesis $H$. In essence Bayes’s rule is a mapping from a priori probability $P(H)$ to a posteriori probability $\Pr(H|D)$ determined by data $D$.

Continuing to obtain more and more data, this way the total inferred probability will concentrate more and more on the ‘true’ hypothesis. We can draw the same conclusion of course, using more examples, by the law of large numbers. In general, the problem is not so much that in the limit the inferred probability would not concentrate on the true hypothesis, but that the inferred probability gives as much information as possible about the possible hypotheses from only a limited number of data. Given the prior probability of the hypotheses, it is easy to obtain the inferred probability, and therefore to make informed decisions. However, in general we don’t know the prior probabilities. The following MDL approach in some sense replaces an unknown prior probability by a fixed ‘universal’ probability.

---

1Some Bayesians prefer replacing $\Pr(D|H)P(H)$ by a joint probability of data and hypotheses together, the prior $P(D, H) = \Pr(D|H)P(H)$. 

4
2.2 Ideal Minimum Description Length

Theoretically the idea of predicting time sequences using shortest effective descriptions was first formulated by R. Solomonoff, [15]. He uses Bayes’s formula equipped with a fixed ‘universal’ prior distribution. In accordance with Occam’s dictum, it tells us to go for the explanation that compresses the data the most. In various forms aimed at practical applications the idea of doing induction or data modelling in statistical hypothesis identification or prediction was proposed by C. Wallace and co-authors [20, 21], who formulated the Minimum Message Length (MML) principle and J. Rissanen [11, 12] who formulated the Minimum Description Length (MDL) principle. Here we abstract away from epistemological and technical differences between MML, and MDL, and other variants, and their concessions to reality in the name of feasibility and practicability. We focus only on the following central ideal version involved. Indeed, we do not even care about whether we deal with statistical or deterministic hypotheses. All effectively describable hypotheses are involved.

Definition 1 Given a sample of data, and an effective enumeration of models, ideal MDL selects the model which minimizes the sum of

- the length, in bits, of an effective description of the model; and
- the length, in bits, of an effective description of the data when encoded with the help of the model.

Intuitively, with a more complex description of the hypothesis $H$, it may fit the data better and therefore decreases the misclassified data. If $H$ describes all the data, then it does not allow for measuring errors. A simpler description of $H$ may be penalized by increasing the number of misclassified data. If $H$ is a trivial hypothesis that contains nothing, then all data are described literally and there is no generalization. The rationale of the method is that a balance in between seems required. However, straightforward application of the idea may lead to obviously false results in pathological cases, as the following analysis demonstrates.

2.3 Derivation from Bayes’s Rule

We analyze Ideal MDL using Kolmogorov complexity. While this does not rigorously prove anything about applied versions of MDL as in [12], the relations with Bayesian inference we establish for our version are corroborated by empirical evidence for applied MDL as in [9, 14, 2].

How to derive MDL from first principles, in casu Bayes’s rule as given by Equation 1? In the latter equation we are only concerned with maximizing the term $\Pr(H|D)$ over $H$. Taking the negative logarithm at both sides of the equation, this is equivalent to minimizing the expression $-\log \Pr(H|D)$ over $H$:

$$-\log \Pr(H|D) = -\log \Pr(D|H) - \log P(H) + \log \Pr(D).$$

Since the probability $\Pr(D)$ is constant under varying $H$, this means we want to find $H$ which minimizes:

$$-\log \Pr(D|H) - \log P(H).$$

(2)
In MML as in [21] or MDL as in [12] one roughly speaking interprets these negative logarithms
$-\log P(x)$ as the corresponding Shannon-Fano (or Huffman) code word lengths. But why should
one use the Shannon-Fano code (or Huffman code) and no other code reaching an expected code
word length equal to the entropy? In particular, ignoring feasibility, why not use the objective
shortest effective code, the shortest effective descriptions with code word length set equal to the
Kolmogorov complexities. This also has an expected code word length equal to the entropy, [8], but
additionally it compresses each object by effectively squeezing out and accounting for all regularities
in it. The resulting code word is maximally random, that is, it has maximal Kolmogorov complexity.

For clarity of treatment, we refer the reader to the Appendices or [8] for all definitions and
analysis of auxiliary notions. This way we also do not deviate from the main argument, do not
obstruct the knowledgeable reader, and do not confuse or discourage the reader who is unfamiliar
with Kolmogorov complexity theory.

To obtain the Ideal MDL principle it suffices to replace the probabilities involved in Equation 2
by the so-called universal probability $m(\cdot)$, Appendix B. The analysis of the conditions under
which this substitution is justified, or, conversely, how application of MDL is equivalent to Bayesian
inference using admissible probabilities, is deferred to the next section. We substitute probabilities
according to

$$\log P(H) = \log m(H) + O(1),$$
$$\log \Pr(D|H) = \log m(D|H) + O(1).$$

According to Appendix B,

$$-\log m(H) = K(H) + O(1),$$
$$-\log m(D|H) = K(D|H) + O(1),$$

Where $K(\cdot)$ is the prefix complexity of Appendix A. Therefore, using the substitution of Equation 3
we can replace the sum of Equation 2 by the sum of the minimum lengths of effective self-delimiting
programs which compute descriptions of $H$ and $D|H$. That is, we look for the $H$ that minimizes

$$K(D|H) + K(H),$$

which is the code-independent, recursively invariant, absolute form of the MDL principle.

We mention that the term $-\log \Pr(D|H)$ is also known as the self-information in information
theory and the negative log-likelihood in statistics. It can now be regarded as the number of bits it
takes to redescribe or encode $D$ with an ideal code relative to $H$.

The discussion seems to have arrived at its goal, but a danger of triviality lurks nearby. Yet
it is exactly the solution how to prevent trivialities, which gives us the key to the very meaning
of Ideal MDL, and by extension some insight in applied versions. In [4] (see also [8], Theorem 3.5
on p. 195) it is proved that symmetry of information holds for individual objects in the following
sense

$$K(H,D) = K(H) + K(D|H, K(H)) + O(1)$$
$$= K(D) + K(H|D, K(D)) + O(1).$$
Since it is not more difficult to describe some object if we get more conditional information, we have \( K(D|H, K(H)) \leq K(D|H) + O(1) \). Thus, by Equation 5 the quantity in Equation 4 satisfies
\[
K(H) + K(D|H) \geq K(H, D) + O(1)
\]
\[
\geq K(D) + O(1).
\]
For the trivial hypothesis \( H_0 = D \) or \( H_0 = \emptyset \) equalities hold. At first glance this would mean that the hypothesis \( H_{\text{mdl}} \) which minimizes the sum of Equation 4 could be set to \( D \) or \( \emptyset \), which is absurd in general. However, we have only derived the validity of Equation 4 under the condition that Equation 3 holds. The crucial part of our justification of MDL is to establish precisely when Equation 3 is valid, which is the main thrust of this paper.

### 2.4 The Rôle of Universal Probability

It is well known, see for example [8], that the so-called Shannon-Fano code for an ensemble of source words distributed according to probability \( Q \) is a prefix code \( E_Q \) with \( l(E_Q(x)) = -\log Q(x) \) satisfying
\[
\sum_x Q(x)l(E_Q(x)) \leq \min_{E'}\{\sum_x Q(x)l(E'(x)) : E' \text{ is a prefix code}\} + 1,
\]
that is, it realizes the least expected code word length among all prefix codes (the entropy of \( Q(\cdot) \) by Shannon’s Noiseless Coding Theorem). Therefore, the \( H \) which minimizes Equation 2, that is,
\[
\text{l}(E_{Pr(\cdot|H)}(D)) + \text{l}(E_P(H))
\]
minimizes the sum of two prefix codes which both have shortest expected code word lengths. This is more or less what MML [21] does and similarly MDL [12].

But there are many prefix codes which have expected code word length equal to the entropy. Among those prefix codes there is one which gives the shortest effective code word to each individual source word: the prefix code with code word length \( K(x) \) for object \( x \), see [8]. In Ideal MDL we want minimize the sum of the effective description lengths of the individual elements \( H, D \) involved. This means using the shortest effective description lengths, as in Equation 4. However, we cannot simply replace negative logarithms in Equation 2 by corresponding \( K(\cdot) \) terms. We can only do so if Equation 3 holds.

To satisfy Equation 3 we are free to make the new assumption that the prior probability \( P(\cdot) \) in Bayes’s rule Equation 1 is fixed as \( m(\cdot) \). Whether this can be justified or not is a question which we address in Section 2.8.

However, we cannot assume that the probability \( Pr(\cdot|H) \) equals \( m(\cdot|H) \). Namely, as explained at length in Section 2.1, probability \( Pr(\cdot|H) \) may be totally determined by the hypothesis \( H \). Depending on \( H \) therefore, \( l(E_{Pr(\cdot|H)}(D)) \) may be very different from \( K(D|H) \). This holds especially for ‘simple’ data \( D \) which have low probability under assumption of hypothesis \( H \).

**Example 1** Let us look at a simple example evidencing this discrepancy. Suppose we flip a coin of unknown bias \( n \) times. Let hypothesis \( H \) and data \( D \) be defined by:
\[
H := \text{Probability ‘head’ is 1/2}
\]
\[
D := \overbrace{\text{h}_1 \ldots \text{h}_n}^{n \text{ times ‘h’ (ead)s}}
\]
Then we have $\Pr(D|H) = 1/2^n$ and
\[
i(E_{\Pr(\cdot|H)}(D)) = -\log \Pr(D|H) = n.
\]
In contrast,
\[
K(D|H) \leq \log n + 2 \log \log n + O(1).
\]

However, the theory dealing with randomness of individual objects states that under certain conditions $-\log \Pr(D|H)$ and $K(D|H)$ are close.

