Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity*

Paul Vitányi†
CWI and University of Amsterdam

Ming Li‡
University of Waterloo

August 17, 1999

Abstract

The relationship between the Bayesian approach and the minimum description length approach is established. We sharpen and clarify the general modeling principles MDL and MML, abstracted as the ideal MDL principle and defined from Bayes's rule by means of Kolmogorov complexity. The basic condition under which the ideal principle should be applied is encapsulated as the Fundamental Inequality, which in broad terms states that the principle is valid when the data are random, relative to every contemplated hypothesis and also these hypotheses are random relative to the (universal) prior. Basically, the ideal principle states that the prior probability associated with the hypothesis should be given by the algorithmic universal probability, and the sum of the log universal probability of the model plus the log of the probability of the data given the model should be minimized. If we restrict the model class to the finite sets then application of the ideal principle turns into Kolmogorov's minimal sufficient statistic. In general we show that data compression is almost always the best strategy, both in hypothesis identification and prediction.

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

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

1 Introduction

It is widely believed that the better a theory compresses the data concerning some phenomenon under investigation, the better we have learned, generalized, and the better the theory predicts unknown data. This belief is vindicated in practice and is a form of "Occam's razor" paradigm

*Parts of the ideas in this paper were announced 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, 518-535; 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. Address: Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: mli@math.uwaterloo.ca

about "simplicity" but apparently has not been rigorously proved in a general setting. Here we show that data compression is almost always the best strategy, both in hypotheses identification by using an ideal form of the minimum description length (ideal MDL) principle and in prediction of sequences. To demonstrate these benificial aspects of compression we use the Kolmogorov theory of complexity [15] to express the optimal effective compression. The general idea to use compressed data for prediction was apparently first put forward by Solomonoff [30]. Independently, Wallace and co-authors formulated in [38] the idea of minimum message length (MML) as compressed two-part codes for the data corresponding to replacing negative-log probabilities in Bayes's rule by Shannon-Fano code lengths. Rissanen [25], independent of Wallace but inspired by Solomonoff's and Kolmogorov's idea of ultimate effective compression of data, formulated the miniumum description length idea using essentially the formal equivalent of negative-log probabilities. In 1989 the current authors formulated ideal MDL [20], and in 1991 Barron and Cover [4] analyzed statistical MDL and hinted at (but didn't formulate) the idea of ideal MDL. Here we extend and complete the analysis of [20] and identify precisely the situations in which ideal MDL and Bayesianism coincide and where they differ. We indirectly validate MDL by showing that in a 'typical' situation its decisions coincide with those of Bayesianism: With probability rising to one with increasing sample size, both the MDL principle and Bayesian induction select the same hypothesis or prediction. In fact, we identify the class of hypothesis-sample pairs for which this happens: the "individually random" ones. Consequently, all results about convergence of Bayesian methods carry over to convergence with probablity one for MDL induction.

Hypothesis identification To demonstrate that compression is good for hypothesis identification we use the ideal MDL principle defined from Bayes's rule by means of Kolmogorov complexity, Section 2. This transformation is valid only for individually random objects in computable distributions; if the contemplated objects are nonrandom or the distributions are not computable then MDL and Bayes's rule may part company. Basing MDL on first principles we probe below the customary presentation of MDL as being justified in and of itself by philosophical persuasion [25, 26]. The minimum message length (MML) approach, while relying on priors, in practice is a related approach [38, 39]. Such approaches balance the complexity of the model (and its tendency for overfitting) against the preciseness of fitting the data (the error of the hypothesis). Our analysis gives evidence why in practice Bayesianism is prone to overfitting and MDL isn't.

Ideal MDL We are only interested in the following common idea shared between all MDL-like 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." We take this to mean that every contemplated individual hypothesis and every contemplated individual data sample is to be maximally compressed: the description lengths involved should be the shortest effective description lengths. We use "effective" in the sense of "Turing computable," [34]. Shortest effective description length is asymptotically unique and objective and known as the *Kolmogorov complexity* [15] of the object being described. Thus, "ideal MDL" is a Kolmogorov complexity based form of the minimum description length principle. In order to define ideal MDL from Bayes's rule we require some deep results due to L.A. Levin [19] and P. Gács [13] based on the novel notion of individual randomness of objects as expressed by P. Martin-Löf's randomness tests [22]. We show that the principle is valid when a basic condition encapsulated as the "Fundamental Inequality" (12) in Section 2 is satisfied. Broadly speaking this happens when the data are random, relative to each contemplated hypothesis, and also these hypotheses are random relative to the contemplated prior. The latter requirement is always satisfied for the so-called "universal" prior. Under those conditions ideal MDL, Bayesianism, MDL, and MML, select pretty much the same hypothesis. Theorem 6 states that minimum description length reasoning using shortest effective descriptions coincides with Bayesian reasoning using the universal prior distribution [19, 12, 9], provided the minimum description length is achieved for those hypotheses with respect to which the data sample is individually random (in the sense of Martin-Löf). If we restrict the model class to finite sets then this procedure specializes to Kolmogorov's minimal sufficient statistics, [8, 21].

Kolmogorov complexity We recapitulate the basic definitions in Appendix A in order to establish notations. Shortest effective descriptions are "effective" in the sense that we can compute the described objects from them. Unfortunately, [15, 41], there is no general method to compute the length of a shortest description (the Kolmogorov complexity) from the object being described. This obviously impedes actual use. Instead, one needs to consider recursive 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 $l_x := -\log P(x)$ 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.

Universal probability distribution Just as the Kolmogorov complexity measures the shortest effective description length of an object, the universal probability measures the greatest effective probability. Both notions are objective and absolute in the sense of being recursively invariant by Church's thesis, [21]. We give definitions in Appendix B. We use universal probability as a universal prior in Bayes's rule to analyze ideal MDL.

Martin-Löf randomness The common meaning of a "random object" is an outcome of a random source. Such outcomes have expected properties but particular outcomes may or may not possess these expected properties. In contrast, we use the notion of randomness of individual objects. This elusive notion's long history goes back to the initial attempts by von Mises, [35], 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, [15], and Martin-Löf tests for randomness, [22], finally succeeded in formally expressing the novel notion of individual randomness in a correct manner, see [21]. Every individually random object possesses individually all effectively testable properties that are only expected for outcomes of the random source concerned. It will satisfy *all* effective tests for randomness— known and unknown alike. In Appendix C we recapitulate the basics.

Two-part codes The prefix-code of the shortest effective descriptions gives an expected code word length close to the entropy and also compresses the regular objects until all regularity is squeezed out. All shortest effective descriptions are completely random themselves, without any regularity whatsoever. The MDL 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 the hypothesis is precisely that. 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 intuition finds its basis in the Definitions 10 and 11 in Appendix A. However, in general there are many ways to make the division into meaningful information and remaining random information. In a painting the represented image, the brush strokes, or even finer detail can be the relevant information, depending on what we are interested in. What we require is a rigorous mathematical condition to force a sensible division of the information at hand in a meaningful part and a meaningless part. One way to do this in a restricted setting where the hypotheses are finite sets was suggested by Kolmogorov at a Tallin conference in 1973 and published in [16]. See [8, 21] and Section 2.1. Given data D, the goal is to identify the "most likely" finite set A of which D is a "typical" element. For this purpose we consider sets A such that $D \in A$ and we represent A by the *shortest* program A^* that computes the characteristic function of A. The *Kolmogorov minimal sufficient statistic* is the shortest A^*, say A_0^* associated with the set A_0, over all A containing D such that the two-part description consisting of A_0^* and $\log d(A_0)$ is as as short as the shortest *single* program that computes D without input. This definition is non-vacuous since there is a two-part code (based on hypothesis $A_D = \{D\}$) that is as concise as the shortest single code.

The shortest two-part code must be at least as long as the shortest one-part code. Therefore, the description of D given A_0^* cannot be significantly shorter than $\log d(A_0)$. By the theory of Martin-Löf randomness in Appendix C this means that D is a "typical" element of A. The ideal MDL principle expounded in this paper is essentially a generalization of the Kolmogorov minimal sufficient statistic.

Note that in general finding a minimal sufficient statistic is not recursive. Similarly, even computing the MDL optimum in a much more restricted class of models may run in computation difficulties since it involves finding an optimum in a large set of candidates. In some cases one can approximate this optimum, [36, 40].

Prediction The most likely single hypothesis does not necessarily give the prediction with the best expectation. For example, consider a situation where we are given a coin of unknown bias p of coming up "heads" which is either $p_1 = \frac{1}{3}$ or $p_2 = \frac{2}{3}$. Suppose we have determined that there is probability $\frac{2}{3}$ that $p = p_1$ and probability $\frac{1}{3}$ that $p = p_2$. Then the most likely hypothesis is: $p = p_1$ which predicts a next outcome "heads" as having probability $\frac{1}{3}$. Yet the expectation of throwing "heads" is given by the mixture:

$$\frac{2}{3}p_1 + \frac{1}{3}p_2 = \frac{4}{9}.$$

Thus, the fact that compression is good for hypothesis identification problems does not imply that compression is good for prediction. In Section 3 we analyze the relation between compression of the data sample and prediction in the very general setting of R. Solomonoff [30, 31]. We explain Solomonoff's prediction method using the universal distribution. We show that this method is not equivalent to the use of shortest descriptions. Nonetheless, we demonstrate that compression of descriptions almost always gives optimal prediction.

Scientific inference The philosopher D. Hume (1711–1776) argued [14] that true induction is impossible because we can only reach conclusions by using known data and methods. Therefore, the conclusion is logically already contained in the start configuration. Consequently, the only form of induction possible is deduction. Philosophers have tried to find a way out of this deterministic conundrum by appealing to probabilistic reasoning such as using Bayes's rule [2]. One problem with this is where the "prior probability" one uses has to come from. Unsatisfactory solutions have been proposed by philosophers like R. Carnap [5] and K. Popper [24].

Essentially, combining the ideas of Epicurus, Ockham, Bayes, and modern computability theory, Solomonoff [30, 31] has successfully invented a "perfect" theory of induction. It incorporates Epicurus's multiple explanations idea, [1], since no hypothesis that is still consistent with the data will be eliminated. It incorporates Ockham's simplest explanation idea since the hypotheses with low Kolmogorov complexity are more probable. The inductive reasoning is performed by means of the mathematically sound rule of Bayes.

