Algorithmic complexity and stochastic properties of finite binary sequences

V.V. V'yugin
Institute for Information Transmission Problems,
Russian Academy of Sciences,
Bol'shoi Karetnyi per. 19, Moscow GSP-4, 101447, Russia

Abstract

This paper is a survey of concepts and results related to simple Kolmogorov complexity, prefix complexity and resource bounded complexity. We consider also a new type of complexity - statistical complexity closely related to mathematical statistics.

Unlike other discoverers of algorithmic complexity A.N.Kolmogorov's leading motive was developing on its basis a mathematical theory more adequately substantiating applications of the probability theory, mathematical statistics and information theory. Kolmogorov wanted to deduce properties of random object from its complexity characteristics without use the notion of probability. In the first part of this paper we present several results in this direction.

Though the following development of algorithmic complexity and randomness was different algorithmic complexity has successful applications in the traditional probabilistic framework. The second part of the paper is a survey of applications to parameters estimation and definition of Bernoulli sequences.

All considerations have finite combinatorial character.

1 Introduction

In the 60-s A.N.Kolmogorov [13] (see also [14]) proposed a program of developing the theory of information and the theory of probability based on the general theory of algorithms. Under this approach, the principal concept is that of complexity, or entropy, of finite objects. By means of it Kolmogorov defined the concept of amount of information in a finite object about another finite object. The need to define randomness for individual objects was the leading motive to introduce the notion of complexity. A thorough historical analysis of Kolmogorov's ideas is given in [7].

Independently, R.J.Solomonoff [35] published analogous ideas on the existence of an optimal, to within an additive constant, way of encoding of finite objects (but he did not introduced complexity as an original notion). Similar ideas were published by G.Chaitin [2], [3].

Since that time several surveys and books on related topics were published. We distinguish [47], [45], [17], [25], [37], [39] and monographs [6], [18], [27]. This paper contains well known results as well as results not covered by these publications.

The concept of algorithmic entropy, or complexity, is applicable to finite objects, such as words in a finite alphabet, finite sequences of integer numbers, etc. The complexity $K(x)$ of a finite object x equals to the length of the shortest binary program that describes x. This is a principal difference from the concept of probabilistic entropy

$$H(\xi) = -\sum_x p(\xi = x) \log p(\xi = x),$$

which is applicable to "random" objects (random variables) or, equivalently, to probability distributions in classes of objects. The probabilistic entropy $H(\xi)$ is the quantity of information sufficient to describe random variable ξ on the average.

Assertions about probabilities are usually interpreted statistically, so in practice, the definition of $H(\xi)$ can be used when applied to bulks of objects large enough for statistical laws to manifest themselves. The need to use concepts of entropy and mutual information (defined via entropy) in case of individual objects not considered as realizations of random variables requires theoretical study of the corresponding concept of entropy - complexity.

Kolmogorov [13], [14] proposed to develop probability theory on the basis of information theory. This means that the algorithmic complexity is the leading concept and that laws of probability theory, or asymptotic properties of convergence to special limiting distributions and other probabilistic properties must hold for individual objects with maximal possible value of their complexity.

By the very essence of this discipline, the foundations of information theory have a finite combinatorial character [12].

2 Kolmogorov complexity

Kolmogorov [13] proposes to measure the conditional complexity of a finite object x given a finite object y by the length of the shortest sequence p (a program for computing x) which consists of 0s and 1s and makes it possible to reconstruct x given y. Mathematically, this is explicated as

$$K_B(x|y) = \min\{l(p) \mid B(p, y) = x\},$$

where $l(p)$ is the length of the sequence p, $B(p,y)$ is some function, maybe partial - a way of decoding, for which there is an algorithm computing its values (we mean also that $\min \emptyset = \infty$).

This definition of complexity is very natural, but depends on the choice of the computable function $B(p,y)$, a "mode of description" of finite objects [39]. But using the idea of universality from the general theory of algorithms Kolmogorov managed to define the concept of complexity independent of the choice of the mode of description $B(p,y)$. So, the notion of complexity becomes an intrinsic property of finite object independent of the ways of its description.

We need some elements of the general theory of algorithms. This theory is systematically treated in Rogers [32]. We make only some remarks in this connection. Algorithms are fed with constructive (finite) objects and produce also constructive objects. A thorough analysis of all these notions is given in Uspensky and Semenov [38]. We will consider the following sets of constructive objects – the set Ξ of all finite sequences consisting of 0s and 1s (the empty sequence $\emptyset$ is also considered), the sets $\mathbb{Z}$ and $\mathbb{N}$ of all integer numbers and all non-negative integer numbers, respectively, the set $\mathbb{Q}$ of all rational numbers (but not the set of all real numbers). We also will generate additional sets of constructive objects – sets of all finite subsets of any previously defined set and Cartesian products of such sets. We will consider constructive real numbers as follows. A real number θ is called computable if there is an algorithm computing some its rational approximation r such that $|\theta - r| < \epsilon$ given positive rational ϵ.

As usual, we consider the natural ordering of the set Ξ such that all sequences of the same length are ordered lexicographically and all sequences of the smaller length precede to all sequences of greater length. The natural structure of the set Ξ is determined by the relation $x \subseteq y$ which means that the sequence y continues the sequence x, sequences x and y are incomparable if $x \not\subseteq y$ and $y \not\subseteq x$. We will consider also a discrete structure on Ξ. When it is convenient we will identify Ξ and $\mathbb{N}$ according to their natural orderings.

If $x \in \Xi$ then $l(x)$ is the length of the sequence x, for each $1 \leq i \leq l(x)$ x_i is the i-th bit of x. For any two sequences $x = x_1 \dots x_n$ and $y = y_1 \dots y_m$ we denote $xy = x_1 \dots x_n y_1 \dots y_m$ the concatenation of x and y. We also define $\overline{z} = z_1 z_1 \dots z_n z_n$ for each sequence $z = z_1 \dots z_n$ of the length n. Let Ξ_n be the set of all finite binary sequences of the length n.

We encode the ordered pair of finite binary sequences (x,y) by the sequence $\overline{l(x)}01xy$. Evidently, there is an algorithm computing x and y given the code of the pair (x,y). So we can identify $\Xi \times \Xi$ and Ξ.

The integer part of a number r is denoted as $\lfloor r \rfloor$, $\lceil r \rceil$ is the least integer number n such that $n \geq r$. $\log r$ denotes the binary logarithm of r, by $\ln r$ we mean the natural logarithm of r.

There are several explications of the intuitive idea of an algorithm. We will not need the details of any of them, besides the conventional one (with the exception of Section 14). Following tradition, we call partial (i.e. not necessary everywhere defined) computable functions *partial recursive* (in accordance

with the well-known explication, see [32]). Everywhere defined partial recursive functions are called *recursive* functions. A set is called *recursively enumerable* if it is the range of some partial recursive function.

The following well-known theorem on the existence of an universal function is the main distinctive feature of the general theory of algorithms. Let X and Y be sets of constructive objects. This theorem asserts that there exists a partial recursive function $U(i,x)$ called universal such that each partial recursive function f from X to Y can be represented as $f(x) = U(i,x)$ for some i. The proof is based on the possibility to arrange all programs (which are words in some finite alphabet; the meaningless words are considered as programs of everywhere undefined functions) like elements of Ξ. Then the algorithm computing $U(i,x)$ goes to the program with ordinal number i (for convenience we can identify the program and its ordinal number) and apply this program to input x by simulating the work of this program on x. The explicit construction is based on a specific formalization of the concept of algorithms, see [38].

Kolmogorov's definition of complexity is based on the *invariance property*, which says that the notion of algorithmic complexity can be made independent of the choice of the mode of description.

Theorem 1 *There exists a partial recursive optimal function $A(p,y)$ such that for each partial recursive function $B(p,y)$ a positive integer constant c exists such that inequality*

$$K_A(x|y) \le K_B(x|y) + c$$

holds.

Proof. As follows from the universality of $U(i,x,y)$, for each partial recursive function $B(p,y)$ there exists a program q such that $B(p,y) = U(q,p,y)$ (recall, that we identify Ξ and $\mathbb{N}$ then it is convenient). We can define the mode of description $A(u,y)$ such that for any p,q,y

$$A(\overline{l(q)}01qp, y) = U(q,p,y).$$

Then $K_A(x|y) \le K_B(x|y) + l(q) + 2\log l(q) + 2$ for all x. This means that $K_A(x|y) \le K_B(x|y) + c$ for any partial recursive function B. □

The code $\overline{l(q)}01qp$ is similar to a self-extracting archive. The algorithm A when fed with a complex code $\overline{l(q)}01qp$ simulates the work of a program q (of the algorithm B) on its inputs p and y.

In the sequel $f(x_1,\dots,x_n) \le^+ g(x_1,\dots,x_n)$ will mean that there exists a non-negative constant c such that

$$f(x_1,\dots,x_n) \le g(x_1,\dots,x_n) + c$$

for all $x_1,\dots,x_n$. If $f(x_1,\dots,x_n) \le^+ g(x_1,\dots,x_n)$ and $g(x_1,\dots,x_n) \le^+ f(x_1,\dots,x_n)$ we write $f(x_1,\dots,x_n) =^+ g(x_1,\dots,x_n)$.

Analogously $f(x_1,\dots,x_n) \leq^{\cdot} g(x_1,\dots,x_n)$ means that there exists a positive constant c such that $f(x_1,\dots,x_n) \leq cg(x_1,\dots,x_n)$ for all $x_1,\dots,x_n$. $f(x_1,\dots,x_n) =^{\cdot} g(x_1,\dots,x_n)$ means that $f(x_1,\dots,x_n) \leq^{\cdot} g(x_1,\dots,x_n)$ and $f(x_1,\dots,x_n) \geq^{\cdot} g(x_1,\dots,x_n)$. In particular, $f(x_1,\dots,x_n) \leq^{\cdot} 1$ means that there is a positive constant c such that $f(x_1,\dots,x_n) \leq^{\cdot} c$ for all $x_1,\dots,x_n$.

Any function $A(p,y)$ satisfying this Theorem is called *optimal* (mode of description). As follows from this theorem $K_A(x|y) =^+ K_B(x|y)$ for any two optimal functions A and B. Hence, the complexity of finite object x does not depend in any way of the decoding of finite objects up to additive constant. This is an intrinsic property of x.

Let us fix some optimal function $A(p,y)$ and denote by $K(x|y)$ the corresponding complexity function $K_A(x,y)$. We call it *the conditional complexity* of x. The function $K(x) = K(x|\emptyset)$ is called *(unconditional) complexity* of x.

We have $K(x) \leq^+ l(x)$, since we can consider the mode of description $F(x,y) = x$. Obviously,

$$K(x) \leq^+ K_F(x|\emptyset) = l(x).$$

By definition the complexity of an object $\alpha = \alpha_1 \dots \alpha_n$ having some regularity may be essentially less than its length. For instance,

$$K(\alpha_1 \dots \alpha_n) \leq^+ \log n \text{ and } K(\alpha_1 \dots \alpha_n|n) \leq^+ 0$$

if there is an algorithm computing i-th bit of α given i.

Comparing the modes of description we obtain $K(x|y) \leq^+ K(x)$. On the other hand, we have

$$K(x) \leq^+ K(x|y) + K(y) + 2\log K(y),$$

since we can define a partial recursive function B such that

$$B(\overline{l(q)}01qp,\emptyset) = A(p,q) = x,$$

where A is the optimal mode of description, p is the shortest program for computing x given y, and q is the shortest program for computing y. From this it follows that

$$K(x) \leq^+ K_B(x|\emptyset) \leq^+ l(p)+l(q)+2\log l(q)+2 = K(x|y)+K(y)+2\log K(y)+2.$$

This inequality implies

$$K(x) \leq^+ K(x|n) + \log n + 2\log\log n$$

for each $x \in \Xi_n$. For any recursive function $\psi(x)$ we have

$$K(\psi(x)) \leq^+ K(x),$$

since we can consider the function $B(p, y) = \psi(A(p, y))$ as a mode of description, where A is the optimal function.

Comparing modes of description we obtain

$$K(x|y) \leq^+ K(x|\psi(y)) \tag{1}$$

for any recursive function $\psi(y)$. Indeed, a function $B(p, y) = A(p, \psi(y))$ defines the needed mode of description.

As follows from the definition, the complexity function $K(x)$ is unbounded, i.e. $\liminf_{n\to\infty} K(n) = \infty$. The function $K(x)$ is not computable. Moreover (see [47]),

Theorem 2 *There is no unbounded computable function $\psi(n)$ such that $K(n) \geq \psi(n)$ for all n.*

Proof. Suppose that the contrary statement holds. Since $\psi(n)$ is unbounded, the function $t(m) = \min\{n \mid \psi(n) \geq m\}$ is also computable. By definition of t we have $K(t(m)) \geq \psi(t(m)) \geq m$. On the other hand, $K(t(m)) \leq^+ K(m)$. We have also, $K(m) \leq^+ l(m) =^+ \log m$. Hence, $m \leq \log m + c$ for all m, where c is some constant. This contradiction proves the statement. □

Nevertheless, $K(x)$ possesses the property of enumerability from above. As follows from the definition, the set of all pairs (m, x), such that $m > K(x)$ (or $m \geq K(x)$) and m is an integer number, is recursively enumerable. In other words, if $m > K(x)$ (or $m \geq K(x)$) this fact will sooner or later be learned, whereas, if $m < K(x)$ we may be for ever uncertain.

Kolmogorov's following theorem [15] shows that Shannon's entropy is a computable upper bound for algorithmic complexity $K(x)$.

Let a binary sequence $x = x(1)x(2)\dots x(n)$ be divided in n blocks $x(i)$ of equal length m. Let p_k be the frequency of occurrence in x of the block with ordinal number k (under the lexicographical ordering of all binary sequences of the length m), $k = 1, 2, \dots, 2^m$.

Let $D^n_{p_1,p_2,\dots p_{2^m}}$ be a set of all sequences y of the length nm such that any block of the length m with an ordinal number k occurs in y with the frequency p_k, where $k = 1, 2, \dots 2^m$.

The Shannon entropy of a random block of the length m is defined as

$$H = -\sum_{k=1}^{2^m} p_k \log p_k.$$

Theorem 3 *For any sequence $x \in D^n_{p_1,p_2,\dots p_{2^m}}$*

$$K(x) \leq n\left(-\sum_{k=1}^{2^m} p_k \log p_k + \alpha(n)\right)$$

holds, where $\alpha(n) = C(m)\frac{\log n}{n}$.

Proof. Let x have an ordinal number t among all elements of the set $D^n_{p_1,p_2,\ldots p_{2^m}}$. The total number of occurrences in x of a block with ordinal number k is $s_k = p_k n$. We have $\sum_{k=1}^{2^m} s_k = n$. Then

$$K(x) \leq^+ K(x|D^n_{p_1,p_2,\ldots p_{2^m}}) + K(D^n_{p_1,p_2,\ldots p_{2^m}}) + 2\log K(D^n_{p_1,p_2,\ldots p_{2^m}}) \leq^+ l(t) + 2(\log s_1 + \ldots + \log s_{2^m}).$$

By definition $s_k \leq n$ for each k. The number t cannot exceed the total number of elements of $D^n_{p_1,p_2,\ldots p_{2^m}}$, which is equal to $\binom{n}{s_1 \ldots s_{2^m}}$.

By Stirling's formula $n! = \sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{\theta_n}{12n}}$, where $|\theta_n| \leq 1$, we obtain

$$K(x) \leq^+ nH + 2^{m+1}(\log n + c),$$

where c is a constant. □

In the particular case, when the length m of the block is equal to 1 and k is the total number of 1s in x of the length n, we have

$$K(x|n,k) \leq^+ \log \binom{n}{k} =^+ nH(\frac{k}{n}) - \frac{1}{2}\log\frac{k(n-k)}{n},$$

where $H(p) = -p\log p - (1-p)\log(1-p)$.

