Resenhas IME-USP 1998, Vol. 3, No. 4, 359 - 390.

On Algorithmic Complexity, Universal Priors and Ockham's Razor [1]

Brani Vidakovic

Abstract: The first part of this paper is a review of basic notions and results connected with Kolmogorov complexity theory. A few original results are presented in Sections 3 and 4; they are not of a statistical nature. Emphasis is given to the so called *universal prior*. Though the prior itself is not a calculable measure, it has highly interesting properties from the Bayesian viewpoint. In the second part of the paper we discuss the principles that emerge from algorithmic complexity theory in the context of statistical prediction and estimation. It is argued that, as a rule, the principles are Bayesian in nature.

Key words: Recursive functions, Kolmogorov complexity, Schnorr complexity, Universal prior, Bayesian analysis, Ockham's razor

Contents

[1] *AMS* 1980 *subject classifications.* Primary 68F15; Secondary 94A17.

1 Introduction

Even before the nineteen-sixties, when algorithmic complexity theory was born, it was felt that the formal definition of the randomness of a binary sequence should depend on some precisely defined measure of disorder in the sequence. For example, von Mises' *Kollektiv* [71], a formal counterpart of a random infinite sequence, was defined to satisfy two requirements: (1) stability of relative frequences in any finite initial part of the sequence, and (2) stability of the relative frequencies in the algorithmically chosen finite part of the sequence.

Von Mises was unable to give a rigorous definition of what he called an *admissible* algorithm for choosing a subsequence. Wald (1937), and later Church (1940), made the notion of the *Kollektiv* mathematically precise. Ville (1939) proved that a *Kollektiv* (in the Wald sense) can be constructed so that the law of iterated logarithms fails.

Using the ideas of von Mises, Kolmogorov and others proposed a definition of the randomness of an infinite binary sequence through the algorithmically defined measure of entropy. Consider a pair of sequences of length twenty obtained by independent flips of a fair coin:

$$11111111111111111111$$

and

$$10101010101010101010.$$

These examples seem to lack randomness. At the same time, the sequence

$$01101110010010110100$$

looks "random." We know that all three sequences have the same probability of $\frac{1}{2^{20}}$, but only the third one intuitively looks like an outcome the above experiment should produce. The first two sequences are easy to describe: "twenty ones" or "ten pairs of 10." The third one requires a longer description.

Thus, the randomness of a sequence is intuitively connected to the difficulty of description, irregularity, and to some measure of disorder.

A series of papers in the mid-sixties by Solomonoff (1964), Kolmogorov (1965) and Chaitin (1966) introduce a formal measure of the complexity of binary words. Their definition needs a formal counterpart of the notion of effectively calculable functions, such as Recursive functions, Turing machines, Post machines, λ-calculable functions, normal algorithms, HG calculable functions, etc.

The measure of the complexity of a finite binary word x was defined as the length of the shortest program (argument) which when input to a "universal computer" prints x. The ultimate goal was to define the randomness of an infinite word ω. The word ω will be considered as random if the complexity of its initial parts is close to their length.

Yet another algorithmic approach to definition of randomness of an infinite binary sequence was proposed by Martin-Löf (1966a and 1966b) in the form of

ML tests. In brief, an ML test is an effective function defined on the class of all infinite words. It "collects" all (algorithmically describable) regularities of the word it tests. If there are too many irregularities, the word is rejected as nonrandom by the test. A word is considered random (in Martin-Löf sense) if it passes *any* ML test. The mathematical apparatus behind ML tests is constructive measure theory. There are many connections between complexity theory and ML tests. We will make these notions precise in Section 4.

The reader interested in the theory of ML tests may see [46], [47], and [14].

By chosing different classes of functions (machines) it is possible to define a variety of complexity measures. Namely, any class of partial recursive functions, for which a calculable numeration exists, can serve as a basis for defining an algorithmic complexity measure.

In addition to the Kolmogorov measure of complexity, we will define and list the basic properties of the Schnorr (monotone) complexity measure, introduced in [58]. The Schnorr measure has two nice properties. First, an infinite word is random (in ML sense) if and only if the Schnorr complexity of its initial fragments of length n is equal, up to a constant, to n. Second, the Schnorr complexity is connected with *the universal prior* on the space of all infinite binary words.

Our goal is to introduce the reader to the problems of complexity theory, as well as to point out the deep connection between complexity theory and the Bayesian paradigm.

In Section 2 we give the necessary notation and prerequisites. Some original results on properties of Kolmogorov and Schnorr complexities are given in Section 3. Section 4 introduces the universal prior and discusses some of its properties. Some Bayesian applications of the universal prior and, in general, of Ockham's razor are given in Section 5. The paper contains an extensive bibliography on the subject.

2 Notation and Prerequisites

The following notation will be used.

- $\mathcal{A}$ - a finite alphabet. Without loss of generality it may be taken as $\{0, 1\}$.
- $x = x_1 x_2 \ldots x_n$ - a word of the length n in the alphabet $\mathcal{A}$.
- Λ - the empty word
- X^n- the set of all words of the length n.
- $X = \cup_n X^n$ - the set of all finite words.
- $|A|$- the cardinal number of the set A.
- One-to-one correspondence between words in X and integers $\{0, 1, \ldots, n, \ldots\}$ can be defined as $x = x_1 x_2 \ldots x_n \to 2^n - 1 + \sum_{i=1}^{n} x_i 2^{n-i}$. For example, the word 010110 corresponds to the integer 85.

- $\ell(x)$ - the length of a word x.
- $\bar{x} = x_1x_1x_2x_2\dots x_nx_n01$ - a code of a word x so that it can be decoded from concatenated words. For example, words x and y can be coded as one word $\bar{x}y$. There exist effective functions ξ_1, and ξ_2 such that $x = \xi_1(\bar{x}y)$, and $y = \xi_2(\bar{x}y)$.
- $x \subset y$ - the word x is an initial part (beginning) of the word y.
- $\omega = \omega_1\omega_2\dots\omega_n\dots$ - an infinite word in the alphabet $\mathcal{A}$.
- Ω - set of all infinite words ω.
- $X^* = \Omega \cup X$ - set of all finite and infinite words.
- ω_{n_1,n_2} - part $\omega_{n_1}\omega_{n_1+1}\dots\omega_{n_2}$ of word ω.
- $f(x) \preceq g(x)$ means $(\exists C)(\forall x) \;\; f(x) \leq g(x) + C$.
- $f(x) \asymp g(x)$ means $f(x) \preceq g(x)$ and $g(x) \preceq f(x)$. For example, $\ell(x) \asymp \log_2 x$.
- $clim_{n\to\infty} f(n) = A$ is a constructive limit. In other words, there is an effective nonnegative function $g(n)$ that tends to zero (often taken 2^{-n}) such that $|f(n) - A| \leq g(n)$; (We know how close $f(n)$ and A are for each n).

2.1 Recursive functions

Recursive functions are a formal, mathematical analogue of the notion *effectively calculable* functions. We give the necessary definitions and results. Good references are [55], [21], and [49], among others.

Definition 2.1 *The functions*

$$\begin{aligned} \mathcal{Z}(x_1,\dots,x_n) &= 0, \\ \mathcal{I}_k(x_1,\dots,x_n) &= x_k, \;\; 1 \leq k \leq n, \\ \mathcal{S}_k(x_1,\dots,x_n) &= x_k + 1, \;\; 1 \leq k \leq n. \end{aligned} \tag{1}$$

are the **initial** *functions.*

Definition 2.2 *(Dedekind 1888) A function $F(x_1,\dots,x_n,x_{n+1})$ is defined from $G(x_1,\dots,x_n)$ and $H(x_1,\dots,x_{n+2})$ by* **primitive recursion** *if*

$$\begin{aligned} F(x_1,\dots,x_n,0) &= G(x_1,\dots,x_n), \\ F(x_1,\dots,x_n,y+1) &= H(x_1,\dots,x_n,y,F(x_1,\dots,x_n,y)). \end{aligned}$$

Definition 2.3 *A function $F(x_1, \ldots, x_n)$ is defined from functions $H(x_1, \ldots, x_m)$ and $G_1(x_1, \ldots, x_n), \ldots, G_m(x_1, \ldots, x_n)$ by* **composition** *if*

$$F(x_1, \ldots, x_n) = H(G_1(x_1, \ldots, x_n), \ldots, G_m(x_1, \ldots, x_n)).$$

Definition 2.4 *(Kleene 1936) A function $F(x_1, \ldots, x_n)$ is defined from $G(x_1, \ldots, x_n)$ by* **μ-recursion** *if*

$$F(x_1, \ldots, x_n) = \mu_y(G(x_1, \ldots, x_{n-1}, y) = x_n),$$

where $\mu_y(G(x_1, \ldots, x_{n-1}, y) = x_n)$ is the least number a such that $G(x_1, \ldots, x_{n-1}, y) = a$ holds.