Comparison with Related Work Kolmogorov's minimal sufficient statistics deals with hypothesis selection where the considered hypotheses are finite sets of bounded cardinality. Ideal MDL hypothesis selection generalizes this procedure to arbitrary settings. It is satisfying that our findings on ideal MDL confirm the validity of the "real" MDL principle which rests on the idea of stochastic complexity. The latter is defined in such a way that it represents the shortest code length only for almost all data samples (stochastically speaking the "typical" ones) for all models with real parameters in certain classes of probabilistic models except for a set of Lebesgue measure zero, [26, 10, 23]. Similar results concerning probability density estimation by MDL are given in [4]. These references consider probabilistic models and conditions. We believe that in many current situations the models are inherently non-probabilistic as, for example, in the transmission of compressed images over noisy channels, [32]. Our algorithmic analysis of ideal MDL is about such non-probabilistic model settings as well as probabilistic ones (provided they are recursive). The results are derived in a nonprobabilistic manner entirely different from the cited papers. It is remarkable that there is a close agreement between the real properly articulated MDL principle and our ideal one. The ideal MDL principle is valid in case the data is individually random with respect to the contemplated hypothesis and the latter is an individually random element of the contemplated prior. Individually random objects are in a rigorous formal sense "typical" objects in a probability ensemble and together they constitute allmost all such objects (all objects except for a set of Lebesgue measure zero in the continuous case). The nonprobabilistic expression of the range of validity of "ideal MDL." implies the probabilistic expressions of the range of validity of the "real MDL" principle.

Our results are more precise than the earlier probabilistic ones in that they explicitly identify the "excepted set of Lebesgue measure zero" for which the principle may not be valid as the set of "individually nonrandom elements." The principle selects models such that the presented data are individually random with respect to these models: if there is a true model and the data are not random with respect to it then the principle avoids this model. 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.

2 Ideal MDL

The idea of predicting sequences using shortest effective descriptions was first formulated by R. Solomonoff, [30]. He uses Bayes's formula equipped with a fixed "universal" prior distribution. In accordance with Occam's dictum, that distribution gives most weight to the explanation that compresses the data the most. This approach inspired Rissanen [25, 26] to formulate the MDL principle. The applied principle and the associated one-part code length (the stochastic complexity) have evolved; the latest and most complete survey is [3]. Unaware of Solomonoff's work Wallace and his co-authors [38, 39] formulated a related but somewhat different *Minimum Message Length (MML)* principle.

We focus only on the following central ideal version which we believe is the essence of the matter. Indeed, we do not even care about whether we deal with statistical or deterministic hypotheses.

DEFINITION 1 *Given a sample of data, and an effective enumeration of models,* ideal MDL *selects the model with the shortest effective description that 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 help of the model.*

Intuitively, a more complex hypothesis H 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 to be required.

One way derive the MDL approach is to start from *Bayes's rule* written as

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

If the hypotheses space $\mathcal{H}$ is countable and the hypotheses H are exhausive and mutually exclusive, then $\sum_{H \in \mathcal{H}} P(H) = 1$ $\Pr(D) = \sum_{H \in \mathcal{H}} \Pr(D|H)P(H)$. For clarity and because it is relevant for the sequel we distinguish notationally between the given *prior probability* "$P(\cdot)$" and the probabilities "$\Pr(\cdot)$" that are induced by $P(\cdot)$ and the hypotheses H. Bayes's rule maps input $(P(H), D)$ to output $\Pr(H|D)$—the *posterior* probability. For many model classes (Bernoulli processes, Markov chains), as the number n of data generated by a true model in the class increases the total inferred probability can be expected to concentrate on the 'true' hypothesis (with probability one for $n \to \infty$). That is, as n grows the weight of the factor $\Pr(D|H)/\Pr(D)$ dominates the influence of the prior $P(\cdot)$ for typical data—by the law of large numbers. The importance of Bayes's rule is that the inferred probability gives us as much information as possible about the possible hypotheses from only a small number of (typical) data and the prior probability.

In general we don't know the prior probabilities. The MDL approach in a sense replaces the unknown prior probability that depends on the phenomenon being investigated by a fixed probability that depends on the coding used to encode the hypotheses. In ideal MDL the fixed "universal" probability (Appendix B) is based on Kolmogorov complexity—the length of the shortest effective code (Appendix A).

In Bayes's rule we are 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:

$$\begin{aligned} -\log \Pr(H|D) &= -\log \Pr(D|H) - \log P(H) \\ &\quad + \log \Pr(D). \end{aligned}$$

Since the probability $\Pr(D)$ is constant under varying H, we want to find the H_0 that [1]

$$H_0 := \operatorname{minarg}_{H \in \mathcal{H}} \{-\log \Pr(D|H) - \log P(H)\}. \tag{2}$$

[1]Notation $x_0 = \operatorname{minarg}_x \{f(x)\}$ assigns to x_0 the argument that minimizes $f(x)$.

In MML as in [39] or MDL as in [26] one roughly interprets these negative logarithms of probabilities as the corresponding Shannon-Fano code word lengths. [2] But we can also use descriptions such that the lengths of the codewords equals the Kolmogorov complexities. Then the expected codeword length is close to the entropy, but additionally each object is compressed so that all its regularities are accounted for, [21]. The resulting code word is maximally random, that is, it has maximal Kolmogorov complexity. [3]

Under certain restrictions to be determined later, the probabilities involved in (2) can be substituted by the corresponding *universal probabilities* $\mathbf{m}(\cdot)$ (Appendix B):

$$\begin{aligned} \log P(H) &:= \log \mathbf{m}(H), \\ \log \Pr(D|H) &:= \log \mathbf{m}(D|H). \end{aligned} \tag{3}$$

According to the Coding Theorem 12 in Appendix B [19, 12, 9] we can substitute

$$\begin{aligned} -\log \mathbf{m}(H) &= K(H), \\ -\log \mathbf{m}(D|H) &= K(D|H), \end{aligned} \tag{4}$$

where $K(\cdot)$ is the prefix complexity of Appendix A. This way we replace the sum of (2) by the sum of the minimum lengths of effective self-delimiting programs that compute descriptions of H and $D|H$. The result is the code-independent, recursively invariant, absolute form of the MDL principle [20]:

DEFINITION 2 *Given an hypothesis class $\mathcal{H}$ and a data sample D, the* ideal MDL *principle selects the hypothesis*

$$H_0 := \text{minarg}_{H \in \mathcal{H}} \{K(D|H) + K(H)\}. \tag{5}$$

If there is more than one H that minimizes (5) then we break the tie by selecting the one of least complexity $K(H)$.

The key question of Bayesianism versus ideal MDL is: When is the substitution (3) valid? We show that in a simple setting were the hypotheses are finite sets the ideal MDL principle and Bayesianism using the universal prior $\mathbf{m}(x)$ coincide with each other and with the Kolmogorov minimal sufficient statistic. We generalize this to probabilistic hypothesis classes. In full generality however, ideal MDL and Bayesianism may diverge due to the distinction between the $-\log P(\cdot)$ (the Shannon-Fano code length) and the Kolmogorov complexity $K(\cdot)$ (the shortest effective code length). We establish the Fundamental Inequality defining the range of coincidence of the two principles.

From now on, we will denote by $\stackrel{+}{<}$ an inequality to within an additive constant, and by $\stackrel{+}{=}$ the situation when both $\stackrel{+}{<}$ and $\stackrel{+}{>}$ hold.

[2]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. For the Shannon-Fano code see Section 2.3.

[3]The relation between the Shannon-Fano code and Kolmogorov complexity is treated in Section 2.3. For clarity of treatment, we refer the reader to the Appendices or [21] 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. The bulk of the material is Appendix C on Martin-Löf's theory of randomness tests. In particular the explicit expressions of universal randomness tests for arbitrary recursive distributions is unpublished apart from [21] and partially in [13].

2.1 Kolmogorov Minimal Sufficient Statistic

Considering only hypotheses that are finite sets of binary strings of finite lengths the hypothesis selection principle known as "Kolmogorov's minimal sufficient statistic" [16] has a crisp formulation in terms of Kolmogorov complexity, Shen' [29], V'yugin [37], and Cover and Thomas [8]. We use prefix complexity instead of plain complexity as in [21]. For this restricted hypothesis class we show that the Kolmogorov minimal sufficient statistic is actually Bayesian hypothesis selection using the universal distribution $\mathbf{m}(\cdot)$ as prior distribution and it also coincides with the ideal MDL principle. Let k and δ be natural numbers. A binary string D representing a data sample is called (k, δ)-*stochastic* if there is a finite set $H \subseteq \{0,1\}^*$ and $D \in H$ such that

$$D \in H, \quad K(H) \leq k, \quad K(D|H) \geq \log d(H) - \delta.$$

The first inequality (with k not too large) means that H is sufficiently simple. The second inequality (with the randomness deficiency δ not too large) means that D is an undistinguished (typical) element of H. Indeed, if D had properties defining a very small subset H' of H, then these could be used to obtain a simple description of D by determining its ordinal number in H', which would require $\log d(H')$ bits, which is much less than $\log d(H)$.

Suppose we carry out some experiment of which the outcome can a priori be every binary string. Suppose the outcome is D. Knowing D, we want to recover the appropriate hypothesis H. It seems reasonable to require that first, H has a simple description, and second, that D would be a "typical" outcome of an experiment under hypothesis H—that is, D is maximally random with respect to H. Suppose that under the hypothesis H every outcome is a binary string of length n with $n/2$ 1's. This set has cardinality at most $\binom{n}{n/2} = \Theta(2^n/\sqrt{n})$. To describe an element $D \in H$ requires $\stackrel{+}{<} n - \frac{1}{2}\log n$ bits. To describe $H \subseteq \{0,1\}^n$ given n requires $O(1)$ bits (that is, k is small in (6) below). Conditioning everything on the length n, we have

$$K(D|n) \stackrel{+}{<} K(D|H,n) + K(H|n) \stackrel{+}{<} n - \frac{1}{2}\log n,$$

and for the overwhelming majority of the D's in H we have

$$K(D|n) \stackrel{+}{>} n - \frac{1}{2}\log n.$$

Such D's are called $(O(1), O(1))$-stochastic.

The *Kolmogorov structure function* $K_k(D|n)$ of $D \in \{0,1\}^n$ is defined in [8] as

$$K_k(D|n) = \min\{\log d(H) : D \in H, \quad K(H|n) \leq k\}.$$

For a given small constant c, let k_0 be the least k such that

$$K_k(D|n) + k \leq K(D|n) + c. \tag{6}$$

Let H_0 be the corresponding set, and let H_0^* be its shortest program. This k_0 with $K(H_0|n) \leq k_0$ is the least k for which the two-part description of D is as parsimonious as the best single part description of D.

For this approach to be meaningful we need to show that there always exists a k satisfying (6). For example, consider the hypothesis $H_D := \{D\}$. Then, $\log d(H_D) = 0$ and $K(H_D|n) \stackrel{\pm}{=} K(D|n)$ which shows that setting $k :\stackrel{\pm}{=} K(D|n)$ satisfies (6) since $K_k(D|n) \stackrel{\pm}{=} 0$.

