Jan Šindelář; Pavel Boček

Kolmogorov complexity, pseudorandom generators and statistical models testing

Kybernetika, Vol. 38 (2002), No. 6, [747]--759

Persistent URL: http://dml.cz/dmlcz/135500

KYBERNETIKA — VOLUME 38 (2002), NUMBER 6, PAGES 747 - 759

KOLMOGOROV COMPLEXITY, PSEUDORANDOM GENERATORS AND STATISTICAL MODELS TESTING

JAN ŠINDELÁŘ AND PAVEL BOČEK

An attempt to formalize heuristic concepts like strings (sequences resp.) "typical" for a probability measure is stated in the paper. Both generating and testing of such strings is considered. Kolmogorov complexity theory is used as a tool.

Classes of strings "typical" for a given probability measure are introduced. It is shown that no pseudorandom generator can produce long strings from the classes. The time complexity of pseudorandom generators with oracles capable to recognize "typical" strings is shown to be at least exponential with respect to the length of the output.

Tests proclaiming some strings "typical" are introduced. We show that the problem of testing strings to be "typical" is undecidable. As a consequence, the problem of correspondence between probability measures and data is undecidable too. If the Lebesgue measure is considered, then the conditional probability of failure of a test is shown to exceed a positive lower bound almost surely.

1. INTRODUCTION

The problem of describing single strings (sequences resp.) which are "typical" or "characteristic" for a given probability measure is an old, important and difficult one.[1] Various approaches to its solution are summarized in [2]. Probably the first attempt to solve it formally was done by Von Mises (see [2]). Significant progress in the solution was achieved by Kolmogorov complexity theory and theory of Martin-Löf tests.

Kolmogorov complexity theory was originated by Kolmogorov in [3].[2] Exposition of the theory could be found e. g. in [1, 6]. Strings (sequences resp.) which are "characteristic", or "typical" for a given probability measure are called random (m–random, asymptotic random) with respect to the measure.

Theory of Martin-Löf tests was initiated by Martin-Löf [7]. Its explanation and the relationship between Kolmogorov complexity theory and theory of Martin-Löf tests can be found in [1, 6]. Basic attempt of theory of Martin-Löf tests is to char-

[1] Probability theory and statistics deals with classes of such strings (sequences resp.).

[2] A similar approach to the program size complexity was initiated independently by Solomonoff and Chaitin.

acterize strings and sequences which, with respect to a given probability measure, possess all possible properties of stochasticity ([1], p. 313, [7]).

We deal with strings and sequences "characteristic" or "typical" for a probability measure. To avoid misunderstanding and confusion with classical terminology, we call such strings (sequences resp.) *"typical"* instead of random.

The paper is organized as follows.

Auxiliary results on Kolmogorov complexity used in the paper are summarized in Section 1.

Classes of strings (sequences resp.) which are *"typical"* for a given probability measure are introduced in Section 2. It is shown that each class of "typical" strings constitutes an immune set.

Section 3 is devoted to *pseudorandom generators.* It is shown that no pseudorandom generator can produce long "typical" strings. Pseudorandom generators with oracles capable to recognize "typical" strings are then introduced. We prove that the time complexity of such generators grows at least exponentially with respect to the length of the output strings. A relationship of these results with applied Monte–Carlo methods is mentioned.

Section 4 is devoted to *testing* of strings to be "typical". We show that this problem is undecidable. As a consequence, one of the basic problems of applied statistics, the problem of correspondence between statistical models and data, is undecidable too. After that, the Lebesgue measure is considered. We introduce the conditional probability that a string of low Kolmogorov complexity is proclaimed "typical". We show that such probabilities are bounded from below by a positive constant almost surely.

Basic results of the paper concern pseudorandom generators with oracles (Section 3) and testing of strings to be "typical" (Section 4).

NOTATION

We shortly describe the notation used in the paper.

The set $\{0, 1, 2, \ldots\}$ of natural numbers is denoted by N, the symbols n, t denote natural numbers.

The symbol Σ denotes a finite alphabet of cardinality $c \geq 2$. The symbol Σ^* denotes the set of all strings over Σ, $l(x)$ denotes the length of a string x. The symbol Σ^n denotes the set of all strings over Σ having the length n.