As will be noted in Section 4 $K(x) \geq n(-\sum_{k=1}^{2^m} p_k \log p_k + \alpha(n)) - r - 1$ for a portion $(1 - 2^{-r})$ of all sequences $x \in D^n_{p_1,p_2,\ldots p_{2^m}}$, see (8).

3 Information, I

Using the concept of algorithmic complexity $K(x|y)$ Kolmogorov [13], [14] defined the *amount of information* in a sequence y about a sequence x analogously to the probabilistic notion

$$I(\xi : \theta) = H(\theta) - H(\theta|\xi),$$

namely,

$$I(y : x) = K(x) - K(x|y).$$

The value $K(x)$ can be interpreted as the amount of information needed to produce x, and $K(x|y)$ can be interpreted as the amount of information which must be added to y to produce x. So we interpret the difference between these two quantities as the amount of information in y about x.

By definition $I(x : y) \geq^+ 0$, $I(y : x) \leq^+ K(x)$ and $I(x : x) =^+ K(x)$.

In contrast to the probabilistic concept, function $I(x : y)$ is not commutative, even up to an additive constant. To show this we reproduce here an example from [47]. For each m we can find x of length m such that $K(x|m) \geq m$.

Indeed, if such x does not exist then for each y of length m there exists p such that $A(p,m) = y$ and $l(p) < m$. The number of such p is $\leq 2^m - 1$. This is in contradiction with the fact that the total number of sequences of length m is equal to 2^m. Analogously, there exist arbitrarily large m such that $K(m) \geq l(m)$. It is easy to see that $K(l(z)|z) \leq^+ 0$. For any such m and any x of length m such that $K(x|m) \geq m$ we obtain

$$I(x : m) = K(m) - K(m|x) \geq^+ l(m), \tag{2}$$

and

$$I(m : x) = K(x) - K(x|m) \leq^+ l(x) - m = 0. \tag{3}$$

The function $I(y : x)$ is commutative up to the logarithm of $K(x, y)$.

It is proved in [47] that

$$|I(y : x) - (K(x) + K(y) - K(x, y))| = O(\log K(x, y)).$$

From this it follows

$$|I(y : x) - I(x : y)| = O(\log K(x, y)).$$

By this reason $I(y : x)$ is called the *mutual information* of y and x.

These inequalities are analogous to the equalities

$$I(\xi : \theta) = I(\theta : \xi)$$

and

$$I(\xi : \theta) = H(\xi) + H(\theta) - H(\xi, \theta),$$

which hold for the probabilistic mutual information and entropy.

We will prove analogous properties for a slightly different notion of the amount of information considered in Section 6.

The following natural problem was considered by Gács and Körner [9]: can we materialize the mutual information $I(y : x)$? More correctly, we shall say that z represents some *common* information of x and y if $K(z|x) \approx 0$ and $K(z|y) \approx 0$, where by $\approx$ we mean equality up to an additive term $O(K(z, x))$, $O(K(z, y))$ or $O(K(x, y, z))$. In this case

$$I(x : z) \approx K(z) \tag{4}$$

and

$$I(y : z) \approx K(z). \tag{5}$$

Then the question was whether

$$K(z) \approx I(x : y) \tag{6}$$

for some common information z of x and y? In [9] a negative answer was obtained using probabilistic methods. An. Muchnik [29] obtained a complexity-theoretic proof of the result of Gács and Körner and gave its generalization. We present the corresponding result without proof.

We call a pair (x, y) bad if there is no z satisfying (4)-(6). The following theorem shows that there are bad (x, y) with arbitrary large $K(x)$, $K(y)$ and arbitrary ratios $I(x : y)/K(y)$, $I(x : y)/K(x)$ belonging to the interval $(0, 1)$.

Theorem 4 *Let $0 \leq \alpha, \beta \leq 1$ and $0 < \gamma < \alpha, \beta$. Then for all sufficiently large j there is a bad pair (x_j, y_j) such that $K(x_j) =^+ \alpha j$, $K(y_j) =^+ \beta j$, $|I(x_j : y_j) - \gamma j| \leq^+ 14 \log j$.*

4 Randomness of finite objects

Kolmogorov wanted to use algorithmic complexity to eliminate the need for a direct interpretation of probabilities. He proposed the notion of randomness of an element x with respect to a finite set D containing it. Given D we can effectively generate all its elements. The corresponding ordinal number under this generation can serve as a code of any $x \in D$. Therefore, we need $\leq \lceil \log \#D \rceil$ bits to encode any $x \in D$, where $\#D$ is the number of elements of D. Then by definition,

$$K(x|D) \leq^+ \log \#D. \tag{7}$$

Let the optimal function $A(p, D)$ define the conditional complexity $K(x|D)$. Then the total number of $x \in D$ for which we can find p such that $l(p) < \lceil \log \#D \rceil - m - 1$ and $A(p, D) = x$, does not exceed the total number of all p such that $l(p) < \log \#D - m$, i.e. $2^{-m}\#D - 1$. The portion of such x in D does not exceed

$$\frac{2^{-m}\#D - 1}{\#D} < 2^{-m}. \tag{8}$$

In other words, for each m there are at least $(1 - 2^{-m})\#D$ sequences $x \in D$ for which $K(x|D) \geq \log \#D - m - 1$. This property is called the *incompressibility property*.

The inequalities (7) and (8) show that for most elements of D the complexity $K(x|D)$ is close to $\log \#D$. Kolmogorov's idea is that randomness of a finite sequence $x \in D$ manifests itself in the absence of regularities in x, which can be interpreted as the absence of a description of x much shorter than the description of a "typical" element of D. Of course, for finite sequences the concept of randomness is relative. The degree of randomness of $x \in D$ can be measured by the value

$$d(x|D) = \log \#D - K(x|D).$$

We call this value the *deficiency of randomness* of a finite object (binary sequence) x with respect to a finite set D [17].

Kolmogorov [15] wrote that finite binary sequences with sufficiently small deficiency of randomness with respect to $D = \Xi_n$ must possess the property of stability of the frequency of 1s in their subsequences. We present a theorem of Asarin [1] realizing this hypothesis.

It is naturally to consider subsequences of a finite sequence selected by computable selection rules. A selection rule R is three partial recursive functions f, g and h on Ξ defined as follows ([12] and [37], Section 6.1). Let $x = x_1 \dots x_n$. The process of selection starts with empty sequence $\emptyset$. The function f makes choice of the following element: $f(\emptyset) = i_1$ and if elements $x_{i_1}, \dots, x_{i_k}$ are formed we compute the index of the following examined element $f(x_{i_1} \dots x_{i_k}) = i$, where $i \notin \{i_1, \dots, i_k\}$ (notice, that maybe $i < i_j, 1 \le j \le k$). The two-valued function g selects this element x_i in the subsequence as the next element if and only if $g(x_{i_1} \dots x_{i_k}) = 1$. The two-valued function h decides when this process must be terminated. The selection process terminates if $h(x_{i_1} \dots x_{i_k}) = 1$ or $f(x_{i_1} \dots x_{i_k}) > n$. Let $R[x]$ denote the selected subsequence.

If in the process of selection one of these functions will be undefined then selected subsequence is also undefined. We put $R[x] = \emptyset$ in this case.

By $K(R|n)$ we mean the length of the shortest program computing values of f, g and h given n. Let $d(x|n) = n - K(x|n)$.

Theorem 5 *For each $\epsilon > 0$ a positive integer number N and $0 < \mu < 1$ exist such that for each n and $x \in \Xi_n$, and selection rule R, such that the selection process terminates on x, $l(R[x]) = n_1$, $\sum_{i=1}^{n_1} R[x]_i = m$ and*

$$n_1 > N, d(x|n) + K(R|n) + 2\log K(R|n) < \mu n_1, \tag{9}$$

the following inequality holds

$$\left|\frac{m}{n_1} - \frac{1}{2}\right| < \left(\frac{d(x|n) + K(R|n) + 2\log K(R|n) + (3+\epsilon)\log n_1}{2n_1(1-\epsilon)\log e}\right)^{\frac{1}{2}}. \tag{10}$$

The proof of this theorem is given in Section 15.1.

As follows from the proof of Theorem 3 $K(x|n,k) \le^+ \log \binom{n}{k}$, where n is the length of x and k is the total number of 1s in x. Kolmogorov mentioned that the property of stability of frequencies in subsequences must hold also for any finite *m-Bernoulli sequence* x, i.e. such that $K(x|n,k) \ge^+ \log \binom{n}{k} - m$ [15].

In [1] a class of finite sets has been defined such that each x from a set of this class has probabilistic properties of normal distribution.

This approach has no essential development (with the exception [26], [27], Section 2.6). But Dawid's and Vovk's prequential (martingale or game theoretic) approach to probability theory it should be noted which does not use algorithms but arose on ideas of algorithmic approach [41], [8].

5 Non-stochastic finite sequences

Let α and β be non-negative integer numbers. By Kolmogorov a finite binary sequence x of the length n is called (α, β)-*stochastic* if there is a finite set D such that $x \in D$, $K(D|n) \leq \alpha$ and

$$K(x|n, D) \geq \log \#D - \beta.$$

(A difference with [34] and Section 4 is that n is assumed to be given in advance, i.e. all complexities are conditional with respect to n.) This means that in the case, where α and β are sufficiently small, x is an element of the "general position" of a "simple" set D. Such elements can be interpreted as the objects appearing as results of random experiments.

The following Shen's [34] theorem is a part of the answer for a corresponding problem posed by Kolmogorov in 1982 at Seminar in the Moscow State University. It shows that "absolutely non-random" objects exist.

Theorem 6 *For any positive α and β satisfying*

$$2\alpha + 2\log\alpha + \beta \leq n - c,$$

where c is some positive constant, there exists a finite binary sequence x of length n which is not (α, β)-stochastic.

Proof. For any n let $D_1, \dots D_s$ be all finite sets D consisting of finite binary sequences such that $K(D|n) \leq \alpha$ and $\#D \leq 2^{n-\alpha-1}$. To compute the list of all such sets we can fix n, α and the program p, $l(p) \leq \alpha$ such that the computation of $A(p, n) = D_i$ requires the maximal number of steps among such computations for all $D_1, \dots D_s$. By definition $s < 2^{\alpha+1}$. Then $\#\bigcup_{i=1}^{s} D_i < 2^{\alpha+1}2^{n-\alpha-1} = 2^n$ and there exists $x \notin \bigcup_{i=1}^{s} D_i$ of length n. To encode the minimal such x we need n, α and the list of all D as above. Hence, $K(x|n, \alpha) \leq^+ \alpha$ and

$$K(x|n) \leq^+ \alpha + 2\log\alpha. \tag{11}$$

Suppose that x is (α, β)-stochastic. Then there exists a finite D such that $K(D|n) \leq \alpha$, $K(x|n, D) \geq \log \#D - \beta$ and $x \in D$. By definition $\#D > 2^{n-\alpha-1}$, and so, $K(x|n, D) > n - \alpha - 1 - \beta$. Combining this inequality with (11) and $K(x|n, D) \leq^+ K(x)$ we obtain $\beta + 2\alpha + 2\log\alpha > n - c$ for some positive constant c. Now the assertion of the theorem follows immediately. □

As noted in [7], Kolmogorov proposed in 1973 at Information Theory Symposium, Tallin, Estonia a variant of the function

$$\beta_x(\alpha) = \min_{x \in D, K(D|n) \leq \alpha} d(x|n, D),$$

where n is the length of x and $d(x|n, D) = \log \#D - K(x|n, D)$ is a conditional variant of the deficiency of randomness.

The function $\beta_x(\alpha)$ characterizes stochastic properties of a finite object x. (A difference with [7] is that n is assumed to be given in advance.)

For any x the function $\beta_x(\alpha)$ is non-increasing and $\beta_x(\alpha) \geq^+ 0$. It represents the tradeoff between the size of the explanation of x and its value. If x of the length n is some sequence of experimental results, then the set D can be considered to be extraction of all features in x that point to non-random regularities. Let k_0 be some non-negative integer number. At the minimal point $k^*(x)$ where $\beta_x(k^*(x)) \leq k_0$, we can say that it is useless to explain x in greater detail than by giving D such that $d(x|n, D) = \beta_x(k^*(x))$ and $K(D|n) = k^*(x)$, see [7], Section 3. Evidently, $k^*(x) \leq K(\{x\}|n) =^+ K(x|n)$ (since we can consider $D = \{x\}$).

The set D plays the role of a "universal minimal sufficient statistics' for x and can be considered as an analog of the corresponding concept in statistics [7], [27]. The set D defined above is such that x is conditionally maximally random given D, that is, $K(x|n, D) \geq \log \#D - k_0$.

For any sequence x we evidently have $\beta_x(\alpha) \leq^+ l(x)$ for each α. A more refined estimate is

$$\beta_x(\alpha) \leq n - \alpha + c$$

for all positive $\alpha \leq n$, where $n = l(x)$ and c is a non-negative constant. To prove this inequality we divide the set of all binary sequences of length n into 2^α equal parts, each of $2^{n-\alpha}$ elements. The conditional complexity of the part D containing x is $K(D|n) \leq^+ \alpha$ and $d(x|D, n) = n - \alpha - K(x|n, D) \leq^+ n - \alpha$.

Usually in statistics, given a data x and a critical value β of a test we try to find a simple model D explaining x, i.e. such that $x \in D$ and $d(x|n, D) \leq \beta$.

Kolmogorov asked whether for any non-increasing function $f(k)$ there are objects x for which $\beta_x(k)$ is close to $f(k)$, and whether there are "absolutely non-random" strings x, for which $k^*(x)$ is close to $K(x)$, see [7].

As follows from the proof of Theorem 6 for any positive α, such that $2\alpha + 2\log\alpha < n - k_0 - c$, there exists an x such that $K(x|n) \leq^+ \alpha + 2\log\alpha$ and $\beta_x(\alpha) \geq n - 2\alpha - 2\log\alpha - c > k_0$. From this it follows that $k^*(x) \geq \alpha$, and so,

$$k^*(x) - c \leq K(x|n) \leq k^*(x) + 2\log k^*(x) + c,$$

for some non-negative constant c. We have also $d(x|n, \{x\}) \leq 0$. Hence, $\beta_x(\alpha + 2\log\alpha) \leq 0$ and $\alpha \leq k^*(x) \leq \alpha + 2\log\alpha$.

For any finite set J of parameters we define

$$\beta_x(\alpha|J) = \min_{x \in D, K(D|J) \leq \alpha} d(x|J, D),$$

where $d(x|J, D) = \log \#D - K(x|J, D)$.

The following result in slightly different form at first time was obtained by Levin and discussed with Kolmogorov in the seventies. Levin never published his proof (Levin's (1998) personal communication). Independently, the function

$\beta_x(\alpha)$ was considered later by V'yugin [44] and the corresponding result was obtained.

We present the general description of all possible forms of the function $\beta_x(\alpha)$. We prove that for any finite simple function $f(\alpha)$ (i.e. function, whose domain is a union of finite number of intervals, and the function is constant on each of them) there is a finite binary sequence x such that $\beta_x(\alpha)$ is close to $f(\alpha)$.

Theorem 7 *For any finite sequence of positive integer numbers*

$$J = (n, \alpha_1, \ldots, \alpha_k, \beta_1 \ldots, \beta_k),$$

such that $k \geq 1$,

$$c_1 \leq \alpha_1 < \ldots < \alpha_k$$

and

$$n > \beta_1 > \ldots > \beta_k,$$

there exists a binary sequence x *of the length* n *such that for* $1 \leq j \leq k+1$ *and all* α *satisfying the inequality*

$$\alpha_{j-1} + 2\log j + c_2 \leq \alpha \leq \alpha_j$$

the following estimate holds

$$\beta_j - 7\alpha_k - k\log k - c_3 \leq \beta_x(\alpha|J) \leq \beta_j + j,$$

where we put $\alpha_0 = 0$, $\alpha_{k+1} = \infty$, $\beta_{k+1} = 0$, c_1, c_2, c_3 *are positive constants.*

The proof of this theorem is given in Section 15.2.