We can view this as selecting H's until we found one, say H_0, such that D is maximally complex in it—rather similar to aiming a magnifying glass at an object choosing different positions until the object is sharply in focus. The optimal position H_0 represents all non-accidental structure in D. The $K_{k_0}(D|n)$-part of the description just provides an index for x in H_0—essentially the description of the randomness or accidental structure of the string.

DEFINITION 3 *Let* $\mathcal{H} := \{H : H \subseteq \{0,1\}^n\}$ *and* $D \in \{0,1\}^n$. *Define*

$$H_0 := \text{minarg}_{H \in \mathcal{H}} \{K(H|n) : K(H|n) + \log d(H) \stackrel{+}{=} K(D|n)\}. \tag{7}$$

The shortest program H_0^* *that prints out the characteristic sequence of* $H_0 \in \{0,1\}^n$ *is called the* Kolmogorov minimal sufficient statistic (KMSS) *for* D, *given* n.

All programs describing sets H with $K(H|n) \leq k_0$ such that $K_{k_0}(D|n) + k_0 \stackrel{+}{=} K(D|n)$ are sufficient statistics. But the "minimal" sufficient statistic is induced by the set H_0 having the shortest description among them. [4] The Kolmogorov minimal sufficient statistic is related to Bayesian inference and ideal MDL.

THEOREM 1 *Let* n *be a large enough positive integer. Consider the hypotheses class* $\mathcal{H} := \{H : H \subseteq \{0,1\}^n\}$ *and a data sample* $D \in \{0,1\}^n$. *All of the following principles select the same hypothesis:*

(i) *Bayes's rule to select the least complexity hypothesis among the hypotheses of maximal a posterior probability using both (a) the universal distribution* $\mathbf{m}(\cdot)$ *as prior distribution, and (b)* $\Pr(D|H)$ *is the uniform probability* $1/d(H)$ *for* $D \in H$ *and 0 otherwise;*

(ii) *Kolmogorov minimal sufficient statistic; and*

(iii) *ideal MDL.*

PROOF. (i) $\leftrightarrow$ (ii). Substitute probabilities as in the statement of the theorem in (2).

(ii) $\leftrightarrow$ (iii). Let H_0 be the Kolmogorov minimal sufficient statistic for D so that (7) holds. It is straightforward that $K(H_0|n) + K(D|H_0, n) \stackrel{+}{>} K(D|n)$, and $K(D|H_0, n) \stackrel{+}{<} \log d(H_0)$ because we can describe D by its index in the set H_0. Altogether it follows that $K(D|H_0, n) \stackrel{+}{=} \log d(H_0)$ and

$$K(H_0|n) + K(D|H_0, n) \stackrel{+}{=} K(D|n).$$

Since $K(H|n) + K(D|H, n) \stackrel{+}{>} K(D|n)$ for all $H \in \mathcal{H}$, if

$$\mathcal{H}_0 = \{H' : H' = \text{minarg}_{H \in \mathcal{H}} \{K(H|n) + K(D|H, n)\}\}$$

then

$$H_0 = \text{minarg}_H \{K(H) : H \in \mathcal{H}_0\},$$

which is what we had to prove. $\square$

EXAMPLE 1 Let us look at a coin toss example. If the probability p of tossing "1" is unknown, then we can give a two-part description of a string D representing the sequence of n outcomes by describing the number k of 1's in D first, followed by the index $j \leq d(H)$ of D in in the set H of strings with k 1's. In this way "$k|n$" functions as the model. If k is incompressible with $K(k|n) \stackrel{+}{=} \log n$ and $K(j|k, n) \stackrel{+}{=} \log \binom{n}{k}$ then the Kolmogorov minimal sufficient statistic is described by $k|n$ in $\log n$ bits. However if p is a simple value like $\frac{1}{2}$ (or $1/\pi$), then with overwhelming probability we obtain a much simpler Kolmogorov minimal sufficient characteristic by a description of $p = \frac{1}{2}$ and $k = \frac{n}{2} + O(\sqrt{n})$ so that $K(k|n) \stackrel{+}{<} \frac{1}{2} \log n$. $\diamond$

[4] Explicit forms of KMSS will be given in a future paper [33].

2.2 Probabilistic Generalization of KMSS

Comparison of (2), (7), and Theorem 1 suggests a more general probabilistic version of Kolmogorov minimal sufficient statistic proposed by V'yugin [37]. This version turns out to coincide with maximum a posteriori Bayesian inference. Our analysis comparing it with ideal MDL roceeds more smoothly if we restrict consideration to hypotheses classes satsfying some mild conditions.

DEFINITION 4 *We call a hypothesis class* $\mathcal{H}$ complete *if it is an enumerable class of probabilistic hypotheses and* $\mathcal{D}$ *is an enumerable domain of data samples such that for every* $H \in \mathcal{H}$ *the probability density function* $\Pr(\cdot|H)$ *over the domain of data samples is recursive. We require additionally that* $\mathcal{H}$ *contains a trivial hypothesis* $H_\emptyset$ *satisfying* $K(H_\emptyset) \stackrel{\pm}{=} 0$ *and also for every data sample* D *in the domain there is an* $H_D \in \mathcal{H}$ *such that* $\Pr(D|H_D) = 1$ *and* $K(D|H_D) \stackrel{\pm}{=} 0$ *(the hypothesis forces the data sample).*

Let $\mathcal{H}$ be a complete hypothesis class and let

$$H_0 := \operatorname{minarg}_{H \in \mathcal{H}} \{K(H) : K(H) - \log \Pr(D|H) \stackrel{\pm}{=} K(D)\}. \tag{8}$$

The hypothesis H_0—rather the shortest program H_0^* that prints out the characteristic sequence of $H_0 \in \{0,1\}^n$—is called the *generalized Kolmogorov minimal sufficient statistic (GKMSS)* for D. Since $\mathcal{H}$ is complete the GKMSS exists. [5] [6]

THEOREM 2 *The least complexity maximum a posteriori probability hypothesis* H_0 *in Bayes's rule using prior* $P(H) := \mathbf{m}(H)$ *coincides with the generalized Kolmogorov minimal sufficient statistic.*

PROOF. Substitute $P(x) := \mathbf{m}(x)$ in (2). Using (4) the least complexity hypothesis satisfying the optimization problem is:

$$H_0 := \operatorname{minarg}_{H'} \{K(H') : H' := \operatorname{minarg}_{H \in \mathcal{H}} \{K(H) - \log \Pr(D|H)\}\}. \tag{9}$$

By assumptions in Definition 4 there is an H_D such that $K(H_D) - \log \Pr(D|H_D) \stackrel{\pm}{=} K(D)$. It remains to show that $K(H) - \log \Pr(D|H) \stackrel{+}{>} K(D)$ for all H, D.

It is straightforward that $K(H) + K(D|H) \stackrel{+}{>} K(D)$. For recursive $\Pr(\cdot|\cdot)$ it holds that $l_D \stackrel{\pm}{=} -\log \Pr(D|H)$ is the code length of the effective Shannon-Fano prefix code (see Section 2.3 or [8, 21]) to recover D given H. Since the prefix complexity is the length of the shortest effective prefix code we have $-\log \Pr(D|H) \stackrel{+}{>} K(D|H)$. □

2.3 Shannon-Fano Code Versus Shortest Programs

It is well-known that there is a tight connection between prefix codes, probabilities, and notions of optimal codes. A key distinction between Bayesianism and ideal MDL lies in the distinction between of codes that are optimal on average and codes that are optimal in every individual case.

[5] The equivalent hypothesis for a data sample D in the setting of the Kolmogorov minimal sufficient statistic was $H_D = \{D\}$.

[6] The Kolmogorov minimal sufficient statistic of Section 2.1 is the special case of the generalized version for hypotheses that are finite sets and with "$\Pr(D|H)$" is the uniform probability "$1/d(H)$".

The Shannon-Fano prefix code [8] for an ensemble of source words with probability density q has code word length $l_q(x) := -\log q(x)$ (up to rounding) for source word x. This code satisfies

$$H(q) \leq \sum_x q(x) l_q(x) \leq H(q) + 1$$

where $H(q)$ is the entropy of q. By the Noiseless Coding Theorem this is the least expected code word length among all prefix codes. Therefore, the hypothesis H which minimizes (2) written as

$$l_{\Pr(\cdot|H)}(D) + l_P(H))$$

minimizes the sum of lengths of two prefix codes that both have shortest *expected* code-word lengths. This is more or less what MML [39] and MDL [26] do.

But there are many prefix codes that have expected code word length *almost* equal to the entropy. Consider only the class of prefix codes that can be decoded by Turing machines (other codes are not practical). There is an optimal code in that class with code word length $K(x)$ for object x. "Optimality" means that for every prefix code in the class there is a constant c such that for all x the length of the code for x is at least $K(x) - c$, see Appendix A or [21].

In ideal MDL we minimize the sum of the effective description lengths of the *individual* elements H, D involved as in (5). This is validated by Bayes's rule provided (3) holds. To satisfy one part of (3) we are free to make the new *assumption* that the prior probability $P(\cdot)$ in Bayes's rule (1) is fixed as $\mathbf{m}(\cdot)$. However, with respect to the other part of (3) we *cannot assume* that the probability $\Pr(\cdot|H)$ equals $\mathbf{m}(\cdot|H)$. The reason is that the probability $\Pr(\cdot|H)$ may be totally determined by the hypothesis H. Depending on H therefore, $l_{\Pr(\cdot|H)}(D)$ 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 2 Suppose we flip a coin of unknown bias n times. Let hypothesis H and data D be defined by:

$$\begin{aligned} H &:= [\text{ Probability 'head' is } \tfrac{1}{2}] \\ D &:= \underbrace{hh\ldots h}_{n \text{ times '}h\text{'(ead)s}} \end{aligned}$$

Then we have $\Pr(D|H) = 1/2^n$ and

$$l_{\Pr(\cdot|H)}(D) = -\log \Pr(D|H) = n.$$

In contrast,

$$K(D|H) \stackrel{+}{<} \log n + 2 \log \log n.$$

◇

2.4 Individually Random Strings

The question arises: when is $-\log P(x) \stackrel{\pm}{=} K(x)$? This is answered by the theory of individual randomness. Consider a probabilistic ensemble consisting of the set $\{0,1\}^*$ endowed with a recursive probability density function $P : \{0,1\}^* \to [0,1]$. [7] [8] By Theorem 15 (Appendix C) an element x is

[7]A real-valued function is *recursive* if there is a Turing machine that for every argument and precision parameter b computes the function value within precision 2^{-b} and halts.

[8]We can identify $\mathcal{N}$ (the natural numbers) and $\{0,1\}^*$ as in Appendix A.