The set of all (infinite) sequences over Σ is denoted by Σ^∞. The symbol S_n denotes the initial segment of a sequence S having the length n. Consider a set $\mathcal{X}$ of sequences. The symbol $\mathcal{SX}$ denotes the set of all initial segments of sequences from $\mathcal{X}$, i.e. $\mathcal{SX} = \{S_n | S \in \mathcal{X} \ \& \ n \in N\}$.

The symbol Ψ denotes a universal Kolmogorov algorithm (see [1], p. 309) with inputs from the set $\Sigma^* \times N$ and with outputs from the set Σ^*.

We consider the σ-algebra of subsets of Σ^∞ generated by the set of cylinders. The symbol P denote a probability measure on Σ^∞, while P_n denotes the corresponding marginal probability measure on Σ^n. Hence

$$P_{l(x)}\{x\} = P\{S \in \Sigma^\infty | S_{l(x)} = x\}$$

holds for each string x.

The symbol $\mathbf{f}$ denotes a sequence $\langle f_0, f_1, f_2, \ldots\rangle$ of nonnegative reals.

2. KOLMOGOROV COMPLEXITY AND PROBABILITY MEASURES

Concepts and results on Kolmogorov complexity applied below are summarized in this section (see [9] for a detailed explanation).

Assume that x is a string. (Conditional) Kolmogorov complexity is defined by

$$K_\Psi(x|n) := \inf\{l(p)|p \in \Sigma^* \ \& \ \Psi(p, n) = x\}.$$

The number n represents our prior information about the string x. *The number of strings of low Kolmogorov complexity is estimated by*

$$\text{card}\ \{x \in \Sigma^*|K_\Psi(x|n) < f\} \leq \frac{c^{f+1} - 1}{c - 1} \tag{2.1}$$

(cf. Lemma 1.1. in [9]); *here f is a nonnegative real.*

Strings from the class

$$\text{Cstr}_\mathbf{f} := \{x \in \Sigma^*|K_\Psi(x|l(x)) \geq f_{l(x)}\}$$

are called $\mathbf{f}$-complex strings. If an $\mathbf{f}$-complex string x has the length n, then we have

$$K_\Psi(x|n) \geq f_n. \tag{2.2}$$

Sequences from the class

$$\text{Cseq}_\mathbf{f} := \{S \in \Sigma^\infty|\exists t\ \forall n \geq t : K_\Psi(S_n|n) \geq f_n\}$$

are called $\mathbf{f}$-*complex* sequences. Such classes were studied e. g. in [4, 5, 9].

Let S be an $\mathbf{f}$-*complex* sequence. If n is sufficiently large, then the Kolmogorov complexity $K_\Psi(S_n|n)$ of the initial segment S_n is greater than or equal to the lower bound f_n.

Sequences from the set

$$\text{Cseq}_{\mathbf{f},t} := \{S \in \Sigma^\infty|\forall n \geq t : K_\Psi(S_n|n) \geq f_n\}$$

are called $(\mathbf{f}, t)$-*complex* sequences. If $n \geq t$, then the initial segments of $(\mathbf{f}, t)$-complex sequences having the length n are $\mathbf{f}$-complex strings. Both sets $\text{Cseq}_{\mathbf{f},t}$ and $\text{Cseq}_\mathbf{f}$ are measurable.

We have

$$x \in \mathcal{S}\text{Cseq}_{\mathbf{f},t} \implies x \in \text{Cstr}_\mathbf{f} \quad \text{a. s.}, \tag{2.3}$$

as is shown in [9]. "Almost surely" means "up to a finite number of cases".

Proposition 1.1. Let P be a probability measure on Σ^∞. For each n we define

$$\pi_n(P) := \max_{x \in \Sigma^n} P_n\{x\}. \tag{2.4}$$

a. We have

$$P(\mathrm{Cseq}_{\mathbf{f},t}) > 1 - 2 \cdot \Sigma_{n=t}^{\infty} \cdot \pi_n(P) \cdot c^{f_n}. \tag{2.5}$$

b. If

$$\Sigma_{n=0}^{\infty} \pi_n(P) \cdot c^{f_n} < \infty \tag{2.6}$$

takes place, then we have

$$P(\mathrm{Cseq}_{\mathbf{f}}) = 1. \tag{2.7}$$

A dependence of the probabilities $P_n(\mathrm{Cstr}_{\mathbf{f}} \cap \Sigma^n)$ of the class of $\mathbf{f}$-complex strings having the length n as well as of the probabilities $P(\mathrm{Cseq}_{\mathbf{f}})$ of the class of $\mathbf{f}$-complex sequences on the lower bounds $\mathbf{f}$ of Kolmogorov complexity is discussed in [9].