We will consider that $\mu_y(G(x_1, \ldots, x_{n-1}, y) = x_n)$ is not defined when:

(i) $F(x_1, \ldots, x_{n-1}, y)$ is defined for all $y < a$, but different than x_n, and $F(x_1, \ldots, x_{n-1}, a)$ is not defined,

(ii) $F(x_1, \ldots, x_{n-1}, y)$ is defined for all values of y, but is different than x_n.

Definition 2.5 *(Skolem, Gödel 1931). The class of* **primitive recursive functions** *is the smallest class of functions*

(i) containing the initial functions,

(ii) closed under primitive recursion and composition.

Definition 2.6 *(Kleene 1936) The class of* **partial recursive functions** *$\mathcal{P}$ is the smallest class of functions*

(i) containing the initial functions,

(ii) closed under primitive recursion, composition and μ-recursion.

The class of everywhere defined partial recursive functions is called **total functions** and are denoted by $\mathcal{O}$.

Church Thesis (Church 1936) The class of effectively computable functions coincides with the class of partial recursive functions.

Theorem 2.1 *(Kleene 1938) There exists a partial recursive function U of $n+1$ arguments,* **universal** *for the class of all n-dimensional partial recursive functions $\mathcal{P}^{(n)}$ with the property*

$$(\forall F \in \mathcal{P}^{(n)})(\exists n_F) \quad F(x_1, \ldots, x_n) = U(n_F, x_1, \ldots, x_n). \tag{2}$$

n_F is the number of function F with respect to U.

The function U is often called an **enumeration** of the class $\mathcal{P}^n$. An enumeration of the set S^n is any n-tuple of total functions $(F_1, \ldots, F_n)$ that map N to S^n. The number of $(x_1, \ldots, x_n) \in S^n$ is k if $F_i(k) = x_i$.

Definition 2.7 *The set A is* **enumerable** *if the set of its numbers (in a fixed enumeration) is a domain of some partial recursive function F. It is said that F enumerates A.*

Theorem 2.2 *The predicate $P^n(a_1, \ldots, a_n)$ is a partial recursive (total) if there is a partial recursive (total) function F taking the value 0 at all and only the n-tuples $(a_1, \ldots, a_n)$ satisfying the predicate.*

Theorem 2.3 *For any partial recursive predicate P^{n+m} the set*

$$\{(x_1, \ldots, x_n) |\ (\exists (a_1, \ldots, a_m)) P^{n+m}(x_1, \ldots, x_n, a_1, \ldots, a_m) \text{ is true}\}$$

is enumerable.

2.2 Kolmogorov Complexity

Definition 2.8 *[30] Let $F \in \mathcal{P}$ and let $x \in X$. The algorithmic (Kolmogorov) complexity of the word x with respect to F is*

$$K_F(x) = \min_{p \in X} \ell(p) : F(p) = x, \tag{3}$$

with $\min \emptyset = \infty$.

The dependence on a particular function F in the previous definition is eliminated by the following *optimality* theorem.

Theorem 2.4 *[62], [30]*

$$(\exists F_0 \in \mathcal{P})(\forall G \in \mathcal{P})(\forall x \in X) \quad K_{F_0}(x) \preceq K_G(x). \tag{4}$$

F_0 is called an optimal partial recursive function and $K_{F_0}(x)$ is denoted simply by $K(x)$.

Proof: Let $F_0(x) = U^{(2)}(\xi_1(x), \xi_2(x))$, where U is the universal function from Theorem 2.1. Let G be any partial recursive function, and let n_G be the number of G in the numeration U. Let $K_G(x) = l_0$. That means that there exists a program p_x of length l_0 such that $G(p_x) = x$. Then the function F_0 applied to the program $q = \bar{n}_G p_x$ prints x as well. Also, $K(x) = K_{F_0}(x) \leq \ell(\bar{n}_G p_x) = 2n_G + 2 + l_0 = C + K_G(x)$. The constant C does not depend on x; it depends only on the choice of the universal numeration U and the number of the function G in the chosen numeration U. □

The optimal function F_0 is not unique. Nevertheless, this poses no difficulty since for any other optimal function F_0'

$$|K_{F_0}(x) - K_{F_0'}(x)| \asymp 0.$$

Remark: The conditional Kolmogorov complexity $K_G(x|y)$ is defined similarly. Namely

$$K_F(x|y) = \min_{p \in X} \ell(p) | F(p, y) = x. \tag{5}$$

As analogy with unconditional complexity, there is an optimal function $F_0^{(2)}$ such that a result equivalent to Theorem 2.4 holds. Also $K(x|\Lambda) = K(x)$.

Since $K(x)$ is an *algorithmic measure of entropy*, it is possible to introduce the algorithmic measure of information that the word y carries about the word x, $I(y : x)$, as

$$I(y : x) = K(x) - K(x|y). \tag{6}$$

For the properties of the algorithmic measure of information, the reader may see [30],[31], [80],[24], and [66], among others.

We give here some basic properties of the Kolmogorov complexity measure.

- $K(x) \preceq \ell(x)$

  The identity function $I(x) = x$ needs a program of length exactly $\ell(x)$.

- [80] The proportion of words $x \in X^n$ for which

  $$K(x) < n - m \tag{7}$$

  is not bigger than 2^{-m}. (For most of the words $K(x) \sim \ell(x)$.)

- [80] $\lim_{x\to\infty} K(x) = \infty$.

- [80] Define the function

  $$m(x) = \inf_{y \geq x} K(y), \tag{8}$$

  i.e. $m(x)$ is the largest nondecreasing function that is a lower bound on $K(x)$ (Figure 1). No recursive function exists that goes to ∞ more slowly than $m(x)$.

- [80] The function $K(x)$ is 'smooth', i.e.

  $$|K(x + h) - K(x)| \preceq \ell(h). \tag{9}$$

- The function $K(x)$ is not recursive.

- [80] There exists a monotone nondecreasing total function $H(t, x)$ such that

  $$\lim_{t\to\infty} H(t, x) = K(x), \tag{10}$$

  but the limit is not constructive.

- [4] $\Pi(n) = \min\{K(x) : \ell(x) = n\} \asymp K(n) \preceq \log n$.

- [4] $n \leq \max\{K(x) : \ell(x) = n\} \preceq n$

- [4] $\max\{K(x|\ell(x)) : \ell(x) = n\} \asymp n$.

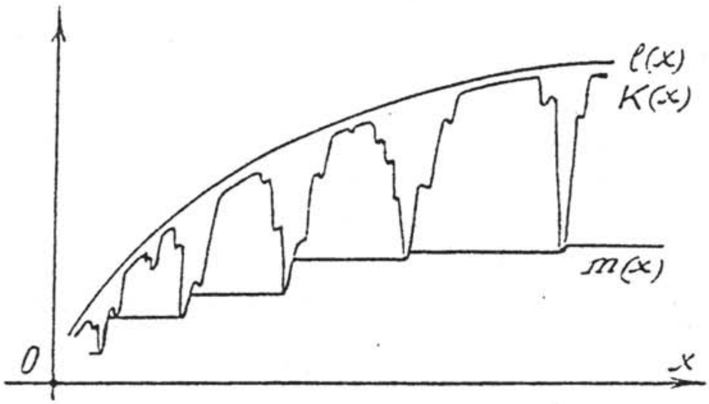

Figure 1: Functions $K(x)$ and $m(x)$

- [5] For any set A
  $\max\{K(x|y) : x \in A\} \geq \ell(|A|) - 1 \asymp \log|A|$,

- [4] Let p_x (p_x^y) be any program for which $F_0(p_x) = x$ $(F_0^2(p_x^y, y) = x)$. The program p_x (p_x^y) can be defined uniquely, but the procedure is not recursive. Then

$$K(p_x^y) \asymp K(x|y).$$

- [4] $\lim_{y\to\infty} K(x|y) \preceq 0$ is not true, but

$$\liminf_{y\to\infty} K(x|y) \preceq 0$$

  holds.

- [5] Let $A_m = \{x : K(x) \leq m\}$ and $B_m = \{x : K(x|m) \leq m\}$ then

$$m - 2\log m \preceq \log|A_m| \preceq m,$$

  and

$$\log|B_m| \asymp m.$$

- [24] Let F and G be any functions.

$$K(F(x)|y) \preceq K(x|G(y))$$

and

$$K(F(x,y)|y) \preceq K(x|y).$$

2.3 Schnorr Complexity

Definition 2.9 *[58] The function $\mathcal{F} \in \mathcal{P}$ is a monotone process (or simply a process) on X if for $x \subset y$ and $\mathcal{F}(y)$ defined, then $\mathcal{F}(x)$ is also defined and $\mathcal{F}(x) \subset \mathcal{F}(y)$. The class of all processes is denoted by $\mathcal{PR}$.*

Examples.

(i) The identity function $\mathcal{I}(x) = x$ is a process.