Martin-Löf random with randomness deficiency k iff the universal test $\kappa_0(x|P) = \log(\mathbf{m}(x)/P(x)) = k$. [9]

Definition 5 *Let $P : \{0,1\}^* \to [0,1]$ be a recursive probability density distribution. Then, the prefix complexity $K(P)$ of P is defined as 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: it is the shortest effective self-delimiting description of P, (Appendix A).*

Definition 6 *Let $P : \{0,1\}^* \to [0,1]$ be a recursive probability density function. An object $x \in \{0,1\}^*$ is* Martin-Löf random *with respect to P if the randomness deficiency $\kappa_0(x|P) = \log(\mathbf{m}(x)/P(x)) \stackrel{+}{<} K(P)$.*

If $P(x)$ is a recursive distribution then we can consider the sequence of recursive conditional probability distributions $P_n(x) := P(x|l(x) = n)$ for $n = 1, 2, \ldots$.

Lemma 1 *Let $R(n)$ denote the probability that a binary string of length n is a Martin-Löf random element of the distribution $P_n(\cdot)$ as in Definition 6. Then $R(n) = 1 - O(1/2^{K(P,n)})$. The probability $R(n)$ goes to 1 for $n \to \infty$. Moreover, $\limsup_{n\to\infty} R(n) = 1 - O(1/n)$.*

Proof. Let $P : \{0,1\}^* \to [0,1]$ be a recursive probability density distribution. The conditional probability on binary strings of length n is $P_n(x)$ as above. We want to estimate the probability

$$P_n\{x : \log \frac{\mathbf{m}(x)}{P_n(x)} \stackrel{+}{<} K(P_n(\cdot))\}. \tag{10}$$

We can describe the distribution $P_n(\cdot)$ by $P(\cdot)$ and n. This shows $K(P_n(\cdot) \stackrel{+}{<} K(P,n)$. Conversely, given the recursive distribution $P_n(\cdot)$ as a Turing machine computing the probability values for all x's we can compute n by simply computing the probabilities for all arguments x in lexicographical nondecreasing order. Since $\sum_{l(x)=n} P_n(x) = 1$ there is an x with probability $P_n(x) \geq 1/2^n$. The first x in the enumeration that satisfies $P(x) > 0$ determines $n := l(x)$. Since P is recursive its description takes $O(1)$ bits. Hence $K(P_n(\cdot)) \stackrel{+}{>} K(P,n)$ and therefore $K(P_n(\cdot)) \stackrel{+}{=} K(P,n)$. By the Coding Theorem 12 in Appendix B we have $K(x) \stackrel{+}{=} -\log \mathbf{m}(x)$. Altogether we can rewrite (10) in exponential form as:

$$A := P_n\{x : \frac{\mathbf{m}(x)}{P_n(x)} = O(2^{K(P,n)})\}. \tag{11}$$

Since the expectation

$$\sum_x P_n(x) \frac{\mathbf{m}(x)}{P_n(x)} \leq 1$$

we find by Markov's Inequality ((26) in Appendix C) that

$$1 - A = O\left(\frac{1}{2^{K(P,n)}}\right).$$

This shows the first part of the theorem. Since $K(P,n) \stackrel{+}{>} K(n)$ and $K(n) \to \infty$ for $n \to \infty$ [21] the second part of the theorem follows immediately. Since also $\limsup_{n\to\infty} K(n) > \log n$ (in fact, $K(n) > \log n + \log\log n$ for infinitely many n [21]) the third part of the theorem follows as well. □

[9] This means that *every* x is random (has randomness deficiency 0) with respect to the universal distribution $\mathbf{m}(x)$ (substitute $P(x) := \mathbf{m}(x)$ above).

2.5 The Fundamental Inequality

In ideal MDL as applied to a complete hypothesis class $\mathcal{H}$ ("complete" in the sense of Definition 4) there are two boundary cases: Using the trivial hypothesis $H_\emptyset$ with $K(H_\emptyset) \stackrel{\pm}{=} 0$ we always obtain $K(D|H_\emptyset) \stackrel{\pm}{=} K(D)$ and therefore $K(H_\emptyset) + K(D|H_\emptyset) \stackrel{\pm}{=} K(D)$. The hypothesis H_D of Definition 4 also yields $K(H_D) + K(D|H_D) \stackrel{\pm}{=} K(D)$. Since always $K(H) + K(D|H) \stackrel{+}{>} K(D)$ both of these hypotheses minimize the ideal MDL description.

But for trivial hypotheses only Kolmogorov random data are typical. In fact, ideal MDL correctly selects the trivial hypothesis for individually random data. But in general "meaningful" data are "nonrandom" in the sense that $K(D) \ll l(D)$. But then D is typical only for nontrivial hypotheses, and a trivial hypothesis selected by ideal MDL is not one for which the data are typical. We need to identify the conditions under which ideal MDL restricts itself to selection among hypotheses for which the given data are typical—it performs as the generalized Kolmogorov minimal sufficient statistic.

Note that hypotheses satisfying (8) may not always exist if we don't require that every data sample in the domain is forced by some hypothesis in the hypothesis space we consider, as we did in Definition 4.

EXAMPLE 3 We look at a situation where the three optimization principles (2), (5), (8) act differently. Again, consider the outcome of n coin flips of a coin with probability p of flipping '1' and probability $1-p$ of flipping '0'. There are two possible hypotheses $\mathcal{H} = \{H_0, H_1\}$ where

$$H_0 = [p = \frac{1}{2}]$$
$$H_1 = [p = 0]$$

The prior P is $P(H_0) = \frac{1}{2}$ and $P(H_1) = \frac{1}{2}$. Consider the data sample $D = 0^n$ with n Kolmogorov random (also with respect to H_0 and H_1) so that $\log n \leq K(D), K(D|H_0), K(D|H_1) \leq \log n + 2\log\log n$. Now

$$-\log P(H_0) - \log \Pr(D|H_0) \stackrel{\pm}{=} n$$
$$-\log P(H_1) - \log \Pr(D|H_1) \stackrel{\pm}{=} 0.$$

Therefore Bayesianism selects H_1 which is intuitively correct. Both hypotheses have complexity $\stackrel{\pm}{=} 0$. Therefore we can substitute $-\log P(H) := K(H)$ to obtain

$$K(H_0) - \log \Pr(D|H_0) \stackrel{\pm}{=} n$$
$$K(H_1) - \log \Pr(D|H_1) \stackrel{\pm}{=} 0.$$

Now generalized Kolmogorov minimal statistic doesn't select any hypothesis at all because the right-hand side is unequal $K(D)$. Ideal MDL on the other hand has the ex equo choice

$$K(H_0) + K(D|H_0) = \log n + O(\log\log n)$$
$$K(H_1) + K(D|H_1) = \log n + O(\log\log n),$$

which intuitively seems incorrect. So we need to identify the conditions under which ideal MDL draws correct conclusions. $\diamond$

While we can set the prior $P(\cdot) := \mathbf{m}(\cdot)$ in Bayes's rule to obtain the generalized Kolmogorov minimal sufficient statistic for a complete hypothesis class, we cannot in general also set $-\log \Pr(D|H_0) := K(D|H_0)$ to obtain ideal MDL. For this we need the theory dealing with randomness of individual objects that states the conditions for $-\log \Pr(D|H)$ and $K(D|H)$ to be close. [10]

DEFINITION 7 (FUNDAMENTAL INEQUALITY) *Let* $\Pr(\cdot|\cdot)$ *and* $P(\cdot)$ *be recursive probability density functions. The* Fundamental Inequality *is the following relation that possibly holds between data* D *and hypothesis* H*:*

$$|K(D|H) + K(H) + \log \Pr(D|H) + \log P(H)| \leq \alpha(P, H), \tag{12}$$

with

$$\alpha(P, H) = K(\Pr(\cdot|H)) + K(P).$$

THEOREM 3 *Assume the terminology above. If both hypothesis* H *is* $P(\cdot)$*-random and data* D *are* $\Pr(\cdot|H)$*-random then the Fundamental Inequality is satisfied.*

PROOF. We first consider the data part: Because $\Pr(\cdot|H)$ is recursive we have $\mathbf{m}(D|H) \geq 2^{-K(\Pr(\cdot|H))} \Pr(D|H)$, (27). Therefore,

$$\log \frac{\mathbf{m}(D|H)}{\Pr(D|H)} \geq -K(\Pr(\cdot|H)). \tag{13}$$

Note that $K(\Pr(\cdot|H)) \stackrel{+}{<} K(H)$ because from H we can compute $\Pr(\cdot|H)$ by assumption on $\Pr(\cdot|\cdot)$. Secondly, D is a Martin-Löf random element of the distribution $\Pr(\cdot|H)$ iff

$$\log \frac{\mathbf{m}(D|H)}{\Pr(D|H)} \leq K(\Pr(\cdot|H)) \tag{14}$$

by Theorem 15 and Definition 6. If D is a Martin-Löf random element of the distribution $\Pr(\cdot|H)$ then by (13, 14, 3) we obtain:

$$|K(D|H) + \log \Pr(D|H)| \leq K(\Pr(\cdot|H)). \tag{15}$$

We now consider the hypothesis part: If we set the *a priori* probability $P(H)$ of hypothesis H to the universal probability then we obtain directly $-\log P(H) = \mathbf{m}(x)$. However, we do not need to make this assumption. For a recursive prior $P(\cdot)$, we can analyze the situation when H is random with respect to $P(\cdot)$.

If H is a Martin-Löf random element of $P(\cdot)$ then we obtain analogous to the proof of (15):

$$|K(H) + \log P(H)| \leq K(H). \tag{16}$$

Together (15, 16) yield the theorem. $\square$

THEOREM 4 *The probability that hypotheses of binary length* m *and data of binary length* n *satisfy the Fundamental Inequality goes to one for* m *and* n *grow unboundedly. Moreover, the* $\limsup$ *of that probability exceeds* $1 - O(1/\min\{m, n\})$.

[10]This follows from (27) in Appendix C.

PROOF. By Theorem 3 the Fundamental Inequality is satisfied for Martin-Löf random hypotheses H (with respect to prior P) together with Martin-Löf random data D (with respect to the density $\Pr(\cdot|H)$ of the corresponding H). The probability of satisfying the fundamental Inequality is estimated by picking a pair from the joint probability $P(D,H) = \Pr(D|H)P(H)$ by first picking a random m-bit hypothesis H from the prior and then picking a random n-bit data sample from the $\Pr(\cdot|H)$ distribution. By Lemma 1 the contemplated probability is at least

$$\left(1 - O\left(\frac{1}{2^{K(m)}}\right)\right)\left(1 - O\left(\frac{1}{2^{K(n)}}\right)\right).$$

Since the Kolmogorov complexity grows unboundedly with increasing arguments [21] this proves the first statement. The lim sup of the displayed expression exceeds $(1 - 1/m)(1 - 1/n)$ since $\limsup_{x\to\infty} K(x) > \log x + \log\log x$ [21]. $\square$

What does the fact that the Fundamental Inequality is satisfied say about the typicality (randomness) of the hypothesis and data concerned?