Assume that $\Pr(\cdot|\cdot)$ is a recursive function. That is, it can be computed to any required precision for each argument $D$ and conditional $H$. Then we can appeal to the following known facts, Appendix C. Firstly, it is known that $m(D|H) \geq 2^{-K(\Pr(\cdot|H))} \Pr(D|H)$, Equation 20. Therefore,
\[
\log \frac{m(D|H)}{\Pr(D|H)} \geq -K(\Pr(\cdot|H)) \geq -K(H) + O(1). \tag{6}
\]
The last inequality arises since from $H$ we can compute $\Pr(\cdot|H)$ by assumption on $\Pr(\cdot|\cdot)$. Second, for $D$ which are sufficiently random with respect to the distribution $\Pr(\cdot|H)$ (with respect to $H$ therefore) in the sense of Martin-Löf one can show, Theorem 5,
\[
\log \frac{m(D|H)}{\Pr(D|H)} \leq 0. \tag{7}
\]
The overwhelming majority of $D$’s are random in this sense because for each $H$ we have
\[
\sum_D \Pr(D|H) 2^{\log \frac{m(D|H)}{\Pr(D|H)}} = \sum_D \log m(D, H) \leq 1,
\]
since $m(\cdot|H)$ is a probability distribution. For $D$’s which are random in the appropriate sense, Equations 6, 7 mean by Equation 3 that
\[
K(D|H) - K(\Pr(\cdot|H)) \leq -\log \Pr(D|H) \leq K(D|H). \tag{8}
\]
Above, by assuming that the $a$ priori probability $P(H)$ of hypothesis $H$ is in fact the universal probability we obtained $-\log P(H) = m(x)$. However, we do not need to make this assumption. For recursive $P(\cdot)$, we can analyze the situation when $H$ is random in the required sense with respect to $P(\cdot)$.

This means that hypothesis $H$ is a ‘typical’ to prior distribution $P(\cdot)$ and not ‘extraordinary’. ‘Typicality’ means that it does not have effectively testable properties which distinguish it from the overwhelming majority.

The first inequality below holds since $m(\cdot)$ majorizes $P(\cdot)$ this way; the second inequality expresses the assumption of randomness of $H$,
\[
m(H) \geq 2^{-K(P)} P(H) \quad \log(m(H)/P(H)) \leq 0.
\]
Here \( K(P) \) is the length of the shortest self-delimiting program for the reference universal prefix machine to simulate the Turing machine computing the probability density function \( P : \mathcal{N} \rightarrow [0,1] \). That is, it is the shortest effective self-delimiting description of \( P \), Appendix A. Then,

\[
K(H) - K(P) \leq -\log P(H) \leq K(H).
\]

Altogether we find

\[
K(D|H) + K(H) - \alpha(P,H) \leq -\log \Pr(D|H) - \log P(H) \leq K(D|H) + K(H),
\]

with

\[
\alpha(P,H) = K(\Pr(\cdot|H)) + K(P),
\]

and we note that \( K(\Pr(\cdot|H)) \leq K(H) + O(1) \). We call Equation 10 the Fundamental Inequality (FI) because it describes the fundamental relation between Bayes’s Rule and MDL, and it is left for us to interpret it.

### 2.5 Validity Range of FI

We begin by stressing that Equation 10 holds only in case simultaneously \( H \) is \( P \)-random and \( D \) is \( \Pr(\cdot|H) \)-random. What is the meaning of this?

\( H \) is \( P \)-random means that the true hypothesis must be ‘typical’ for the prior distribution \( P \) in the sense that it must belong to all effective majorities (sets on which the majority of \( P \)-probability is concentrated). In [8] it is shown that this is the set of \( H \)’s such that \( K(H) \approx -\log P(H) \). In case \( P(H) = m(H) \), that is, the prior distribution equals the universal distribution, then for all \( H \) we have \( K(H) = -\log P(H) \), that is, all hypotheses are random with respect to the universal distribution.

Let us look at an example of a distribution where some hypotheses are random, and some other hypotheses are nonrandom. Let the possible hypotheses correspond to the binary strings of length \( n \), and \( P \) is the uniform distribution that assigns probability \( P(H) = 1/2^n \) to each hypothesis \( H \in \{0,1\}^n \). Then \( H = 00\ldots0 \) has low complexity \( K(H) \leq \log n + 2\log \log n \). However, \( -\log P(H) = n \). Therefore, by Equation 9, \( H \) is not \( P \)-random. If we obtain \( H \) by \( n \) flips of a fair coin however, then with overwhelming probability we will have that \( K(H) = n + O(\log n) \) and therefore \( -\log P(H) \approx K(H) \) and \( H \) is \( P \)-random.

\( D \) is \( \Pr(\cdot|H) \)-random means that the data are random with respect to the probability distribution \( \Pr(\cdot|H) \) induced by the hypothesis \( H \). Therefore, we require that the sample data \( D \) are ‘typical’, that is, ‘randomly distributed’ with respect to \( \Pr(\cdot|H) \). If, for example, \( H = (\mu, \sigma) \) induces the Gaussian distribution \( \Pr(\cdot|(\mu, \sigma)) = N(\mu, \sigma) \) and the data \( D \) are concentrated in a tail of this distribution, like \( D = 00\ldots0 \), then \( D \) is atypical with respect to \( \Pr(\cdot|H) \) in the sense of being non-random because it violates the \( \Pr(\cdot|H) \)-randomness test Equation 7.

### 2.6 If Optimal Hypothesis Violates FI

The only way to violate the Fundamental Inequality is that either \( D \) is not \( \Pr(\cdot|H) \)-random and by Equation 7, \( -\log \Pr(D|H) > K(D,H) \), or that \( H \) is not \( P \)-random and by Equation 9, \( -\log P(H) > K(H) \). We give an example of the first case.
We sample a polynomial \( H_2 = ax^2 + bx + c \) at \( n \) arguments chosen uniformly at random from the interval \([0, 1]\). The sampling process introduces Gaussian errors in the function values obtained. The set of possible hypotheses is the set of polynomials. Assume that all numbers involved are of fixed bounded accuracy.

Because of the Gaussian error in the measuring process, with overwhelming probability the only polynomials \( H_{n-1} \) which fit the sample precisely are of degree \( n - 1 \). Denote the data by \( D \). Now this hypothesis \( H_{n-1} \) is likely to minimize \( K(D|H) \) since we have just to describe the \( n \) lengths of the intervals between the sample points along the graph of \( H_{n-1} \). However, for this \( H_{n-1} \) data \( D \) is certainly not \( \Pr(\cdot|H_{n-1}) \)-random, since it is extremely unlikely, and hence atypical, that \( D \) arises when sampling \( H_{n-1} \) with Gaussian error. Therefore, Equation 7 is violated which means that \(-\log \Pr(D|H_{n-1}) > K(D|H_{n-1})\) contrary to what we used in deriving the Fundamental Inequality. With prior probability \( P(\cdot) := m(\cdot) \), which means \(-\log P(\cdot) = K(\cdot) + O(1)\), this moreover violates the Fundamental Inequality.

In contrast, with overwhelming likelihood, \( H_2 \) will show the data sample \( D \) random to it. That being so, the Fundamental Inequality holds. Now what happens if \( H_{n-1} \) is the true hypothesis and the data sample \( D \) by chance is as above? In that case the Fundamental Inequality is violated and Bayes’s Rule and MDL may select very different hypotheses as the most likely ones, respectively.

### 2.7 If Optimal Hypothesis Satisfies FI

Given data sample \( D \) and prior probability \( P \), we call a hypothesis \( H \) admissible if the Fundamental Inequality Equation 10 holds.

Restriction to the set of admissible hypotheses excludes setting \( K(D|H) + K(H) \approx K(D) \) for trivial hypothesis (like \( H = \emptyset \) or \( H = D \) which are not admissible).

**Theorem 1** Let \( \alpha(P,H) \) as in Equation 10 be small. Then Bayes’s Rule and Ideal MDL are optimized (or almost optimized) by the same hypothesis among the admissible \( H \)’s. That is, there is one admissible \( H \) that simultaneously (almost) minimizes both \(-\log \Pr(D|H) - \log P(H)\) (selection according to Bayes’s Rule) and \( K(D|H) + K(H) \) (selection according to MDL).

**Proof.** The smallness of \( \alpha(P,H) \) means that both the prior distribution \( P \) is simple, and that the probability distribution \( \Pr(\cdot|H) \) over the data samples induced by hypothesis \( H \) simple. In contrast, if \( \alpha(P,H) \) is large, which means that either of the mentioned distributions is not simple, for example when \( K(\Pr(\cdot|H)) = K(H) \) for complex \( H \), then there may be some discrepancy. Namely, in Bayes’s Rule our purpose is to maximize \( \Pr(H|D) \), and the hypothesis \( H \) which minimizes \( K(D|H) + K(H) \) also maximizes \( \Pr(H|D) \) up to a \( 2^{-\alpha(P,H)} \) multiplicative factor. Conversely, the \( H \) that maximizes \( \Pr(H|D) \) also minimizes \( K(D|H) + K(H) \) up to an additive term \( \alpha(P,H) \). That is, with

\[
H_{\text{mdl}} := \min_{H} \{K(D|H) + K(H)\}
\]

\[
H_{\text{bayes}} := \max_{H} \{\Pr(H|D)\}
\]

we have

\[
2^{-\alpha(P,H)} \leq \frac{\Pr(H_{\text{mdl}}|D)}{\Pr(H_{\text{bayes}}|D)} \leq 1
\]

\[
\alpha(P,H) \geq K(D|H_{\text{mdl}}) + K(H_{\text{mdl}}) - K(D|H_{\text{bayes}}) - K(H_{\text{bayes}}) \geq 0.
\]
Therefore, if \( \alpha(P,H) \) is small enough and Bayes’s rule selects an admissible hypothesis, and so does Ideal MDL, then both criteria are (almost) optimized by both selected hypotheses.

### 2.8 What MDL Does

We can now assess what prior distributions and assumptions about the relation between the data sample and selected hypothesis MDL assumes. That is, how we can translate MDL in terms of Bayes’s Rule. Identifying application of MDL with application of Bayes’s rule on some prior distribution \( P \) we must assume that, given \( D \), the Fundamental Inequality is satisfied for \( H_{mdl} \) as defined in Equation 11. This means that \( H_{mdl} \) is \( P \)-random for the used prior distribution \( P \). One choice to guarantee this is to choose

\[
P(\cdot) := m(\cdot)(= 2^{-K(\cdot)}).
\]

This is a valid choice even though \( m \) is not recursive, since the latter requirement arose from the requirement that \( m(\cdot)/P(\cdot) \) be enumerable, which is certainly guaranteed by choice of \( P(\cdot) := m(\cdot) \). This choice of prior distribution over the hypotheses is an objective and recursively invariant quantified form of Occam’s razor: simple hypothesis \( H \) (with \( K(H) \ll l(H) \) have high probability, and complex or random hypothesis \( H \) (with \( K(H) \approx l(H) \)) have low probability, namely, \( 2^{-l(H)} \). This choice of prior distribution is most convenient, since the randomness test \( \log m(H)/P(H) = 0 \) for each hypothesis \( H \). This means that all hypotheses \( H \) are random with respect to distribution \( m(\cdot) \).