We can learn the asymptotic behaviour of $\beta_x(\alpha)$, $l(x) = n$, in the rectangle $0 \leq \alpha \leq \alpha(n)$ and $0 \leq \beta \leq \beta(n)$ in the case where $\alpha(n) = o(\beta(n))$ as n tends to infinity.

Let $\nu(n)$ and $\mu(n)$ be two non-negative non-decreasing unbounded integer-valued functions such that $\mu(n) < n$ for all n and $\nu(n) = o(\mu(n))$ as n tends to infinity. Then we can consider the family of "normed" functions

$$f_x(t) = \frac{\beta_x(\nu(n)t + c)}{\mu(n)},$$

where $n = l(x)$, c is a positive constant and $0 \leq t \leq 1$.

We consider a space $L_\infty = L_\infty([0,1])$ with a norm $||f|| = \max\limits_{x\in[0,1]} |f(x)|$.

Then from Theorem 7 it follows

Corollary 1. *Each non-increasing function whose graph is in the unit square is a limit point in metrics* L_∞ *of the family functions* $\{f_x\}$, *where* $x \in \Xi$.

The interesting case in statistics is $\nu(n) = c\lfloor \log n \rfloor$ for arbitrary positive rational number c.

Notice, that Theorem 7 cannot explain the situation in the case where $\nu(n) = o(\mu(n))$ does not hold.

6 Prefix complexity

In this section we consider a special type of complexity based on some specific way of encoding of finite objects. This complexity was introduced at first time by Levin in the beginning of the seventies (first publication was later in [22], see also [10], [23]), and by Chaitin [4], [5]. This definition involves a prefix encoding of all finite binary sequences, i.e. the decoding functions are required to respect the discrete structure of the set Ξ.

A partial recursive function $B(p, y)$ is called *prefix* with respect to the first argument if it satisfies the following condition: $B(p, y) = B(p', y)$ if $p \subseteq p'$ and $(p, y), (p', y)$ are in the domain of B. The second argument y is a parameter. Such a function defines complexity

$$KP_B(x|y) = \min \{l(p)|B(p, y) = x\} .$$

The invariance property also takes place. The following theorem on the existence of an *optimal prefix function* holds.

Theorem 8 *There exists an optimal partial recursive prefix function $A(p, y)$ such that for each partial recursive prefix function $B(p, y)$ the inequality*

$$KP_A(x|y) \leq^+ KP_B(x|y)$$

holds.

The proof of this theorem is given in Section 15.3.

Any two optimal prefix complexities are equal up to an additive constant. As usual, we fix any such optimal A and denote $KP(x|y) = KP_A(x|y)$.

The connection between simple Kolmogorov complexity $K(x)$ and prefix complexity $KP(x)$ is given by the following inequalities:

$$K(x) \leq^+ KP(x) \leq^+ K(x) + \log K(x) + 2 \log \log K(x). \tag{12}$$

The first inequality follows from comparison of the definitions of $K(x)$ and $KP(x)$. To prove the second inequality it is sufficient to improve the way of encoding generating $K(x)$ to be prefix. Let $A(p, \emptyset) = x$, where $A(p, y)$ defines complexity $K(x|y)$. Then it is easy to reconstruct this function to a computable function $B(q)$, which on code $\overline{l(l(p))}01l(p)p$ computes x. The last way of coding is prefix. From this the right-hand side inequality (12) follows.

7 Algorithmic probability and prefix complexity

Prefix complexity can be also described in terms of probability distributions enumerable from below in the set Ξ with discrete structure.

R.J.Solomonoff [35] proposed the thoughtful philosophical ideas on defining of the a priory probability distribution on the basis of the general theory of

algorithms. A procedure of optimal inductive inference can be constructed using this distribution.

We briefly describe these ideas as follows. Let us attribute to any sequence x the probability $M(x) = L\{p \mid U(p) = x\}$ that an universal computer U will print out x when fed a random program p. Here $L(p) = 2^{-l(p)}$ is the uniform probability of a finite binary sequence p.

According to general conception, such a priory probability distribution $M(x)$ must be in some sense maximal among analogous probability distributions. Solomonoff [35] defined the shortest description of x as the shortest string p such that $U(p) = x$. Then the probability $M(x)$ will be approximately $2^{-l(p)}$ (see Section 8).

More correctly, $M(x)$ is not a probability distribution because $\sum 2^{-l(p)}$ diverges, which Solomonoff did notice. Solomonoff did not have the tool of prefix less algorithms but he tried to correct the problem in other ways. Probably it is impossible to combine the two features he insisted on: the normalization of measures and optimality [36]. However the intuitive concept was clear and important.

Levin gives a precise form of these ideas in [47] in a concept of the maximal semimeasure enumerable from below. We give an exposition for set Ξ with discrete structure and prefix machine U.

Let $P(x|y)$ be a function defined on the set $\Xi \times \Xi$ and taking non-negative real values. We will consider the second argument y as parameter. A function $P(x|y)$ is called *enumerable from below* if the set

$$\Big\{(r, x, y) \mid r < P(x|y), r \text{ is rational}\Big\}$$

is recursively enumerable. A function $P(x|y)$ is called (conditional) *semimeasure* if

$$\sum_x P(x|y) \le 1$$

for each y.

Levin's theorem on the existence of a maximal (up to a multiplicative constant) factor semimeasure enumerable from below also holds.

Theorem 9 *There exists a semimeasure P enumerable from below such that for each semimeasure Q enumerable from below a positive constant c exists such that the inequality*

$$cP(x|y) \ge Q(x|y)$$

holds for all x and y.

The proof of this theorem is given in Section 15.4.

Choose some semimeasure P enumerable from below satisfying Theorem 9. We will call it *a priory* probability in the set Ξ. For any other maximal semimeasure P' enumerable from below we have $P(x) =^{\cdot} P'(x)$.

The following Levin's [23] theorem shows that prefix entropy and the a priory probability are closely connected.

Theorem 10 $KP(x|y) =^+ -\log P(x|y)$.

Proof. For simplicity we omit parameter y. If $x \neq y$ then $KP(x)$ and $KP(y)$ are lengths of two incomparable finite sequences. This implies Kraft inequality [27] (Section 1.11.2)

$$\sum_x 2^{-KP(x)} \leq 1.$$

The function $Q(x) = 2^{-KP(x)}$ is enumerable from below, and so we have

$$cP(x) \geq 2^{-KP(x)}$$

for some positive constant c. Therefore,

$$-\log P(x) \leq^+ KP(x).$$

.

To prove the converse inequality we will define a prefix function B such that $KP_B(x) \leq^+ -\log P(x)$. We will use the fact that P is enumerable from below. Let us enumerate without repetition all pairs (m, y) of $m \in \mathbb{N}$ and $y \in \Xi$ such that $2^{-m} < \frac{1}{2}P(y)$. Let (m_k, y_k) be the k-th pair under this enumeration. Then we have

$$\sum_k 2^{-m_k} = \sum_y \sum_{y_k=y} 2^{-m_k} \leq \sum_y 2^{-s(y)+1} \leq \sum_y P(y) \leq 1,$$

where

$$s(y) = \min\{m_k \mid y_k = y\}.$$

By definition,

$$\frac{1}{4}P(y) \leq 2^{-s(y)} < \frac{1}{2}P(y).$$

We will use well known extension of Kraft inequality condition for the existence of an instantaneous code

Lemma 1 *Let $l_1, l_2, \ldots$ be recursively enumerable sequence of positive integer numbers such that $\sum 2^{-l_k} \leq 1$. Then there exists a recursively enumerable sequence of pairwise incomparable binary sequences $x_1, x_2, \ldots$ such that $l(x_k) = l_k$ for all k.*

Notice, that analogous assertion also holds for finite sequence $l_1, l_2, \ldots$.

The proof of this lemma is given in Section 15.5.

Lemma 1 can be found in [4], where its proof is attributed to N.J.Pippenger.

(The situation in the case of finite sequence $l_1, l_2, \ldots$ is much simpler: the proof can be based on the well known construction of the Shannon – Fano (or Huffman) codes [27], Section 1.11.)

We will also use the relativized variant of Lemma 1, in which the sequence $l_1, l_2, \ldots$ is replaced with $l_{n,1}, l_{n,2}, \ldots$, where n is a parameter. The resulted sequence $x_{n,1}, x_{n,2}, \ldots$ will also effectively depend on parameter n.

By Lemma 1 for $l_1 = m_1, l_2 = m_2 \ldots$ there exists a recursively enumerable sequence of pairwise incomparable sequences $x_1, x_2, \ldots$ such that $l(x_k) = m_k$ for all k. Now we can define partial recursive prefix function B such that $B(x_k) = y_k$ for all k and $B(x)$ is undefined for all other x. Then

$$KP_B(y) = \min\{m_k \mid y_k = y\} = s(y).$$

The above bound for $2^{-s(y)}$ can be rewritten as

$$-\log P(y) + 1 < s(y) \leq -\log P(y) + 2.$$

Then

$$KP(y) \leq^+ KP_B(y) =^+ -\log P(y).$$

□

Now we can show that the average code-word $KP(x|n)$ is equal to the Shannon entropy of the corresponding probability distribution [19] and [7].

For any n let Q^n be a computable probability distribution in Ξ_n. The computability means that there is an algorithm computing $Q^n(x)$ given n and a finite sequence $x \in \Xi_n$ with arbitrary degree of accuracy.

The mathematical expectation of a real function $f(x)$ on Ξ_n with respect to a measure Q^n is defined as

$$\mathbb{E}^n_{Q^n}(f(x)) = \sum_{l(x)=n} f(x)Q^n(x).$$

The Shannon entropy $H(Q^n)$ of probability distribution Q^n is defined as

$$H(Q^n) = \mathbb{E}^n_{Q^n}(-\log Q^n(x)) = -\sum_{l(x)=n} Q^n(x)\log Q^n(x).$$

Theorem 11 *For any computable measure Q^n in Ξ_n it holds*

$$\mathbb{E}^n_{Q^n}(KP(x|n)) =^+ H(Q^n).$$

Proof. Since $KP(x|n)$ satisfies $\sum_{l(x)=n} 2^{-KP(x|n)} \leq 1$ the left half of the inequality holds

$$H(Q^n) \leq \mathbb{E}^n_{Q^n}(KP(x|n)).$$

To prove the right half of the inequality define $Q(x|n) = Q^n(x)$ if $l(x) = n$, and $Q(x|n) = 0$, otherwise. By definition, $\sum_x Q(x|n) = 1$ for each n. Then we have

$P(x|n) \geq^{\cdot} Q(x|n)$, where $P(x|n)$ is the a priory probability. From this it follows that

$$H(Q^n) = -\sum_{l(x)=n} Q^n(x) \log Q^n(x) \geq^+ -\sum_{l(x)=n} Q^n(x) \log P(x|n) =^+ \sum_{l(x)=n} Q^n(x) KP(x|n) = \mathbb{E}^n_{Q^n}(KP(x|n)).$$

□

8 Computational model

A set consisting of finite binary sequences is called *prefix-free* if each two sequences from it are pairwise incomparable. Notice, that the function B from the second part of the proof of Theorem 10 has a prefix-free domain. This shows that we can define prefix complexity starting from a narrower class of prefix functions, viz, from the functions whose domains are prefix-free. As in Theorem 8 it can be proved that among such functions there is an optimal one. The proof of the second part of Theorem 10 shows that $KP'(x) \leq^+ KP(x)$, where KP' is the complexity defined in this way. The converse inequality follows from the definition of $KP(x)$. Hence, we have

$$KP'(x) =^+ KP(x).$$

The type of encoding used in the definition of $KP'(x)$ is well known in the classical theory of information. This unique decodable code is called Huffman code or prefix-code [27] (Sections 1.4, 1.11).

There is a convenient computational model, viz, a special class of multitape Turing machines such that $KP'(x)$ is equal to the length of the shortest program from which an optimal machine in this class can calculate x. Any machine of this class has three tapes: on the first tape the program is written (any symbol of the end of program do not used), the second is the work tape, and on the third tape the result is printed. The heads of the first and the third tapes can move only to the right. The initial position of the first head is on the leftmost symbol of the program. The peculiar feature of a machine in this class is that if this machine finishes a computation and prints out the result, the head of the first tape (moving only rightward during the computation) never intersects the right bound of the program. In other words, the machine does not use any special mark for the end of the program and decides itself when to stop reading the program. Any program for machine of this class is also called *self-delimiting program*. Additional input tapes may be used when we consider conditional prefix complexity. These tapes have no restriction on reading information.

It is easy to see that each function computable on such machine is a function with prefix-free domain and each partial recursive function with prefix-free

domain can be computed on such machine. To prove the last assertion, suppose that a function $B(p)$ with prefix-free domain is computed on some Turing machine T. We proceed in stages. At each stage we use some initial segment $p_1 \dots p_i$ of the argument p (we start with empty sequence $\emptyset$ on the first stage) and simulate on the work tape all computations of machine T for all q, where $p_1 \dots p_i \subseteq q$, until one of these computations will terminate. Such terminated computation exists since $B(p)$ is defined. If $p_1 \dots p_i = q$ then the process is finished, $p = q$ and $T(q) = B(p)$ is a result. Otherwise, we replace $p_1 \dots p_i$ on $p_1 \dots p_i p_{i+1}$ and go to the next stage. Evidently, such computation terminates when $p_1 \dots p_i = p$ for $i = l(p)$ (the detailed algorithm is given in [27], Section 3.1). So the defined machine does not use any right bound on the program p.

Now, we can make more precise informal considerations to the beginning of Section 7.

Theorem 12 *Let F be an optimal machine with prefix-free domain defining complexity $KP(x)$. Then the probability that F will output x when fed with an uniformly distributed program p is $2^{-KP(x)}$ up to a multiplicative constant.*

Proof. This probability is

$$M(x) = L\{p \mid F(p) = x\} = \sum_{F(p)=x} 2^{-l(p)}.$$

We have $M(x) \leq^{\cdot} P(x) =^{\cdot} 2^{-KP(x)}$ since $M(x)$ is the semimeasure enumerable from below . Evidently, $M(x) \geq 2^{-KP(x)}$. □

Notice, that this theorem also holds for an arbitrary machine defining prefix function F (we could take the sum in the proof only for all shortest p such that $F(p) = x$).

From this theorem and estimates below it will follow

$$M(x) \geq^{\cdot} \frac{1}{n \log n \log\log^2 n},$$

where $n = l(x)$.

9 Inequalities for prefix complexity

There are very convenient inequalities for the prefix complexity. In this section we shall prove the most widely used of them.

To illustrate using of the model from the previous section, we prove an inequality [23]

$$KP(\phi(x, y)) \leq^{+} KP(x) + KP(y),$$

where $\phi(x, y)$ is an arbitrary recursive function (for instance, some effective enumeration of all pairs (x, y)). Let B be an optimal machine with prefix-free

domain, p be the shortest program for x, q be the shortest program for y. Then we describe a machine with prefix-free domain which, when applied to the program pq, acts as follows. First it simulates the work of B on pq. By the choice of p, machine B outputs $B(p)$; moreover, the head of its input tape stops over the last symbol of p and thus indicates the first symbol of q. After that our machine simulates the work of B on q, and computes and prints out $\phi(x, y)$ on the output tape. We have

$$KP(\phi(x,y)) \leq^+ l(pq) = l(p) + l(q) =^+ KP(x) + KP(y).$$

Analogously we can prove

$$KP(x|z) \leq^+ KP(x|z,y) + KP(y|z) \tag{13}$$

Theorem 10 also allows us to construct various upper bounds for the complexity $KP(x)$ using slowly convergent series. Thus we can obtain estimate from [5]

$$KP(x) \leq^+ l(x) + KP(l(x)). \tag{14}$$

Indeed, since there are 2^m different x of length m, there holds

$$\sum_x 2^{-l(x)-KP(l(x))} = \sum_m 2^{-m-KP(m)} 2^m = \sum_m 2^{-KP(m)} \leq 1.$$

Since the function $Q(x) = 2^{l(x)-KP(l(x))}$ is enumerable from below we have $P(x) \geq^\cdot Q(x)$, where P is the a priory probability. Then inequality (14) follows immediately from Theorem 10.