THEOREM 5 *Assume the above notation. If*

$$-\log \Pr(D|H) - \log P(H) \stackrel{+}{=} K(D|H) + K(H) \tag{17}$$

then H is $P(\cdot)$-random up to randomness deficiency $\stackrel{+}{=} K(\Pr(\cdot|H)) - K(P)$, and D is $\Pr(\cdot|H)$-random up to randomness deficiency $\stackrel{+}{<} K(P) - K(\Pr(\cdot|H))$. (Negative randomness deficiencies correspond to 0.)

PROOF. By Theorem 12 in Appendix B and (27) in Appendix C we know that always $-\log \Pr(D|H) \geq K(D|H) - K(\Pr(\cdot|H))$. Similarly, always $-\log P(H) \geq K(H) - K(P)$. Suppose D is *not* $\Pr(\cdot|H)$-random. Then, by (14) $-\log \Pr(D|H) - K(D|H) = K(\Pr(\cdot|H)) + \Delta$ for randomness deficiency $\Delta > 0$. By our assumption this implies $-\log P(H) - K(H) \stackrel{+}{=} -K(\Pr(\cdot|H)) - \Delta$ which is only possible if $\Delta \stackrel{+}{<} K(P) - K(\Pr(\cdot|H)$. The case that H is not P-random is handled similarly. $\square$

REMARK 1 The only way to violate the Fundamental Inequality is that either the data D are not $\Pr(\cdot|H)$-random and therefore $-\log \Pr(D|H) \gg K(D|H)$, or that H is not P-random and therefore $-\log P(H) \gg K(H)$. Hypothesis H is P-random means that H is "typical" for the prior distribution $P(\cdot)$ in the sense that it must not belong to any effective minority (sets on which a minority of P-probability is concentrated). That is, hypothesis H does not have any effectively testable properties that distinguish it from a majority. In the present case of *finite* hypotheses this statement must be less general and can be taken to mean that the hypothesis H does not have any *simple* effectively testable properties—here "simple" means properties that can be expressed by tests that have Kolmogorov complexity that is low in terms of the complexity of H. This matter is treated in some detail in Appendix C and more completely in [21]. The set of "typical" hypotheses H are those that satisfy $K(H) \stackrel{+}{=} -\log P(H)$ up to an additive term of $K(P)$. In case $P(H) = \mathbf{m}(H)$, that is, the prior distribution equals the universal distribution, then for *all* H we have $K(H) \stackrel{+}{=} -\log P(H)$, that is, *every* hypotheses is random with respect to the universal distribution.

For other prior distributions hypotheses can be random or nonrandom. For example, let the possible hypotheses correspond to the binary strings of length n, and let L_n be the uniform distribution that assigns probability $L_n(H) = 1/2^n$ to every hypothesis H. Let us assume that the hypotheses are coded as binary strings of length n, so that $H \in \{0,1\}^n$. Then, $H := 00\ldots0$ has low complexity: $K(H|n) \stackrel{+}{<} \log n$. However, $-\log L_n(H) = n$. Therefore, by (16), H is not L_n-random. If we obtain H by n flips of a fair coin however, then with overwhelming probability we will have that $K(H|n) \stackrel{\pm}{=} n$ and therefore $-\log L_n(H) \stackrel{\pm}{=} K(H|n)$ and H is L_n-random.

That data sample 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. This is illustrated easiest using the identification of a Bernoulli process $B_p = (p, 1-p)$ $(0 < p < 1)$ that generates a given data sample $D \in \{0,1\}^n$. Let $\Pr(D|B_p, n)$ denote the distribution of the outcome D of n trials of the process B_p. If the data D are "atypical" like $D = 00\ldots0$ (n failures) for $p = \frac{1}{2}$ and n large, then it violates the $\Pr(\cdot|B_{1/2}, n)$-randomness test (14) by having $-\log \Pr(D|B_{1/2}) = n$ and $-\log \mathbf{m}(D|B_{1/2}) \stackrel{\pm}{=} K(D|B_{1/2}) \stackrel{+}{<} \log n + 2\log\log n$. $\diamond$

2.6 Ideal MDL and Bayesianism

The best model or hypothesis to explain the data should a "typical" element of the prior distribution such that the data are "typical" for the contemplated hypothesis—as prescribed by Kolmogorov's minimum sufficient statistics. Thus, it is reasonable to admit only hypotheses as defined below in selecting the best one.

DEFINITION 8 *Given data sample D and prior probability P, we call a hypothesis H* admissible *if H is P-random and D is $\Pr(\cdot|H)$-random.*

By Theorem 3 admissible hypotheses satisfy the Fundamental Inequality (12). By Theorem 4 admissible hypotheses have high probability.

THEOREM 6 *Let the data sample be D and let the corresponding set of admissible hypotheses be $\mathcal{H}_D \subseteq \mathcal{H}$. Then the maximum a posteriori probability hypothesis $H_{\text{bayes}} \in \mathcal{H}_D$ in Bayes's rule and the hypothesis $H_{\text{mdl}} \in \mathcal{H}_D$ selected by ideal MDL are roughly equal:*

$$
\begin{aligned}
&|\log \frac{\Pr(H_{\text{mdl}}|D)}{\Pr(H_{\text{bayes}}|D)}| \leq 2\alpha(P,H) \\
&|K(D|H_{\text{mdl}}) + K(H_{\text{mdl}}) - K(D|H_{\text{bayes}}) - K(H_{\text{bayes}})| \leq 2\alpha(P,H).
\end{aligned}
\tag{18}
$$

PROOF. Assume the Fundamental Inequality (12). By definition $H = H_{\text{mdl}}$ minimizes $K(H|D) + K(H)$. Denote the minimal value by A. Then in (12) the H' that minimizes $-\log \Pr(D|H) - \log P(H)$ yields a value B with $|A - B| \leq \alpha(P,H)$. This is easy to see since if $A - B > \alpha(P,H)$ then $K(H'|D) + K(H') < A$ by (12) contradicting that A is the minimum for the sum of the complexities, and if $B - A > \alpha(P,H)$ then $-\log \Pr(D|H_{\text{mdl}}) - \log P(H_{\text{mdl}}) < B$ contradicting that B is the minimum for the sum of the negative log probabilities. Now $H = H_{\text{bayes}}$ maximizes $\Pr(H|D) = \Pr(D|H)P(H)/\Pr(D)$ with $\Pr(D)$ constant and therefore H_{bayes} is a H' that minimizes $-\log \Pr(D|H) - \log P(H)$. Denote $-\log \Pr(D|H_{\text{mdl}}) - \log P(H_{\text{mdl}})$ by B'. Then by (12) we have $|A - B'| \leq \alpha(P,H)$ and therefore

$$|B - B'| \leq 2\alpha(P,H). \tag{19}$$

By Bayes's rule $B + \log \Pr(D) = -\log \Pr(H_{\mathrm{bayes}}|D)$ and $B' + \log \Pr(D) = -\log \Pr(H_{\mathrm{mdl}}|D)$ where $\log \Pr(D)$ is constant. Substitution in (19) yields the first inequality of the theorem. The second inequality follows by the same argument with the roles of complexities and negative log probablities interchanged. $\square$

REMARK 2 Let us first interpret what the theorem says: If in the Fundamental Inequality $\alpha(P, H)$ is small then this 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 is simple. The theorem states that the hypothesis selected by MDL and by Bayesianism are close both in the sense of a postiori probability (the Bayesian criterion) and in sum of the complexities of the hypothesis and the data encoded with help of the hypothesis (the MDL criterion). 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 the discrepancy can widen for both criteria. $\diamond$

As a consequence, if $\alpha(P, H)$ is small enough and Bayes's rule selects an admissible hypothesis, and so does ideal MDL, then both criteria are (approximately) optimized by both selected hypotheses. Is this a likely situation?

THEOREM 7 *The probability that for data of binary length n hypotheses of binary length m selected by Bayesian maximum a posteriori and minimum description length principle, respectively, are close in the sense of satisfying relations* (18) *of Theorem 6 goes to one for m and n grow unboundedly. Moreover, the* $\limsup$ *of that probability exceeds* $1 - O(1/\min\{m, n\})$.

PROOF. By Theorems 4 and 6. $\square$

The choice of $\mathbf{m}(\cdot)$ as prior fulfills part of the requirements for the Fundamental Inequality (and the above theorem). This prior is an objective and recursively invariant form of Occam's razor: a simple hypothesis H (with $K(H) \ll l(H)$ has high $\mathbf{m}$-probability, and a complex or random hypothesis H (with $K(H) \approx l(H)$) has low $\mathbf{m}$-probability $2^{-l(H)}$. The randomness test $\log(\mathbf{m}(H)/P(H))$ evaluates to 0 for *every* H, which means that all hypotheses are random with respect to distribution $\mathbf{m}(\cdot)$. [11]

THEOREM 8 *Let $\alpha(P, H)$ in the FI (12) be small (for example $\alpha \stackrel{\pm}{=} 0$) and prior $P(\cdot) := \mathbf{m}(\cdot)$. Then the Fundamental Inequality (12) is satisfied iff data sample D is $\Pr(\cdot|H_{\mathrm{mdl}})$-random. This has probability going to one for the binary length n of the data rising unboundedly (and the* $\limsup$ *of the probability exceeds* $1 - O(1/n)$*).*

PROOF. With $\alpha(P, H) \stackrel{\pm}{=} 0$ and $P(\cdot) := \mathbf{m}(\cdot)$ (so $-\log P(H) = K(H)$ by the Coding Theorem 12 in Appendix B. we can rewrite (12) as

$$-\log \Pr(D|H) \stackrel{\pm}{=} K(D|H).$$

This defines the admissible hypotheses for data D. By Definition 6 D is $\Pr(\cdot|H)$-random for these admissible hypotheses. In particular the $H := H_{\mathrm{mdl}}$ minimizing $K(D|H) + K(H)$ is admissible iff D is $\Pr(\cdot|H)$-random. This happens with probability going to one by Lemma 1. $\square$

[11] The choice of $\mathbf{m}$ as prior agrees with the preceding analysis even though it is not recursive. This is because the randomness test $\theta_0(H|\mathbf{m})$ is identically 0.

Corollary 1 *Ideal MDL is an application of Bayes's rule with the universal prior distribution* $\mathbf{m}(\cdot)$ *and selection of an optimal admissible hypothesis* H_{mdl} *(that is, the data sample* D *is* $\Pr(\cdot|H_{\mathrm{mdl}})$*-random) with probability going to one for increasing data.*

Since the notion of individual randomness incorporates all effectively testable properties of randomness (in the finite case only to some degree), application of ideal MDL will select the simplest hypothesis H that balances $K(D|H)$ and $K(H)$ such that the data sample D is random to it—as far as can effectively be ascertained. Restricted to the class of admissible hypotheses, ideal MDL does not simply select the hypothesis that precisely fits the data but it selects an hypothesis that would typically generate the data.