**Corollary 1** Let \( \alpha(P,H) \) in de FI Equation 10 be small (for example \( \alpha = O(1) \)). With prior \( P(\cdot) \) set to \( m(\cdot) \), the Fundamental Inequality Equation 10 is satisfied iff data sample \( D \) is \( \Pr(\cdot|H_{mdl}) \)-random.

With the chosen prior, and given data sample \( D \), this \( \Pr(\cdot|H_{mdl}) \)-randomness of \( D \) therefore constrains the domain of hypotheses from which we can choose \( H_{mdl} \). Hence we can interprete Ideal MDL as an application of Bayes’s Rule with as prior distribution the universal distribution \( m(\cdot) \) and selection of a hypothesis \( H_{mdl} \) which shows the given data sample random to it in the precise sense of \( \Pr(\cdot|H_{mdl}) \)-randomness of individual objects as developed in [8].

Since the notion of individual randomness incorporates all effectively testable properties of randomness, application of Ideal MDL will select the simplest hypothesis which balances the \( K(D|H) \) and \( K(H) \) and also shows the data sample \( D \) random (as far as we are able to tell) with respect to the selected hypothesis \( H_{mdl} \).

This is the ‘reason’ why the hypothesis selected by Ideal MDL is not simply the one which perfectly fits the data. With some amount of overstatement on can say that if one obtains perfect data for a true hypothesis, then Ideal MDL interprets these data as data obtained from a simpler hypothesis subject to measuring errors. Consequently, in this case Ideal MDL is going to give you the false simple hypothesis and not the complex true hypothesis.

- Ideal MDL only gives us the true hypothesis if the data satisfies certain conditions relative to the true hypothesis. Stated differently: there are only data and no true hypothesis for Ideal MDL. The principle simply obtains the hypothesis which is suggested by the data and it assumes that the data are random with respect to the hypothesis.
2.9 Applying MDL

Unfortunately, the function $K$ is not computable, [8]. In this paper we will not address the question which encoding one uses in practice, but refer to references [12, 20, 22, 19]. For practical applications one must settle for easily computable approximations.

In statistical applications, $H$ is some statistical distribution (or model) $H = P(\theta)$ with a list of parameters $\theta = (\theta_1, \ldots, \theta_k)$, where the number $k$ may vary and influence the (descriptive) complexity of $\theta$. (For example, $H$ can be a normal distribution $N(\mu, \sigma)$ described by $\theta = (\mu, \sigma)$.) Each parameter $\theta_i$ is truncated to fixed finite precision. The data sample consists of $n$ outcomes $y = (y_1, \ldots, y_n)$ of $n$ trials $x = (x_1, \ldots, x_n)$ for distribution $P(\theta)$. The data $D$ in the above formulas is given as $D = (x, y)$. By expansion of conditional probabilities we have therefore

$$
\Pr(D|H) = \Pr(x, y|H) = \Pr(x|H) \cdot \Pr(y|H, x).
$$

In the argument above we take the negative logarithm of $\Pr(D|H)$, that is,

$$
- \log \Pr(D|H) = - \log \Pr(x|H) - \log \Pr(y|H, x).
$$

Taking the negative logarithm in Bayes's rule and the analysis of the previous section now yields that MDL selects the hypothesis with highest inferred probability satisfying $x$ is $\Pr(\cdot|H)$-random and $y$ is $\Pr(\cdot|H, x)$-random. Bayesian reasoning selects the same hypothesis provided the hypothesis with maximal inferred probability has $x, y$ satisfy the same conditions.

Example: Learning Polynomials.

We wish to fit a polynomial $f$ of unknown degree to a set of data points $D$ such that it can predict future data $y$ given $x$. Even if the data did come from a polynomial curve of degree, say two, because of measurement errors and noise, we still cannot find a polynomial of degree two fitting all $n$ points exactly. In general, the higher the degree of fitting polynomial, the greater the precision of the fit. For $n$ data points, a polynomial of degree $n - 1$ can be made to fit exactly, but probably has no predictive value. Applying Ideal MDL we look for $H_{\text{mdl}} := \min_{g \in H} \{K(x, y|H) + K(H)\}$.

Let us apply the Ideal MDL principle where we describe all $(k - 1)$-degree polynomials by a vector of $k$ entries, each entry with a precision of $d$ bits. Then, the entire polynomial is described by

$$
k d + O(\log kd) \text{ bits.}
$$

We have to describe $k$, $d$, and account for self-delimiting encoding of the separate items.) For example, $ax^2 + bx + c$ is described by $(a, b, c)$ and can be encoded by about $3d$ bits. Each datapoint $(x_i, y_i)$ which needs to be encoded separately with precision of $d$ bits per coordinate costs about $2d$ bits.

For simplicity assume that probability $\Pr(x|H) = 1$ (because $x$ is prescribed). To apply the Ideal MDL principle we must trade the cost of hypothesis $H$ (Equation 13) against the cost of describing $y$ with help of $H$ and $x$. As a trivial example, suppose the $n - 1$ out of $n$ datapoints fit a polynomial of degree 2 exactly, but only 2 points lie on any polynomial of degree 1 (a straight line). Of course, there is a polynomial of degree $n - 1$ which fits the data precisely (up to precision). Then the Ideal MDL cost is $3d + 2d$ for the 2nd degree polynomial, $2d + (n - 2)d$ for the 1st degree polynomial, and $nd$ for the $(n - 1)$th degree polynomial. Given the choice among those three options, we select the 2nd degree polynomial for all $n > 5$. For a more sophisticated approach accounting for the average encoding cost of exceptions assuming they are Gaussian distributed, see [12].
2.10 Modern MDL

It is only to within terms of order \(O(1)\) that the MDL and the Bayesian techniques are equivalent. In modern forms of MDL one departs from the straight correspondence with Bayes’s rule and takes \(-\log m(D \mid H) / \sum_{D'} m(D' \mid H)\), instead of \(-\log [m(D \mid H)]\), where \(H(D)\) is the minimizing hypothesis and the summation runs through all data \(D'\) such that \(H(D') = H(D)\), [13]. The probability in the denominator gets absorbed by the term \(O(1)\), however, but for smaller amounts of data it is claimed to make a difference.

2.11 Alternative MDL-Like Principle

In the above interpretation of Ideal MDL we essentially look for a hypothesis \(H\) minimizing \(K(D \mid H) + K(H)\). This always satisfies
\[
K(D \mid H) + K(H) \geq K(D).
\]

An incorrect interpretation is sometimes confused with MDL. In the new approach the idea is that we define \(E := D - D_H\), where \(D_H\) is the data set classified according to \(H\). We want to minimize
\[
K(H, E) \approx K(H) + K(E \mid H)
\]
over \(H\). That is, \(E\) denotes the subset of the data sample \(D\) which are exceptions to \(H\) in the sense of being ‘not covered’ by \(H\). We want to find \(H\) such that the description of \(H\) and the exception data \(E\) not covered by \(H\) is minimized. Note that in this case always
\[
\min_H [K(H) + K(E \mid H)] \leq K(\emptyset) + K(D) = K(D),
\]
in contrast to the standard interpretation of MDL above. This incarnation of MDL is not straightforwardly derived by our approach above. We may interpret it that we look for the shortest description of an accepting program for the data consisting of a classification rule \(H\) and an exception list \(E\). While this principle sometimes gives good results, application may lead to absurdity as the following shows.

In many problems our data sample \(D\) consists of only positive examples. Suppose we want to learn (a grammar for) English language, given a corpus of data \(D\). Here, data \(D\) can be the Bible or an even larger sample like the so-called ‘Brown Corpus’. Then, according to our new MDL rule the best hypothesis is the trivial grammar \(H\) generating all sentences over the alphabet. Namely, this grammar gives \(K(H) = O(1)\) independent of \(D\) and also \(E := \emptyset\). Consequently,
\[
\min_H [K(H) + K(E \mid H)] = K(H) = O(1),
\]
which is absurd. The principle is vindicated and reduces to the standard one in the context of interpreting \(D = (x, y)\) as in Section 2.9 with \(x\) fixed as in ‘supervised learning’. We want to find \(H\) minimizing
\[
K(H) + K(y \mid H, x) + K(x \mid H)
\]
which is the same as minimizing
\[
K(H) + K(y \mid H, x)
\]
if we take \(K(x \mid H)\) constant. Now, \(K(y \mid H, x)\) corresponds to \(K(E \mid H)\) if we ignore the constant \(x\) in the conditional.
Here we briefly discuss the case where the aim is to predict outcomes concerning a phenomenon $\mu$ under investigation. In this case we have some prior evidence (prior distribution over the hypotheses, experimental data) and we want to predict future events. This situation can be modelled by considering a sample space $S$ of one-way infinite sequences of basic elements $B$ defined by $S = B^\infty$. We assume a prior distribution $\mu$ over $S$ with $\mu(x)$ denoting the probability of a sequence starting with $x$. Here $\mu(\cdot)$ is a semimeasure\(^2\) satisfying

$$
\begin{align*}
\mu(\epsilon) & \leq 1, \\
\mu(x) & \geq \sum_{a \in B} \mu(xa).
\end{align*}
$$

Given a previously observed data string $x$, the inference problem is to predict the next symbol in the output sequence, that is, to extrapolate the sequence $x$. In terms of the variables in formula 1, $H_{xy}$ is the hypothesis that the sequence starts with initial segment $xy$. Data $D_x$ consists of the fact that the sequence starts with initial segment $x$. Then, $\Pr(D_x|H_{xy}) = 1$, that is, the data is forced by the hypothesis, or $\Pr(D_z|H_{xy}) = 0$ for $z$ is not a prefix of $xy$, that is, the hypothesis contradicts the data. For $P(H_{xy})$ and $\Pr(D_x)$ in formula 1 we substitute $\mu(xy)$ and $\mu(x)$, respectively. For $P(H_{xy}|D_x)$ we substitute $\mu(y|x)$. This way the formula is rewritten as

$$
\mu(y|x) = \frac{\mu(xy)}{\mu(x)}. \quad (14)
$$

The final probability $\mu(y|x)$ is the probability of the next symbol string being $y$, given the initial string $x$. Obviously we now only need the prior probability $\mu$ to evaluate $\mu(y|x)$. The goal of inductive inference in general is to be able to either (i) predict, or extrapolate, the next element after $x$ or (ii) to infer an underlying effective process that generated $x$, and hence to be able to predict the next symbol. In the most general deterministic case such an effective process is a Turing machine, but it can also be a probabilistic Turing machine or, say, a Markov process (which makes its brief and single appearance here). The central task of inductive inference is to find a universally valid approximation to $\mu$ which is good at estimating the conditional probability that a given segment $x$ will be followed by a segment $y$.