3. ON "TYPICAL" STRINGS AND SEQUENCES

An attempt to formalize heuristic concepts like strings and sequences "typical" for a given probability measure is stated in the section. Classes of such strings and sequences are introduced. It is shown that the classes of "typical" strings constitute immune sets.

3.1. On "typical" strings

Let us think over "typical" strings. The set of strings "typical" for a given probability measure P is denoted by

$$\mathrm{Typstr}_P.$$

We assume, that

$$\mathrm{Typstr}_P \subseteq \Sigma^*,$$

because our marginal probability measures operate on subsets Σ^n of Σ^*.

Of course, we give no definition of the set of "typical" strings! Instead, we shall introduce a global property relating the set of "typical" strings with the set of $\mathbf{f}$-complex strings. Our analysis of "typical" strings is based on this property.

The property just mentioned is illustrated on ergodic measures. Consider a coding of "typical" strings. It is a well-known fact that the "typical" strings having the length n can be compressed by the coding up to the length approximately equal $C_0 \cdot n$ (where C_0 is a positive constant related with entropy of the probability measure), but not much shorter. It means that they can be compressed up to the length greater than or equal to $(C_0 - \varepsilon_0) \cdot n$, where ε_0 is a small positive constant, at least for large values of n. From the viewpoint of Kolmogorov complexity it means that the Kolmogorov complexity $K_\Psi(x|n)$ of long "typical" strings x should be greater than or equal to $(C_0 - \varepsilon_0) \cdot n + C$, where C is a constant and n is the length of x. Take ε positive such that $(C_0 - \varepsilon_0) \cdot n + C \geq \varepsilon \cdot n$ holds almost surely, put $f_n = \varepsilon \cdot n$.

We claim that $K_\Psi(x|n) \geq f_n$ should hold for the "typical" strings x of the length n, at least for large values of n. It results that long "typical" strings should be **f**-complex, i.e. that "typical" strings should be **f**-complex almost surely.

In general, our property reads:

(∗) *there is a sequence* **f** *such that almost all "typical" strings are* **f***-complex, i. e. that*

$$x \in \mathrm{Typstr}_P \implies x \in \mathrm{Cstr}_{\mathbf{f}} \tag{3.1}$$

holds almost surely[3].

Our property gives a heuristic upper bound of the set of "typical" strings. In fact, it states that

$$\mathrm{Typstr}_P \subseteq \mathrm{Cstr}_{\mathbf{f}} \cup X \quad \text{for a finite set } X. \tag{3.2}$$

Almost all probability measures used in practice are covered by (∗). For instance, ergodic measures are covered by (∗) with a single sequence **f** (see Example 1 in [9] for detailed discussion of the topic).

The set Typstr_P is usually immune[4], as is shown in

Theorem 2.1. Let (∗) take place, let $\lim_{n\to\infty} f_n = \infty$.
If the set Typstr_P is infinite, then it is immune.

Proof. Let Typstr_P be infinite. Then $\mathrm{Cstr}_{\mathbf{f}}$ is infinite too. This set is immune, which is an easy consequence of Theorem (4.3) from [1], pp. 332–333. The set $\mathrm{Cstr}_{\mathbf{f}} \setminus \mathrm{Typstr}_P$ is finite, hence Typstr_P is immune. □

3.2. On "typical" sequences

Let us turn to "typical" sequences. The set of sequences "typical" for a given probability measure P is denoted by

$$\mathrm{Typseq}_P.$$

We assume that

$$\mathrm{Typseq}_P \subseteq \Sigma^\infty,$$

because our probability measures operate on Borel subsets of Σ^∞.

We can assume, that there is a lot of "typical" sequences, i.e. that (the set Typseq_P is measurable and)

$$P(\mathrm{Typseq}_P) = 1.$$

[3]Like above, "almost surely" means "up to a finite number of cases".