2.7 Applications

Unfortunately, the function K is not computable, [21]. For practical applications one must settle for easily computable approximations, for example, restricted model classes and particular codings of hypotheses. In this paper we will not address the question which encoding one uses in practice, but refer to references [26, 38, 40, 36].

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 (descriptional) complexity of θ. (For example, H can be a normal distribution $N(\mu, \sigma)$ described by $\theta = (\mu, \sigma)$.) Each parameter θ_i is truncated to fixed finite precision. The data sample consists of n outcomes $\mathbf{y} = (y_1, \ldots, x_n)$ of n trials $\mathbf{x} = (x_1, \ldots, x_n)$ for distribution $P(\theta)$. The data D in the above formulas is given as $D = (\mathbf{x}, \mathbf{y})$. By expansion of conditional probabilities we have therefore

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

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

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

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 $\mathbf{x}$ is $\Pr(\cdot|H)$-random and $\mathbf{y}$ is $\Pr(\cdot|H, \mathbf{x})$-random. Bayesian reasoning selects the same hypothesis provided the hypothesis with maximal inferred probability has $\mathbf{x}, \mathbf{y}$ satisfy the same conditions. Illustrative examples related to learning polynomials are given by Rissanen [26].

Remark 3 (Exception-Based MDL) A hypothesis H minimizing $K(D|H) + K(H)$ always satisfies

$$K(D|H) + K(H) \geq K(D).$$

Let $E \subseteq D$ denote the subset of the data that are *exceptions* to H in the sense of not being classified correctly by H. The following *exception-based* MDL (E-MDL) is sometimes confused with MDL: With $E := D - D_H$ and D_H is the data set classified according to H, select

$$H_{\text{e-mdl}} = \mathrm{minarg}_{H'}\{K(H') : H' := \mathrm{minarg}_H\{K(H) + K(E|H)\}\}.$$

In E-MDL 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 the data sample consists of positive examples only. For example, in learning (a grammar for) English language, given the *Oxford English Dictionary*. According to E-MDL the best hypothesis is the trivial grammar H generating *all* sentences over the alphabet. This grammar gives $K(H) \stackrel{+}{=} 0$ independent of D and also $E := \emptyset$. Consequently,

$$\min_H \{K(H) + K(E|H)\} = K(H) \stackrel{+}{=} 0,$$

which is absurd. The E-MDL principle is vindicated and reduces to standard MDL in the context of interpreting $D = (\mathbf{x}, \mathbf{y})$ with $\mathbf{x}$ fixed as in "supervised learning." Now for constant $K(\mathbf{x}|H)$

$$H_{\text{e-mdl}} = \text{minarg}_{H'} \{K(H') : H' := \text{minarg}_H \{K(H) + K(\mathbf{y}|H, \mathbf{x}) + K(\mathbf{x}|H)\}\}$$

is the same as

$$H_{\text{mdl}} = \text{minarg}_{H'} \{K(H') : H' := \text{minarg}_H \{K(H) + K(\mathbf{y}|H, \mathbf{x})\}\}.$$

Ignoring the constant $\mathbf{x}$ in the conditional $K(\mathbf{y}|H, \mathbf{x})$ corresponds to $K(E|H)$. $\diamond$

3 Prediction by Minimum Description Length

Let us consider theory formation in science as the process of obtaining a compact description of past observations together with predictions of future ones. Ray Solomonoff [30, 31] argues that the preliminary data of the investigator, the hypotheses he proposes, the experimental setup he designs, the trials he performs, the outcomes he obtains, the new hypotheses he formulates, and so on, can all be encoded as the initial segment of a potentially infinite binary sequence. The investigator obtains increasingly longer initial segments of an infinite binary sequence ω by performing more and more experiments on some aspect of nature. To describe the underlying regularity of ω, the investigator tries to formulate a theory that governs ω on the basis of the outcome of past experiments. Candidate theories (hypotheses) are identified with computer programs that compute binary sequences starting with the observed initial segment.

There are many different possible infinite sequences (histories) on which the investigator can embark. The phenomenon he wants to understand or the strategy he uses can be stochastic. Each such sequence corresponds to one never-ending sequential history of conjectures and refutations and confirmations and each initial segment has different continuations governed by certain probabilities. In this view each phenomenon can be identified with a measure μ on the continuous sample space of infinite sequences over a basic description alphabet. This distribution μ can be said to be the concept or phenomenon involved. Now the aim is to *predict* outcomes concerning a phenomenon μ 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 $\mathcal{B}$ defined by $S = \mathcal{B}^\infty$. We assume a prior distribution μ over S with $\mu(x)$ denoting the probability of a sequence starting with x. Here $\mu(\cdot)$ is a *semimeasure*[12] satisfying

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

[12]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. μ is a *measure* if equalities hold.

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 (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 (1) we substitute $\mu(xy)$ and $\mu(x)$, respectively. For $\Pr(H_{xy}|D_x)$ we substitute $\mu(y|x)$. This way (1) is rewritten as

$$\mu(y|x) = \frac{\mu(xy)}{\mu(x)}. \tag{20}$$

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 μ 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. The central task of inductive inference is to find a universally valid approximation to μ 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 μ to the *recursive* semimeasures[13] and restrict the set of basic elements to $\{0,1\}$. Under this relatively mild restriction on the admissible semimeasures μ, it turns out that we can use the *universal semimeasure* $\mathbf{M}$ as a "universal prior" (replacing the real prior μ) for prediction. The theory of the universal semimeasure $\mathbf{M}$, the analogue in the sample space $\{0,1\}^\infty$ of $\mathbf{m}$ in the sample space $\{0,1\}^*$ equivalent to $\mathcal{N}$, is developed in [21], Chapter 4, and Chapter 5. It is defined with respect to a special type Turing machine called *monotone* Turing machine. The universal semimeasure $\mathbf{M}$ multiplicatively dominates all enumerable (Definition 12, Appendix B) semimeasures. It can be shown that 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 $\mathbf{M}(x)$, [41].

The universal probability $\mathbf{M}(\cdot)$ allows us to explicitly express a universal randomness test for the elements in $\{0,1\}^\infty$ analogous to the universal randomness tests for the finite elements of $\{0,1\}^*$ developed in Appendix C. This notion of randomness with respect to a recursive semimeasure μ satisfies the following explicit characterization of a universal (sequential) randomness test (for proof see [21], Chapter 4):

LEMMA 2 *Let μ be a recursive semimeasure on $\{0,1\}^\infty$. An infinite binary sequence ω is μ-random iff*

$$\sup_n \mathbf{M}(\omega_1 \ldots \omega_n)/\mu(\omega_1 \ldots \omega_n) < \infty,$$

and the set of μ-random sequences has μ-measure one,

In contrast with the discrete case, the elements of $\{0,1\}^\infty$ can be sharply divided into the random ones that pass *all* effective (sequential) randomness tests and the nonrandom ones that do not.

[13] there is a Turing machine that for every x and b computes $\mu(x)$ within precision 2^{-b}.

We start by demonstrating convergence of $\mathbf{M}(y|x)$ and $\mu(y|x)$ for $x \to \infty$, with μ-probability 1. [14]

THEOREM 10 *Let μ be a positive recursive measure. If the length of y is fixed and the length of x grows to infinity, then*

$$\frac{\mathbf{M}(y|x)}{\mu(y|x)} \to 1,$$

with μ-probability one. The infinite sequences ω with prefixes x satisfying the displayed asymptotics are precisely the μ-random sequences.

PROOF. We use an approach based on the Submartingale Convergence Theorem, [11] 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 μ-submartingale, and the μ-expectation $\mathbf{E}|f(\omega_{1:n})| < \infty$, then it follows that $\lim_{n\to\infty} f(\omega_{1:n})$ exists with μ-probability one.

In our case,

$$t(\omega_{1:n}|\mu) = \frac{\mathbf{M}(\omega_{1:n})}{\mu(\omega_{1:n})}$$

is a μ-submartingale, and the μ-expectation $\mathbf{E}t(\omega_{1:n}|\mu) \leq 1$. Therefore, there is a set $A \subseteq \{0,1\}^\infty$ 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 μ-random ω's by Corollary 4.5.5 in [21]. Consequently, for fixed m, for each ω in A, we have

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

provided the limit of the denominator is not zero. The latter fact is guarantied by the universality of $\mathbf{M}$: for every $x \in \{0,1\}^*$ we have $\mathbf{M}(x)/\mu(x) \geq 2^{-K(\mu)}$ by Theorem 4.5.1 and Equation 4.11 in [21]. □

EXAMPLE 4 Suppose we are given an infinite decimal sequence ω. 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, $\mathbf{M}(a|\omega_{1:2i}) \to 1/10$ for $a = 0, 1, \ldots, 9$, while $\mathbf{M}(a|\omega_{1:2i+1}) \to 1$ if a is the ith digit of π, and to 0 otherwise. ◇

The universal distribution combines a weighted version of the predictions of all enumerable semimeasures, including the prediction of the semimeasure with the shortest program. It is not a priori

[14] We can express the "goodness" of predictions according to $\mathbf{M}$ with respect to a true μ as follows: Let S_n be the μ-expected value of the square of the difference in μ-probability and $\mathbf{M}$-probability of 0 occurring at the nth prediction

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

We may call S_n the *expected squared error at the nth prediction*. The following celebrated result of Solomonoff, [31], says that $\mathbf{M}$ is very suitable for prediction (a proof using Kulback-Leibler divergence is given in [21]):

THEOREM 9 *Let μ 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 $1/n$.)*

However, Solomonoff's result is not strong enough to give the required convergence of conditional probabilities with μ-probability 1.

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 use shortest programs for prediction.

Taking the negative logarithm on both sides of (20), we want to determine y with $l(y) = n$ that minimizes

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

This y is the most probable extrapolation of x.

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

$$Km(x) = \min\{l(p) : U(p) = x\omega, \omega \in \{0,1\}^{\infty}\}.$$

We omit the Invariance Theorem for Km complexity, stated and proven completely analogous to the Theorems with respect to the C and K varieties.

THEOREM 11 *Let μ be a recursive semimeasure, and let ω be a μ-random infinite binary sequence and xy be a finite prefix of ω. For $l(x)$ growing unboundedly and $l(y)$ fixed,*

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

where $Km(xy)$ and $Km(x)$ grow unboundedly.