In general this is impossible. But suppose we restrict the class of priors $\mu$ to the recursive semimeasures and restrict the set of basic elements $B$ to $\{0, 1\}$. Under this relatively mild restriction on the admissible semimeasures $\mu$, it turns out that we can use the single universal semimeasure $M$ as a ‘universal prior’ (replacing the real prior $\mu$) for prediction. The notion of universal semimeasure $M$ is a continuous version of $m$ we saw before, and which is explained in the Appendix D, defined with respect to a special type Turing machine called monotone Turing machine. The universal semimeasure $M$ multiplicatively dominates all enumerable (computable from below) semimeasures. If we flip a fair coin to generate the successive bits on the input tape of the universal reference monotone Turing machine, then the probability that it outputs $x \alpha$ ($x$ followed by something) is $M(x)$. For definitions and properties we use below see Appendix D.

There it is shown that the universal distribution itself is directly suited for prediction. The universal distribution combines a weighted version of the predictions of all enumerable semimeasures,

\(^2\)Traditional notation is ‘$\mu(\Gamma_x)$’ instead of ‘$\mu(x)$’ where cylinder $\Gamma_x = \{\omega \in S : \omega \text{ starts with } x\}$. We use ‘$\mu(x)$’ for convenience. $\mu$ is a measure if equalities hold.
including the prediction of the semimeasure with the shortest program. It is not a priori clear that
the shortest program dominates in all cases—and as we shall see it does not. However, we show
that in the overwhelming majority of cases—the typical cases—the shortest program dominates
sufficiently to validate the approach that only uses shortest programs for prediction. The prop-
erties of $M(x)$ allow us to demonstrate that a minimum description length procedure is almost
always optimal for prediction.

Given a semimeasure on $\{0,1\}^\infty$ and an initial binary string $x$ our goal is to find the most
probable extrapolation of $x$. That is, taking the negative logarithm on both sides of Equation 14,
we want to determine $y$ with $l(y) = n$ that minimizes

$$-\log \mu(y|x) = -\log \mu(xy) + \log \mu(x).$$

Let $\mu$ be a recursive semimeasure. An infinite binary sequence $\omega$ is $\mu$-random iff

$$\sup_n M(\omega_1 \cdots \omega_n) / \mu(\omega_1 \cdots \omega_n) < \infty,$$

and the set of $\mu$-random sequences has $\mu$-measure one, see [8], Chapter 4. Let $\omega$ be a $\mu$-random
infinite binary sequence and $xy$ be a finite prefix of $\omega$. For $l(x)$ grows unboundedly with $l(y)$ fixed,
we have by Theorem 7

$$\lim_{l(x) \to \infty} \log \mu(y|x) - \log M(y|x) = 0. \quad (15)$$

Therefore, if $x$ and $y$ satisfy above conditions, then maximizing $\mu(y|x)$ over $y$ means minimizing
$-\log M(y|x)$. It is shown in Lemma 5 in the Appendix D that $-\log M(x)$ is slightly smaller than
$Km(x)$, the length of the shortest program for $x$ on the reference universal monotonic machine.
For binary programs this difference is very small, Lemma 4, but can be unboundedly in the length
of $x$.

Together this shows the following. Given $xy$ that is a prefix of a (possibly not $\mu$-random) $\omega$,
optimal prediction of fixed length extrapolation $y$ from an unboundedly growing prefix $x$ of $\omega$ need
not necessarily be reached by the shortest programs for $xy$ and $x$ minimizing $Km(xy) - Km(x)$,
but is reached by considering the weighted version of all programs for $xy$ and $x$ which is represented by

$$-\log M(xy) + \log M(x) = (Km(xy) - g(xy)) - (Km(x) - g(x)).$$

Here $g(x)$ is a function which can rise to in between the inverse of the Ackermann function and
$Km(l(x)) \leq \log \log x$—but only in case $x$ is not $\mu$-random.

Therefore, for certain $x$ and $y$ which are not $\mu$-random, optimization using the minimum
length programs may result in very incorrect predictions. However, for $\mu$-random $x$ we have
that $-\log M(x)$ and $Km(x)$ coincide up to an additional constant independent of $x$, that is,
$g(xy) = g(x) = O(1)$, Lemma 5. Hence, together with Equation 15, we find the following.

**Theorem 2.** Let $\mu$ be a recursive semimeasure, and let $\omega$ be a $\mu$-random infinite binary sequence
and $xy$ be a finite prefix of $\omega$. For $l(x)$ grows unboundedly and $l(y)$ fixed,

$$\lim_{l(x) \to \infty} -\log \mu(y|x) = Km(xy) - Km(x) \pm O(1) < \infty,$$

where $Km(xy)$ and $Km(x)$ grow unboundedly.
By its definition $Km$ is monotone in the sense that always $Km(xy) - Km(x) \geq 0$. The closer this difference is to zero, the better the shortest effective monotone program for $x$ is also a shortest effective monotone program for $xy$ and hence predicts $y$ given $x$. Therefore, for all large enough $\mu$-random $x$, predicting by determining $y$ which minimizes the difference of the minimum program lengths of $xy$ and $x$ gives a good prediction. Here $y$ should be preferably large enough to eliminate the influence of the $O(1)$ term.

**Corollary 2 (Prediction by Data Compression)** Assume the conditions of Theorem 2. With $\mu$-probability going to one as $l(x)$ grows unboundedly, a fixed-length $y$ extrapolation from $x$ maximizes $\mu(y|x)$ iff $y$ can be maximally compressed with respect to $x$ in the sense that it minimizes $Km(xy) - Km(x)$. That is, $y$ is the string that minimizes the length difference between the shortest program that outputs $xy, \ldots$ and the shortest program that outputs $x$.

### 4 Conclusion

The analysis of both hypothesis identification by Ideal MDL and prediction shows that maximally compressed descriptions give good results on the data samples which are random with respect to probabilistic hypotheses. These data samples form the overwhelming majority and occur with probability going to one when the length of the data sample grows unboundedly.

**Acknowledgement**

Ray Solomonoff suggested to also analyse the case of prediction. Jorma Rissanen, Chris Wallace, David Dowe, and Kevin Korb gave valuable comments. Part of this work was performed during a stay of the first author at Monash University, Melbourne, Australia.

### A Appendix: Kolmogorov Complexity

The Kolmogorov complexity, $[6, 23, 8]$, of $x$ is simply the length of the shortest effective binary description of $x$. Formally, this is defined as follows. Let $x, y, z \in \mathcal{N}$, where $\mathcal{N}$ denotes the natural numbers and we identify $\mathcal{N}$ and $\{0, 1\}^*$ according to the correspondence

$$(0, \epsilon), (1, 0), (2, 1), (3, 00), (4, 01), \ldots$$

Here $\epsilon$ denotes the empty word $\epsilon$ with no letters. The length $l(x)$ of $x$ is the number of bits in the binary string $x$. For example, $l(010) = 3$ and $l(\epsilon) = 0$.

The emphasis is on binary sequences only for convenience; observations in any alphabet can be so encoded in a way that is 'theory neutral'.

A binary string $x$ is a proper prefix of a binary string $y$ if we can write $x = yz$ for $z \neq \epsilon$. A set $\{x, y, \ldots\} \subseteq \{0, 1\}^*$ is prefix-free if for any pair of distinct elements in the set neither is a proper prefix of the other. A prefix-free set is also called a prefix code. Each binary string $x = x_1x_2 \ldots x_n$ has a special type of prefix code, called a self-delimiting code,

$$\bar{x} = x_1x_2x_2 \ldots x_n \bar{\epsilon},$$

16
where $-x_n = 0$ if $x_n = 1$ and $-x_n = 1$ otherwise. This code is self-delimiting because we can determine where the code word $x$ ends by reading it from left to right without backing up. Using this code we define the standard self-delimiting code for $x$ to be $x' = \overline{l(x)x}$. It is easy to check that $l(\overline{x}) = 2n$ and $l(x') = n + 2 \log n$.

Let $T_1, T_2, \ldots$ be a standard enumeration of all Turing machines, and let $\phi_1, \phi_2, \ldots$ be the enumeration of corresponding functions which are computed by the respective Turing machines. That is, $T_i$ computes $\phi_i$. These functions are the partial recursive functions or computable functions. The Kolmogorov complexity $C(x)$ of $x$ is the length of the shortest binary program from which $x$ is computed. Formally, we define this as follows.

**Definition 2** The Kolmogorov complexity of $x$ given $y$ (for free on a special input tape) is

$$C(x|y) = \min_{p,i} \{l(p'p) : \phi_i(p, y) = x, p \in \{0, 1\}^*, i \in \mathcal{N}\}.$$  

Define $C(x) = C(x|\epsilon)$.

The Kolmogorov complexity is absolute in the sense of being recursively invariant by Church’s Thesis and the ability of universal machines to simulate one another, [8]. For technical reasons we also need a variant of complexity, so-called prefix complexity, which associated with Turing machines for which the set of programs resulting in a halting computation is prefix free. We can realize this by equipping the Turing machine with a one-way input tape, a separate work tape, and a one-way output tape. Such Turing machines are called prefix machines since the halting programs for anyone of them form a prefix free set.