[4]The set of strings is called *immune* iff it has no infinite recursively enumerable subset ([8], p. 107). If $\mathcal{X}$ is immune and $G : N \to \Sigma^*$ is a recursive function with infinite range, then $G(n)$ lies outside the set $\mathcal{X}$ for infinitely many n's.

In general, we use this condition at some specific places only.

Finally, we consider the sets

$$\mathcal{I}_{P,\mathbf{f},t} := \mathrm{Typseq}_P \cap \mathrm{Cseq}_{\mathbf{f},t} \tag{3.3}$$

of sequences which are both "typical" and $(\mathbf{f}, t)$-complex.

As a rule, probabilities of the sets $\mathcal{I}_{P,\mathbf{f},t}$ converge to probability of the set Typseq_P of typical sequences, as is shown in

Proposition 2.1. Let P be a probability measure on Σ^∞.
Assume that $\Sigma_{n=0}^{\infty}\pi_n(P)\cdot c^{f_n} < \infty$ (i.e. (2.6)) is fulfilled.
If Typseq_P is a measurable set, then we have

$$\lim_{t\to\infty} P(\mathcal{I}_{P,\mathbf{f},t}) = P(\mathrm{Typseq}_P). \tag{3.4}$$

Proof. The sets $\mathrm{Cseq}_{\mathbf{f},t}$ are measurable (see Section 1). Hence the sets $\mathcal{I}_{P,\mathbf{f},t}$ are measurable by (3.3). We have

$$\lim_{t\to\infty} P(\mathrm{Cseq}_{\mathbf{f},t}) = 1 \tag{3.5}$$

according to (2.5), hence (3.4) follows from (3.3). □

3.3. Extending results of the paper

Basic results of the paper are formulated by means of the sets Typstr_P of "typical" strings and the sets $\mathcal{I}_{P,\mathbf{f},t}$ of "typical" and $(\mathbf{f}, t)$-complex sequences. *Analogical results are true for the other sets introduced in the paper.* We does not formulate them explicitly because of space limitations. Instead, *we introduce and prove them by means of the following*

Metatheorem 2.1. (extending results of the paper)

a. All results on the sets Typstr_P hold for the sets $\mathcal{SI}_{P,\mathbf{f},t}$ in the following sense. Assume that some statement $\mathcal{M}_1$ concerning the set Typstr_P is true. Replace Typstr_P by $\mathcal{SI}_{P,\mathbf{f},t}$ in the statement, exclude the assumption $(*)$ from the statement. Then the new statement $\mathcal{M}_2$ is true.

b. All results on the sets Typstr_P hold for the sets $\mathrm{Cstr}_{\mathbf{f}}$.

c. All results on the sets $\mathcal{I}_{P,\mathbf{f},t}$ hold for the sets $\mathrm{Cseq}_{\mathbf{f},t}$.

Proof.

a. Consider the statement $\mathcal{M}_1$ without the assumption $(*)$. Put $\mathrm{Typstr}_P := \mathcal{SI}_{P,\mathbf{f},t}$. Then $(*)$ takes place, as follows from (2.3) and (3.3). Hence the statement $\mathcal{M}_2$ is true.

b. Clearly, (*) is true for $\mathrm{Typstr}_P := \mathrm{Cstr}_{\mathbf{f}}$.

c. Put $\mathrm{Typseq}_P := \mathrm{Cseq}_{\mathbf{f}}$; then the equality $\mathcal{I}_{P,\mathbf{f},t} = \mathrm{Cseq}_{\mathbf{f},t}$ takes place. □

It follows from Metatheorem 2.1 b, c, that our results on sets Typstr_P and $\mathcal{I}_{P,\mathbf{f},t}$ can be transformed into results on sets of **f**-complex strings and (**f**, t)-complex sequences, i. e. into those formulated purely in terms of Kolmogorov complexity theory.

4. ON PSEUDORANDOM GENERATORS

We show in this section that no pseudorandom generator can produce long strings "typical" for probability measures used in practice. Pseudorandom generators with oracles capable to recognize "typical" strings are introduced. We show that the time complexity of such generators grows at least exponentially with the length of the output. A relationship of these results with applied Monte–Carlo methods is mentioned at the end of the section.

Assume that G is a *pseudorandom generator*. It means that G represents an effectively computable function ascribing strings from Σ^* to natural numbers. According to Church's thesis (see [1], p. 92) we can suppose that G is a recursive function. Moreover, we assume that the length of the output string $G(n)$ equals the value of the input number n.