(ii) The word function defined by $\mathcal{F}(x0) = \mathcal{F}(x)00$ and $\mathcal{F}(x1) = \mathcal{F}(x)1$ is a process.

The following theorem shows the class of processes to be a basis for defining a measure of complexity:

Theorem 2.5 *(Schnorr 1970) There exists a universal process $\mathcal{U}^{(2)}$ (enumeration of $\mathcal{PR}$) such that*

$$(\forall \mathcal{F} \in \mathcal{PR})(\exists n_{\mathcal{F}})\ \mathcal{U}^{(2)}(n_{\mathcal{F}}, x) = \mathcal{F}(x). \tag{11}$$

Definition 2.10 *The process complexity of $x \in X$, with respect to $\mathcal{F} \in \mathcal{PR}$ is the quantity*

$$KP_{\mathcal{F}}(x) = \min_{p \in X} l(p) : \mathcal{F}(p) = x. \tag{12}$$

The existence of the universal process (Theorem 2.5) gives the following optimality theorem:

Theorem 2.6 *[58],[80]*

$$(\exists \mathcal{F}_0 \in \mathcal{PR})(\forall \mathcal{G} \in \mathcal{PR})(\forall x) \quad KP(x) \asymp KP_{\mathcal{F}_0}(x) \preceq KP_{\mathcal{G}}(x). \tag{13}$$

We give without proof some basic properties of the measure $KP(x)$.

- $KP(x) \preceq \ell(x)$.
- [58] $K(x) \preceq KP(x) \preceq K(x) + 2\ell(K(x))$. The constant 2 can be improved to $1+\epsilon$ by more compact coding, for arbitrary $\epsilon > 0$.
- [66] For any $F \in \mathcal{P}^{(2)}$
  $KP(F(x,y)) \preceq KP(x) + KP(y) + 2\ell(K(x)K(y))$.
- [65] $KP(x)$ is not in the class $\mathcal{P}$.
- $\lim_{x\to\infty} KP(x) = \infty$.

3 More Properties of K(x) and KP(x)

The function $m(x)$ defined by (8) has interesting properties. We already mentioned the fact that it is unbounded.

Theorem 3.1

$$\lim_{x\to\infty} m(x) = \infty. \tag{14}$$

Proof: Suppose the opposite. Then there is a constant C for which $\liminf_{x\to\infty} m(x) \leq C$, and we can find an infinite sequence $x_1, x_2, \ldots$, with the property $K(x_i) \leq C$. This is impossible since there are at most $2^{C+1} - 1$ distinct words with complexity less than or equal to C. $\square$

Some other functions connected with $K(x)$ and $m(x)$ can be defined.

Definition 3.1

$$M(x) = \max_{K(y)\leq x} y \tag{15}$$

$$P(x) = \min_{m(y)>x} y \tag{16}$$

Since the functions $K(x)$ and $m(x)$ are defined on a set of integers, the functions $M(x)$ and $P(x)$ are integers and $M(x) + 1 = P(x)$. Figure 2 shows the connection between the four functions $K(x), m(x), M(x)$, and $P(x)$.

Theorem 3.2 *$P(x)$ ($M(x)$) is not a recursive function. It tends to infinity faster than any other partial recursive function that tends to infinity, i.e.*

$$(\forall F \in \mathcal{P})(\exists x_0)(\forall x > x_0)\ F(x) < P(x). \tag{17}$$

Proof: Suppose the opposite. Then there is an infinite set S for which $(\forall x \in S)\ F(x) \geq P(x)$. This set is enumerable (Lemma 3.1), and there is an infinite set $S_0 \subset S$ on which $F(x)$ is a total function. Define

$$G(x) = \begin{cases} F(x) + 1, & x \in S_0 \\ F(\min_{y\geq x, y\in S_0} y) + 1, & x \in (S_0)^c \end{cases}$$

$G(x)$ is total and $G(x) > F(x) > P(x),\ x \in S_0$.

On the other hand, one has $K(G(x)) > K(P(x)) > x$ by the definition of the function $P(x)$. Therefore, $x < K(G(x)) \asymp K_{G\circ F_0}(x) \preceq K(x) \preceq \ell(x)$, which is a contradiction. $\square$

Corollary 3.1 *The function $m(x)$ is not recursive. It tends to infinity more slowly than any other recursive function that tends to infinity.*

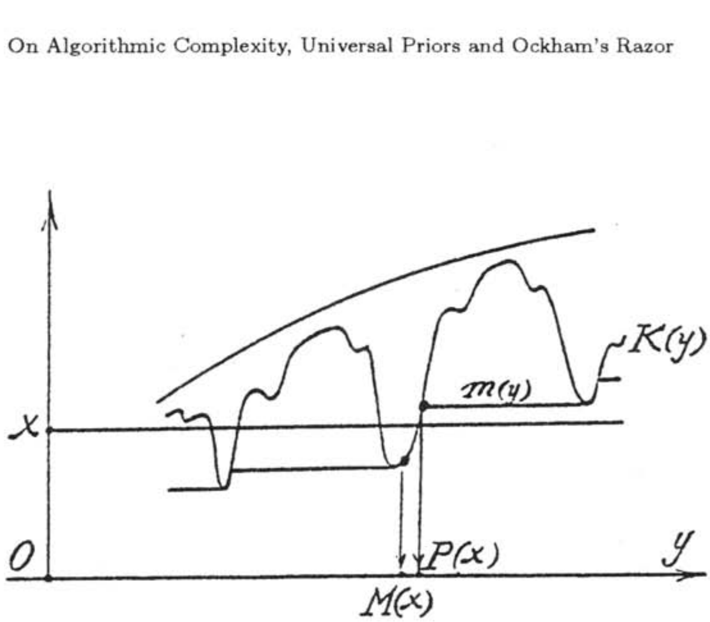

Figure 2: Relations between the functions K, m, M, and P

Theorem 3.3 *There is a total function* $\underset{\sim}{m}(x,t)$ *such that* $\lim_{t\to\infty} \underset{\sim}{m}(x,t) = m(x)$, *but the limit is not constructive.*

Proof: Since $(\exists C)\ K(x) < \ell(x) + C$, we can take an algorithm that calculates F_0, and performs t steps on all words of length less than $\ell(x) + C$, taken in a natural (lexicographic) ordering. Let

$$\underset{\sim}{K}(x,t) = \begin{cases} \ell(p), & \text{if the above procedure yielded } x. \\ \ell(x) + C, & \text{otherwise} \end{cases} \tag{18}$$

Define

$$\underset{\sim}{m}(x,t) = \inf_{y>x} \underset{\sim}{K}(y,t). \tag{19}$$

The function $\underset{\sim}{m}(x,t)$ is a total function and $\lim_{t\to\infty} \underset{\sim}{m}(x,t) = m(x)$.

Similarly, one can define total counterparts $\underset{\sim}{P}(x,t)$ and $\underset{\sim}{M}(x,t)$ of the functions $P(x)$ and $M(x)$, such that $\lim_{t\to\infty} \underset{\sim}{P}(x,t) = P(x)$ and $\lim_{t\to\infty} \underset{\sim}{M}(x,t) = M(x)$.

Lemma 3.1 *For a fixed* $F \in \mathcal{P}$, *the set* $\{x | F(x) \geq P(x), F \in \mathcal{P}\}$ *is enumerable.*

Proof of the Lemma: The predicate $[F(x) \geq \underset{\sim}{P}(x,t)]$ is total. Consequently, the set $\{x |\ (\exists t) F(x) \geq \underset{\sim}{P}(x,t)\} = \{x |\ F(x) \geq P(x)\}$ is enumerable. $\square$

Example: By an elementary argument we can check that $m(x)$ is slower than n times a repeated logarithm, for arbitrary n.

Let $N = 2^{2^{\cdot^{\cdot^{2}}}}$. One can define a function $F(n)$ which for the input n prints N zeroes. Then, $K(x) \preceq K_F(x) = n = \overbrace{\ell(\ell \ldots \ell}^{n}(x))\ldots) + C$.

Therefore, the following result holds

Theorem 3.4

$$(\forall n)\ m(x) \preceq \overbrace{\ell(\ell(\ldots \ell}^{n}(x)\ldots) \tag{20}$$

Banjević (1981) proved that there is no partial recursive function $F^{(2)}$ for which $K(F(x,y)) \asymp K(x) + K(y)$, thus the relation $K(F(x,y)) \succeq K(x) + K(y)$ is true.

We now give a few upper bounds on $K(F(x,y))$.