In prefix machines language we can explain inequality (14) as follows. To compute x on some machine with prefix-free domain it is sufficient to take as a program the shortest program for $l(x)$ concatenated with all symbols of x.

Analogously we can obtain the following estimate for any positive integer number n: for each $\epsilon > 0$ there holds

$$KP(n) \leq^+ \log n + (1+\epsilon) \log\log n.$$

Inequality (12) between simple Kolmogorov complexity $K(x)$ and prefix complexity $KP(x)$ can be improved in the following form: for each $\epsilon > 0$

$$KP(x) \leq^+ K(x) + \log K(x) + (1+\epsilon) \log\log K(x).$$

It is sufficient to show that the series

$$\sum_x 2^{-K(x)} / K(x) \log^{1+\epsilon} K(x)$$

is convergent. Since the number of x satisfying $K(x) = m$ does not exceed 2^m, we have

$$\sum_x 2^{-K(x)} / K(x) \log^{1+\epsilon} K(x) \leq \sum_m \frac{2^{-m}}{m \log^{1+\epsilon} m} 2^m = \sum_m \frac{1}{m \log^{1+\epsilon} m} < \infty.$$

Analogously we can obtain estimate

$$KP(x) \leq^+ K(x) + \log K(x) + \log\log K(x) + (1+\epsilon)\log\log\log K(x),$$

and so on.

The additional term in the right of the inequality for $KP(x)$ appears since we must add to the information about x additional information about the final position of the head on input tape.

Since $\sum_n 2^{-KP(n)} < \infty$ for any function $f(n)$ such that $\sum_n 2^{-f(n)} = \infty$ we have $KP(n) > f(n)$ for infinitely many n. So, for instance, we have

$$KP(n) > \log n + \log\log n$$

for infinitely many n.

The following nontrivial and fundamental result is due to Levin (see Gács [10]).

Theorem 13 $KP(x,y) =^+ KP(x) + KP(y|x, KP(x))$.

Proof. The proof of the first inequality ($\leq^+$) is based on the inequality (13) and the following elegant lemma of Gács [10].

Lemma 2 $KP(x, KP(x)) =^+ KP(x)$.

Proof. Let p be the shortest code for x. Obviously, both x and $KP(x) = l(p)$ are computable from p. Therefore, $KP(x, KP(x)) \leq^+ KP(x)$. The converse inequality $KP(x) \leq^+ KP(x, KP(x))$ is trivial. □

By (13) and Lemma 2 we have

$$\begin{aligned} KP(x,y) \leq^+ KP(y,x,KP(x)) \leq^+ \\ KP(x,KP(x)) + KP(y|x,KP(x)) =^+ \\ KP(x) + KP(y|x,KP(x)). \end{aligned}$$

To prove the converse inequality we shall use an equivalent representation of the prefix complexity by semimeasures enumerable from below. So, we must prove that

$$P(y|x, KP(x)) \geq^\cdot P(x,y)/P(x) =^\cdot 2^{KP(x)}P(x,y),$$

where P is the a priory probability. Since the function $Q(x) = \sum_y P(x,y)$ is a semimeasure enumerable from below (recall the correspondence between $\Xi \otimes \Xi$ and Ξ), $c_1 2^{-KP(x)} \geq \sum_y P(x,y)$ for some positive constant c_1 (recall that $P(x) =^\cdot 2^{-KP(x)}$).

By definition, the set

$$W = \{(r,x,z) | r \in \mathbb{Q}, r < P(x,z)\}$$

is recursively enumerable. Let W^t be a finite subset of W enumerated within t steps. Define

$$P^t(x,z) = \max(\{r|(r,x,z) \in W^t\} \bigcup \{0\}).$$

To continue the proof we must get over some technical problem, since $KP(x)$ is not computable function. To avoid this difficulty let us define a conditional semimeasure $Q(y|x,m)$ enumerable from below as follows. Given x, m for any positive integer number s define the maximal $t \leq s$ such that

$$c_1^{-1} 2^m \sum_{z \leq t} P^t(x,z) \leq 1.$$

Define

$$Q^s(y|x,m) = c_1^{-1} 2^m P^t(x,y)$$

if $y \leq t$ and $Q^s(y|x,m) = 0$, otherwise. By definition, $Q^s(y|x,m) \leq Q^{s+1}(y|x,m)$ for all s. Then we can define

$$Q(y|x,m) = \sup_s Q^s(y|x,m).$$

As follows from the definition,

$$\sum_y Q(y|x,m) \leq 1$$

and, since Q is enumerable from below, we have $P(y|x,m) \geq^{\cdot} Q(y|x,m)$.

Let us compute $Q(y|x,m)$ for $m = KP(x)$. Since

$$c_1^{-1} 2^{KP(x)} \sum_{z \leq t} P^t(x,z) \leq c_1^{-1} 2^{KP(x)} \sum_{z \leq t} P(x,z) \leq 1$$

for each t, we have

$$Q(y|x, KP(x)) = c_1^{-1} 2^{KP(x)} P(x,y).$$

Hence, we have

$$P(y|x, KP(x)) \geq^{\cdot} 2^{KP(x)} P(x,y) =^{\cdot} \frac{P(x,y)}{P(x)}.$$

□

Corollary 2. $KP(x) + KP(y|x) - \log KP(x) - 2\log\log KP(x) \leq^+ KP(x,y) \leq^+ KP(x) + KP(y|x)$.

Proof. Using inequality (13) we obtain

$$KP(y|x, KP(x)) \leq^+ KP(y|x) \leq^+ KP(y|x, KP(x)) + KP(KP(x)) \leq^+ KP(y|x, KP(x)) + \log KP(x) + 2\log\log KP(x).$$

Now we can apply Theorem 13. □

10 Information, II

The amount of information in y about x also can be defined on the base of prefix complexity

$$IP(y:x) = KP(x) - KP(x|y).$$

As follows from Theorem 13

$$KP(y) - KP(y|x, KP(x)) =^+ KP(x) - KP(x|y, KP(y)).$$

This equality can be rewritten as an equality of strict symmetry of information [10]

$$I((x, KP(x)) : y) =^+ I((y, KP(y)) : x).$$

We have connection of $IP(x:y)$ with the symmetric expression

$$IP^*(x:y) = KP(x) + KP(y) - KP(x,y).$$

By Theorem 13 we have

$$|IP^*(x:y) - IP(x:y)| = |KP(x) - KP(x,y) + KP(y|x)| \leq^+ \log KP(x) + 2\log\log KP(x).$$

Analogously we obtain an estimate

$$|IP^*(x:y) - IP(y:x)| \leq^+ \log KP(y) + 2\log\log KP(y).$$

These inequalities also lead to the symmetry estimate of $IP(x:y)$

$$|IP(x:y) - IP(y:x)| \leq^+ \log KP(x) + 2\log\log KP(x) + \log KP(y) + 2\log\log KP(y).$$

As proved by Gács [10] (see also [27], Section 3.9) $IP(x:y)$ is no commutative up to an additive constant.

11 Statistical complexity, I

The systematic variation within a set of data, as represented by some statistical model, may be used to encode the data using a way of generating data in the model and description of this model. An estimate

$$KP(x) \leq^+ KP(x|A) + KP(A) \tag{15}$$

based on a model A, becomes trivial then $A = \emptyset$ or $A = \{x\}$. Normally, A is selected from a limited set of possible hypotheses. In statistics we usually confine ourselves to parametric models.

The upper estimate (15) is impractical since $KP(x|A)$ and $KP(A)$ are non-computable. Wallace's and Freeman's MML (Minimum Message Length) [46] and Rissanen's MDL (Minimum Description Length) [30], [31] principles avoid this drawback by considering computable upper bounds of these complexities based on Shannon information theory and arithmetical encoding.

Vovk [42], [43] considered a similar combined upper bound. Although his upper bound is non-computable, for some parametric families its minimum is attained at the parameter computable by the data. Besides, this minimum is very close to Kolmogorov complexity of the data generated by an arbitrary Bernoulli probability distribution.

In this section we present the main results from [42] and [43]. Let, according to the traditional statistical framework, for any n a finite sequence $x \in \Xi_n$ be generated by a computable measure P^n in Ξ_n. The computability means that there is an algorithm computing the value $P^n(x)$ given n and x with arbitrary degree of accuracy. By definition $\sum\limits_{l(x)=n} P^n(x) = 1$ for each n. Then by Lemma 1 (relativized with respect to n) we can define a computable function $C(x|n)$ such that for any x and x' of the length n the sequence $C(x|n)$ is incomparable with the sequence $C(x'|n)$ and

$$l(C(x|n)) =^+ -\log P^n(x), \tag{16}$$

where $x \in \Xi_n$. Then by definition

$$KP(x|n) \leq^+ l(C(x|n)) \tag{17}$$

and by the Kraft inequality for $KP(x|n)$

$$\mathbb{E}^n_{P^n} 2^{l(C(x|n))-KP(x|n)} \leq^\cdot 1, \tag{18}$$

where $\mathbb{E}^n_{P^n}$ denotes the mathematical expectation with respect to P^n, namely, $\mathbb{E}^n_{P^n}(f(x)) = \sum\limits_{l(x)=n} f(x)P^n(x)$.

By Chebyshev's inequality, (18) implies

$$P^n\{x \in \Xi_n | KP(x|n) \leq l(C(x|n)) - m\} \leq 2^{-m}, \tag{19}$$

so we can be practically sure that the length of code $C(x|n)$ is close to the length of the optimal code $KP(x|n)$.

Let us consider a case, where the probability distribution depends on a real parameter θ. For simplicity we assume that θ is a computable real number. In this case for each parameter value θ we can efficiently encode x using about $-\log P^n_\theta(x)$ bits, namely, there exists a computable function $C(x|n,\theta)$ such that the inequalities (16)-(18) are transformed to

$$l(C(x|n,\theta)) =^+ -\log P^n_\theta(x), \tag{20}$$

$$KP(x|n,\theta) \leq^+ l(C(x|n,\theta)), \tag{21}$$

$$\mathbb{E}^n_{P^n_\theta} 2^{l(C(x|n,\theta))-KP(x|n,\theta)} \leq^\cdot 1, \tag{22}$$

where $n = l(x)$. Strictly speaking, these functions depend not on θ but on a program computing parameter θ, which is a finite binary sequence.

In the following we will consider the standard Bernoulli statistical model $\{B_\theta^n(x) \mid \theta \in [0,1]\}$. The *Bernoulli measure* in Ξ_n with parameters (n, θ) is defined as $B_\theta^n(x) = \theta^k(1-\theta)^{n-k}$, where $x \in \Xi_n$ and $k = \sum_{i=1}^{n} x_i$ is the total number of 1s in x.

We are given data $x \in \Xi_n$, and we suppose that for some θ the probability distribution B_θ^n in Ξ_n is a good description of x. As we know, $KP(x|n)$ is all information contained in x. The main problem of parametric statistics, is what we can learn from x about θ? We shall extract from x all "useful information" it contains. In other words, we shall split all the information in x into the useful information and useless noise.

The total code length of $x \in \Xi_n$ corresponding to θ is defined as

$$DL_\theta(x) = -\log B_\theta^n(x) + KP(\theta|n). \tag{23}$$

According to [43] the *statistical coding scheme* is defined as follows. Each codeword for $x \in \Xi_n$, consists of two parts: the preamble, which is description of some θ, and the body, $C(x|n,\theta)$. The preamble encodes the useful information, and the body is the noise, which is incompressible. The minimal possible code length under this coding scheme is defined as

$$DL(x) = \inf_\theta DL_\theta(x).$$

This function is analogous to the minimum description length (MDL) function for Rissanen's ([30], [31]) coding scheme and is called *statistical complexity*.

A connection with $KP(x|n)$ can be easily obtained. By (13) we have

$$KP(x|n) \leq^+ KP(x|n,\theta) + KP(\theta|n).$$

Then by (20) and (21) any parametric computable statistical model $P_\theta(x)$ for any θ provides an upper bound for complexity $KP(x|n)$, since we have the inequality

$$KP(x|n) \leq^+ -\log_2 P_\theta(x) + KP(\theta|n).$$

Taking minimum by θ we obtain

$$KP(x|n) \leq^+ DL(x).$$

By MDL/MML principle, given a data x, we select a model P_θ giving a minimum to

$$-\log_2 P_\theta(x) + KP(\theta|n).$$

The following Vovk's [43] theorem (given here without proof) shows that statistical complexity on average is close to the prefix complexity.

Theorem 14 *There exists a positive constant c such that*

$$\mathbb{E}^n_{B_\theta} 2^{DL(x)-KP(x|n)} \le c$$

for each n and θ.

This theorem strengthens inequality (18). By Jensen's inequality, this theorem implies

Corollary 3. $\mathbb{E}^n_{B_\theta} KP(x|n) =^+ \mathbb{E}^n_{B_\theta} DL(x)$.

Simplest example shows that, for some x, $DL(x)$ and $KP(x|n)$ can be very different: when x is the sequence $0101\ldots$ of alternating 0s and 1s of length n, we have $DL_\theta(x) \ge -\log(\theta^{\frac{n}{2}}(1-\theta)^{\frac{n}{2}}) \ge^+ n$ for each θ. From this it follows

$$DL(x) \ge^+ n, \quad KP(x|n) =^+ 0$$

(this sequence is untypical in the highest degree under the Bernoulli model).

A *point estimator* is a real-valued function $E(x)$ defined on Ξ such that $0 \le E(x) \le 1$, for all finite binary sequences x. It is computable if there exists an algorithm which transforms each x into a program computing rational approximation of the real number $E(x)$ with given degree of accuracy.

Although $DL(x)$ is non-computable (since $KP(\theta|n)$ is non-computable) the infinum $\inf\limits_\theta DL_\theta(x)$ can be attained (to within an additive constant) for θ efficiently computed by the data x. Vovk [42] proved the existence of such estimators for some parametric models: Bernoulli and two Gauss families.

The next Vovk's [43] theorem asserts that in the case of the Bernoulli family the useful information in x can be extracted efficiently.

Theorem 15 *There exists a computable point estimator E such that*

$$DL(x) =^+ DL_{E(x)}(x).$$

Proof. The scheme of the proof is as follows. The main purpose is to minimize the sum (23). The maximum likelihood estimate $\hat\theta(x) = \frac{k}{n}$ minimizes only the first addend. Since by Lemma 5 (Section 15.5) the changes of likelihood function near an extremum are small, we can decrease the sum (15) using not the whole $\hat\theta(x)$ but only the most significant digits of $\hat\theta(x)$.

The estimator $E(x)$ is defined on the base of the following *net* in the interval $[0,1]$ of the real line

$$\theta_n(a) = \sin^2(an^{-1/2}), a = 1, \ldots, \left\lfloor \pi n^{1/2}/2 \right\rfloor - 1. \tag{24}$$

The choice of this net is justified by Lemmas 4, 5 and Corollary 4 (Section 15.6).