Our pseudorandom generators are "purely deterministic" ones. Hence random side affects, like random seeds performed by means of physical entities at the beginning of the process or periodically in the course of the process, are not considered here.

No pseudorandom generator can produce long "typical" strings, as follows from

Theorem 3.1. Let (*) take place, let $\lim_{n\to\infty} f_n = \infty$. Then we have

$$G(n) \notin \mathrm{Typstr}_P \quad \text{a. s.}$$

Proof. Clearly, the conditional Kolmogorov complexities $K_\Psi(G(n)|n)$ are bounded from above by some constant. Moreover $\lim_{n\to\infty} f_n = \infty$, hence there is a constant $n(\Psi, \mathbf{f}, G)$ such that

$$K_\Psi(G(n)|n) < f_n \tag{4.1}$$

is true for all $n \geq n(\Psi, \mathbf{f}, G)$.

There is some n_0 such that (3.1) is true for all strings x with $l(x) \geq n_0$, as follows from (*). Consider a "typical" string x with $n := l(x) \geq \max\{n(\Psi, \mathbf{f}, G), n_0\}$. It suffices to prove that $G(n) \neq x$. The string x lies in $\mathrm{Cstr}_{\mathbf{f}}$ by (3.1), hence (2.2) is true, so that $G(n) \neq x$ by (4.1). □

No pseudorandom generator can produce long initial segments of sequences from the set $\mathcal{SI}_{P,\mathbf{f},t}$ (by Theorem 3.1 and Metatheorem 2.1 a). Moreover, the length of these segments is independent of the probability measure under consideration, as follows from

Theorem 3.2. Let $\lim_{n\to\infty} f_n = \infty$.
Then there is a constant $n(\Psi, \mathbf{f}, G, t)$ independent of P such that we have

$$G(n) \notin \mathcal{SI}_{P,\mathbf{f},t} \quad \forall\, n \geq n(\Psi, \mathbf{f}, G, t).$$

P r o o f. Choose n_0 equal t in the proof of Theorem 3.1. Apply Metatheorem 2.1 a. to the proof. □

Assume for a moment that there is a lot of "typical" sequences, i. e. that

$$P(\mathrm{Typseq}_P) = 1. \tag{4.2}$$

Moreover, suppose that the probabilities $\pi_n(P)$ of a most probable string of the length n are of $O(n^{-2})$ type (there may be several most probable strings). Majority of probability measures used in practice satisfies this condition, like the ergodic measures do (see Example 1 in [9] for details). Finally, take a sequence $\mathbf{f}$ of lower bounds converging slowly to infinity, e. g. like $\log_c^{1/2} n$ does. Then $\lim_{t\to\infty} P(\mathrm{Cseq}_{\mathbf{f},t}) = 1$ holds, as follows from (2.5).

Let ε be a small positive constant. Consider a probability measure P satisfying

$$P(\mathrm{Cseq}_{\mathbf{f},t}) \geq 1 - \varepsilon. \tag{4.3}$$

The class of such measures, say $\mathcal{P}$, is very large, at least for large t. We have

$$P_n(\mathcal{SI}_{P,\mathbf{f},t} \cap \Sigma^n) \geq 1 - \varepsilon$$

for all n (by (4.2), (3.3) and (4.3)). Therefore, majority of strings "typical" for any probability measure from the class $\mathcal{P}$ lies in the set $\mathcal{SI}_{P,\mathbf{f},t}$. But no pseudorandom generator can produce long strings from this large and heterogeneous set by Theorem 3.2. It reflects our intuitive feeling that no pseudorandom generator can produce long strings "typical" for any of the probability measures used in practice.

Let us turn to pseudorandom generators with oracles. Before going ahead, we limit the class of probability measures under consideration. Nevertheless, it remains substantially general for practical purposes.

In the rest of the section we assume, that

$$\varepsilon \cdot n \leq f_n \quad \text{a. s.}$$

takes place, where $0 < \varepsilon < 1$ *is a positive real.* Majority of probability measures used in practice, e. g. ergodic measures, is taken into account in this case, as follows from Example 1 in [9]. (But different measures may be covered by different ε's.)