Let $\langle \cdot \rangle$ be a standard invertible effective one-one encoding from $\mathcal{N} \times \mathcal{N}$ to prefix-free recursive subset of $\mathcal{N}$. For example, we can set $\langle x, y \rangle = x'y'$. We insist on prefix-freeness and recursiveness because we want a universal Turing machine to be able to read an image under $\langle \cdot \rangle$ from left to right and determine where it ends.

**Definition 3** The prefix Kolmogorov complexity of $x$ given $y$ (for free) is

$$K(x|y) = \min_{p,i} \{l(p,i) : \phi_i((p,y)) = x, p \in \{0, 1\}^*, i \in \mathcal{N}\}.$$  

Define $K(x) = K(x|\epsilon)$.

## B Appendix: Universal Distribution

A Turing machine $T$ computes a function on the natural numbers. However, we can also consider the computation of real valued functions. For this purpose we consider both the argument of $\phi$ and the value of $\phi$ as a pair of natural numbers according to the standard pairing function $\langle \cdot \rangle$. We define a function from $\mathcal{N}$ to the reals $\mathcal{R}$ by a Turing machine $T$ computing a function $\phi$ as follows. Interpret the computation $\phi(\langle x,t \rangle) = \langle p,q \rangle$ to mean that the quotient $p/q$ is the rational valued $t$th approximation of $f(x)$.

**Definition 4** A function $f : \mathcal{N} \to \mathcal{R}$ is enumerable if there is a Turing machine $T$ computing a total function $\phi$ such that $\phi(x, t + 1) \geq \phi(x, t)$ and $\lim_{t \to \infty} \phi(x, t) = f(x)$. This means that $f$ can be computably approximated from below. If $f$ can also be computably approximated from above then we call $f$ recursive.
A function $P : \mathcal{N} \to [0, 1]$ is a probability distribution if $\sum_{x \in \mathcal{N}} P(x) \leq 1$. (The inequality is a technical convenience. We can consider the surplus probability to be concentrated on the undefined element $u \notin \mathcal{N}$.)

Consider the family $\mathcal{EP}$ of enumerable probability distributions on the sample space $\mathcal{N}$ (equivalently, $\{0, 1\}^\ast$). It is known, [8], that $\mathcal{EP}$ contains an element $m$ that multiplicatively dominates all elements of $\mathcal{EP}$. That is, for each $P \in \mathcal{EP}$ there is a constant $c$ such that $c m(x) > P(x)$ for all $x \in \mathcal{N}$. We call $m$ a universal distribution.

The family $\mathcal{EP}$ contains all distributions with computable parameters which have a name, or in which we could conceivably be interested, or which have ever been considered. The dominating property means that $m$ assigns at least as much probability to each object as any other distribution in the family $\mathcal{EP}$ does. In this sense it is a universal a priori by accounting for maximal ignorance. It turns out that if the true a priori distribution in Bayes’s Rule is recursive, then using the single distribution $m$, or its continuous analogue the measure $M$ on the sample space $\{0, 1\}^\infty$ (defined later), is provably as good as using the true a priori distribution.

We also know, [8], that

**Lemma 1**

$$- \log m(x) = K(x) \pm O(1).$$

That means that $m$ assigns high probability to simple objects and low probability to complex or random objects. For example, for $x = 00\ldots 0$ ($n$ 0’s) we have $K(x) = K(n) + O(1) \leq \log n + 2\log \log n + O(1)$ since the program

```
print n.times a '0'
```

prints $x$. (The additional $2\log \log n$ term is the penalty term for a self-delimiting encoding.) Then, $1/(n \log^2 n) = O(m(x))$. But if we flip a coin to obtain a string $y$ of $n$ bits, then with overwhelming probability $K(y) \geq n - O(1)$ (because $y$ does not contain effective regularities which allow compression), and hence $m(y) = O(1/2^n)$.

### C Appendix: Randomness Tests

One can consider those objects as nonrandom in which one can find sufficiently many regularities. In other words, we would like to identify ‘incompressibility’ with ‘randomness’. This is proper if the sequences that are incompressible can be shown to possess the various properties of randomness (stochasticity) known from the theory of probability. That this is possible is the substance of the celebrated theory developed by the Swedish mathematician Per Martin-Löf.

There are many properties known which probability theory attributes to random objects. To give an example, consider sequences of $n$ tosses with a fair coin. Each sequence of $n$ zeros and ones is equiprobable as an outcome: its probability is $2^{-n}$. If such a sequence is to be random in the sense of a proposed new definition, then the number of ones in $x$ should be near to $n/2$, the number of occurrences of blocks ‘00’ should be close to $n/4$, and so on.

It is not difficult to show that each such single property separately holds for all incompressible binary strings. But we want to demonstrate that incompressibility implies all conceivable effectively testable properties of randomness (both the known ones and the as yet unknown ones). This way, the various theorems in probability theory about random sequences carry over automatically to incompressible sequences.
In the case of finite strings we cannot hope to distinguish sharply between random and nonrandom strings. For instance, considering the set of binary strings of a fixed length, it would not be natural to to fix an \( m \) and call a string with \( m \) zeros random and a string with \( m + 1 \) zeros nonrandom.

Let us borrow some ideas from statistics. We are given a certain sample space \( S \) with an associated distribution \( P \). Given an element \( x \) of the sample space, we want to test the hypothesis ‘\( x \) is a typical outcome’. Practically speaking, the property of being typical is the property of belonging to any reasonable majority. In choosing an object at random, we have confidence that this object will fall precisely in the intersection of all such majorities. The latter condition we identify with being random.

To ascertain whether a given element of the sample space belongs to a particular reasonable majority we introduce the notion of a test. Generally, a test is given by a prescription which, for every level of significance \( \epsilon \), tells us for what elements \( x \) of \( S \) the hypothesis ‘\( x \) belongs to majority \( M \) in \( S' \) should be rejected, where \( \epsilon = 1 - P(M) \). Taking \( \epsilon = 2^{-m} \), \( m = 1, 2, \ldots \), this amounts to saying that we have a description of the set \( V \subseteq \mathcal{N} \times S \) of nested critical regions

\[
V_m = \{ x : (m, x) \in V \} \\
V_m \supseteq V_{m+1}, \quad m = 1, 2, \ldots
\]

The condition that \( V_m \) be a critical region on the significance level \( \epsilon = 2^{-m} \) amounts to requiring, for all \( n \)

\[
\sum_x \{ P(x) : l(x) = n, x \in V_m \} \leq \epsilon.
\]

The complement of a critical region \( V_m \) is called the \((1 - \epsilon)\) confidence interval. If \( x \in V_m \), then the hypothesis ‘\( x \) belongs to majority \( M' \), and therefore the stronger hypothesis ‘\( x \) is random’, is rejected with significance level \( \epsilon \). We can say that \( x \) fails the test at the level of critical region \( V_m \).

**Example 2** A string \( x_1x_2\ldots x_n \) with many initial zeros is not very random. We can test this aspect as follows. The special test \( V \) has critical regions \( V_1, V_2, \ldots \). Consider \( x = 0.x_1x_2\ldots x_n \) as a rational number, and each critical region as a half-open interval \( V_m = [0, 2^{-m}) \) in \([0, 1), \)

\( m = 1, 2, \ldots \). Then the subsequent critical regions test the hypothesis ‘\( x \) is random’ by considering the subsequent digits in the binary expansion of \( x \). We reject the hypothesis on the significance level \( \epsilon = 2^{-m} \) provided \( x_1 = x_2 = \cdots = x_m = 0 \).

**Example 3** Another test for randomness of finite binary strings rejects when the relative frequency of ones differs too much from \( 1/2 \). This particular test can be implemented by rejecting the hypothesis of randomness of \( x = x_1x_2\ldots x_n \) at level \( \epsilon = 2^{-m} \) provided \( |2f_n - n| > g(n, m) \), where \( f_n = \sum_{i=1}^{n} x_i \), and \( g(n, m) \) is the least number determined by the requirement that the number of binary strings \( x \) of length \( n \) for which this inequality holds is at most \( 2^{n-m} \).

In practice, statistical tests are effective prescriptions such that we can compute, at each level of significance, for what strings the associated hypothesis should be rejected. It would be hard to imagine what use it would be in statistics to have tests that are not effective in the sense of computability theory.

**Definition 5** Let \( P \) be a recursive probability distribution on the sample space \( \mathcal{N} \). A total function \( \delta: \mathcal{N} \to \mathcal{N} \) is a \( P \)-test (Martin-Löf test for randomness) if:
1. \( \delta \) is enumerable (the set \( V = \{(m,x) : \delta(x) \geq m \} \) is recursively enumerable); and

2. \( \sum\{P(x) : \delta(x) \geq m, l(x) = n\} \leq 2^{-m} \), for all \( n \).

The critical regions associated with the common statistical tests are present in the form of the sequence \( V_1 \supseteq V_2 \supseteq \cdots \), where \( V_m = \{x : \delta(x) \geq m\} \), for \( m \geq 1 \). Nesting is assured since \( \delta(x) \geq m+1 \) implies \( \delta(x) \geq m \). Each set \( V_m \) is recursively enumerable because of Item 1.

A particularly important case is \( \mathcal{P} \) is the uniform distribution, defined by \( L(x) = 2^{-2l(x)} \). The restriction of \( L \) to strings of length \( n \) is defined by \( L_n(x) = 2^{-n} \) for \( l(x) = n \) and 0 otherwise. (By definition, \( L_n(x) = L(x|l(x) = n) \).) Then, Item 2 can be rewritten as \( \sum_{x \in V_m} L_n(x) \leq 2^{-m} \) which is the same as

\[
d(\{x : l(x) = n, x \in V_m\}) \leq 2^{n-m}.
\]

In this case we often speak simply of a test, with the uniform distribution \( L \) understood.

In statistical tests membership of \( (m,x) \) in \( V \) can usually be determined in time polynomial in \( l(m) + l(x) \).