Let x be a sequence of the length n and k be the total number of 1s in it. Then $E_n(k) = E(x)$ is defined as the element of the net $\sin^2(an^{-1/2})$ closest to $\frac{k}{n}$ (we need also some convention to avoid a contradiction when $\frac{k}{n}$ is exactly

halfway between two adjacent elements of the net, but it is a simple technical problem, see [43]).

We use the notation $G_{n,k}$ for the log-likelihood function expressed through the variable a (which will no longer be assumed to be integer) introduced by $\theta_n(a) = \sin^2(an^{-1/2})$:

$$G_{n,k}(a) = \ln\left(\sin^{2k}(an^{-1/2})\cos^{2(n-k)}(an^{-1/2})\right),$$

a ranging over $[0, \pi n^{1/2}/2]$.

We use the notation $\hat{a}(n,k)$ for the maximum likelihood estimate of the parameter a:

$$\hat{a}(n,k) = \arg\max_a G_{n,k}(a)$$

(therefore, $\sin^2(\hat{a}(n,k)n^{-1/2}) = k/n$). By definition $\hat{a}(n,k)$ is a computable (by n and k) real number.

When $k = 0$ the assertion of the theorem reduces to $0 =^+ -\log((1 - \sin^2 n^{-1/2})^n)$ which is easy to validate. The case $k = n$ is considered analogously.

The remaining part of proof of the theorem is based on notations and lemmas from Section 15.6. We must prove that for any computable θ

$$-\log((E_n(k))^k(1 - E_n(k))^{n-k}) + KP(E_n(k)|n) \leq^+ \tag{25}$$
$$-\log(\theta^k(1-\theta)^{n-k}) + KP(\theta|n). \tag{26}$$

It is easy to see that

$$KP(E_n(k)|n) =^+ KP\left(\lfloor\hat{a}(n,k)\rfloor\,|n\right)$$
$$\leq^+ KP\left(\lfloor\hat{a}(n,k)\rfloor, \theta|n\right) \leq^+ KP(\theta|n) + KP\left(\lfloor\hat{a}(n,k)\rfloor\,|n,\theta\right),$$

so to ensure (25) and (26) it suffices to prove

$$-\log(E_n(k)^k(1 - E_n(k))^{n-k}) + KP\left(\lfloor\hat{a}(n,k)\rfloor\,|n,\theta\right) \leq^+ -\log(\theta^k(1-\theta)^{n-k}).$$

Lemma 5 (Section 15.6) asserts that $G_{n,k}(a) =^+ G_{n,k}(\hat{a}(n,k))$ if $|a - \hat{a}(n,k)| < 1$. Then, using the parameterization $\theta = \theta_n(a)$, we must prove

$$KP\left(\lfloor\hat{a}(n,k)\rfloor\,|n,a\right) \leq^+ (\ln^{-1} 2)(G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a)).$$

This inequality immediately follows from Lemma 7 (Section 15.6). □

MDL-principle is based on a straightforward coding scheme for θ.

$$\Theta_n = [\epsilon, 1-\epsilon]\bigcap\{an^{-1/2}|a \in \mathbb{Z}\} \tag{27}$$

with a description of length at most $\lceil\frac{1}{2}\log n\rceil$ (other θ may have no description). Replacing $KP(\theta|n)$ with this coding scheme in the definition of the statistical

coding scheme Rissanen's (see, e.g., [30]) a coding scheme is obtained. The minimum description length function for Rissanen's coding scheme is within an additive constant from

$$DL^*(x) = \min_{\theta\in\Theta_n}\left(-\log B^n_\theta\{x\} + \frac{1}{2}\log n\right),$$

n being the length of x.

It is easy to see that

$$KP(x|n) \leq^+ DL(x) \leq^+ DL^*(x), \tag{28}$$

The right-hand inequality of (28) is not an equality even on the average, as the next Vovk's [43] theorem shows.

Theorem 16

$$\mathbb{E}^n_{B^n_\theta} DL(x) =^+ H_n(\theta) + KP(\theta_n|n), \tag{29}$$

$$\mathbb{E}^n_{B^n_\theta} DL^*(x) =^+ H_n(\theta) + \tfrac{1}{2}\log n, \tag{30}$$

where θ_n is an element of Θ_n closest to θ and $H_n(\theta)$ is the entropy of B^n_θ:

$$H_n(\theta) = \mathbb{E}^n_{B^n_\theta}\left(-\log B^n_\theta\{x\}\right).$$

We have

$$KP(\theta_n|n) \leq^+ \frac{1}{2}\log n;$$

besides, for the most θ $KP(\theta_n|n)$ is close to $\frac{1}{2}\log n$. However, it is easy to find θ for which $KP(\theta_n|n)$ is very different from $\frac{1}{2}\log n$: say, for $\theta = \frac{1}{2}$ we have $KP(\theta_n|n) =^+ 0$.

Theorems 14, 15, and 16 immediately imply

$$\mathbb{E}^n_{B^n_\theta} KP(x|n) + d(\theta_n|\Theta_n) \geq^+ H_n(\theta) + \frac{1}{2}\log n \geq^+ \mathbb{E}^n_{B^n_\theta} DL^*(x),$$

where

$$d(\theta_n|\Theta_n) = \log \#\Theta_n - KP(\theta_n|n)$$

(notice that $\#\Theta_n =^\cdot n^{1/2}$) is the prefix randomness deficiency of θ_n in Θ_n.

12 Bernoulli sequences

Let us consider the notion of randomness of an individual object with respect to computable probability distribution P^n in the set Ξ_n.

Recall, that with any probability distribution P^n in Ξ_n an efficient code $C(x|n)$, where $x \in \Xi_n$, is associated such that inequalities (16)–(18) hold. Inequality (18) can be rewritten as

$$\mathbb{E}^n_{P^n} 2^{l(C(x|n))-KP(x|n,p)} \leq^{\cdot} 1$$

(where we add in the condition a program p computing the measure P^n).

These inequalities and Chebyshev's inequality (19) show that the quantity

$$d(x|n, P^n) = -\log P^n(x) - KP(x|n, p), \tag{31}$$

can be considered as *deficiency of randomness* (or *test of randomness*) of the sequence x of the length n with respect to P^n. Here p denotes a program computing P^n (more correctly, this notion is defined with respect to a program p).

By (16) and (31) we have

$$\mathbb{E}^n_{P^n}(2^{d(x|n,P^n)}) \leq^{\cdot} 1.$$

This inequality and (19) show that deficiency of randomness $d(x|n, P^n)$ is small for most (with respect to P^n) sequences of the length n. This value can be considered as a measure of disagreement between the measure P^n and an outcome x. Outcomes x with large value of $d(x|n, P^n)$ are interpreted as almost impossible from the viewpoint of the holder of the measure P^n.

In the case of the Bernoulli family a transition from this notion to the notion of the deficiency of randomness with respect to a finite set will be given by theorems 18 and 19 below.

The Bernoulli measure $B^n_\theta(x)$ in Ξ_n with parameters (n, θ) was defined in Section 11.

Let $b_{n,\theta}(x)$ be corresponding deficiency of randomness with respect to B^n_θ

$$b_{n,\theta}(x) = -\log B^n_\theta(x) - KP(x|n, \theta).$$

As in the previous section we suppose that θ is a computable real number (strictly speaking, we must replace θ with its program).

Usually in statistics a precise probability distribution generating data is unknown. We have only information that this probability distribution belongs to some class of similar distributions. We will use a concept of randomness with respect to a class of probability distributions. In [21] a definition of the deficiency of randomness of an object x with respect a class Φ of probability distribution was defined

$$d_\Phi(x) = \inf_{P\in\Phi} d(x|P),$$

where $d(x|P)$ is a deficiency of randomness with respect to a probability distribution P.

The class of Bernoulli measures is very natural class of probability distributions, which realizes the hypothesis that our data are generated in a process of independent trials with the same probability distribution (like coin flipping). We consider a concept of randomness with respect to the this class of probability distributions.

The concept of Bernoulli sequence at first was introduced in [14] and studied in [28] (see Section 4) and [21].

The *Bernoulli deficiency* is the function

$$b_n(x) = \inf_\theta b_{n,\theta}(x),$$

where x is a finite sequence of the length n.

The *binomial measure* in the set $\{0, 1, \dots, n\}$ with parameters (n, θ) is defined as

$$Bin_{n,\theta}(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

The corresponding deficiency of randomness is defined as

$$bin_{n,\theta}(k) = -\log Bin_{n,\theta}(k) - KP(k|n,\theta)$$

and has the analogous properties

$$KP(k|n,\theta) \leq^+ -\log Bin_{n,\theta}(k)$$

and

$$\mathbb{E}^n_{Bin_{n,\theta}} 2^{bin_{n,\theta}(k)} \leq^\cdot 1.$$

The *binomial deficiency* is defined as

$$bin_n(k) = \inf_\theta bin_{n,\theta}(k).$$

Vovk's net (24) (Section (11)) explains the definition of the binomial deficiency [40]. The definitions of the net $\theta_n(a)$ and corresponding estimator $E_n(k)$ are given in the proof of Theorem 15.

Theorem 17 $bin_n(k) =^+ bin_{n,E_n(k)}(k)$.

Proof. The case where $k = 0$ or $k = n$ is considered analogously to the proof of Theorem 15. When $1 \leq k \leq n-1$ it is sufficient to prove that

$$-\log(\theta^k(1-\theta)^{n-k}) - KP(k|n,\theta) \geq^+ \tag{32}$$

$$-\log((E_n(k))^k(1-E_n(k))^{n-k}) - KP(k|n,E_n(k)), \tag{33}$$

where θ is a computable real number. By Lemma 5 (Section 15.6)

$$-\log((E_n(k))^k(1-E_n(k))^{n-k}) =^+ (\ln^{-1} 2) G_{n,k}(\hat{a}(n,k)).$$

Then inequality (32)-(33) can be rewritten as

$$(\ln^{-1} 2)(G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a)) \geq^+ KP(k|n,\theta) - KP(k|n,E_n(k)).$$

By Corollary 2 (Section 9) and Lemma 7 (Section 15.6) we have

$$\begin{aligned} KP(k|n,\theta_n(a)) - KP(k|n,E_n(k)) \leq^+ \\ KP(k|E_n(k),n,\theta_n(a)) + KP(E_n(k)|n,\theta_n(a)) - KP(k|n,E_n(k)) \leq^+ \\ KP(E_n(k)|n,\theta_n(a)) =^+ KP(\lfloor \hat{a}(n,k) \rfloor \,|n,a) \leq^+ \\ (\ln^{-1} 2)(G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a)). \end{aligned}$$

□

As mentioned at the end of Section 2 for most sequences x of the length n with k ones the difference $\log \binom{n}{k} - KP(x|n,k)$ is small (here we replace complexity K with KP). Vovk's [40] following theorem shows that the Bernoulli deficiency of x can be decomposed into the sum of this difference and the binomial deficiency of k.

Theorem 18 *For any $x \in \Xi_n$ with k 1s*

$$\begin{aligned} bin_n(k) \leq^+ b_n(x) - (\log \binom{n}{k} - KP(x|n,k)) \leq^+ \\ bin_n(k) + 2\log|bin_n(k)|. \end{aligned}$$

Proof. From

$$KP(x|\theta,n) \leq^+ KP(x|k,n) + KP(k|n,\theta)$$

we obtain

$$b_n(x) \geq^+ \log \binom{n}{k} - KP(x|n,k) + bin_n(k).$$

The proof of the converse inequality is based on Theorem 17 and results of Section 15.6.

Since $k = k(x)$ is a computable function $KP(x,k(x)|n,\theta) =^+ KP(x|n,\theta)$ and by Theorem 13

$$\begin{aligned} b_{n,\theta}(x) =^+ -\log B^n_\theta(x) - KP(x|n,\theta) =^+ \\ \log \binom{n}{k} - KP(x|n,k,KP(k|n,\theta),\theta) - \log Bin_{n,\theta}(k) - KP(k|n,\theta). \end{aligned}$$

Since $KP(k|n,E_n(k))$ is computed by $\lfloor bin_{n,E_n(k)}(k) \rfloor$, n and k we have by (1)

$$KP(x|n,k,KP(k|n,E_n(k)),E_n(k)) \geq^+ KP(x|n,k,\lfloor bin_{n,E_n(k)}(k) \rfloor).$$

Hence,

$$b_n(x) \leq^+ \log \binom{n}{k} - KP(x|n,k,\lfloor bin_{n,E_n(k)}(k) \rfloor) + bin_{n,E_n(k)}(k).$$

From this by Corollary 2 (Section 9) we obtain

$$b_n(k) \leq^+ \log \binom{n}{k} - KP(x|n,k) + bin_{n,E_n(k)}(k) + 2\log|bin_{n,E_n(k)}(k)|.$$

Now, the needed inequality follows from Theorem 17. □

The randomness of k with respect to the binomial measure can be also arranged in layers according to different finite sets [40]. Let us consider the net (24) from Section 11 and corresponding division of the set $\{0,1,\ldots,n\}$ on subsets $[n\theta_n(s), n\theta_n(s+1)) \bigcap \{0,1,\ldots,n\}$, where $\theta_n(s) = \sin^2(sn^{-1/2})$, $s = 1,\ldots,\lfloor \pi n^{1/2}/2 \rfloor - 1$, and $\theta_n(\lfloor \pi n^{1/2}/2 \rfloor) = +\infty$. For any $0 < k < n$ denote $U(k)$ the element of the division containing k. By Corollary 4 (Section 15.6)

$$\#U(k) =^+ \sqrt{k(n-k)/n}$$

(This value is an estimate of the standard deviation of the number of 1s in x).

Theorem 19 *For* $0 < k < n$

$$bin_n(k) =^+ \log \#U(k) - KP(k|n,U(k)).$$

Proof. By Stirling's formula we obtain

$$\begin{aligned} -\log Bin_{n,E_n(k)}(k) &= -\log \tbinom{n}{k} - \log((E_n(k))^k(1-E_n(k))^{n-k}) \\ &=^+ -\tfrac{1}{2}\log n - n\log\tfrac{n}{e} + \tfrac{1}{2}\log k + k\log\tfrac{k}{e} + \tfrac{1}{2}\log(n-k) + (n-k)\log\tfrac{n-k}{e} \\ &\quad -k\log E_n(k) - (n-k)\log(1-E_n(k)) \\ &= \tfrac{1}{2}\log\tfrac{k(n-k)}{n} + k\log\tfrac{k}{n} + (n-k)\log\tfrac{n-k}{n} \\ &\quad -k\log E_n(k) - (n-k)\log(1-E_n(k)). \end{aligned}$$

By Lemma 5 (Section 15.6), we further obtain

$$-\log Bin_{n,E_n(k)}(k) =^+ \frac{1}{2}\log\frac{k(n-k)}{n}.$$

Now, the theorem follows from Corollary 4 and Theorem 17. □

By Theorem 19 $bin_n(k) \leq^+ \frac{1}{2}\log n$.

Following the general approach of this section we can consider a new definition of m-Bernoulli sequence, namely, a sequence x of the length n is called *m-Bernoulli* if $b_n(x) \leq m$. Theorem 18 shows the difference between this definition and Kolmogorov's definition of m-Bernoulli sequence from Section 4. It

says that the Bernoulli sequence x with unknown parameter θ must have not only complexity close to $\log \binom{n}{k}$, where k is the total number of 1s in x, but also k must have sufficiently large complexity.

This difference is based on two different underlaying statistical models. Kolmogorov's approach is based on the assumption that most sequences $x = x_1 \dots x_n$ of the same length n and with the same number $k = \sum_{i=1}^{n} x_i$ of 1s have identical statistical properties. The independence of $x_1, \dots, x_n$ is not assumed.

Levin's definition formalizes the requirements to a sequence $x = x_1 \dots x_n$ be typical with respect to some random experiment consisting in repetition of n independent trials $x_1, \dots, x_n$, i.e. some *i.i.d model* is considered.