- [65] $K(F(x,y)) \preceq 2K(x) + K(y)$ is a straightforward bound.
- [65] $K(F(x,y)) \leq K(x) + K(y) + \frac{1}{2}\ell(K(x)K(y)) + \ell(\ell(K(x)K(y)))$.
- [65] For any $s, 0 < s \leq \ell(x)$,

$$K(F(x,y)) \preceq (1 + \frac{1}{2s})(K(x) + K(y)) + s. \tag{21}$$

and

$$KP(x) \preceq (1 + \frac{1}{s})K(x) + s. \tag{22}$$

- [65] If $F(x,y) \in \mathcal{P}^{(2)}$ is such that x and y are decodable, i.e. $(\exists G, H \in \mathcal{P})G(F(x,y)) = x,\ H(F(x,y)) = y$ then

$$K(F(x,y)) \succeq \frac{1}{2}(K(x) + K(y)). \tag{23}$$

A consequence of (23) is $K(\bar{x}y) \geq \frac{1}{2}(K(x) + K(y))$.

4 Measures on Ω and Martin-Löf's tests

The set $\Gamma_x \subset \Omega$ defined as

$$\Gamma_x = \{\omega \in \Omega |\ \omega_{1,l(x)} = x\} \tag{24}$$

is called a cylinder centered at x. To define a measure on the space Ω it is enough to define the measure on each of the sets $\Gamma_x,\ x \in X$. (Sets Γ_x form a basis for topology on the space Ω and they are Borel subsets of Ω.) Moreover, for an arbitrary function $m : X \to R$, that for any $x \in X$ satisfies

(i) $m(\Lambda) = 1$,

(ii) $m(x0) + m(x1) = m(x)$, and

(iii) $m(x) \geq 0$,

there exists a unique measure μ on Ω for which $(\forall x) \ \ \mu(\Gamma_x) = m(x)$. Sometimes we will write $\mu(x)$ instead of $\mu(\Gamma_x)$, when there is no danger of misunderstanding. The measure μ of a single word x will be denoted by $\mu(\{x\})$.

Let $\Sigma_x = \Gamma_x \cup \{xy | \ y \in X\}$. The measure ν on X^* can be defined by assigning $m(x)$ to each of the Σ_x (It is possible to introduce a topology on X^* with the sets Σ_x as a basis, but the resulting topological space is very poor – it is a T_0 space.)

Restricted to Ω, ν defines a semi-measure, i.e. a set function with the properties:

(i) $\nu(\Omega) \leq 1$,

(ii) $\nu(\Gamma_x 0) + \nu(\Gamma_x 1) \leq \nu(\Gamma_x)$,

(iii) $\nu(\Gamma_x) \geq 0$.

Definition 4.1 *[80] Probability (semi) measure μ on Ω is called* **calculable** *if*

$$(\exists F^{(2)}, G^{(2)} \in \mathcal{O}) \ \ r_\mu(x,t) = \frac{F(x,t)}{G(x,t)} \tag{25}$$

is nondecreasing and

$$clim_{t \to \infty} r_\mu(x,t) = \mu(x). \tag{26}$$

Examples:

- Uniform probability measure on Ω is defined as

$$\lambda(x) = 2^{-\ell(x)}. \tag{27}$$

  In a natural transformation of Ω to the interval $[0,1]$, (by $\omega \to (0.\omega)_2$), the measure λ corresponds to the Lebesgue measure. Obviously, the measure λ is calculable.

- Another example of a calculable measure on Ω is Bernoulli measure. If $w(x) = \sum_{i=1}^{\ell(x)} x_i$ is the "weight" of $x \in X$, then the measure β_p defined through the function $b : X \to R$ as

$$b(x) = p^{w(x)}(1-p)^{\ell(x)-w(x)}; \ \ 0 < p < 1, \tag{28}$$

  is called *Bernoulli(p)* measure. Note that $\beta_{1/2} = \lambda$.

- [14] Let $n \geq 2$ be a fixed number. Define measure ι on Ω as follows.

  (i) $\iota(\Lambda) = 1$,

  (ii)

$$\begin{cases} \iota(\Gamma_{x0}) = \iota(\Gamma_x)(1 - \frac{1}{(\ell(x)+2)^n}), \\ \iota(\Gamma_{x1}) = \iota(\Gamma_x)\frac{1}{(\ell(x)+2)^n}. \end{cases}$$

  For example, for $n = 2$, $\iota(\{000\ldots 0\ldots\}) = \frac{1}{2}$.

4.1 Martin-Löf's tests

Definition 4.2 *[46],[80] A total function V defined on finite words is called Martin-Löf's test (ML test) with respect to a calculable measure μ if*

$$clim_{m \to} \mu(\omega | \; V(\omega) \geq m) = 0, \tag{29}$$

where $V(\omega) = \sup_n V(\omega^n)$. A word ω is ML-nonrandom with respect to test V (does not withhold ML test V) if $V(\omega) = \infty$.

Theorem 4.1 *[46] There exists a universal ML test U, such that for any other ML test V:*

$$(\forall x) \; U(x) \succeq V(x). \tag{30}$$

Definition 4.3 *[46] A word ω is ML-random if it passes the universal ML test.*

Example: The function

$$V_\epsilon(x) = \sum_{i=1}^{\ell(x)} \mathbf{1}(|\frac{w(x^i)}{i} - \frac{1}{2}| > \epsilon), \tag{31}$$

where x^i is the first i symbols of x, and w is the weight of x, is an ML test under the uniform measure, for any fixed ϵ.

(i) V_ϵ is a total function,

(ii) $\lambda(\omega | V(\omega) \geq m) \to 0$, because of the Borel strong law of large numbers.

In other words, V_ϵ is a ML test that rejects all $\omega \in \Omega$ for which the relative frequency of ones is different than $\frac{1}{2}$.

Example: This example is an adaptation of the result of Erdös and Revesz (1976). If the word ω is ML-random with respect to the uniform measure, then the length of the longest 1 run (the longest piece consisting only of ones) in ω^n has to be between

$$\log n - \log\log\log n + \log\log e - 2 - \epsilon, \tag{32}$$

and

$$\log n + 1.1 \log\log\log n, \tag{33}$$

for any ϵ. If $n = 2^{2^{20}}$, the length of the longest 1-run is between 1,048,569 and 1,048,598. Amazingly, the difference is only 29, so that the random sequences are almost deterministic in some aspects of their stochastic behavior.

Example: [69] The sign-test can be interpreted as an ML test. Suppose we have two samples, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, for which we want to test the hypothesis that they come from the same continuous population. Form the finite word $x = x_1 x_2 \ldots x_n$ as follows:

$$x_i = \begin{cases} 1, & X_i > Y_i \\ 0, & X_i \leq Y_i \end{cases}$$

Then the function

$$F(x) = |\frac{2w(x) - n}{\sqrt{n}}| \tag{34}$$

is an ML test with respect to the uniform measure.

(i) F is a total function,

(ii) For large m, if Φ is the cdf of the standard normal law, $\lambda(\omega |\, F(\omega) \geq m) \leq 2\Phi(-m) \leq 2^{-m}$.

In the paper [46] Martin Löf introduces the **measure of randomness** of a word x, with respect to an ML test F as

$$KB_F(x) = \ell(x) - \inf_{x \subset z} F(z). \tag{35}$$

The measure $KB(x)$ resembles closely the measures of complexity $K(x), KP(x)$, etc. Some properties of the measure $KB(x)$ are given here.

- [46] There is a universal test U so that for any other test F

$$KB(x) = KB_U(x) \leq KB_F(x). \tag{36}$$

- [68] Let $G_x(i, y)$ be the result of the application of $\ell(x)$ steps of the algorithm that calculates $\mathcal{U}^2(i, y)$. Then

$$\ell(x) - \max_{i \leq \ell(x), y \subset x} G_x(i, y) \leq KB(x) \leq \ell(x). \tag{37}$$

  KB is a smooth function,

$$KB(xy) - KB(x) \leq \ell(y) \tag{38}$$

- There is an increasing total function $\Phi(t, x)$ such that

  (i) $\Phi(t, x) \leq KB(x)$,

  (ii) $\lim_{t \to \infty} \Phi(t, x) = KB(x)$.

- [68] $KB(x)$ is not recursive, but the predicate

$$\Pi(x,a) = [KB(x) < a]$$

is a partial recursive and the set

$$\{x|\ (\exists a) KB(x) < a\}$$

is recursively enumerable.

-

$$\lambda(\Gamma_x|\ KB(x) \leq \ell(x) - m) \leq 2^{-m}.$$

- [46]

$$|KB(x) - K(x)| \preceq (2-\epsilon)\ell(\ell(x)).$$

- [68] If $\mathcal{F}$ is a process for which $\delta(\mathcal{F}(x)) = \ell(x) - \ell(\mathcal{F}(x))$, then

$$KB(x) - KB(\mathcal{F}(x)) \preceq \delta(\mathcal{F}(x)).$$

4.2 Measure Transformations and Universal Prior

Definition 4.4 *[80] Let $\mathcal{F}$ be a process. We say that the process $\mathcal{F}$ is* **applicable** *to an infinite word ω if the result is also an infinite word. We will call process $\mathcal{F}$ μ-***regular** *if the μ measure of words to which it is applicable is 1.*

Define a measure ν on X^* (semi-measure on Ω) as follows:

$$\nu(\Sigma_x) = \mu(\bigcup_{x:\mathcal{F}(x)=y} \Gamma_x). \tag{39}$$

We will say that the measure ν is a process transformation of the measure μ and write $\nu = \mathcal{F}(\mu)$.