**Example 4** The previous test examples can be rephrased in terms of Martin-Löf tests. Let us try a more subtle example. A real number such that all bits in odd positions in its binary representation are 1’s is not random with respect to the uniform distribution. To show this we need a test which detects sequences of the form \( x = 1x_21x_41x_61x_8 \ldots \). Define a test \( \delta \) by

\[
\delta(x) = \max\{i : x_1 = x_3 = \cdots = x_{2i-1} = 1\},
\]

and \( \delta(x) = 0 \) if \( x_1 = 0 \). For example: \( \delta(01111) = 0 \); \( \delta(10011) = 1 \); \( \delta(11011) = 1 \); \( \delta(10100) = 2 \); \( \delta(11111) = 3 \). To show that \( \delta \) is a test we have to show that \( \delta \) satisfies the definition of a test. Clearly, \( \delta \) is enumerable (even recursive). If \( \delta(x) \geq m \) where \( l(x) = n \geq 2m \), then there are \( 2^{m-1} \) possibilities for the \((2m-1)\)-length prefix of \( x \), and \( 2^n-(2^{m-1}) \) possibilities for the remainder of \( x \). Therefore, \( d\{x : \delta(x) \geq m, l(x) = n\} \leq 2^{n-m} \).

**Definition 6** A universal Martin-Löf test for randomness with respect to distribution \( P \), a universal \( P \)-test for short, is a test \( \delta_0(x|P) \) such that for each \( P \)-test \( \delta \), there is a constant \( c \), such that for all \( x \), we have \( \delta_0(x|P) \geq \delta(x) - c \).

We say that \( \delta_0(x|P) \) (additively) majorizes \( \delta \). Intuitively, \( \delta_0(x|P) \) constitutes a test for randomness which incorporates all particular tests \( \delta \) in a single test. No test for randomness \( \delta \) other than \( \delta_0(x|P) \) can discover more than a constant amount more deficiency of randomness in any string \( x \). In terms of critical regions, a universal test is a test such that if a binary sequence is random with respect to that test, then it is random with respect to any conceivable test, neglecting a change in significance level. Namely, with \( \delta_0(x|P) \) a universal \( P \)-test, let \( U = \{(m,x) : \delta_0(x|P) \geq m\} \), and, for any test \( \delta \), let \( V = \{(m,x) : \delta(x) \geq m\} \). Then, defining the associated critical zones as before, we find

\[
V_{m+c} \subseteq U_m, \quad m = 1, 2, \ldots,
\]

where \( c \) is a constant (dependent only on \( U \) and \( V \)).

It is a major result that there exists a universal \( P \)-test. The proof goes by first showing that the set of all tests is enumerable.
LEMMA 2 We can effectively enumerate all $P$-tests.

Proof. We start with the standard enumeration $\phi_1, \phi_2, \ldots$ of partial recursive functions from $\mathcal{N}$ into $\mathcal{N} \times \mathcal{N}$, and turn this into an enumeration $\delta_1, \delta_2, \ldots$ of all and only $P$-tests. The list $\phi_1, \phi_2, \ldots$ enumerates all and only recursively enumerable sets of pairs of integers as $\{ \phi_i(x) : x \geq 1 \}$ for $i = 1, 2, \ldots$. In particular, for any $P$-test $\delta$, the set $\{(m, x) : \delta(x) \geq m\}$ occurs in this list. The only thing we have to do is to eliminate those $\phi_i$ of which the range does not correspond to a $P$-test.

First, we effectively modify each $\phi$ (we drop the subscript for convenience) to a function $\psi$ such that range $\phi$ equals range $\psi$, and $\psi$ has the special property that if $\psi(n)$ is defined, then $\psi(1), \psi(2), \ldots, \psi(n-1)$ are also defined. This can be done by dovetailing the computations of $\phi$ on the different arguments: in the first phase do one step of the computation of $\phi(1)$, in the second phase do the second step of the computation of $\phi(1)$ and the first step of the computation of $\phi(2)$. In general, in the $n$th phase we execute the $n_1$th step of the computation of $\phi(n_2)$, for all $n_1, n_2$ satisfying $n_1 + n_2 = n$. We now define $\psi$ as follows. If the first computation that halts is that of $\phi(i)$, then set $\psi(1) := \phi(i)$. If the second computation that halts is that of $\phi(j)$, then set $\psi(2) := \phi(j)$, and so on.

Secondly, use each $\psi$ to construct a test $\delta$ by approximation from below. In the algorithm, at each stage of the computation the local variable array $\delta(1 : \infty)$ contains the current approximation to the list of function values $\delta(1), \delta(2), \ldots$. This is doable because the nonzero part of the approximation is always finite.

**Step 1** Initialize $\delta$ by setting $\delta(x) := 0$ for all $x$. (If the range of $\psi$ is empty, then this assignment will not be changed in the remainder of the procedure. That is, $\delta$ stays identically zero and it is trivially a test.) Initialize $i := 0$.

**Step 2** Set $i := i + 1$; compute $\psi(i)$ and let its value be $(x, m)$.

**Step 3** If $\delta(x) \geq m$ then go to Step 2. else set $\delta(x) := m$.

**Step 4** If $\sum \{ P(y) : \delta(y) \geq k, l(y) = l(x)) > 2^{-k}$ for some $k, k = 1, \ldots, m$ (since $P$ is a recursive function we can effectively test whether the new value of $\delta(x)$ violates Definition 6) then set $\delta(x) := 0$ and terminate (the computation of $\delta$ is finished) else go to Step 2.

(With $P$ the uniform distribution, for $i = 1$ the conditional in Step 4 simplifies to $m > l(x)$.) In case the range of $\psi$ is already a test, then the algorithm never finishes but forever approximates $\delta$ from below. If $\psi$ diverges for some argument then the computation goes on forever and does not change $\delta$ any more. The resulting $\delta$ is an enumerable test. If the range of $\psi$ is not a test, then at some point the conditional in Step 4 is violated and the approximation of $\delta$ terminates. The resulting $\delta$ is a test, even a recursive one. Executing this procedure on all functions in the list $\phi_1, \phi_2, \ldots$, we obtain an effective enumeration $\delta_1, \delta_2, \ldots$ of all $P$-tests (and only $P$-tests). We are now in the position to define a universal $P$-test. \hfill \Box

**Theorem 3** Let $\delta_1, \delta_2, \ldots$ be an enumeration of above $P$-tests. Then, $\delta_0(x | P) = \max \{ \delta_y(x) - y : y \geq 1 \}$ is a universal $P$-test.

Proof. Note first that $\delta_0(\cdot | P)$ is a total function on $\mathcal{N}$ because of Item 2 in Definition 6.

21
(1) The enumeration \( \delta_1, \delta_2, \ldots \) in Lemma 2 yields an enumeration of recursively enumerable sets:
\[
\{ (m, x) : \delta_1(x) \geq m \}, \{ (m, x) : \delta_2(x) \geq m \}, \ldots
\]
Therefore, \( V = \{ (m, x) : \delta_0(x|P) \geq m \} \) is recursively enumerable.

(2) Let us verify that the critical regions are small enough: for each \( n \),
\[
\sum_{l(x)=n} \{ P(x) : \delta_0(x P) \geq m \} \leq \sum_{y=1}^{\infty} \sum_{l(x)=n} \{ P(x) : \delta_y(x) \geq m + y \}
\]
\[
\leq \sum_{y=1}^{\infty} 2^{-m-y} = 2^{-m}.
\]

(3) By its definition, \( \delta_0(P) \) majorizes each \( \delta \) additively. Hence, it is universal. \( \square \)

By definition of \( \delta_0(P) \) as a universal \( P \)-test, any particular \( P \)-test \( \delta \) can discover at most a constant amount more regularity in a sequence \( x \) than does \( \delta_0(P) \), in the sense that for each \( \delta_y \) we have \( \delta_y(x) \leq \delta_0(x|P) + y \) for all \( x \).

For any two universal \( P \)-tests \( \delta_0(P) \) and \( \delta_0'(P) \), there is a constant \( c \geq 0 \), such that for all \( x \), we have \( \delta_0(x|P) - \delta_0'(x|P) \leq c \).

We started out with the objective to establish in what sense incompressible strings may be called random.

**Theorem 4** The function \( f(x) = l(x) - C(x|l(x)) - 1 \) is a universal \( L \)-test with \( L \) the uniform distribution.

**Proof.** (1) We first show that \( f(x) \) is a test with respect to the uniform distribution. The set \( \{ (m, x) : f(x) \geq m \} \) is recursively enumerable since \( C() \) can be approximated from above by a recursive process.

(2) We verify the condition on the critical regions. Since the number of \( x \)'s with \( C(x|l(x)) \leq l(x) - m - 1 \) cannot exceed the number of programs of length at most \( l(x) - m - 1 \), we have
\[
d(\{ x : f(x) \geq m \}) \leq 2^{l(x) - m - 1}.
\]

(3) We show that for each test \( \delta \), there is a constant \( c \), such that \( f(x) \geq \delta(x) - c \). The main idea is to bound \( C(x|l(x)) \) by exhibiting a description of \( x \), given \( l(x) \). Fix \( x \). Let the set \( A \) be defined as
\[
A = \{ z : \delta(z) \geq \delta(x), l(z) = l(x) \}.
\]

We have defined \( A \) such that \( x \in A \) and \( d(A) \leq 2^{l(x) - \delta(x)} \). Let \( \delta = \delta_y \) in the standard enumeration \( \delta_1, \delta_2, \ldots \) of tests. Given \( y, l(x) \), and \( \delta(x) \), we can enumerate all elements of \( A \). Together with \( x \)'s index \( j \) in enumeration order of \( A \), this suffices to find \( x \). We pad the standard binary representation of \( j \) with nonsignificant zeros to a string \( s = 00 \ldots 0j \) of length \( l(x) - \delta(x) \). This is possible since \( l(s) \geq l(d(A)) \). The purpose of changing \( j \) to \( s \) is that now the number \( \delta(x) \) can be deduced from \( l(s) \) and \( l(x) \). In particular, there is a Turing machine which computes \( x \) from input \( y \), when \( l(x) \) is given for free. Consequently, since \( C() \) is the shortest effective description, \( C(x|l(x)) \leq l(x) - \delta(x) + 2l(y) + 1 \). Since \( y \) is a constant depending only on \( \delta \), we can set \( c = 2l(y) + 2 \). \( \square \)

In Theorem 3, we have exhibited a universal \( P \)-test for randomness of a string \( x \) of length \( n \) with respect to an arbitrary recursive distribution \( P \) over the sample set \( S = B^n \) with \( B = \{ 0, 1 \} \).
The universal $P$-test measures how justified is the assumption that $x$ is the outcome of an experiment with distribution $P$. We now use $m$ to investigate alternative characterizations of random elements of the sample set $S = B^*$ (equivalently, $S = \mathcal{N}$).