Vovk (personal communication (1998)) mentioned that Kolmogorov's definition of the Bernoulli sequence is equivalent to the notion of randomness with respect to a natural class of *exchangeable* measures.

A probability distribution P^n in the set Ξ_n is called *exchangeable* if

$$P^n(x_1 x_2 \dots x_n) = P^n(x_{\pi(1)} x_{\pi(2)} \dots x_{\pi(n)}) \tag{34}$$

for any $x = x_1 x_2 \dots x_n \in \Xi_n$ and for any permutation π on the set $\{1, 2 \dots n\}$ (see [33]).

Let Φ_n^{ex} be the class of all computable exchangeable measures in the set Ξ_n. The *deficiency of exchangeability* is defined as

$$e_n(x) = \inf_{P \in \Phi_n^{ex}} d(x|n, P).$$

The following version of Vovk's theorem for binary case shows that Kolmogorov's deficiency and deficiency of exchangeability are coincide up to an additive constant.

Theorem 20 *$e_n(x) =^+ \log \binom{n}{k} - KP(x|n,k)$, where $n = l(x)$ is the length of x and $k = \sum_{i=1}^{n} x_i$ is the total number of 1s in x.*

Proof. By (34) $k = k(x)$ is a sufficient statistics for Φ_n^{ex} and there is one to one correspondence between computable exchangeable measures P^n and computable probability distributions $r(n,k)$ in the set $\{1, 2 \dots n\}$ such that

$$\sum_{k=0}^{n} r(n,k) = 1. \tag{35}$$

This correspondence is defined by the following equality

$$P^n(x) = \frac{r(n,k)}{\binom{n}{k}}, \tag{36}$$

where x is a binary sequence of the length n and k is the total number of 1s in x. By (36)

$$d(x|n,P^n) =^+ \log\binom{n}{k} - \log r(n,k) - KP(x|n,r(n,0),\dots,r(n,n)),$$

and then for any $x \in \Xi_n$

$$e_n(x) = \log\binom{n}{k} + \inf_r \{-\log r(n,k) - KP(x|n,r(n,0),\dots,r(n,n))\}, \qquad (37)$$

where $k = \sum_{i=1}^{n} x_i$ is the total number of 1s in x.

For any $x \in \Xi_n$ let $k = \sum_{i=1}^{n} x_i$. There is an exchangeable measure P^n such that $r(n,k) = 1$ and $r(n,k') = 0$ for all $k' \neq k$. Then $KP(x|n,r(n,0),\dots,r(n,n)) =^+ KP(x|n,k)$ and we have by (37)

$$e_n(x) \leq^+ \log\binom{n}{k} - KP(x|n,k).$$

To obtain the converse inequality we must prove that for any computable function $r(n,k)$ satisfying (35)

$$\log r(n,k) + KP(x|n,r(n,0),\dots,r(n,n)) \leq^+ KP(x|n,k).$$

By Theorem 10 we rewrite this inequality as

$$P(x|n,r(n,0),\dots,r(n,n)) \geq^\cdot r(n,k)P(x|n,k), \qquad (38)$$

where P is the a priory probability.

To prove (38) we define a conditional semimeasure Q semicomputable from below satisfying

$$Q(x|n,r_0,r_1,\dots r_n) = \sum_{i=0}^{n} P(x|n,i)r_i,$$

for each sequence $r_0,r_1,\dots r_n$ of computable non-negative real numbers such that $\sum_{i=0}^{n} r_i \leq 1$. We omit technical details of this definition.

Then by Theorem 9

$$P(x|n,r_0,r_1,\dots r_n) \geq^\cdot Q(x|n,r_0,r_1,\dots r_n).$$

In particular, for any computable function $r(n,k)$ satisfying (35) we have

$$P(x|n,r(n,0),\dots,r(n,n)) \geq^\cdot \sum_{i=0}^{n} P(x|n,i)r(n,i) \geq r(n,k)P(x|n,k),$$

where $k = \sum_{i=1}^{n} x_i$. This completes the proof of the theorem. $\Box$

By theorems 18 and 20 we have

$$b_n(x) =^+ e_n(x) + bin_n(k) + O(\log |bin_n(k)|).$$

13 Statistical complexity, II

In this section we present a connection between $KP(x|n)$ and $DL(x)$ in a pointwise form.

Theorem 21 *For $x \in \Xi_n$*

$$KP(x|n) + b_n(x) \leq^+ DL(x) \leq^+ KP(x|n) + 2\log KP(x|n) + b_n(x) + 2\log |b_n(x)|.$$

Proof. Comparing definitions of $DL_\theta(x)$ and $b_{n,\theta}(x)$ we have

$$DL_\theta(x) - b_{n,\theta}(x) = KP(x|n,\theta) + KP(\theta|n). \tag{39}$$

By (13) we have

$$KP(x|n) \leq^+ KP(x|n,\theta) + KP(\theta|n).$$

Hence, taking minimum by θ in (39), we obtain

$$DL(x) \geq^+ b_n(x) + KP(x|n).$$

To obtain the converse inequality we put $\theta = E(x) = E_n(k)$ and obtain from (39)

$$DL(x) = DL_{E(x)}(x) = b_{n,E_n(k)}(x) + KP(x|n, E_n(k)) + KP(E_n(k)|n).$$

As follows from Theorem 18 and its proof

$$b_{n,E_n(k)}(x) \leq^+ (\log \binom{n}{k} - KP(x|n,k)) + bin_n(k) + 2\log |bin_n(k)| \leq^+ b_n(x) + 2\log |bin_n(k)| \leq^+ b_n(x) + 2\log |b_n(x)|.$$

Since $k = k(x)$ is computable function, $KP(x|n) =^+ KP(x, E_n(k)|n)$ and $KP(E_n(k)|n) \leq^+ KP(x|n)$. Then by Corollary 2 we obtain

$$KP(x|n) \geq^+ KP(E_n(k)|n) + KP(x|n, E_n(k)) - 2\log KP(E_n(k)|n)$$

Hence,

$$DL(x) \leq^+ b_n(x) + 2\log |b_n(x)| + KP(x|n) + 2\log KP(x|n).$$

□

This theorem asserts that the corresponding minimal upper bound $DL(x)$ for complexity $KP(x|n)$ is almost the best possible, since it holds

$$\min_{\theta}\{-\log_2 B_\theta(x)+KP(\theta|n)\} = KP(x|n)+b_n(x)+O(\log_2 KP(x|n)+\log_2 |b_n(x)|).$$

We must suppose that the value $b_n(x)$ is small if we believe that some probability distribution from the class $B_{n,\theta}(x)$ generates x.

14 Time of decoding and descriptional complexity

The speed of computing of decoding algorithms was ignored in this paper. Let us consider an example from [24]. Let $A(p)$ be an optimal algorithm defining the Kolmogorov complexity $K(x)$:

$$K(x) = \min\{l(p)|A(p) = x\},$$

and let $f(p) = (A(p), l(p))$. Then for any x there is an p such that $f(p) = (x, K(x))$. Exhaustive search for such p takes exponential time, even when $f(p)$ is fast. Levin in [20] proposed a fastest algorithm finding p. The corresponding complexity was defined by Levin in the beginning of the seventies and used in his *optimal search* algorithm, see [20], [24], Section 1.3, and [27], Section 7.5.

Let $A(p, y)$ be an optimal function (machine) and $T_A(p, y)$ be the time of computation of the value $A(p, y)$, if such computation is terminated, and $T_A(p, y) = \infty$, otherwise. Define

$$Kt_A(x|y) = \min\{l(p) + \log T_A(p, y) \mid A(p, y) = x\}. \tag{40}$$

A theorem on the existence of an optimal complexity holds.

Theorem 22 *There exists a partial recursive optimal function $A(p, y)$ such that for each partial recursive function $B(p, y)$ the inequality*

$$Kt_A(x|y) \leq^+ Kt_B(x|y)$$

holds.

The scheme of the proof is analogous to that of Theorem 1. The logarithmic term in (40) requires a linear time of simulation.

To prove Theorem 23 below we also need to specify the model of computation. It is not known and probably not true that this theorem holds in Turing model (see a popular discussion in [11]). Usually, it is used the Kolmogorov – Uspensky machines, which are algorithms with semi-local transformation of

information, see [16], [38], Section 1.2. An additional requirement is that any machine from this class does not use any special mark for the end of the program and decide itself when to stop reading the program, i.e. it must never even read the digits following the program. In Section 8 these programs are called self-delimiting. Any machine from this class has a prefix-free domain.

It can be proved that Theorem 22 holds for Kolmogorov – Uspensky machines of this type. We fix some optimal complexity $Kt_A(x|y)$ and omit index A.

Using Levin's complexity $Kt(x|y)$ it is possible to solve any problem of a certain class of inverting problems in time that is optimal up to a constant factor. Let $f(x)$ be a partial recursive function of the type $\Xi \to \Xi$. The *search problem* uses a recursive $f(x)$ computable in polynomial time. An algorithm A inverts the function f, if given y it computes an x such that $f(x) = y$, if y is in the range of f, and this algorithm diverges, otherwise. An example of such problem is the following: given positive integer number to find some its factorization. To solve the inverting problem naively often requires exhaustive search through exponentially many candidates, that takes exponential time.

Now we can present the main result of this section. The following theorem was proved by Levin in the beginning of the seventies (see [20]).

Theorem 23 *An algorithm U (described below) exists such that the following holds. Let f be an arbitrary recursive function (not necessary polynomial time computable), and assume that a self-delimiting program q inverts the function f on any y and checks the result in time $t(y)$, i.e. produces an x, and finds out whether $f(x) = y$, actually applying the given algorithm for f to check the result (it is not permitted to use a faster algorithm to compute f).*

Then the algorithm U inverts f on y and checks the result in time $c2^{l(q)}t(y)$, where c is a constant.

Proof. We slightly modify the definition of Kt -

$$Kt(x|y, f) = \min \{l(p) + \log T_A(p, y) \mid A(p, y)) = x, f(x) = y\},$$

where $A(p, y)$ is an optimal Kolmogorov – Uspensky machine.

Given y the algorithm U runs all self-delimiting programs in succession according to the value of $Kt(x|y, f)$ for all possible candidates x. More correctly, for each $i = 1, 2, \ldots$, this algorithm lexicographically runs all self-delimiting programs p such that

$$l(p) + \log T(p, y) \leq i, \tag{41}$$

and checks for $x = A(p, y)$, whether $y = f(x)$ (applying the given algorithm for f), until a positive answer will be found.

The crucial point here is that we must check the inequality (41) in time $O(T(p, y))$, i.e. to solve the bounded (by time t) halting problem in time $O(t)$. A solution for Turing machines is unknown now. For Kolmogorov – Uspensky

machines the bounded (by time t) halting problem can be easily solved in time $O(t)$ by creating a clock and running it simultaneously with the computation.

Since the domain of $A(p,y)$ is prefix-free by p, we have for any y

$$\sum_{A(p,y)\ terminates} 2^{-l(p)} \le 1.$$

By (41) we have $T(p,y) \le 2^{i-l(p)}$ for all i. Therefore, the total time of terminated computation for any y is bounded by

$$c\sum_{i\le k}\ \sum_{l(p)\le i, A(p,y)\ terminates} 2^{i-l(p)} \le$$
$$c\sum_{i\le k} 2^i \sum_{l(p)\le i, A(p,y)\ terminates} 2^{-l(p)} \le c2^{k+1},$$

where $k = \min\{l(p) + \log T(p,y) \mid f(A(p,y)) = y\}$ and c is a positive constant.

Let some self-delimiting program q for an universal Kolmogorov – Uspensky machine B inverts f on y and finds out whether $f(x) = y$ in time $t(y) = T_B(q,y)$. Then by Theorem 22 we have

$$k \le^+ \{l(q) + \log T_B(q,y) \mid f(A(q,y)) = y\},$$

and $2^k \le^\cdot t(y)2^{l(q)}$. Therefore, the time of computation of the algorithm U on y is $\le c2^{l(q)}t(y)$ for some positive constant c. □

15 Some proofs and auxiliary results

15.1 Proof of Theorem 6

Let $x^1 = R[x]$, $n_1 = l(x^1)$ and x^2 be the remaining part of the x represented as concatenation of two binary sequences $x^2 = uv$. Here u are examined (but not selected in x^1) elements of x put in order in accordance with the examining procedure, v are the elements of x not examined put in order according their indices in x. Then the sequence x can be restored by n, x^1, x^2 and by the program computing R given n. Indeed, $f(\emptyset) = i_1$ gives the number of the first examined bit of x, and by the value of $g(\emptyset)$ we define where this element locates, in x^1, or in u. So, we can define the i_1-th bit of x (and the first examined element). Analogously, using previously defined bits of x, say, $x_{i_1}, x_{i_2}, \dots x_{i_k}$, we can define the following examined bit and its ordinal number in x (replacing $\emptyset$ on $x_{i_1}, x_{i_2}, \dots x_{i_k}$ in the previously defined part of procedure). The remaining bits of x can be taken from v in consecutive order. From this it follows

$$K(x|n) \le K(x^1) + \log K(x^1) + 2\log\log K(x^1) + n - n_1 +$$
$$K(R|n) + 2\log K(R|n) + c,$$

for some constant c. For all sufficiently large n_1 this inequality can be simplified

$$K(x|n) \leq K(x^1) + n - n_1 + (1 + \frac{1}{2}\epsilon)\log n_1 + K(R|n) + 2\log K(R|n). \quad (42)$$

Let

$$A_{n_1,m} = \left\{ x \in \Xi_{n_1} \mid \sum_{i=1}^{n_1} x_i = m \right\}.$$

Since $x^1 \in A_{n_1,m}$ we can encode x^1 by m, n_1 and its ordinal number in this set,

$$K(x^1) \leq \log m + 2\log\log m + \log n_1 + 2\log\log n_1 + \log \# A_{n_1,m} + c_1,$$

where c_1 is a constant. For all sufficiently large n_1 we have

$$K(x^1) \leq (2 + \frac{1}{2}\epsilon)\log n_1 + \log \binom{n_1}{m}. \quad (43)$$

From (42) and (43) we obtain

$$n_1 - \log \binom{n_1}{m} \leq d(x|n) + K(R|n) + 2\log K(R|n) + (3+\epsilon)\log n_1. \quad (44)$$

By Stirling's formula we obtain

$$\log \binom{n_1}{m} =^+ n_1 H(\frac{m}{n_1}) - \frac{1}{2}\log \frac{m(n_1-m)}{n_1}, \quad (45)$$

where $H(p) = -p\log p - (1-p)\log(1-p)$. Then by (44)

$$n_1(1 - H(\frac{m}{n_1})) + \frac{1}{2}\log \frac{m(n_1-m)}{n_1} \leq^+$$
$$d(x|n) + K(R|n) + 2\log K(R|n) + (3+\epsilon)\log n_1.$$

If $m = 0$ or $m = n_1$ then for $0 < \mu < 1$ we obtain a contradiction with (44) for all sufficiently large n_1 (we put $\binom{n_1}{0} = 1$). Then if $1 \leq m \leq n_1 - 1$ and $n_1 \geq 2$

$$\log \sqrt{\frac{m(n_1-m)}{n_1}} \geq -\frac{1}{2}.$$

By (9), (44) and (45) we have

$$n_1(1 - H(\frac{m}{n_1})) + \frac{1}{2}\log \frac{m(n_1-m)}{n_1} \leq^+ (3+\epsilon)\log n_1 + \mu n_1. \quad (46)$$

Then by properties of the function $H(p)$ nearby its maximum for each $\epsilon' > 0$ there is sufficiently small μ such that for all sufficiently large n_1 inequality (46) implies $|\frac{m}{n_1} - \frac{1}{2}| < \epsilon'$.