It was shown above, that the pseudorandom generators fail to produce long "typical" strings. For this reason *we add an oracle to a pseudorandom generator,* namely the oracle capable to recognize "typical" strings. We investigate the time complexity of such generators.

A Turing machine equipped by an oracle is considered. Inputs of our machine are natural numbers, outputs are strings. Starting on the input n, the machine works

as follows. It subsequently generates auxiliary strings. Whenever some auxiliary string is obtained, the machine asks the oracle whether it is a "typical" string of the length n, or not. If the answer is positive, the string is placed on the output tape of the machine and the machine halts. If the answer is negative, new auxiliary string is generated, etc. *Each of Turing machines just described is called a pseudorandom generator with oracle.*

Clearly, producing of an auxiliary string takes at least one unit of time. Therefore, if i auxiliary strings are generated until the machine halts, then the time complexity of the procedure is greater than or equal to i.

Theorem 3.3. Let $(*)$ take place. Assume that $\varepsilon \cdot n \leq f_n$ is true for almost all n, where $0 < \varepsilon < 1$.
Then the time complexity of a pseudorandom generator with oracle grows at least exponentially with the length of the output.

Proof. Consider an input n of the generator. Suppose that the auxiliary strings $x_1, x_2, \ldots, x_i$ were generated until the machine halts. Hence the "typical" string $x_i \in \mathrm{Typstr}_P$ is produced as the output. If n is sufficiently large, then $\varepsilon \cdot n \leq f_n$ and $x_i \in \mathrm{Cstr}_{\mathbf{f}}$ are true. So that we have

$$\varepsilon \cdot n \leq f_n \leq K_\Psi(x_i|n)$$

by (2.2). Moreover, there is a constant C such that

$$K_\Psi(x_i|n) \leq l(i) + C$$

takes place. Therefore, we have

$$\varepsilon \cdot n \leq l(i) + C \leq \log_c(i) + C. \tag{4.4}$$

If n is sufficiently large, then $\frac{\varepsilon}{2} \cdot n \leq \log_c(i)$ is valid by (4.4), i.e. $c^{\frac{\varepsilon}{2} \cdot n} \leq i$ is true. The time complexity of producing x_i is greater than or equal to i, which finishes the proof. □

An exponential upper bound of time complexity can be obtained too. Consider a pseudorandom generator with oracle performing the following steps. Starting on the input n, it subsequently generates strings of the length n in a prescribed lexicographical order. Whenever it generates a string, it asks the oracle. If the string is "typical", the generator outputs the string and halts. Otherwise it generates the next string. Clearly, generating of one string and asking the oracle once can be done in a polynomial amount of time. Hence the whole procedure of producing of a "typical" string takes at most $p(n) \cdot \mathrm{card}\,\Sigma^n = p(n) \cdot c^n$ units of time, where $p(\cdot)$ is a polynomial. Finally, $p(n) \cdot c^n \leq c^{2n}$ holds almost surely.

Our considerations on pseudorandom generators turn some light on applied Monte–Carlo methods. They suggest that safe "purely deterministic" pseudorandom generators cannot be obtained in the frame of contemporary computer science.

5. ON TESTING STRINGS TO BE "TYPICAL"

Tests proclaiming some strings as "typical" are considered in this section. We show that the problem of testing the strings to be "typical" is undecidable. As a consequence, one of the fundamental problems of applied statistics, the problem of correspondence between statistical models and data, is undecidable. Finally, we prove that the conditional probability that a string which is not **f**-complex is proclaimed "typical" is bounded from below by a positive constant.

Consider a probability measure P on Σ^∞.

A test is a recursively enumerable[5] *set of strings,*[6] i. e. a recursively enumerable subset of Σ^*. *It is denoted by*

$$T_P.$$

We assume that just the strings from the test T_P are proclaimed "typical" for the probability measure P.

No test T_P can proclaim exactly the "typical" strings as "typical", which follows from

Theorem 4.1. Let $(*)$ take place, let $\lim_{n\to\infty} f_n = \infty$.
If T_P is an infinite test, then $T_P \setminus \mathrm{Typstr}_P$ is an infinite set.