(i) If μ is calculable, then ν is also a calculable measure.

(ii) For *any* calculable measure ν there is λ-regular process $\mathcal{F}$ such that

$$\mathcal{F}(\lambda) = \nu. \tag{40}$$

The process $\mathcal{F}$ can be chosen in a such way that $\mathcal{G} = \mathcal{F}^{-1}$ is ν-regular.

Definition 4.5 *[80] The measure μ is called* **semi-calculable** *if there is a process $\mathcal{F}$, such that*

$$\mu = \mathcal{F}(\lambda). \tag{41}$$

It can be proved that a semi-calculable measure μ can be approximated by the ratio (25) in which the functions F and G are partial recursive.

Definition 4.6 *[80] The probability measure π defined as*

$$\pi = \mathcal{F}_0(\lambda), \tag{42}$$

where $\mathcal{F}_0$ is an optimal process, is called the **universal prior.**

The universal prior is "larger" (up to a multiplicative constant) than any other semi-calculable measure – thus the name *prior.* In the absence of any information about the distribution on Ω, the most noninformative assumption is that the distribution is π. It has the "fattest tails."

Theorem 4.2 *[80]*

$$(\exists C_\mu)(\forall x)\ C_\mu \cdot \pi(x) \geq \mu(x). \tag{43}$$

Proof: Let μ be an arbitrary semi-calculable measure and let $\mathcal{J}$ be the process generating the measure μ from the uniform measure λ. If

$$A = \{\bigcup \Gamma_p | x \subset \mathcal{J}(p)\}$$

then $\mu(\Gamma_x) = \lambda(A)$. Let

$$B = \{\bar{i}a|\ i \text{ is the number of the process } \mathcal{J} \text{ w.r.t. } \mathcal{U}^{(2)}; a \in A\}.$$

Then $\mathcal{F}_0(x) = \mathcal{U}(\xi_1(x), \xi_2(x))$ transforms B into Γ_x:

$$\mathcal{F}_0(\bar{i}a) = \mathcal{U}(i, a) = \mathcal{J}(a) \in \Gamma_x.$$

Finally,

$$\pi(\Gamma_x) \geq \lambda(B) = 2^{-\ell(\bar{i}a)} = 2^{-\ell(\bar{i})}\lambda(A).$$

The constant C_μ in the statement of the theorem is $2^{-\ell(\bar{i})}$, where i is the number of the process $\mathcal{J}$, in the numeration $\mathcal{U}$. $\square$

Corollary 4.1 $(\forall x)\ \pi(x) > 0$.

Proof: Suppose the opposite, that for some $x_0 \in X$, $\pi(x_0) = 0$ holds. Take any semi-calculable measure μ which is concentrated on Γ_x. Then $0 = \pi(x_0) \geq C_\mu \mu(x_0) > 0$. $\square$

The following theorem connects the measure of Schnorr complexity KP and the universal prior.

Theorem 4.3 *[80]*

$$KP(x) \asymp -\log \pi(x). \tag{44}$$

As Theorem 4.3 says, the prior π gives a large probability to the words with small complexity. The complex words, on the other hand, have a small probability. Another consequence of Theorem 4.3 is:

Corollary 4.2 *(i)* $\pi(\overbrace{000\ldots0}^{n}) \geq \frac{1}{n}\frac{1}{C\log^2 n}$,
(ii) If $K(\omega^n) \geq n + C$, *then*

$$\pi(\omega^n) \leq Const\, 2^{-n}.$$

Proof: Since $KP(\overbrace{000\ldots0}^{n}) \preceq \ell(n) + 2\ell(\ell(n))$, assertion (i) follows. The proof of fact (ii) is easy. □

Example: We can construct a measure on Ω that simulates π in the following sense: it gives high probability to sequences consisting of a large number (close to the length) of zeroes or ones.

Suppose that we know that a measure on Ω is β_p, but p is unknown. If the prior on p is $\mathcal{B}e(a,b)$, then the standard Bayesian calculation gives that the distribution of $p|x$ is $\mathcal{B}e(a + w(x), b + \ell(x) - w(x))$. The predictive (marginal) distribution for x is $m(x) = \mathbf{B}(a + w(x), b + \ell(x) - w(x))$, where $B(.,.)$ is the standard Beta function.

If the prior on p is "noninformative", i.e. $p \sim \mathcal{B}e(1,1)$ then

$$m(\overbrace{000\ldots0}^{n}) = B(1, n+1) = \frac{1}{n+1}.$$

The following theorem connects ML tests and the universal prior.

Theorem 4.4 *An infinite word ω is ML-random with respect to the measure λ, if and only if there are constants C_1 and C_2 such that*

$$(\forall n)\ C_1\, 2^{-n} \leq \pi(\omega^n) \leq C_2\, 2^{-n}. \tag{45}$$

Moreover,

$$\lambda\{\omega|\pi(\omega^n) > 2^{-n+m}\} < 2^{-m}. \tag{46}$$

4.3 Robustness results for the universal prior

Let μ and ν be two semi-calculable measures. Let

$$r(\mu, \nu, x) = \log\frac{\mu(x)}{\nu(x)}, \tag{47}$$

and

$$d(\mu, \nu, n) = E^{\mu} r(\mu, \nu, \omega_{1,n}) - r(\mu, \nu, \omega_{1,n-1}). \tag{48}$$

Theorem 4.5 *$d(\mu, \nu, n)$ is the Kullback-Leibler distance between the predictive measures $\mu(\bullet_n|\bullet_{1,n-1})$ and $\nu(\bullet_n|\bullet_{1,n-1})$.*

Proof: Simple transformations give

$$d(\mu, \nu, n) = E^{\mu} \log \frac{\mu(\omega_n|\omega_{1,n-1})}{\nu(\omega_n|\omega_{1,n-1})}. \quad \square$$

If we do not know μ and use π instead as a conditional measure, then after observing a word long enough, the prediction by the universal measure becomes almost as good as the prediction by μ. The prior π "catches" the measure μ.

Theorem 4.6 $d(\mu, \pi, n) \to 0, \quad n \to \infty.$

Proof:

$$r(\mu, \pi, \omega_{1,n}) = r(\mu, \pi, \omega_1) + r(\mu, \pi, \omega_{1,2}) - r(\mu, \pi, \omega_1) + \ldots + r(\mu, \pi, \omega_{1,n}) - r(\mu, \nu, \omega_{1,n-1}). \tag{49}$$

It follows that

$$E^{\mu} r(\mu, \pi, \omega_{1,n}) = E^{\mu} r(\mu, \pi, \omega_1) + \sum_{i=1}^{n} d(\mu, \pi, i). \tag{50}$$

First, $d(\mu, \pi, i) \geq 0$. It follows from the fact that for two probability vectors $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$

$$\sum_i p_i \log \frac{p_i}{q_i} \geq 0.$$

Second, $E^{\mu} r(\mu, \pi, \omega_{1,n})$ is uniformly bounded in ω and n, because of property (43) of universal measure, namely

$$\log \frac{\mu(\omega_{1,n})}{\pi(\omega_{1,n})} \leq \log \frac{\mu(\omega_{1,n})}{c_\mu \mu(\omega_{1,n})} = -\log c_\mu = c'_\mu.$$

Therefore, $\sum_{i=1}^{\infty} d(\mu, \pi, i)$ is convergent, and $d(\mu, \pi, n) = o(\frac{1}{n})$ $\square$.

Theorem 4.7 *There exists a sequence π_n such that*
(i) π_n is a computable measure for any n,
(ii) $\lim_{n \to \infty} d(\mu, \pi_n, n) = 0$.

Remark: Gacs (1974) proved a stronger result. For any fixed finite word y, and for any semi-computable measure μ:

$$\frac{\pi(y|x)}{\mu(y|x)} \to 1, \quad \text{when } x \to \infty,$$

holds with μ-measure 1.

4.4 Universal word

Definition 4.7 *The infinite word ω is called* **calculable** *if there is a total function G, such that*

$$(\forall n) \ \ \omega_n = G(n). \tag{51}$$

Define the lower frequency of a finite word x with respect to the word ω as

$$\phi_\omega(x) = \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x = x(i)), \tag{52}$$

where $\omega = x(1)x(2)x(3)\ldots$ and $\ell(x(i)) = \ell(x)$, $(\forall i)$.

Theorem 4.8 *There is a universal word ρ so that for any other word ω,*

$$(\forall x) \ \ C\phi_\rho(x) \geq \phi_\omega(x), \tag{53}$$

where the constant C depends only on the word ω.

One can define a measure of complexity of x as

$$C(x) = -\log \phi_\rho(x). \tag{54}$$

Remark: The word ρ is not calculable. This is a consequence of the fact that the universal function for the class of all total functions is only a partial recursive.