Since $\ln(1+r) \geq r - (1+\epsilon)r^2$ for all r with sufficiently small absolute value, we have

$$n_1(1 - H(\frac{m}{n_1})) \geq 4n_1(1-\epsilon)\log e \left(\frac{m}{n_1} - \frac{1}{2}\right)^2.$$

Hence, for all sufficiently large n_1

$$4n_1(1-\epsilon)\log e(\frac{m}{n_1} - \frac{1}{2})^2 \leq d(x|n) + K(R|n) + 2\log K(R|n) + (3+\epsilon)\log n_1.$$

From this we obtain (10). □

15.2 Proof of Theorem 8

Suppose that a finite sequence of positive integer numbers

$$J = (n, \alpha_1, \ldots, \alpha_k, \beta_1 \ldots, \beta_k),$$

where $c_1 \leq \alpha_1 < \ldots < \alpha_k$ and $n > \beta_1 > \ldots > \beta_k$, is given. We suppose also that $\alpha_0 = 0$, $\alpha_{k+1} = \infty$ and $\beta_{k+1} = 0$. Let us define an auxiliary function

$$\tilde{P}_j(U) = t_j^{-1} \sum_{K(D|J) \leq \alpha_j} \frac{\#(U \bigcap D)}{\#D},$$

where $1 \leq j \leq k$, t_j is the number of all addends in the sum and $U \subseteq \Xi_n$. By definition

$$\tilde{P}_j(U \bigcup V) = \tilde{P}_j(U) + \tilde{P}_j(V)$$

for each $U, V \in \Xi_n$ such that $U \bigcap V = \emptyset$.

We will define for any $1 \leq s \leq k$ the set $U_s \subseteq \Xi$ such that the following hold

- (i) $K(U_s|s, J) \leq \alpha_s + c$, where c is a constant;
- (ii) $\tilde{P}_j(U_s) < 2^{-l_j^s+1}$, where $l_j^s = \beta_j - \beta_{s+1} - s\log s - s$, $1 \leq j \leq s$.
- (iii) $\#U_s = 2^{n-(\beta_1 - \beta_{s+1})}$.

To define U_1 we represent $\Xi_n = \bigcup V_i$, where $\#V_i = 2^{n-(\beta_1-\beta_2)}$ and $V_i \bigcap V_{i'} = \emptyset$ for each i and $i' \neq i$. There are $< 2^{\beta_1-\beta_2}$ of i such that $\tilde{P}_1(V_i) > 2^{-(\beta_1-\beta_2)}$. Then there is an V_i such that

$$\tilde{P}_1(V_i) < 2^{-(\beta_1-\beta_2)+1} < 2^{-l_1^1+1}.$$

Put $U_1 = V_i$.

Let q_s, $s = 1, 2, \ldots k$, be a program of the length $\leq \alpha_s$ such that time of computation of $A(q, J) = D$ is maximal among all programs q of the length $\leq \alpha_s$ on which this computation is terminated (A is an optimal mode of description).

We have $K(U_1|J) \leq \alpha_1 + c$, since using q_1 we can compute all finite sets D such that $K(D|J) \leq \alpha_1$ and then run algorithm computing V_i.

Let us show how to reduce the construction of U_s, $s \geq 2$, to the construction of U_{s-1} satisfying (i)-(iii). We will also use s, J and the program q_s. To construct U_{s-1} we use q_{s-1} which can be easy computed by J, s and q_s.

Let us represent $U_{s-1} = \bigcup V_i$, where $\#V_i = 2^{n-(\beta_1-\beta_{s+1})}$ and $V_i \bigcap V_{i'} = \emptyset$ for each i and $i' \neq i$ The number of such V_i is equal to $2^{\beta_s-\beta_{s+1}}$, since $\#U_{s-1} = 2^{n-(\beta_1-\beta_s)}$. Let for $1 \leq j \leq s$ the number k_j is equal to the number of i such what $\tilde{P}_j(V_i) \geq 2^{-l_j^s}$. By (ii) for U_{s-1} we have for $1 \leq j < s$

$$k_j 2^{-(\beta_j-\beta_{s+1})+s\log s+s} \leq \tilde{P}_j(U_{s-1}) < 2^{-(\beta_j-\beta_s)+(s-1)\log(s-1)+s}.$$

Then $k_j < 2^{\beta_s-\beta_{s+1}-\log s}$. Besides, $k_s 2^{-l_s^s} \leq \tilde{P}_s(U_s) \leq 1$, and from (ii) we have $k_s < 2^{\beta_s-\beta_{s+1}-\log s}$. The total number of such V_i is equal to

$$k_1 + \ldots + k_s < s2^{\beta_s-\beta_{s+1}-\log s} \leq 2^{\beta_s-\beta_{s+1}}.$$

From this it follows that there exists V_i such that $\tilde{P}_j(V_i) < 2^{-l_j^s}$ for all $1 \leq j \leq s$. Put $U_s = V_i$. It is easy to verify that properties (i)-(iii) are satisfied.

Let $1 \leq j \leq k$ and D be an arbitrary finite set such that $K(D|J) \leq \alpha_j$. Then by (ii) for U_k

$$\#D \bigcap U_k < 2^{\alpha_j+1} \#D\tilde{P}_j(U_k) \leq \#D2^{-l_j+\alpha_j+2},$$

where $l_j = \beta_j - k\log k - k$.

By (i) we have $K(U_s|J) \leq \alpha_s + 2\log s + c_2$, where c_2 is a positive constant.

We need the following

Lemma 3 *If $V \subseteq D$ and $x \in V$ then $d(x|J,D) \geq^{\cdot} \log \frac{\#D}{\#V} - 2K(V|J)$.*

Proof. We have $K(x|V) \leq^{+} \log \#V$, and so, $K(x|J) \leq^{+} \log \#V + 2K(V|J)$. By definition

$$d(x|J,D) = \log \#D - K(x|J,D) \geq^{+} \log \#D - \log \#V - 2K(V|J).$$

□

By Lemma 3 for all $1 \leq j \leq k$, $K(D|J) \leq \alpha_j$, and $x \in U_k \bigcap D$

$$d(x|D,J) \geq \log \frac{\#D}{\#D \bigcap U_k} - 2K(D \bigcap U_k|J) \geq \tag{47}$$

$$l_j - \alpha_j - 2 - 6\alpha_k - c = l_j - 7\alpha_k - c = \tag{48}$$

$$\beta_j - 7\alpha_k - k\log k - k - c_3, \tag{49}$$

where c, c_3 are positive constants. Here we use $K(D \bigcap U_k|J) \leq^{+} \alpha_j + 2\alpha_k \leq 3\alpha_k$.

As noted earlier, in Section 4

$$\#\{x \mid x \in A, d(x|A,J) > m\} \leq 2^{\lceil \log \#A \rceil - m}$$

for each finite set A. Since $\#U_k = 2^{n-\beta_1}$, there is a set $W_1 \subseteq U_k$ such that

$$\#W_1 \geq 2^{n-\beta_1-1}$$

and

$$d(x|J, \Xi_n) \leq \beta_1 + 1$$

for each $x \in W_1$.

Since the number of $x \in W_1$ such that

$$d(x|U_1, J) = \log \#U_1 - K(x|U_1, J) > \beta_2 + 2$$

is less or equal to

$$2^{-\beta_2-2}\#U_1 = 2^{-\beta_2-2}2^{n-(\beta_1-\beta_2)} = 2^{n-\beta_1-2} \leq \frac{1}{2}\#W_1,$$

we can choose the set $W_2 \subseteq W_1$ such that

$$\#W_2 \geq 2^{n-\beta_1-2}$$

and

$$d(x|U_1, J) \leq \beta_2 + 2$$

for all $x \in W_2$.

Continuing this process, we choose $W_k \subseteq W_{k-1}$ such that

$$\#W_k \geq 2^{n-\beta_1-k}$$

and

$$d(x|U_{k-1}, J) \leq \beta_k + k$$

for all $x \in W_k$. We can do it since the number of $x \in W_{k-1}$ such that $d(x|U_{k-1}, J) > \beta_k + k$ is less or equal to

$$2^{-\beta_k-k}\#U_{k-1} = 2^{-\beta_k-k}2^{n-(\beta_1-\beta_k)} = 2^{n-\beta_1-k} \leq \frac{1}{2}\#W_{k-1}.$$

At the end of the process we choose the set $W_{k+1} \subseteq W_k$ such that $\#W_{k+1} \geq 2^{n-\beta_1-k-1}$ and $d(x|U_k, J) \leq k + 1$ for all $x \in W_{k+1}$.

We prove that any $x \in W_{k+1}$ satisfies the conclusion of the theorem. Since $W_{k+1} \subseteq U_k$, for any $x \in W_{k+1}$ and for all $1 \leq j \leq k + 1$ by (47)-(49) the following inequality

$$\beta_x(\alpha|J) \geq \beta_j - 7\alpha_k - k \log k - k - c_3 \tag{50}$$

holds for $\alpha \leq \alpha_j$.

For each $x \in W_j$ we have

$$d(x|U_{j-1}, J) \leq \beta_j + j$$

and

$$K(U_{j-1}|J) \le \alpha_{j-1} + 2\log j + c_2,$$

where $U_0 = \Xi_n$ end $1 \le j \le k+1$. Then for all $x \in W_{k+1} \subseteq W_j$ and $\alpha \ge \alpha_{j-1} + 2\log j + c_2$

$$\beta_x(\alpha|J) \le \beta_j + j.$$

Choose a constant c_1 such that $K(\Xi_n|n, J) \le c_1$ for each J. We take $c_1 \le \alpha_1$. Then (50) also holds for all $c_1 \le \alpha \le \alpha_1$. □

15.3 Proof of Theorem 9

Proof. In this proof we will omit the parameter y for simplicity of exposition. Let $U(q,p)$ be the universal function as in the proof of Theorem 1. Define $U^s(q,p) = U(q,p)$ if $l(p) \le s$, $l(q) \le s$ and the right hand-side value was computed within $\le s$ steps, and $U^s(q,p)$ undefined, otherwise. Let

$$s(q) = \sup\{s \mid U^s(q,p) \text{ is a prefix with respect to } p \text{ function}\}$$

(may be $s(q) = \infty$ and we suppose that in this case $U^{s(q)}(q,p) = U(q,p)$). So, $U^{s(q)}(q,p)$ is a prefix with respect to p function.

Let us define partial recursive function A such that

$$A(\overline{l(q)}01qp) = U^{s(q)}(q,p).$$

By definition, for each partial recursive function $B(p)$ there exists a program q such that $B(p) = U(q,p)$. If $B(p)$ is a prefix function then $s(q) = \infty$. From this $KP_A(x) \le KP_B(x) + l(q) + 2\log l(q) + 2$ holds for all x. This means that $KP_A(x) \le^+ KP_B(x)$ for any partial recursive prefix function B. □

15.4 Proof of Theorem 10

The proof is based on the possibility to effectively enumerate all semimeasures enumerable from below. More accurately, we can define a sequence $\{P_i\}$ of semimeasures such that

- $\{(i,r,x,y) \mid r < P_i(x|y),\ r \in \mathbb{Q}\}$ is a recursively enumerable set;
- $\sum_x P_i(x|y) \le 1$ for each i and y;
- for each semimeasure Q enumerable from below there is an i such that $Q = P_i$.

Such a sequence can be defined as follows. Each conditional semimeasure $Q(x|y)$ enumerable from below defines a recursively enumerable set

$$\{(r,x,y) \mid r \in \mathbb{Q}, r < Q(x|y)\}.$$

We will effectively enumerate all such sets as follows. Let $U(i,n)$ be a function universal for all partial recursive functions of the type $\mathbb{N} \to \mathbb{Q} \otimes \Xi \otimes \Xi$, i.e. taking as values triples (r,x,y), where $r \in \mathbb{Q}$ and $x,y \in \Xi$.

Let W_i be the range of the function $n \to U(i,n)$, and W_i^k be the finite subset of W_i consisting of the values $U(i,n)$, $n \le k$, which are computed in $\le k$ steps. Then $W_i^k \subseteq W_i^{k+1}$ and $W_i = \bigcup_k W_i^k$. Further, we will pick out of W_i the maximal subset defining semimeasure. Define

$$P_i^k(x|y) = \max(\{r \mid (r,x,y) \in W_i^k\} \bigcup \{0\});$$

$$P_i(x|y) = \sup_k \{P_i^k(x|y) \mid \sum_z P_i^k(z|y) \le 1\}.$$

Let Q be any semimeasure enumerable from below and $U(i,n), n = 1,2,\ldots$ enumerate the set $\{(r,x,y) \mid r \in \mathbb{Q}, r < Q(x|y)\}$. Then this set is equal to W_i and $P_i(x|y) = Q(x|y)$ for all x,y.

Define

$$P(x|y) = \sum_i \frac{1}{2i^2} P_i(x|y).$$

Evidently, this function is enumerable from below. Besides,

$$\sum_x P(x|y) = \sum_x \sum_i \frac{1}{2i^2} P_i(x|y) =$$
$$\sum_i \frac{1}{2i^2} \sum_x P_i(x|y) \le \sum_i \frac{1}{2i^2} < 1.$$

Let $Q(x|y)$ be an arbitrary semimeasure enumerable from below such that $\sum\limits_x Q(x|y) \le 1$ for each y. Then $Q = P_i$ for some i. This implies $2i^2 P(x|y) \ge Q(x|y)$ for each x,y. $\square$

15.5 Proof of Lemma 1

This proof is due to Shen (see [45], Lemma 2.1, see also [39], Section 3.4 and [27], Section 4.3.3).

For any finite binary sequence x we consider an interval

$$\Gamma_x = \{\omega | \omega \in \Omega, x \subseteq \omega\}$$

in the Cantor space Ω of all infinite binary sequences. The Lebesgue measure of such interval is $L(\Gamma_x) = 2^{-l(x)}$.

Suppose $x_1, x_2 \ldots, x_n$ are already chosen and satisfy the following property: the set

$$\Omega - (\Gamma_{x_1} \bigcup \Gamma_{x_2} \ldots \bigcup \Gamma_{x_n})$$

can be represented as the union of pairwise disjoint intervals $\Gamma_{t_1}, \Gamma_{t_2}, \ldots, \Gamma_{t_q}$, where lengths of all t_i are distinct. Since this property holds for $n = 0$, we must only prove that it is possible to find x_{n+1} of length k_{n+1} which is incomparable with all $x_1, x_2 \ldots, x_n$ and such that $x_1, x_2 \ldots, x_{n+1}$ satisfy above property. There is a sequence of length at most k_{n+1} among $t_1, t_2 \ldots, t_q$ since otherwise we would have

$$L(\Gamma_{t_1} \bigcup \Gamma_{t_2} \ldots \bigcup \Gamma_{t_q}) < \sum_{s > k_{n+1}} 2^{-s} = 2^{-k_{n+1}}$$

and

$$L(\Gamma_{x_1} \bigcup \Gamma_{x_2} \ldots \bigcup \Gamma_{x_n}) > 1 - 2^{-k_{n+1}},$$

which would contradict to

$$\sum_n 2^{-k_n} \leq 1.$$

Without loss of generality we assume that $l(t_1) \leq k_{n+1}$ and that t_1 is the longest such sequence. If $l(t_1) = k_{n+1}$ we put $x_{n+1} = t_1$ and have representation

$$\Omega - (\Gamma_{x_1} \bigcup \ldots \bigcup \Gamma_{x_n} \bigcup \Gamma_{x_{n+1}}) = \Gamma_{t_2} \bigcup \ldots \bigcup \Gamma_{t_q}.$$

If $l(t_1) < k_{n+1}$ we represent Γ_{t_1} as the union of pairwise disjoint intervals $\Gamma_{a_1}, \ldots, \Gamma_{a_s}$ such that among $a_i, i = 1, \ldots, s$, there are two sequences, say a_1 and a_2, of length k_{n+1}, one sequence of length $k_{n+1}-1$, one of length $k_{n+1}-2 \ldots$, and one of length $l(t_1) + 1$. Putting $x_{n+1} = a_1$ we get the representation

$$\Omega - (\Gamma_{x_1} \bigcup \ldots \bigcup \Gamma_{x_{n+1}}) = \Gamma_{t_2} \bigcup \ldots \bigcup \Gamma_{t_q} \bigcup \Gamma_{a_2} \bigcup \ldots \bigcup \Gamma_{a_s}.$$

There are no two sequences of the same length among $t_2, \ldots, t_q, a_2, \ldots, a_s$ since the lengths of $t_2, \ldots, t_q$ and the lengths of $a_2, \ldots, a_s$ are pairwise distinct and the lengths of $a_2, \ldots, a_s$ are in the interval between $l(t_1) + 1$ and k_{n+1} whereas the lengths of $t_2, \ldots, t_q$ are outside this interval. □.