PROOF. By definition, $-\log \mathbf{M}(x) \leq Km(x)$ since the left-hand side of the inequality weighs the probability of *all* programs that produce x while the right-hand side weighs the probability of the shortest program only. In the discrete case we have the Coding Theorem 12: $K(x) \stackrel{+}{=} -\log \mathbf{m}(x)$. L.A. Levin, [18] erroneously conjectured that also $Km(x) \stackrel{+}{=} -\log \mathbf{M}(x)$. But P. Gács [13] showed that they are different, although the differences must in some sense be very small:

CLAIM 1

$$\begin{aligned} -\log \mathbf{M}(x) \leq Km(x) &\stackrel{+}{<} -\log \mathbf{M}(x) + Km(l(x)); \qquad (21)\\ \sup_{x\in\{0,1\}^*} |KM(x) - Km(x)| &= \infty. \end{aligned}$$

However, for *a priori* almost all infinite sequences x, the difference between $Km(\cdot)$ and $-\log \mathbf{M}(\cdot)$ is bounded by a constant [13]:

CLAIM 2 *(i) For random strings $x \in \{0,1\}^*$ we have $Km(x) + \log \mathbf{M}(x) \stackrel{+}{=} 0$.*

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

Let ω be a μ-random infinite binary sequence and xy be a finite prefix of ω. For $l(x)$ grows unboundedly with $l(y)$ fixed, we have by Theorem 10:

$$\lim_{l(x)\to\infty} \log \mu(y|x) - \log \mathbf{M}(y|x) = 0. \qquad (22)$$

Therefore, if x and y satisfy above conditions, then maximizing $\mu(y|x)$ over y means minimizing $-\log \mathbf{M}(y|x)$. It is shown in Claim 2 that $-\log \mathbf{M}(x)$ 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, Claim 1, but can be unbounded in the length of x.

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

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

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 μ-random.

Therefore, for certain x and y which are *not* μ-random, optimization using the minimum length programs may result in incorrect predictions. For μ-random x we have that $-\log \mathbf{M}(x)$ and $Km(x)$ coincide up to an additional constant independent of x, that is, $g(xy) = g(x) \stackrel{+}{=} 0$, Claim 2. Hence, together with Equation 22, the theorem is proven. □

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 μ-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 11. With μ-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 $x \ldots$.*

4 Conclusion

The analysis of both hypothesis identification by ideal MDL and prediction shows that maximally compressed descriptions give good results on data that is individually random with respect to the contemplated probabilistic hypothesis. This situation occurs with probability going to one with increasing data.

Acknowledgement

Ray Solomonoff suggested to also analyze the case of prediction. Jorma Rissanen, Chris Wallace, David Dowe, Peter Grünwald, Kevin Korb, and the anonymous referees 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 [15] of a finite object x is the length of the shortest effective binary description of x. We give some definitions to establish notation. For more details see [41, 21]. 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 ϵ denotes the *empty word* '' 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 $x = x_1 x_2 \ldots x_n$ has a special type of prefix code, called a *self-delimiting code*,

$$\bar{x} = x_1 x_1 x_2 x_2 \ldots x_n \neg x_n,$$

where $\neg x_n = 0$ if $x_n = 1$ and $\neg x_n = 1$ otherwise. This code is self-delimiting because we can determine where the code word $\bar{x}$ 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 $l(\bar{x}) = 2n$ 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 ϕ_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 10 *The* Kolmogorov complexity *of x given y (for free on a special input tape) is*

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

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

Though defined in terms of a particular machine model, the Kolmogorov complexity is machine-independent up to an additive constant and acquires an asymptotically universal and absolute character through Church's thesis, from the ability of universal machines to simulate one another and execute any effective process. The Kolmogorov complexity of an object can be viewed as an absolute and objective quantification of the amount of information in it. This leads to a theory of *absolute* information *contents* of *individual* objects in contrast to classic information theory which deals with *average* information *to communicate* objects produced by a *random source* [21].

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 equiping 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. Taking the universal prefix machine U we can define the prefix complexity analogously with the plain Kolmogorov complexity. If x^* is the first shortest program for x then the set $\{x^* : U(x^*) = x, x \in \{0,1\}^*\}$ is a *prefix code*. That is, each x^* is a code word for some x, and if x^* and y^* are code words for x and y with $x \neq y$ then x^* is not a prefix of x.

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 11 *The* prefix Kolmogorov complexity *of x given y (for free) is*

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

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

The nice thing about $K(x)$ is that we can interpret $2^{-K(x)}$ as a probability distribution since $K(x)$ is the length of a shortest prefix-free program for x. By the fundamental Kraft's inequality, see for example [8, 21], we know that if $l_1, l_2, \ldots$ are the code-word lengths of a prefix code, then $\sum_x 2^{-l_x} \leq 1$. This leads to the notion of universal distribution—a rigorous form of Occam's razor– below.

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 ϕ and the value of ϕ 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 ϕ as follows. Interprete the computation $\phi(\langle x, t \rangle) = \langle p, q \rangle$ to mean that the quotient p/q is the rational valued tth approxmation of $f(x)$.

DEFINITION 12 *A function $f : \mathcal{N} \to \mathcal{R}$ is* enumerable *if there is a Turing machine T computing a total function ϕ 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, $\{0,1\}^*$). It is known, [21], that $\mathcal{EP}$ contains an element $\mathbf{m}$ that multiplicatively dominates all elements of $\mathcal{EP}$. That is, for each $P \in \mathcal{EP}$ there is a constant c such that $c\,\mathbf{m}(x) > P(x)$ for all $x \in \mathcal{N}$. We call $\mathbf{m}$ 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 $\mathbf{m}$ 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 $\mathbf{m}$, or its continuous analogue the measure $\mathbf{M}$ on the sample space $\{0,1\}^\infty$ (Section 3), is provably as good as using the true *a priori* distribution.

We also know, [19, 12, 9], that

THEOREM 12

$$-\log \mathbf{m}(x) = K(x) \pm O(1). \tag{23}$$

That means that $\mathbf{m}$ assigns high probability to simple objects and low probability to complex or random objects. For example, for $x = 00\ldots0$ (n 0's) we have $K(x) \stackrel{\pm}{=} K(n) \stackrel{+}{<} \log n + 2\log\log n$ 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, $1/(n\log^2 n) = O(\mathbf{m}(x))$. But if we flip a coin to obtain a string y of n bits, then with overwhelming probability $K(y) \stackrel{+}{>} n$ (because y does not contain effective regularities which allow compression), and hence $\mathbf{m}(y) = O(1/2^n)$.

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 [22]. This theory was further elaborated in [41, 28, 17] and later papers.

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.

We take some ideas from statistics. Consider a 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 x 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 ϵ, 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*

$$\begin{aligned} V_m &= \{x : (m, x) \in V\} \\ V_m &\supseteq V_{m+1}, \quad m = 1, 2, \ldots. \end{aligned}$$

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 ϵ. We can say that x fails the test at the level of critical region V_m.

EXAMPLE 5 A string $x_1 x_2 \ldots x_n$ 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 $x = 0.x_1 x_2 \ldots x_n$ 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 6 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 $x = x_1 x_2 \ldots x_n$ 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 13 *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. *δ 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 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 polynomial time in $l(m) + l(x)$.

EXAMPLE 7 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 $x = 1x_2 1x_4 1x_6 1x_8 \ldots$. Define a test δ 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 δ is a test we have to show that δ satisfies the definition of a test. Clearly, δ 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 $2^{n-(2m-1)}$ possibilities for the remainder of x. Therefore, $d\{x : \delta(x) \geq m,\ l(x) = n\} \leq 2^{n-m}$. $\diamond$

DEFINITION 14 *A universal Martin-Löf test for randomness with respect to distribution P, a universal P-test* for short, is a test $\delta_0(\cdot|P)$ such that for each P-test δ, there is a constant c, such that for all x, we have $\delta_0(x|P) \geq \delta(x) - c$.

> We say that $\delta_0(\cdot|P)$ (additively) majorizes δ. Intuitively, $\delta_0(\cdot|P)$ constitutes a test for randomness which incorporates all particular tests δ in a single test. No test for randomness δ other than $\delta_0(\cdot|P)$ 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. With $\delta_0(\cdot|P)$ a universal P-test, let $U = \{(m, x) : \delta_0(x|P) \geq m\}$, and, for any test δ, let $V = \{(m, x) : \delta(x) \geq m\}$. Then, defining the associated critical zones as before, we find
>
> $$V_{m+c} \subseteq U_m,\ m = 1, 2, \ldots,$$
>
> 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 3 *We can effectively enumerate all P-tests.*

PROOF. (Sketch.) 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 δ, the set $\{(m, x) : \delta(x) \geq m\}$ occurs in this list. The *only* thing we have to do is to eliminate those ϕ_i of which the range does not correspond to a P-test. This gives the idea. For details see [21]. $\square$

THEOREM 13 *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 14.

(1) The enumeration $\delta_1, \delta_2, \ldots$ in Lemma 3 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,

$$\begin{aligned} \sum_{l(x)=n} \{P(x) : \delta_0(x|P) \geq m\} &\leq \sum_{y=1}^{\infty} \sum_{l(x)=n} \{P(x) : \delta_y(x) \geq m + y\} \\ &\leq \sum_{y=1}^{\infty} 2^{-m-y} = 2^{-m}. \end{aligned}$$

(3) By its definition, $\delta_0(\cdot|P)$ majorizes each δ additively. Hence, it is universal. □

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

For any two universal P-tests $\delta_0(\cdot|P)$ and $\delta'_0(\cdot|P)$, 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 14 *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 δ, 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 $s = 00\ldots0j$ 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 $\bar{y}s$, 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 δ, we can set $c = 2l(y) + 2$. $\Box$

In Theorem 13, 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 $S = \mathcal{B}^n$ with $\mathcal{B} = \{0, 1\}$.

The universal P-test measures how justified is the assumption that x is the outcome of an experiment with distribution P. We now use $\mathbf{m}$ to investigate alternative characterizations of random elements of the sample set $S = \mathcal{B}^*$ (equivalently, $S = \mathcal{N}$).

DEFINITION 15 *Let P be a recursive probability distribution on $\mathcal{N}$. A* sum P-test *is a nonnegative enumerable function δ satisfying*

$$\sum_x P(x) 2^{\delta(x)} \leq 1. \tag{24}$$

A universal sum P-test *is a test that additively dominates each sum P-test.*

The sum tests of Definition 15 are slightly stronger than the tests according to Martin-Löf's original Definition 13.

LEMMA 4 *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

$$\sum \{P(x) : \delta(x) > k, l(x) = n\} \leq 2^{-k}. \tag{25}$$

If (25) is false, then we contradict (24) 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 (25) for all n, then for some constant c, the function $\delta(x) - 2\log\delta(x) - c$ satisfies (24). $\Box$

This shows that the sum test is not much stronger than the original test. One advantage of (24) 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 15 *Let P be a recursive probability distribution. The function*

$$\kappa_0(x|P) = \log(\mathbf{m}(x)/P(x))$$

is a universal sum P-test.