5 Minimum Description Length Principles

Pluralitas non est ponenda sine necessitate.

William of Ockham (1290-1349)

As Jeffreys and Berger (1991) pointed out, the idea of measuring complexity and connecting the notions of complexity and prior probability goes back to Sir Harold Jeffreys' pioneering work on statistical inference in the 1920s. On page 47 of his classical work [27], Jeffreys says:

> Precise statement of the prior probabilities of the laws in accordance with the condition of convergence requires that they should actually be put in an order of decreasing prior probability. But this corresponds to actual scientific procedure. A physicist would test first whether the whole variation is random against the existence of a linear trend; than a linear law against a quadratic one, then proceeding in order of increasing complexity. All we have to say is that simpler laws have the greater prior probabilities. This is what Wrinch and I called the simplicity postulate. To make the order definite, however, requires a numerical rule for assessing the complexity law.

> In the case of laws expressible by differential equations this is easy. We would define the complexity of a differential equation, cleared of roots and fractions, by the sum of order, the degree, and the absolute values of the coefficients. Thus $s = a$ would be written as $ds/dt = 0$ with complexity $1+1+1=3$. $s = a+ut+\frac{1}{2}gt^2$ would become $d^2s/dt^2 = 0$ with complexity $2+1+1=4$. Prior probability 2^{-m} of $6/\pi^2m^2$ could be attached to the disjunction of all laws of complexity m and distributed uniformly among them.

In the spirit of Jeffreys' ideas, and building on work of Wallace and Boulton, Akaike, Dawid, Good, Kolmogorov, and others, Rissanen (1978) proposed the Minimum Description Length Principle (MDLP) as a paradigm in statistical inference. Informally, the MDLP can be stated as follows:

The preferred theory H for explaining observed data D is one that minimizes:

- the length of the description of the theory (Ockham's razor principle)
- the length of the description of the data with the help of the chosen theory.

Let C represent *some* measure of complexity. Then the above may be expressed, again informally, as:

Prefer the theory H, for which $C(H) + C(D|H)$ is minimal.

In the above sentence we emphasized the word "some." Aside from the formal algorithmic definitions of complexity, which lack recursiveness, one can define a complexity measure by other means. The following example gives one way.

Example: Let μ be a measure on Ω, and let $\mu(x) = \mu(\Gamma_x)$. Then the Shannon code for a word x uses $\lceil -\log\mu(x)\rceil$ binary symbols. With $\lceil\alpha\rceil$ we denote the smallest integer larger than the number α. The Shannon code is optimal in the sense that it uses the minimum number of symbols for coding. A complexity measure $C(x)$ can be defined as the length of its Shannon code, i.e. $-\log\mu(x)$, rounded up to the next integer.

Many other effective measures of complexity have been proposed. Lempel and Ziv (1976) gave a combinatorial measure of complexity. Their measure is used in the theory of compact coding. Vidakovic (1985) and Stojanovic and Vidakovic (1987) propose a measure of complexity based on the number of $\rceil$, $\bigvee$, and $\bigwedge$ operations in the Boolean function generating the binary word. Though their measure is effective, practically it is impossible to calculate the complexity of words of even moderate length (e.g. 64), because of the exponential calculational complexity.

It is interesting that Bayes rule implies MDLP in the following way. For a Bayesian, the theory H for which

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{55}$$

is maximal, is preferred. Taking negative logarithms on both sides we get

$$\begin{aligned} -\log P(H|D) &= -\log P(D|H) - \log P(H) + \log P(D) \\ &= C(D|H) + C(H) + \text{Const.} \end{aligned} \tag{56}$$

The Maximum Likelihood Principle (MLP) can also be interpreted as a special case of Rissanen's MDL principle. The ML principle says that, given the data, one should prefer the hypothesis that maximizes $P(D|H)$, or that minimizes $-\log P(D|H)$, the first term in the right hand side of (56).

If the complexities of the hypotheses are constant, i.e., if their descriptions have the same length, then the MDL principle becomes the ML principle. The rationale of the ML principle was to *be objective* and independent of prior assumptions. From the MDLP standpoint, the ML is very subjective, having all hypotheses of the same complexity. Berger and Wolpert (1988) give a lucid discussion on the ML principle.

5.1 Algorithmic Complexity Criterion

Let μ be an unknown semi-computable measure on Ω. After observing $x \in X$, we want to estimate μ.

As an estimate of μ, choose a measure $\hat{\mu}$ that minimizes

$$K(\nu) + \log \frac{1}{\nu(x)}. \tag{57}$$

The second part of (57) is minimized for any measure ν for which $\nu(\Gamma_x)$ is 1. The first part of (57) is an algorithmic counterpart of the penalty for choosing measures that are too complex. With no data in hand, $\frac{1}{\log \nu(\Gamma_\Lambda)} = 0$, and the preferred measure is the simplest measure.

A natural enumeration of the class Γ of all semi-computable measures can be defined by the function $T(p)$ as follows. The function T takes an argument p and finds the process $\mathcal{U}_0^{(2)}(p, .)$. The process $\mathcal{U}_0^{(2)}(p, .)$ is a modification of the universal process $\mathcal{U}^2$ for which the number of $\mathcal{F}_0$ is Λ (empty word).

The process $\mathcal{G}(.) = \mathcal{U}_0^2(p, .)$ transforms the uniform measure λ to some semi-calculable measure $\mu \in \Gamma$. In that way, the function T enumerates Γ. Therefore, the following theorem holds:

Theorem 5.1 *In absence of data, the best estimate, with respect to the above described enumeration T, is the universal measure π. It embodies Ockham's razor principle.*

Proof: $K(\pi) \preceq K_T(\pi) \asymp 0$. $\square$

5.2 Bayesian interpretation of the algorithmic complexity criterion (Barron-Cover (1989))

Let $X_1, X_2, \ldots, X_n$ be observed random variables from an unknown probability density we want to estimate. The class of candidates Γ is enumerable, and to each density f in the class Γ, the prior probability $\pi(f)$ is assigned. The "complexity" $C(f)$ of a particular density f is $-\log \pi(f)$.

The minimum over Γ of

$$C(f) + \log \frac{1}{\prod_k f(X_k)}$$

is equivalent to the maximum of $\pi(f)\prod_k f(X_k)$, which as a function of f, is proportional to the Bayes posterior probability of f given $X_1, \ldots, X_n$.

Remark: There is a connection between the Bayesian and the coding interpretations in that if π is a prior on Γ then $\log \frac{1}{\pi(f)}$ is the length (rounded up to integer) of the Shannon code for $f \in \Gamma$ based on the prior π. Conversely, if $C(f)$ is a codelength for a uniquely decodable code for f, then $\pi(f) = 2^{-C(f)}/D$ defines a proper prior probability ($D = \sum_{f\in\Gamma} 2^{-C(f)} \leq 1$ is the normalizing constant).

Let $\hat{f}_n$ be a minimum complexity density estimator. If the true density f is on the list Γ, then

$$(\exists n_0)(\forall n \geq n_0)\hat{f}_n = f.$$

Unfortunately, the number n_0 is not effective, i.e. given Γ that contains the true density and $X_1, \ldots, X_n$, we do not know if $\hat{f}_n$ is equal to f or not. Even when the true density f is not on the list Γ, we have the consistency of $\hat{f}_n$. Let $\bar{\Gamma}$ denote *the information closure* of Γ, i.e. the class of all densities f for which $\inf_{g\in\Gamma} D(f||g) = 0$, where $D(f||g)$ is the Kullback-Leibler distance between f and g. The following result holds [9]: If $\sum_{g\in\Gamma} 2^{-C(g)}$ is finite, and the true density is in $\bar{\Gamma}$, then

$$\lim_n \hat{P}_n(S) = P(S)$$

holds with probability 1, for all Borel sets S.

5.3 Wallace-Freeman Criterion

Wallace and Freeman (1987) propose a criterion similar to the Barron-Cover criterion for the case when Γ is a parametric class of densities.

Let $X_1, X_2, \ldots X_n$ be a sample from the population with density $f(x|\theta)$. Let $\pi(\theta)$ be a prior on θ.

The Minimum Message Length (MML) estimate is defined as

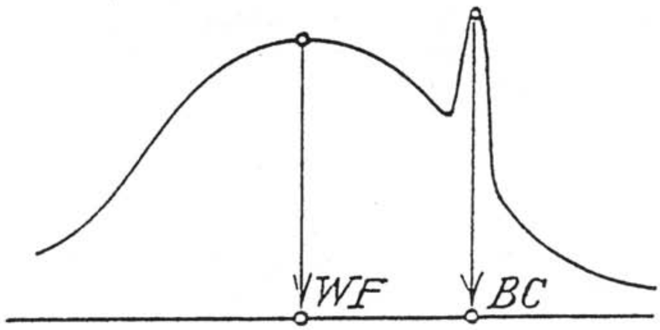

Figure 3: Barron-Cover and Wallace-Freeman estimators

$$arg\min_{\theta}[-\log\pi(\theta)-\log\prod_{i=1}^{n}f(x_i|\theta)+\frac{1}{2}\log|\mathcal{I}(\theta)|]. \tag{58}$$

where $\mathcal{I}(\theta)$ is the appropriate information matrix. Note that this is equivalent to maximizing

$$\frac{\pi(\theta)\prod_{i=1}^{n}f(x_i|\theta)}{|\mathcal{I}(\theta)|^{1/2}}. \tag{59}$$

Interestingly, if the prior on θ is chosen to be the noninformative Jeffreys' prior, then the MML estimator reduces to ML estimator. Another nice property of the MML estimator is its invariance under 1-1 transformations.