Proof. If the set Typstr_P is finite, then the assertion the theorem is true. Let Typstr_P be infinite. Then it constitutes an immune set by Theorem 2.1. Hence $T_P \setminus \mathrm{Typstr}_P$ is infinite. □

Theorem 4.1 *means that the problem of testing the strings to be "typical" is undecidable by means of Turing machines.*

A fundamental problem of applied statistics consists in answering of the question: "Does the collection of the observed data correspond to a given probability measure?" We show that this problem is undecidable. The question should be answered in an effective manner, safely and for infinitely many collections of input data. Therefore, it can be formalized by: "Does an infinite test T_P exist such that $T_P \subseteq \mathrm{Typstr}_P$ takes place?" Theorem 4.1 shows that the answer is negative.

It will be interesting to estimate the conditional probability

$$P_n(T_P \mid \Sigma^n \setminus \mathrm{Typstr}_P)$$

that some string which is not "typical" is proclaimed "typical" by the test. Unfortunately, such an estimate cannot be obtained by means of the assumptions stated above. Namely, the property $(*)$ gives no lower bound of the set Typstr_P of "typical" strings, as follows from the equivalent condition (3.2) on page 751. It means that we have at disposal no upper bound of the probability $P_n(\Sigma^n \setminus \mathrm{Typstr}_P)$.

[5]This is the weakest constructive restriction; see [1], p. 314, Comment b) for details.

[6]Traditionally, a Martin-Löf test for randomness is a specific recursively enumerable subset of $\Sigma^* \times N$. Our approach is different. Classical approach deals with testing a hypothesis about a statistical model and data, while we are dealing with problem of correspondence between statistical models and data.

In the rest of the section we assume, that P is the Lebesgue measure. Hence

$$P_{l(x)}\{x\} = c^{-l(x)}$$

is true for each string $x \in \Sigma^*$.

Clearly, strings of low Kolmogorov complexity are too regular to be "typical" for the Lebesgue measure. Therefore, the conditional probability

$$P_n(T_P \mid \Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) \tag{5.1}$$

that some string which is not **f**-complex is proclaimed "typical" is of interest. We show that if a lot of strings is proclaimed "typical" by a test, then the probabilities (5.1) are bounded from below by a positive constant almost surely.

Theorem 4.2. Assume that P is the Lebesgue measure on Σ^∞, $\lim_{n\to\infty} f_n = \infty$. Moreover, let

$$\liminf_{n\to\infty} P_n(T_P \cap \Sigma^n) > 0. \tag{5.2}$$

Then

$$\liminf_{n\to\infty} P_n(T_P \mid \Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) > 0. \tag{5.3}$$

Proof.

1. There is some $n_0 \in N$ such that the sets $\Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}$ are nonempty for all $n \geq n_0$, because $\lim_{n\to\infty} f_n = \infty$. From now on, let $n \geq n_0$.

   The probability (5.1) equals

   $$\frac{\mathrm{card}\,[T_P \cap (\Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}})]}{\mathrm{card}\,[\Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}]}. \tag{5.4}$$

   The value of the denominator in (5.4) is bounded from above by $\frac{c^{f_n+1}-1}{c-1}$, as follows from (2.1).

2. There is a partial recursive function F from the set $\Sigma^* \times N$ into the set Σ^* satisfying the following properties. The domain of the function $F(\cdot, n)$ contains exactly $\mathrm{card}\,[T_P \cap \Sigma^n]$ shortest strings from the set Σ^*. The range of the function coincides with $T_P \cap \Sigma^n$. Hence for each string z from the domain of the function $F(\cdot, n)$ we have

   $$K_\Psi(F(z, n)|n) \leq l(z) + C, \tag{5.5}$$

   where C is a constant.

   Let us put

   $$r_n := \mathrm{card}\,[T_P \cap \Sigma^n]. \tag{5.6}$$

   We introduce auxiliary sets

   $$\mathcal{X}_n := \{z \in \Sigma^* | l(z) + C < f_n\}. \tag{5.7}$$

3. We show that if z lies in $\mathcal{X}_n$ and $F(z, n)$ is defined, then $F(z, n)$ lies in the set $T_P \cap (\Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}})$. This fact enables us to obtain a lower bound of the numerator in (5.4). Consider the string z. The string $F(z, n)$ lies in the set $T_P \cap \Sigma^n$ according to definition of the function F. At the same time we have $K_\Psi(F(z, n)|n) < f_n$ by (5.5) and (5.7), hence $F(z, n)$ does not lie in the set $\mathrm{Cstr}_{\mathbf{f}}$ by (2.2).