**Definition 7** Let $P$ be a recursive probability distribution on $\mathcal{N}$. A **sum $P$-test** is a nonnegative enumerable function $\delta$ satisfying
\[ \sum_x P(x)2^{\delta(x)} \leq 1. \tag{17} \]
A **universal sum $P$-test** is a test that additively dominates each sum $P$-test.

The sum tests of Definition 7 are slightly stronger than the tests according to Martin-Löf's original Definition 5.

**Lemma 3** Each sum $P$-test is a $P$-test. If $\delta(x)$ is a $P$-test, then there is a constant $c$ such that $\delta'(x) = \delta(x) - 2\log \delta(x) - c$ is a sum $P$-test.

**Proof.** It follows immediately from the new definition that for all $n$
\[ \sum\{P(x) : \delta(x) > k, l(x) = n\} \leq 2^{-k}. \tag{18} \]
Namely, if Equation 18 is false, then we contradict Equation 17 by
\[ \sum_{x \in \mathcal{N}} P(x)2^{\delta(x)} > \sum_{l(x) = n} P(x)2^k \geq 1. \]

Conversely, if $\delta(x)$ satisfies Equation 18 for all $n$, then for some constant $c$, the function $\delta(x) - 2\log \delta(x) - c$ satisfies Equation 17. $\square$

This shows that the sum test is not much stronger than the original test. One advantage of Equation 17 is that it is just one inequality, instead of infinitely many, one for each $n$. We give an exact expression for a universal sum $P$-test in terms of complexity.

**Theorem 5** Let $P$ be a recursive probability distribution. The function
\[ \kappa_0(x|P) = \log(m(x)/P(x)) \]
is a universal sum $P$-test.

**Proof.** Since $m$ is enumerable, and $P$ is recursive, $\kappa_0(x|P)$ is enumerable. We first show that $\kappa_0(x|P)$ is a sum $P$-test:
\[ \sum_x P(x)2^{\kappa_0(x|P)} = \sum_x m(x) \leq 1. \]
It is only left to show that $\kappa_0(x|P)$ additively dominates all sum $P$-tests. For each sum $P$-test $\delta$, the function $P(x)2^{\delta(x)}$ is a semimeasure that is enumerable. It has been shown, Section A, that there is a positive constant $c$ such that $c \cdot m(x) \geq P(x)2^{\delta(x)}$. Hence, there is another constant $c$ such that $c \cdot \kappa_0(x|P) \geq \delta(x)$, for all $x$. $\square$
Example 5  An important case is as follows. If we consider a distribution $P$ restricted to a domain $A \subseteq \mathcal{N}$, then the universal sum $P$-test becomes $\log(m(x|A)/P(x|A))$. For example, if $L_n$ is the uniform distribution on $A = \{0, 1\}^n$, then the universal sum $L_n$-test for $x \in A$ becomes

$$\kappa_0(x|L_n) = \log(m(x|A)/L_n(x)) = n - K(x|n) - O(1).$$

Namely, $L_n(x) = 1/2^n$ and $\log m(x|A) = -K(x|A)$ by the Coding Theorem, Section A. where we can describe $A$ by giving $n$. \hfill \Box

Example 6  The Noiseless Coding Theorem states that the Shannon-Fano code, which codes a source word $x$ straightforwardly as a word of about $-\log P(x)$ bits, Section A, nearly achieves the optimal expected code word length. This code is uniform in the sense that it does not use any characteristics of $x$ itself to associate a code word with a source word $x$. The code that codes each source word $x$ as a code word of length $K(x)$ also achieves the optimal expected code word length. This code is nonuniform in that it uses characteristics of individual $x$'s to obtain shorter code words. Any difference in code word length between these two encodings for a particular object $x$ can only be due to exploitation of the individual regularities in $x$.

Define the randomness deficiency of a finite object $x$ with respect to $P$ as

$$-\log P(x) - K(x) = -\log P(x) + \log m(x) + O(1) = \kappa_0(x|P) + O(1),$$

by the major theorems in Section C. That is, the randomness deficiency is the outcome of the universal sum $P$-test of Theorem 5. \hfill \Box

Example 7  Let us compare the randomness deficiency as measured by $\kappa_0(x|P)$ with that measured by the universal test $\delta_0(x)$, for the uniform distribution. That test consisted actually of tests for a whole family $L_n$ of distributions, where $L_n$ is the uniform distribution such that each $L_n(x) = 2^{-n}$ for $l(x) = n$, and zero otherwise. Rewrite $\delta_0(x)$ as

$$\delta_0(x|L_n) = n - C(x|n),$$

for $l(x) = n$, and $\infty$ otherwise. This is close to the expression for $\kappa_0(x|L_n)$ obtained in Example 5.

From the relations between $C$ and $K$ in [8] it follows that

$$|\delta_0(x|L_n) - \kappa_0(x|L_n)| \leq 2\log C(x) + O(1).$$

The formulation of the universal sum test in Theorem 5 can be interpreted as follows. An element $x$ is random with respect to distribution $P$, that is, $\kappa_0(x|P) = O(1)$, if $P(x)$ is large enough, not in absolute value but relative to $m(x)$. If we did not have this relativization, then we would not be able to distinguish between random and nonrandom outcomes for the uniform distribution $L_n(x)$ above.

Let us look at an example. Let $x = 00\ldots 0$ of length $n$. Then, $\kappa_0(x|L_n) = n - K(x|n) + O(1) = n \pm O(1)$. If we flip a coin $n$ times to generate $y$, then with overwhelming probability $K(y|n) \geq n$ and $\kappa_0(y|L_n) = O(1).$ \hfill \Box

Example 8  According to modern physics, electrons, neutrons and protons satisfy the Fermi-Dirac distribution. We distribute $n$ particles among $k$ cells, for $n \leq k$, such that each cell is occupied by at most one particle; and all distinguished arrangements satisfying this have the same probability.
We can treat each arrangement as a binary string: an empty cell is a zero and a cell with a particle is a one. Since there are \(\binom{n}{k}\) possible arrangements, the probability for each arrangement \(x\) to happen, under the Fermi-Dirac distribution, is \(FD_{n,k}(x) = \binom{k}{n}^{-1}\). According to Theorem 5:

\[
\kappa_0(x | FD_{n,k}) = \log(\mathcal{m}(x, k, n) / FD_{n,k}(x)) = -K(x, n, k) + \log \left(\binom{k}{n}\right) + O(1)
\]

is a universal sum test with respect to the Fermi-Dirac distribution. It is easy to see that a binary string \(x\) of length \(k\) with \(n\) ones has complexity \(K(x, n, k) \leq \log \binom{k}{n}\), and \(K(x, n, k) \geq \log \binom{k}{n} - O(1)\) for most such \(x\). Hence, a string \(x\) with maximal \(K(x, n, k)\) will pass this universal sum test. Each individual such string possesses all effectively testable properties of typical strings under the Fermi-Dirac distribution. Hence, in the limit for \(n\) and \(k\) growing unboundedly, we cannot effectively distinguish one such string from other such strings. \(\diamondsuit\)

**Example 9** Markov’s Inequality says the following. Let \(P\) be any probability distribution, \(f\) any nonnegative function with \(P\)-expected value \(E = \sum_x P(x) f(x) < \infty\). For \(\mathbb{E} \geq 0\) we have \(\sum\{P(x) : f(x)/\mathbb{E} > k\} < 1/k\).

Let \(P\) be any probability distribution (not necessarily recursive). The \(P\)-expected value of \(\mathcal{m}(x)/P(x)\) is

\[
\sum_x P(x) \frac{\mathcal{m}(x)}{P(x)} \leq 1.
\]

Then, by Markov’s Inequality

\[
\sum_x \{P(x) : \mathcal{m}(x) \leq kP(x)\} \geq 1 - \frac{1}{k}.
\]

(19)

Since \(\mathcal{m}\) dominates all enumerable semimeasures multiplicatively, we have for all \(x\),

\[
P(x) \leq c_P \mathcal{m}(x), \text{ and it can be shown } c_P = 2^{K(P)}.
\]

(20)

Equations 19 and 20 have the following consequences.

1. If \(x\) is a random sample from a simple recursive distribution \(P\), where ‘simple’ means that \(K(P)\) is small, then \(\mathcal{m}\) is a good estimate for \(P\). For instance, if \(x\) is randomly drawn from distribution \(P\), then the probability that

\[
c_P^{-1} \mathcal{m}(x) \leq P(x) \leq c_P \mathcal{m}(x)
\]

is at least \(1 - 1/c_P\).

2. If we know or believe that \(x\) is random with respect to \(P\), and we know \(P(x)\), then we can use \(P(x)\) as an estimate of \(\mathcal{m}(x)\).

In both cases the degree of approximation depends on the index of \(P\), and the randomness of \(x\) with respect to \(P\), as measured by the randomness deficiency \(\kappa_0(x | P) = \log(\mathcal{m}(x)/P(x))\). For example, the uniform discrete distribution on \(\mathcal{B}^*\) can be defined by \(L(x) = 2^{-2l(x)}\). Then, for each \(n\) we have \(L_n(x) = L(x | l(x) = n)\). To describe \(L\) takes \(O(1)\) bits, and therefore

\[
\kappa_0(x | L) = l(x) - K(x) + O(1).
\]

25
The randomness deficiency $\kappa_0(x|L) = O(1)$ iff $K(x) \geq l(x) - O(1)$, that is, iff $x$ is random.

The nonrecursive ‘distribution’ $m(x) = 2^{-K(x)}$ has the remarkable property that the test $\kappa_0(x|m) = O(1)$ for all $x$: the test shows all outcomes $x$ random with respect to it. We can interpret Equations 19, 20 as saying that if the real distribution is $P$, then $P(x)$ and $m(x)$ are close to each other with large $P$-probability. Therefore, if $x$ comes from some unknown recursive distribution $P$, then we can use $m(x)$ as an estimate for $P(x)$. In other words, $m(x)$ can be viewed as the universal a priori probability’ of $x$.