Dividing by $|\mathcal{I}(\theta)|^{1/2}$ in (59) may not be what a Bayesian would do. In this case, instead of choosing the highest posterior mode, the MML estimator choses the local posterior mode with the highest probability content, if it exists (Figure 3).

Example: [77] Suppose a Bernoulli experiment gives m successes and $n-m$ failures. Take the $Beta(a,b)$ prior on θ. Then, $\mathcal{I}(\theta)=\frac{n}{\theta(1-\theta)}$.

The MML estimator is a value that maximizes $\theta^{a+m-1/2}(1-\theta)^{b+n-m-1/2}$, i.e.

$$\theta' = \frac{a+m-\frac{1}{2}}{a+b+n-1}. \tag{60}$$

Note that the Bayes estimator $\hat{\theta}_B = \frac{a+m}{a+b+n}$ is slightly different.

Example: [77] Another example of the application of MML criteria is a simple model selection procedure.

Let $\mathcal{P}_\mu = \{N(\mu, \sigma^2), \sigma^2 \text{ known}\}$. Chose the best of the hypotheses: $H_0 : \mu = \mu_0$, and $H_1 : \mu \neq \mu_0$, in light of data $x = (x_1, \dots, x_n)$.

H_0 is parameter free, the message length is $-\log f(x|\mu)$.

Let $\mu \sim Unif[L\sigma, U\sigma]$. Then, assuming equal prior probabilities for H_0 and H_1, the hypothesis H_0 is preferred to H_1 if

$$z = |\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}| < \sqrt{\log \frac{ne(U-L)^2}{12}}. \tag{61}$$

This is in contrast with the usual frequentist significance test in which the right-hand side of (61) has the constant $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

In the case of vague prior information on μ $(U - L \to \infty)$, the above criterion leads to a strong favoring of the simple hypothesis H_0, as Jeffreys (1939) pointed out.

Remark: O'Hagan (1987) proposed a modification of the MML estimator as follows: Estimate θ by the value $\hat{\theta}$ that maximizes

$$\frac{\pi(\theta|x)}{H(\theta, x)^{1/2}} \tag{62}$$

where $H(\theta, x) = -\frac{\partial^2}{\partial\theta^2} \log \pi(\theta|x)$.

O'Hagan's modification is more in the Bayesian spirit, since everything depends only on the posterior. But the maximizing $\hat{\theta}$ may not be at any posterior mode, and in addition, the invariance property of the MML estimator is lost.

5.4 Rissanen's Criteria

Except for motivational purposes, Rissanen does not include algorithmic complexity in his criteria. The "complexities" he refers to are effective measures emerging from the theory of optimal coding. They simulate non-effective complexity measures and give a working, real criteria. The MDLP, which Rissanen discusses in [53], goes as follows: In the case when the parameter $\theta = (\theta_1, \dots, \theta_k)$ of variable dimension k describes the model, chose the model that minimizes

$$-\log P(x|\theta) + \log^*[C(k)(\|\theta\|_{M(\theta)})^k], \tag{63}$$

where

(i) $P(x|\theta)$ is the likelihood;

(ii) $\log^*(z) = \log(z) + \log\log(z) + \log\log\log(z) + \ldots$, where only positive terms are included in the sum;

(iii) $C(k)$ is the volume of the k-dimensional ball;

(iv) $\|\theta\|_{M(\theta)} = \sqrt{\theta' M(\theta)\theta}$, where $M(\theta)$ is the $k \times k$ matrix of second derivatives of $-\log P(x|\theta)$. The second part in (63) is Rissanen's counterpart for the negative complexity of the model (θ).

6 Epilogue

There are few more ways of using the complexity theory ideas in statistics. We may want to produce finite binary words of maximal complexity.

Example: (Parmigiani) Let a finite binary word x of the length n represents an ordered group of n patients. The symbol 1 on kth place in the word x means that the kth patient has received a treatment. Zeroes stand for placebo. The word x can be designed. The response is again a binary word of length n, in which the symbol 1 stands for "survived".

The goal is to test if $\theta = P(1|1)$ is different than $P(1|0)$. It is felt that θ depends on the place of the corresponding 1 in the word x. (The medical staff giving the treatment becomes more experienced, or perhaps, after a while, the staff gets bored and the quality of treatment decreases.)

Theoretically, one should choose the following design. The word x should be of maximal complexity. That ensures that the testing procedure is robust with respect to all *simply describable* dependences $\theta = \theta(k)$, $1 \leq k \leq n$, which we pose as our prior. This choice is in the spirit of Mises' "preserving the randomness" recursive choice of a subsequence.

Vovk (1991) connected the complexity theory results with the theory of asymptotic efficiency of estimators.

The complexity theory approach to statistical inference is far from being a unified theory. The main difficulty is that there is no effective measures of complexity. All working Minimum Description Length Procedures include some calculable counterpart of an algorithmic complexity measure. The compromise is to simulate algorithmic complexity measures as closely as possible and keep the procedure effective.

7 Acknowledgements

I would like to thank colleagues Michael Lavine, Peter Müller, and Giovanni Parmigiani for useful discussions.

References

[1] AGAFONOV, B. N. (1975). *Complexity of Algorithms and Calculations.* NGU Publications, Novosibirsk (in Russian).

[2] AKAIKE, H. (1977). On entropy maximization principle. *In Applications of Statistics (P. R. Krishnaiah, ed) 27-41.* North-Holland.

[3] ASARIN, E. A. (1988). On some properties of finite objects random in an algorithmic sense. *Soviet Math. Dokl.* **36** *109-112.*

[4] BANJEVIĆ, D. (1981). On some basic properties of the Kolmogorov complexity. *Publications De L'Institut Mathématique* **30 (44)** *17-23.*

[5] BANJEVIĆ, D. (1984). Note on the number of sequences with given complexity. *Publications De L'Institut Mathématique* **36 (50)** *107-109.*

[6] BANJEVIĆ, D. and IVKOVIĆ, Z. (1978). A definition of regularity and randomness founded on Kolmogorov's ideas (In Serbian).*Matematički Vesnik* **2** *(15)*

[7] BANJEVIĆ, D. and IVKOVIĆ, Z. (1979). On algorithmical testing tables of random numbers. *Publications De L'Institut Mathématique* **25 (39)** *11-15.*

[8] BARRON A. R. and CLARKE, B. (1988). Information theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory.*

[9] BARRON, A. R. and COVER, T. M. (1989). Minimum complexity density estimation. *University of Illinois, Statistical Department, Technical Report #28.*

[10] BERGER, J. O. (1985). *Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

[11] BERGER, J. O. and JEFFREYS, W. H. (1991). Minimal Bayesian testing of precise hypothesis, model selection and Ockham's razor. *Technical Report, Purdue University.*

[12] BERGER, J. O. and WOLPERT, R. (1988). *The Likelihood Principle, Second Edition.* Institute of Mathematical Statistics Monograph Series, Hayward, California.

[13] BOULTON, D. M. and WALLACE, C. S. (1973). An information measure for hierarchic classification. *Comp. J.* **13** *254-261.*

[14] CALUDE, C. and CHITESCU, I. (1981). On Per Martin Löf random sequences. *Bull. Math. Soc. Math. de Romanie* **26 (74)** *3, 217-221.*

[15] CALUDE, C. and CHITESCU, I. (1983). On representability of P. Martin Löf tests. *Kibernetika* **19** *1 42-47.*

[16] CALUDE, C. and CHITESCU, I. (1983). Representability of recursive P. Martin Löf tests. *Kibernetika* **19** *6 526-535.*

[17] CALUDE, C., CHITESCU, I. and STAIGER, L. (1985). P. Martin Löf tests: Representability and embeddability. *Rev. Roumaine Math. Pures Appl.* **30** *719-732.*

[18] CHAITIN, G. J. (1966). On the length of programs for computing finite binary sequences. *J. Assoc. Comp. Mach.* **13** *547-569.*

[19] CHAITIN, G. J. (1987). *Algorithmic Information Theory.* Cambridge University Press.

[20] CHURCH, A. (1940). On the concept of random sequence. *Bull. Amer. Math. Soc* **46** *130-135.*

[21] DAVIS, M. (1982). *Computability and Unsolvability.* Dover, New York.