PROOF. Since $\mathbf{m}$ 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:

$$\sum_x P(x) 2^{\kappa_0(x|P)} = \sum_x \mathbf{m}(x) \leq 1.$$

It is only left to show that $\kappa_0(x|P)$ additively dominates all sum P-tests. For each sum P-test δ, 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 $c \cdot \mathbf{m}(x) \geq P(x)2^{\delta(x)}$. Hence, there is another constant c such that $c \cdot \kappa_0(x|P) \geq \delta(x)$, for all x. $\Box$

EXAMPLE 8 An important case is as follows. If we consider a distribution P restricted to a domain $A \subset \mathcal{N}$, then we have two choices for the universal sum P-test:

(i) $\log(\mathbf{m}(x|A)/P(x|A))$. 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

$$\log(\mathbf{m}(x|A)/L_n(x)) \stackrel{+}{=} n - K(x|n).$$

We have $L_n(x) = 1/2^n$ and $\log \mathbf{m}(x|A) = -K(x|A)$ by the Coding Theorem, Section A. where we can describe A by giving n.

(ii) $\log(\mathbf{m}(x)/P(x|A))$ where we view $P(x|A)$ as a distribution $P_A(x)$ that is positive only on subdomain A. For the uniform distribution L_n this gives the universal sum L_n-test

$$\log(\mathbf{m}(x)/L_n(x)) \stackrel{+}{=} n - K(x).$$

◇

EXAMPLE 9 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

$$-\lfloor \log P(x) \rfloor - K(x) \stackrel{+}{=} -\log P(x) + \log \mathbf{m}(x) \stackrel{+}{=} \kappa_0(x|P),$$

by the major theorems in Section C. That is, the randomness deficiency is the outcome of the universal sum P-test of Theorem 15. ◇

EXAMPLE 10 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 ∞ otherwise. This is close to the expression for $\kappa_0(x|L_n)$ obtained in Example 8. From the relations between C and K in [21] it follows that

$$|\delta_0(x|L_n) - \kappa_0(x|L_n)| \stackrel{+}{<} 2 \log C(x).$$

◇

The formulation of the universal sum test in Theorem 15 can be interpreted as to express the notion of randomness of objects with respect to a given distribution. While for infinite binary sequences the Martin-Löf theory gives a sharp distinction between the random and the nonrandom sequences, for finite sequences this distinction is blurred. We can't say that a given finite string is random and by changing one bit it becomes nonrandom. Therefore we more or less arbitrarily introduce cutoff points. The real significance lies in the fact that a random sequence should have small randomness deficiency as defined previously. In the main text in Definition 6 we define an element x to be Martin-Löf random with respect to distribution P if $\kappa_0(x|P) \stackrel{+}{<} K(P)$.

This means that for x to be random $P(x)$ should be large enough, not in absolute value but relative to $\mathbf{m}(x)$. 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 $x = 00\ldots0$ of length n. Then, $\kappa_0(x|L_n) \stackrel{+}{=} n - K(x|n) \stackrel{+}{=} n$. If we flip a coin n times to generate y, then with overwhelming probability $K(y|n) \geq n$ and $\kappa_0(y|L_n) = O(1)$.

EXAMPLE 11 *Markov's Inequality* says the following. Let P be any probability distribution, f any nonnegative function with P-expected value $\mathbf{E} = \sum_x P(x)f(x) < \infty$. For $\mathbf{E} \geq 0$ we have $\sum\{P(x) : f(x)/\mathbf{E} > k\} < 1/k$.

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

$$\sum_x P(x)\frac{\mathbf{m}(x)}{P(x)} \leq 1.$$

Then, by Markov's Inequality

$$\sum_x \{P(x) : \mathbf{m}(x) \leq kP(x)\} \geq 1 - \frac{1}{k}. \tag{26}$$

Since $\mathbf{m}$ dominates all enumerable semimeasures multiplicatively, we have for all x,

$$P(x) \leq c_P\mathbf{m}(x), \text{ and it can be shown } c_P = 2^{K(P)}. \tag{27}$$

Equations (26, 27) 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 $\mathbf{m}$ is a good estimate for P. For instance, if x is randomly drawn from distribution P, then the probability that

$$c_P^{-1}\mathbf{m}(x) \leq P(x) \leq c_P\mathbf{m}(x)$$

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 $\mathbf{m}(x)$.

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 $\kappa_0(x|P) = \log(\mathbf{m}(x)/P(x))$. 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

$$\kappa_0(x|L) \stackrel{+}{=} l(x) - K(x).$$

The randomness deficiency $\kappa_0(x|L) \stackrel{+}{=} 0$ iff $K(x) \stackrel{+}{>} l(x)$, that is, iff x is random.

The nonrecursive "distribution" $\mathbf{m}(x) = 2^{-K(x)}$ has the remarkable property that the test $\kappa_0(x|\mathbf{m}) \stackrel{+}{=} 0$ for all x: the test shows all outcomes x random with respect to it. We can interpret (26, 27) as saying that if the real distribution is P, then $P(x)$ and $\mathbf{m}(x)$ are close to each other with large P-probability. Therefore, if x comes from some unknown recursive distribution P, then we can use $\mathbf{m}(x)$ as an estimate for $P(x)$. In other words, $\mathbf{m}(x)$ 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 $\mathbf{m}$. 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. $\diamond$

References

[1] E. Asmis, *Epicurus Scientific Method*, Cornell University Press, 1984.

[2] T. Bayes, An essay towards solving a problem in the doctrine of cha nces, *Philos. Trans. Roy. Soc.*, 53:376-398 and 54:298-31, 1764.

[3] A.R. Barron, J. Rissanen, and B. Yu, The minimum description length principle in coding and modelling, *IEEE Trans. Inform. Theory*, IT-44:6(1998), 2743–2760.

[4] A.R. Barron and T.M. Cover, Minimum complexity density estimation, *IEEE Trans. Inform. Theory*, IT-37(1991), 1034–1054.

[5] R. Carnap, *Logical Foundations of Probability* Univ. Chicago Press, 1950.

[6] T.M. Cover, Kolmogorov complexity, data compression, and inference, pp. 23–33 in: *The Impact of Processing Techniques on Communications*, J.K. Skwirzynski, Ed., Martinus Nijhoff Publishers, 1985.

[7] , T.M. Cover and P. Gács and R.M. Gray, Kolmogorov's contributions to Information Theory and Algorithmic Complexity, *Ann. Probab.*, 17(1989), 840-865.

[8] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

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

[10] A.P. Dawid, Prequential analysis, stochastic complexity, and Bayesian inference, In: *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.), 1992.

[11] J.L. Doob, *Stochastic Processes* , Wiley, 1953.

[12] P. Gács, On the symmetry of algorithmic information, *Soviet Math. Dokl.*, 15 (1974) 1477-1480. Correction: ibid., 15 (1974) 1480.

[13] P. Gács, On the relation between descriptional complexity and algorithmic probability, *Theoret. Comput. Sci.*, 22(1983), 71-93.

[14] D. Hume, *Treatise of Human Nature, Book I*, 1739.

[15] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1-7.

[16] A.N. Kolmogorov, On logical foundations of probability theory, pp. 1–5 in *Probability Theory and Mathematical Statistics*, Lect. Notes Math., Vol. 1021, K. Itô and Yu.V. Prokhorov, Eds., Springer-Verlag, Heidelberg, 1983.

[17] A.N. Kolmogorov and V.A. Uspensky, Algorithms and Randomness, *SIAM J. Theory Probab. Appl.*, 32(1987), 389–412. Without annoying translation errors pp. 3-53 in: Yu.V. Prokhorov and V.V. Sazonov, Eds., *Proc. 1st World Congress of the Bernoulli Society (Tashkent 1986), Vol. 1: Probab. Theory and Appl.*, VNU Science Press, Utrecht, 1987.

[18] L.A. Levin, On the notion of a random sequence, *Soviet Math. Dokl.*, 14(1973), 1413-1416.

[19] L.A. Levin, Laws of information conservation (non-growth) and aspects of the foundation of probability theory, *Problems Inform. Transmission*, 10(1974), 206–210.

[20] M. Li and P.M.B. Vitányi, *Inductive reasoning and Kolmogorov complexity*, Proc. 4th IEEE Structure in Complexity Theory Conference, 1989, pp. 165-185.

[21] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 2nd Edition, 1997.

[22] P. Martin-Löf, The definition of random sequences, *Inform. Contr.*, 9(1966), 602-619.

[23] N. Merhav and M. Feder, A strong version of the redundancy-capacity theorem of universal coding, *IEEE Trans. Inform. Theory*, IT-41(1995), 714–722.

[24] K.R. Popper, *The Logic of Scientific Discovery*, University of Toronto Press, 1959.

[25] J.J. Rissanen, Modeling by the shortest data description, *Automatica-J.IFAC* 14 (1978) 465-471.

[26] J.J. Rissanen, Stochastic Complexity and Modelling, *The Annals of Statistics*, 14(1986), 1080–1100. Also: J.J. Rissanen, *Stochastic Complexity and Statistical Inquiry*, World Scientific Publishers, 1989.

[27] J.J. Rissanen, Fisher information and stochastic complexity, *IEEE Trans. Inform. Theory*, IT-42:1(1996), 40-47.

[28] C.P. Schnorr, *Zufälligkeit und Wahrscheinlichkeit; Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*, Lect. Notes Math., Vol. 218, Springer-Verlag, Heidelberg, 1971. See also: C.P. Schnorr, A survey of the theory of random sequences, pp. 193–210 in: *Basic Problems in Methodology and Linguistics*,, R.E. Butts and J. Hintikka, Eds., Reidel, 1977.

[29] A.Kh. Shen', The concept of (α, β)-stochasticity in the Kolmogorov sense, and its properties, *Soviet Math. Dokl.*, 28:1(1983), 295-299.

[30] R.J. Solomonoff, A formal theory of inductive inference, Part 1 and Part 2, *Inform. Contr.*, 7(1964), 1-22, 224-254.

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

[32] D. Sow and A. Elefteriadis, Complexity distortion theory, Submitted to *IEEE Trans. Inform. Theory*, 1997.

[33] J. Tromp and P.M.B. Vitanyi, Kolmogorov minimal sufficient statistic, Manuscript in preparation.

[34] 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.

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

[36] V. Vovk, Learning about the parameter of the Bernoulli model, *J. Comput. System Sci.*, 55:1(1997), 96-104.

[37] V.V. V'yugin, On the defect of randomness of a finite object with respect to measures with given complexity bounds, *Theory Probab. Appl.*, 32:3(1987), 508–512.

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

[39] 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.

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

[41] 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.