4. Two cases are considered below, $r_n \geq \operatorname{card} \mathcal{X}_n$ and $r_n < \operatorname{card} \mathcal{X}_n$.

Case a. First, let $r_n \geq \operatorname{card} \mathcal{X}_n$. Then $F(z, n)$ is defined for all $z \in \mathcal{X}_n$. Hence the value of the numerator in (5.4) is bounded from below by $\operatorname{card} \mathcal{X}_n$, which is at the same time bounded from below by

$$\frac{c^{f_n - C} - 1}{c - 1}. \tag{5.8}$$

Hence the value of the numerator in (5.4) is bounded from below by the value of (5.8). Therefore, we have

$$P_n(T_P \mid \Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) \geq \frac{c^{f_n - C} - 1}{c^{f_n + 1} - 1} \tag{5.9}$$

by part 1. of the proof.

Case b. Assume that $r_n < \operatorname{card} \mathcal{X}_n$ is true. Then there are r_n strings z in $\mathcal{X}_n$ such that $F(z, n)$ is defined. Let $z_1, \ldots, z_{r_n}$ be that strings. We have

$$\{F(z_i, n) | i = 1, \ldots, r_n\} = T_P \cap \Sigma^n \tag{5.10}$$

in this case according to definition of the function F and (5.6). Moreover, part 3. of the proof gives

$$T_P \cap (\Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) \supseteq \{F(z_i, n) | i = 1, \ldots, r_n\},$$

which together (5.10) proves that

$$T_P \cap (\Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) = T_P \cap \Sigma^n$$

takes place. Therefore, we have

$$P_n(T_P \mid \Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) \geq P_n(T_P \cap \Sigma^n). \tag{5.11}$$

5. Two lower bounds (5.9) and (5.11) of the probability (5.1) show, that we have

$$P_n(T_P \mid \Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) \geq \min \left\{ \frac{c^{f_n - C} - 1}{c^{f_n + 1} - 1}, P_n(T_P \cap \Sigma^n) \right\}. \tag{5.12}$$

This inequality together with $\lim_{n \to \infty} f_n = \infty$ and (5.2) show, that (5.3) is true. □

The assumption "there is a lot of strings proclaimed 'typical by the test' formalized by (5.2) can be replaced by

$$\lim_{n\to\infty} P_n(T_P \cap \Sigma^n) = 1.$$

Then we have

$$\liminf_{n\to\infty} P_n(T_P \mid \Sigma^n \setminus \mathrm{Cstr}_{\mathbf{f}}) \geq c^{-C-1}. \tag{5.13}$$

as follows from (5.12). If the universal Kolmogorov algorithm is chosen appropriately, then the constant C equals 1 and the lower bound in (5.13) equals c^{-2}.

ACKNOWLEDGEMENT

This work has been partially supported by the Grant Agency of the Czech Republic under Grant 102/99/1564 and by Prodactool IST–1999–12058.

(Received October 30, 2000.)

REFERENCES

[1] C. Calude: Theories of Computational Complexity. North–Holland, Amsterdam 1988.
[2] T. L. Fine: Theories of Probability – an Examination of Foundations. Academic Press, New York 1973.
[3] A. N. Kolmogorov: Three approaches to the quantitative definition of information. Problems Inform. Transmission *1* (1965), 1, 1–7.
[4] I. Kramosil and J. Šindelář: A note on the law of iterated logarithm from the viewpoint of Kolmogorov program complexity. Problems Control Inform. Theory *16* (1987), 6, 399–409.
[5] I. Kramosil and J. Šindelář: On pseudo-random sequences and their relation to a class of stochastical laws. Kybernetika *28* (1991), 6, 383–391.
[6] M. Li and P. Vitayi: Introduction to Kolmogorov Complexity and its Applications. Springer, New York 1997.
[7] P. Martin-Löf: The definition of random sequences. Inform. and Control *9* (1966), 602–619.
[8] H. Rogers, Jr.: Theory of Recursive Functions and Effective Computability. McGraw–Hill, New York 1967.
[9] J. Šindelář and P. Boček: Kolmogorov complexity and probability measures. Kybernetika *38* (2002), 729–745.

RNDr. Jan Šindelář, CSc. and Mgr. Pavel Boček, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: sindelar@utia.cs.cz, bocek@utia.cas.cz