[22] DAWID, A. P. (1984). Present Position and Potential Developments: Some Personal Views, Statistical Theory, The Prequential Approach. *J. R. Statist. Soc. A* **147** *278-292.*

[23] ERDÖS, P. and REVESZ, P. (1976). On the length of the longest head run. *Coll. Math. Soc. T. Bolyai* **16**. *Topics in Information Theory. North Holland, Amsterdam.*

[24]. GÁCS, P. (1974). On symmetry of algorithmic information. *Soviet. Math. Dokl.* **15** *p.1477,* (Correction,*Ibid* **15**).

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

[26] JACOBS, K. (1970). Turing Maschinen und Zufällige 0-1 Folgen. *In: Selecta Mathematica* **2**, *Springer-Verlag.*

[27] JEFFREYS, H. (1939). *Theory of probability* Clarendon Press. Oxford 1985.

[28] JEFFREYS, W. H. and BERGER, J. O. (1991). Sharpening Ockham's razor on Bayesian strop. *Technical Report #91-44C, Statistical Department, Purdue University*

[29] KOLMOGOROV, A. N. (1963). On tables of random numbers. *Sankhya, Series A* **25** *369-376.*

[30] KOLMOGOROV, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission* **1** *1 1-7.*

[31] KOLMOGOROV, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Trans. on Information Theory,* **IT-14.5** *662-664.*

[32] KOLMOGOROV, A. N. and USPENSKI, V. (1986). Algorithms and randomness. *Proc. Ist World Congress of the Bernoulli Society, Tashkent, USSR 3-53.*

[33] KRAMOSIL, I. (1985). Non-deterministic prediction based on algorithmic complexity. *Problems of Control and Information Theory* **14** *461-476.*

[34] KRAMOSIL, I. (1985). On pseudo-random sequences over finite automata. *Ann. Soc. Math. Polonae, Ser. IV: Fundamenta Informaticae VIII.1.*

[35] KRAMOSIL, I. (1985). Statistical testing of finite sequences based on algorithmic complexity. In *Lecture Notes in Computer Science* **199.**

[36] LEMPEL, A. and ZIV, J. (1976). On the complexity of finite sequences. *IEEE Transactions on the Information Theory, vol* **IT-22** *75-81.*

[37] LEVIN L. A. (1973). Universal search problems. *Problems in Information Transmission* **9** *265-266.*

[38] LEVIN L. A. (1973). On the notion of random sequence. *Soviet. Math. Dokl.* **14** *1413-1416.*

[39] LEVIN L. A. (1974). Laws of information conservation (non-growth) and aspects of the foundation of probability. *Problems in Information Transmission* **10** *206-210.*

[40] LEVIN L. A. (1984). Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control* **61** *15-37.*

[41] LI, M. and VITÁNYI, P. M. B. (1989). Inductive reasoning and Kolmogorov complexity. *Manuscript.*

[42] LI, M. and VITÁNYI, P. M. B. (1989). Kolmogorov complexity and its applications (Revised version). *CWI Report CS-R8901.*

[43] LI, M. and VITÁNYI, P. M. B. (1989). Average case complexity under the universal distribution equals worst case complexity. *Presented at 30th Annual IEEE Symposium on Foundations of Computer Science.*

[44] LOVELAND, D. W. (1966). A new interpretation of the von Mises concept of random sequence. *Z. Math. Logik und Grundlagen der Math.* **12** *279-294.*

[45] LOVELAND, D. W. (1969). A variant of the Kolmogorov concept of complexity. *Information and Control* **15** *510-526.*

[46] MARTIN-LÖF, P. (1966a). The definition of random sequences. *Information and Control* **9** *602-619.*

[47] MARTIN-LÖF, P. (1966b). On the concept of random sequence. *Theory. Probab. Appl.* **11** *177-179.*

[48] MARTIN-LÖF, P. (1974). The notion of redundancy and its use as a quantitative measure of discrepancy between a statistical hypothesis and a set of observational data. *Scand. J. Statist.* **1** *3-18.*

[49] ODIFREDDI, P. (1989). *Classical Recursion Theory.* Studies in Logic, Volume **125**, North-Holland.

[50] O'HAGAN, A. (1987). Discussion of the papers by Dr Rissanen and Professors Wallace and Freeman. *J. R. Statist. Soc. B.* **49** *3, 256-257.*

[51] PARMIGIANI, G. (1992). Private communication.

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

[53] RISSANEN, J. (1982). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** *416-431.*

[54] RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** *1080-1100.*

[55] ROGERS, H. (1967). *Theory of recursive functions and effective computability.* McGraw Hill.

[56] RYABKO, B. YA. (1988). Prediction of random sequences and universal coding. *Problemy Peredachi Informatsii* **24** *2, 3-14.*

[57] SCHNORR, C. P. (1971). *Zufälligkeit und Wahrscheinlichkeit; Eine algorithmische Begründung der Wahrscheinlichkeitstheorie.* Lecture Notes in Mathematics, Springer Verlag, Berlin.

[58] SCHNORR, C. P. (1973). Process complexity and effective random tests. *J. Comput. System Sciences* **7** *376.*

[59] SHEN, A. (1983). The concept of (α, β)-stochasticity in the Kolmogorov sense, and its properties. *Soviet Math. Doklady* **28** *1, 295-299.*

[60] SHEN, A (1992). Algorithmic complexity and randomness: Recent developments. *Theory Probab. Appl.* **37** *124-131.*

[61] SOLOMONOFF, R. J. (1960). A preliminary report on a general theory of inductive inference. *Techn. Report* **ZTB-138**, *Zator Company, Cambridge, Mass.*

[62] SOLOMONOFF, R. J. (1964). A formal theory of inductive inference. Part 1 and Part 2. *Information and Control* **7** *1-22; 224-254.*

[63] STOJANOVIC, S. and VIDAKOVIC, B. (1987). Some properties of the combinatorial measure of complexity of binary words. *Publications De L'Institut Mathématique* **42 (56)** *143-147.*

[64] USPENSKII, V. A., SEMENOV, A. H. and SHEN', A. H. (1990). Can (the particular) sequence zeroes and ones be random? *Uspekhi Math. Nauk, t. LXV* **1** *(271), 105-162.*

[65] VIDAKOVIC, B. (1981). Complexity and Randomness. *Unpublished Manuscript, University of Belgrade.*

[66] VIDAKOVIC, B. (1983). Some characteristics of the process measure of the amount of information. *Publications De L'Institut Mathématique* **33 (47)** *235-238.*

[67] VIDAKOVIC, B. (1984). On some properties of the Martin-Löf's measure of randomness of finite binary words. *Proceedings of the Conference "Algebra and Logic", Zagreb 1984.*

[68] VIDAKOVIC, B. (1985a). An effective measure of complexity of binary words (In Serbian). *Matematički Vesnik* **37** *327-332.*

[69] VIDAKOVIC, B. (1985b). The sign-test from the algorithmic complexity standpoint. *Communicated at: 5th Pannonian Symposium on Math. Statistics, Visegrad* Statisticka revija **3-4** 289-291 (in Serbian).

[70] VILLE, J. (1939). *Étude critique de la notion de collectif.* Paris, Gauthier-Villars.

[71] VON MISES, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Zeitschrift* **5** *55-99.*

[72] VON MISES, R. (1957). *Probability, Statistics and Truth.* Dover Publications, New York.

[73] VOVK, V. G. (1991). Asymptotic efficiency of estimates: algorithmic approach. *Theory Probab. Appl.* **36** *247-261.*

[74] V'YUGIN, V. V. (1985). Nonstochastic estimates. *Problemy Peredachi Informatsii* **21** *3-9.*

[75] V'YUGIN, V. V. (1986). Some estimates for nonstochastic sequences. *Proc. Ist World Congress of the Bernoulli Society, Tashkent, USSR.*

[76] WALD, A. (1937). Die Widerspruchsfreiheit des Kollektivbegriffs der Wahrscheinlichkeitsrechnung. *Ergebnisse eines Mathemathischen Kolloquiums* **8** *38-72.*

[77] WALLACE, C. S. and FREEMAN, P. R. (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. B* **49** *240-265.*

[78] WALLACE, C. S. and BOULTON, D. M. (1975). An invariant Bayes method for point estimation. *Classification Soc. Bull.* **3** *11-34.*

[79] WRINCH, D. and JEFFREYS, H. (1921). On certain fundamental principles of scientific inquiry. *Phil. Mag.* **42** *369-390.*

[80] ZVONKIN, A. K. and LEVIN, L. A. (1970). Complexity of finite objects and the basing of the concepts of information and randomness on the theory of algorithms (In Russian). *Russian Math. Surveys, t. XXV* **6** *(1956), 85-127.*

Brani Vidakovic
Institute of Statistics and Decision Sciences
Duke University.
P.O. Box 90251, Durham, NC 27708-0251
brani@isds.duke.edu
United States of America