The universal sum $P$-test $\kappa_0(x|P)$ can be interpreted in the framework of hypothesis testing as the likelihood ratio between hypothesis $P$ and the fixed alternative hypothesis $m$. In ordinary statistical hypothesis testing, some properties of an unknown distribution $P$ are taken for granted, and the role of the universal test can probably be reduced to some tests that are used in statistical practice.

\[ S_n = \sum_{l(x)=n-1} \mu(x)(M(0|x) - \mu(0|x))^2. \]

We may call $S_n$ the expected squared error at the $n$th prediction.

**Theorem 6** Let $\mu$ be a recursive semimeasure. Using the notation above, $\sum_n S_n \leq k/2$ with $k = K(\mu)\ln 2$. (Hence, $S_n$ converges to 0 faster than $1/n$.)

A proof using Kulback-Leibler divergence is given in [8]. There it is additionally demonstrated that for almost all unbounded $x$ the conditional probability of $M$ converges to the conditional probability of $\mu$. Note that while the following Theorem does imply the convergence of the conditional probabilities similarly to Theorem 6, it does not imply the speed of convergence estimate. Conversely, Theorem 6 does not imply the following.

**Theorem 7** Let $\mu$ be a positive recursive measure. If the length of $y$ is fixed and the length of $x$ grows to infinity, then

\[ \frac{M(y|x)}{\mu(y|x)} \to 1, \]

with $\mu$-probability one. In infinite sequences $\omega$ with prefixes $x$ satisfying the displayed asymptotics are precisely the $\mu$-random sequences.

**Proof.** We use an approach based on the Submartingale Convergence Theorem, [3] pp. 324-325, which states that the following property holds for each sequence of random variables $\omega_1, \omega_2, \ldots$ If $f(\omega_{1:n})$ is a $\mu$-submartingale, and the $\mu$-expectation $E f(\omega_{1:n}) < \infty$, then it follows that $\lim_{n \to \infty} f(\omega_{1:n})$ exists with $\mu$-probability one.

In our case,

\[ t(\omega_{1:n}|\mu) = \frac{M(\omega_{1:n})}{\mu(\omega_{1:n})} \]
is a $\mu$-submartingale, and the $\mu$-expectation $\mathbb{E}t(\omega_{1:n}|\mu) \leq 1$. Therefore, there is a set $A \subseteq \mathcal{B}^\infty$ with $\mu(A) = 1$, such that for each $\omega \in A$ the limit $\lim_{n \to \infty} t(\omega_{1:n}|\mu) < \infty$. These are the $\mu$-random $\omega$'s by Corollary 4.8 in [8]. Consequently, for fixed $m$, for each $\omega$ in $A$, we have

$$\lim_{n \to \infty} \frac{M(\omega_{1:n+m}|\mu)}{M(\omega_{1:n})} = 1,$$

provided the limit of the denominator is not zero. The latter fact is guaranteed by the universality of $M$: for each $x \in \mathcal{B}^*$ we have $M(x)/\mu(x) \geq 2^{-K(\mu)}$ by Theorem 4.4 and Equation 4.10 in [8]. (This is the continuous analogue of Theorem 5.)

**Example 10** Suppose we are given an infinite decimal sequence $\omega$. The even positions contain the subsequent digits of $\pi = 3.1415\ldots$, and the odd positions contain uniformly distributed, independently drawn random decimal digits. Then, $M(a|\omega_{1:2i}) \to 1/10$ for $a = 0, 1, \ldots, 9$, while $M(a|\omega_{1:2i+1}) \to 1$ if $a$ is the $i$th digit of $\pi$, and to 0 otherwise.

There are two possibilities to associate complexities with machines. The first possibility is to take the length of the shortest program, while the second possibility is to take the negative logarithm of the universal probability. In the discrete case, using prefix machines, these turned out to be the same by the Coding Theorem 1. In the continuous case, using monotone machines, it turns out they are different.

**Definition 8** The complexity $KM$ is defined as

$$KM(x) = -\log M(x).$$

In contrast with $C$ and $K$ complexities, in the above definition the greatest prefix-free subset of all programs which produce output starting with $x$ on the reference monotone machine $U$ are weighed.

**Definition 9** Let $U$ be the reference monotone machine. The complexity $Km$, called monotone complexity, is defined as

$$Km(x) = \min\{l(p) : U(p) = x, \omega \in S_U\}.$$

We omit the Invariance Theorems for $KM$ complexity and $Km$ complexity, stated and proven completely analogous to the Theorems with respect to the $C$ and $K$ varieties. By definition, $KM(x) \leq Km(x)$. In fact, all proper complexities coincide up to a logarithmic additive term. It has been shown that equality does not hold: the difference between $KM(x)$ ($= -\log M(x)$) and $Km(x)$ is very small, but still rises unboundedly. This contrasts with the equality between $-\log m(x)$ and $K(x)$ in Theorem 1. Intuitively, this phenomenon is justified by exposing the relation between $M$ and $m$.

The Coding Theorem 1 states that $K(x) = -\log m(x) + O(1)$. L.A. Levin, [7], conjectured that the analogue would hold for the unrestricted continuous version. But it has been shown, [5], that

$$\sup_{x \in \mathcal{B}^*} |KM(x) - Km(x)| = \infty,$$

There it is shown that the exact relation is (for each particular choice of basis $\mathcal{B}$ such as $\mathcal{B} = \mathcal{N}$, the natural numbers, or $\mathcal{B} = \{0, 1\}$)

27
This shows that the differences between $Km(x)$ and $KM(x)$ must in some sense be very small. The next question to ask is whether the quantities involved are usually different, or whether this is a rare occurrence. In other words, whether for \textit{a priori} almost all infinite sequences $x$, the difference between $Km$ and $KM$ is bounded by a constant. The following facts have been proven, [5].

\textbf{Lemma 5} (i) For random strings $x \in B^*$ we have $Km(x) - KM(x) = O(1)$.

(ii) There exists a function $f(n)$ which goes to infinity with $n \to \infty$ such that $Km(x) - KM(x) \geq f(l(x))$, for infinitely many $x$. If $x$ is a finite binary string ($B = \{0, 1\}$), then we can choose $f(n)$ as the inverse of some version of Ackermann's function

\textbf{References}

[1] L. Breiman, J. Friedman, R. Olshen and C. Stone, \textit{Classification and Regression Trees}. Wadsworth International Group, Belmont, CA, 1984.

[2] W. Buntine, Personal communication, September 1994.

[3] J.L. Doob, \textit{Stochastic Processes }, Wiley, 1953.

[4] P. Gács, On the symmetry of algorithmic information, \textit{Soviet Math. Dokl.}, 15 (1974) 1477-1480. Correction: ibid., 15 (1974) 1480.

[5] P. Gács, On the relation between descriptive complexity and algorithmic probability, \textit{Theoret. Comput. Sci.}, 22(1983), 71-93.

[6] A.N. Kolmogorov, Three approaches to the quantitative definition of information, \textit{Problems Inform. Transmission} 1:1 (1965) 1-7.

[7] L.A. Levin, On the notion of a random sequence, \textit{Soviet Math. Dokl.}, 14(1973), 1413-1416.

[8] M. Li and P.M.B. Vitányi, \textit{An Introduction to Kolmogorov Complexity and its Applications}, Springer-Verlag, New York, 1993.

[9] D. MacKay, pp 59 in \textit{Maximum Entropy and Bayesian Methods}, Kluwer, 1992. (Personal communication W. Buntine.)

[10] P. Martin-Löf, The definition of random sequences, \textit{Inform. Contr.}, 9(1966), 602-619.

[11] J.J. Rissanen, Modeling by the shortest data description, \textit{Automatica-J. IFAC} 14 (1978) 465-471.

[12] J.J. Rissanen, \textit{Stochastic Complexity and Statistical Inquiry}, World Scientific Publishers, 1989.

[13] J.J. Rissanen, Fisher information and stochastic complexity, \textit{IEEE Trans. Inform. Theory}, IT-42:1(1996), 40-47.

[14] J. Segen, \textit{Pattern-Directed Signal Analysis}, PhD Thesis, Carnegie-Mellon University, Pittsburgh, 1980.
[15] R.J. Solomonoff, A formal theory of inductive inference, Part 1 and Part 2, *Inform. Contr.*, 7(1964), 1-22, 224-254.

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

[17] A.M. Turing, On computable numbers with an application to the Entscheidungsproblem, *Proc. London Math. Soc., Ser. 2*, 42(1936), 230-265; Correction, Ibid, 43(1937), 544-546.

[18] R. von Mises, Grundlagen der Wahrcheinlichkeitsrechnung, *Mathemat. Zeitsch.*, 5(1919), 52-99.

[19] V. Vovk, Minimum description length estimators under the universal coding scheme, in: P. Vitányi (Ed.), *Computational Learning Theory*, Proc. 2nd European Conf. (EuroCOLT '95), Lecture Notes in Artificial Intelligence, Vol. 904, Springer-Verlag, Heidelberg, 1995, pp. 237-251; Learning about the parameter of the Bernoulli model, *J. Comput. System Sci.*, to appear.

[20] C.S. Wallace and D.M. Boulton, An information measure for classification, *Computing Journal* 11 (1968) 185-195.

[21] C.S. Wallace and P.R. Freeman, Estimation and inference by compact coding, *J. Royal Stat. Soc., Series B*, 49 (1987) 240-251. Discussion: ibid.,252-265.

[22] K. Yamanishi, A Randomized Approximation of the MDL for Stochastic Models with Hidden Variables, *Proc. 9th ACM Comput. Learning Conference*, ACM Press, 1996.

[23] A.K. Zvonkin and L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, *Russian Math. Surveys* 25:6 (1970) 83-124.

29