15.6 Some properties of the log-likelihood function for the Bernoulli family

In this section we reproduce several technical lemmas on the Bernoulli family from [43].

Lemma 4 *When $n \in \mathbb{N}$, $\alpha \in [1/2, \pi n^{1/2}/2 - 1/2]$, and $a, b \in [0, \pi n^{1/2}/2]$ range so that $a \leq \alpha \leq b$ and $1/2 \leq b - a \leq 2$, we have*

$$\sin^2(bn^{-1/2}) - \sin^2(an^{-1/2}) =^{\cdot} n^{-1/2} \sin(\alpha n^{-1/2}) \cos(\alpha n^{-1/2}). \qquad (51)$$

Proof. Equivalent transformations of (51) yield:

$$\begin{gathered}\left(\sin(bn^{-1/2})-\sin(an^{-1/2})\right)\left(\sin(bn^{-1/2})+\sin(an^{-1/2})\right)\\ =^{\cdot} n^{-1/2}\sin(\alpha n^{-1/2})\cos(\alpha n^{-1/2});\\ \cos\left(\tfrac{b+a}{2}n^{-1/2}\right)\sin\left(\tfrac{b-a}{2}n^{-1/2}\right)\sin\left(\tfrac{b+a}{2}n^{-1/2}\right)\cos\left(\tfrac{b-a}{2}n^{-1/2}\right)\\ =^{\cdot} n^{-1/2}\sin(\alpha n^{-1/2})\cos(\alpha n^{-1/2}).\end{gathered}$$

Our task has reduced to proving that

$$\cos\left(\tfrac{b+a}{2}n^{-1/2}\right) =^{\cdot} \cos(\alpha n^{-1/2}), \tag{52}$$

$$\sin\left(\tfrac{b-a}{2}n^{-1/2}\right) =^{\cdot} n^{-1/2}, \tag{53}$$

$$\sin\left(\tfrac{b+a}{2}n^{-1/2}\right) =^{\cdot} \sin(\alpha n^{-1/2}), \tag{54}$$

$$\cos\left(\tfrac{b-a}{2}n^{-1/2}\right) =^{\cdot} 1. \tag{55}$$

Equalities (53) and (55) immediately follow from $1/2 \le b-a \le 2$; (52) and (54) reduce, in view of $a \le \alpha \le b$ and $\alpha \in [1/2, \pi n^{1/2}/2 - 1/2]$, to

$$\cos\left(\frac{\pi}{2}-\frac{1}{2}n^{-1/2}\right) =^{\cdot} \cos\left(\frac{\pi}{2}-\frac{1}{4}n^{-1/2}\right)$$

and

$$\sin\left(\frac{1}{2}n^{-1/2}\right) =^{\cdot} \sin\left(\frac{1}{4}n^{-1/2}\right),$$

respectively; these two relations are equivalent, and the second of them is obviously true. □

Corollary 4 *When* $k \in \{1, \dots, n-1\}$,

$$\#E_n^{-1}(E_n(k)) =^{\cdot} \sqrt{\frac{k(n-k)}{n}}. \tag{56}$$

Proof. For simplicity, we shall assume that $E_n^{-1}(E_n(k))$ always consists of consecutive elements of the set $\{0, \dots, n\}$. Define a, α, b by the conditions

$$\begin{gathered}\sin^2(an^{-1/2}) = \tfrac{1}{n}\inf E_n^{-1}(E_n(k)),\\ \sin^2(bn^{-1/2}) = \tfrac{1}{n}\sup E_n^{-1}(E_n(k)),\\ \sin^2(\alpha n^{-1/2}) = k/n.\end{gathered}$$

We can see that $a \le \alpha \le b$, $\alpha \in [1, \pi n^{1/2}/2 - 1]$, and $1/2 \le b-a \le 2$. Since

$$\#E_n^{-1}(E_n(k)) =^{\cdot} n\left(\sin^2(bn^{-1/2})-\sin^2(an^{-1/2})\right)$$

and

$$\begin{gathered}\sqrt{\tfrac{k(n-k)}{n}} = \sqrt{n\tfrac{k}{n}\left(1-\tfrac{k}{n}\right)}\\ = \sqrt{n\sin^2(\alpha n^{-1/2})\cos^2(\alpha n^{-1/2})} = \sqrt{n}\sin(\alpha n^{-1/2})\cos(\alpha n^{-1/2}),\end{gathered}$$

we can rewrite (56) as

$$\sin^2(bn^{-1/2}) - \sin^2(an^{-1/2}) =^{\cdot} n^{-1/2}\sin(\alpha n^{-1/2})\cos(\alpha n^{-1/2}),$$

which coincides with (51). □

The following two lemmas describe important properties of the log-likelihood function for the Bernoulli model. We use the notation $G_{n,k}$ for the log-likelihood function expressed through the variable a (which will not longer be assumed to be integer) introduced by $\theta_n(a) = \sin^2(an^{-1/2})$:

$$G_{n,k}(a) = \ln\left(\sin^{2k}(an^{-1/2})\cos^{2(n-k)}(an^{-1/2})\right),$$

a ranging over $[0, \pi n^{1/2}/2]$.

We use the notation $\hat{a}(n,k)$, for the maximum likelihood estimate of the parameter a:

$$\hat{a}(n,k) = \arg\max_a G_{n,k}(a)$$

(therefore, $\sin^2(\hat{a}(n,k)n^{-1/2}) = k/n$).

Lemma 5 *When $n \geq 1$, $a \in [1, \pi n^{1/2}/2 - 1]$, and $k \in \{1, \ldots, n-1\}$ range so that $|a - \hat{a}(n,k)| < 1$,*

$$G_{n,k}(a) =^{+} G_{n,k}(\hat{a}(n,k)).$$

Proof. Denote $\hat{a} = \hat{a}(n,k)$. It suffices to prove that the values

$$\sup_a \left|\frac{d^2 G_{n,k}(a)}{da^2}\right|, \tag{57}$$

where a ranges over

$$[1, \pi n^{1/2}/2 - 1] \cap [\hat{a} - 1, \hat{a} + 1],$$

do not exceed some bound. Calculating the second derivative, we rewrite (57) as

$$2\sup_a\left(\frac{k/n}{\sin^2\left(an^{-1/2}\right)} + \frac{1 - k/n}{\cos^2\left(an^{-1/2}\right)}\right).$$

Note that

$$k/n = \sin^2\left(\hat{a}n^{-1/2}\right), \quad 1 - k/n = \cos^2\left(\hat{a}n^{-1/2}\right),$$

so it suffices to prove that

$$\sup_a \frac{\sin^2\left((a+1)n^{-1/2}\right)}{\sin^2\left(an^{-1/2}\right)} \quad \text{and} \quad \sup_a \frac{\cos^2\left((a-1)n^{-1/2}\right)}{\cos^2\left(an^{-1/2}\right)}$$

are bounded above by some constant. It is easy to see that both suprema are equal to

$$\frac{\sin^2\left(2n^{-1/2}\right)}{\sin^2\left(n^{-1/2}\right)} \to 4 \quad (n \to \infty),$$

so they are indeed bounded. □

It follows from Lemma 5 that

$$\ln((E_n(k))^k(1-E_n(k))^{n-k}) =^+ G_{n,k}(\hat{a}(n,k))$$

(we can see that if $\sin^2(an^{-1/2}) = k/n$ and $1 \le k \le n-1$ then $a \in [1, \pi n^{1/2}/2 - 1]$).

Lemma 6 *Let $n \ge 1$, $a \in [0, \pi n^{1/2}/2]$, and $k \in \{1, \ldots, n-1\}$. For some constant $\epsilon > 0$,*

$$G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a) \ge^+ \epsilon |a - \hat{a}(n,k)|.$$

Proof. Denote $\hat{a} = \hat{a}(n,k)$. By the symmetry of the problem, we can suppose $a > \hat{a}$. Furthermore, we can consider only the case $a \ge \hat{a} + 1/2$. Since $G_k''(a)$ is negative everywhere, it is sufficient to prove that $-G_k'(\hat{a}+1/2)$ is greater than some constant $\epsilon > 0$. We find

$$-G_k'(a) = 2n^{-1/2}\left((n-k)\frac{\sin\left(an^{-1/2}\right)}{\cos\left(an^{-1/2}\right)} - k\frac{\cos\left(an^{-1/2}\right)}{\sin\left(an^{-1/2}\right)}\right),$$

so we are required to prove

$$\begin{gathered}(n-k)\sin^2\left((\hat{a}+1/2)n^{-1/2}\right) - k\cos^2\left((\hat{a}+1/2)n^{-1/2}\right) \\ > \tfrac{\epsilon}{2}n^{1/2}\sin\left((\hat{a}+1/2)n^{-1/2}\right)\cos\left((\hat{a}+1/2)n^{-1/2}\right).\end{gathered}$$

This inequality is equivalent to

$$\begin{gathered}n\sin^2\left((\hat{a}+1/2)n^{-1/2}\right) - k \\ > \tfrac{\epsilon}{2}n^{1/2}\sin\left((\hat{a}+1/2)n^{-1/2}\right)\cos\left((\hat{a}+1/2)n^{-1/2}\right),\end{gathered}$$

or

$$\begin{gathered}\sin^2\left((\hat{a}+1/2)n^{-1/2}\right) - \sin^2\left(\hat{a}n^{-1/2}\right) \\ > \tfrac{\epsilon}{2}n^{-1/2}\sin\left((\hat{a}+1/2)n^{-1/2}\right)\cos\left((\hat{a}+1/2)n^{-1/2}\right).\end{gathered}$$

The last inequality immediately follows from Lemma 4.

The following inequalities

$$\begin{gathered}G_{n,k}(\hat{a}) - G_{n,k}(a) \ge G_{n,k}(\hat{a}+1/2) - G_{n,k}(a) = \\ -G'(\tilde{a})(a - \hat{a} - 1/2) \ge \epsilon(a-\hat{a}) - \epsilon/2,\end{gathered}$$

where $\hat{a} + 1/2 < \tilde{a} < a$, complete the proof. □

Lemma 7. *Let $n \geq 1$, $a \in [0, \pi n^{1/2}/2]$, and $k \in \{1, \ldots, n-1\}$. Then*

$$KP(\lfloor \hat{a}(n,k) \rfloor \, | n, a) \leq^+ (\ln^{-1} 2)(G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a)).$$

Proof. By Lemma 6 $G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a)) \geq^+ \epsilon |a - \hat{a}(n,k)|$ for some $\epsilon > 0$. Then the assertion of the lemma follows from

$$KP(\lfloor \hat{a}(n,k) \rfloor \, | n, a) \leq^+ KP(\lfloor \hat{a}(n,k) \rfloor \, | n, \lfloor a \rfloor) \leq^+$$
$$KP(\lfloor \hat{a}(n,k) \rfloor - \lfloor a \rfloor \, | n) \leq^+ 2 \log | \lfloor \hat{a}(n,k) \rfloor - \lfloor \hat{a} \rfloor \, | \leq^+ \epsilon (\ln^{-1} 2) | \lfloor \hat{a}(n,k) \rfloor - \lfloor \hat{a} \rfloor \, | =^+$$
$$\epsilon (\ln^{-1} 2) | \hat{a}(n,k) - a |.$$

□

Acknowledgments

Leonid Levin explained some subtle points of Section 14 and gave important historical comments. Alexander Shen's critical remarks helped us to improve several proofs. Konstantin Gorbunov and Yura Kalnichkan checked preliminary versions of this paper. The author is deeply grateful to all of them and especially to Volodya Vovk for useful discussions.

References

[1] E.A. Asarin (1987) Some properties of Kolmogorov δ random finite sequences, *SIAM Theory Probab. Appl.* **32**, 507–508.

[2] G.J. Chaitin (1966) On the length of programs for computing binary sequences, *J. Assoc. Comput. Mach.* **13**, 547–569.

[3] G.J. Chaitin (1969) On the length of programs for computing binary sequences: Statistical considerations, *J. Assoc. Comput. Mach.* **16**, 145–159.

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

[5] G.J. Chaitin (1977) Algorithmic information theory, *IBM J. Res. Develop.* **21**, 350–359.

[6] G.J. Chaitin (1987) Algorithmic information theory, Cambridge University Press.

[7] T.M. Cover, P. Gács, R.M. Gray (1989) Kolmogorov's contributions to information theory and algorithmic complexity, *Ann. Probab.* **17**, No.1, 840–865.

[8] A.P.Dawid, V.G. Vovk (1997) Prequential probability: principles and properties, *Bernoulli* **3**, pp. 1–38.

[9] P. Gács, J.Körner. (1973) Common information is far less than mutual information, *Problems. of Control. and Inform. Theory* **2**, 149–162.

[10] P. Gács (1974) On the symmetry of algorithmic information, *Soviet. Math. Dokl.* **15**, 1477–1480.

[11] Y. Gurevich (1988) The logik in computer science column, *Bulletin of European Assoc. for Theor. Comp. Science* **35)**, June 1988, 71–82.

[12] A.N. Kolmogorov (1963) On tables of random numbers, *Sankhyaã, The Indian Journal of Statistics, Ser. A* **25)**, 369–376.

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

[14] A.N. Kolmogorov (1968) The logical basis for information theory and probability theory, *IEEE Trans. Inf. Theory* IT **14**, 662–664.

[15] A.N. Kolmogorov (1983) Combinatorial basis of information theory and probability theory, *Russ. Math. Surveys.* **38**, 29–40.

[16] A.N. Kolmogorov, V.A. Uspensky (1958) To the definition of an Algorithm, *Uspekhi Mat. Nauk* **13:4**, 3–28 (Russian); English translation in AMS Translations, ser. 2, vol. 21 (1963), 217–245.

[17] A.N. Kolmogorov, V.A. Uspensky (1987) Algorithms and randomness, *Theory Probab. Applic.* **32**, 389–412.

[18] M. Lambalgen. (1987) Random sequences, Amsterdam: Academish Proefshri't.

[19] Sik K. Leung-Yan-Chrong, T.M.Cover (1978) Some equivalences between Shannon entropy and Kolmogorov complexity, *IEEE. Trans. Inform. Theory.* **IT-24**, 331–338.

[20] L.A. Levin (1973) Universal search problems, *Problems. Inform. Transmission.* **9**, 265–266.

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

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

[23] L.A. Levin (1976) Various measures of complexity for finite objects (axiomatic description), *Soviet Math. Dokl.* **17**, 522–526.

[24] L.A. Levin (1984) Randomness conservation inequalities; information and independence in mathematical theories, *Inform. and Contr.* **61**, 15–37.

[25] M. Li, P. Vitányi (1988) Kolmogorovskaya slozhnost' dvadsat' let spustia, *Uspekhi Mat. Nauk* **43(6)**, 129–166, in Russian.

[26] M. Li, P. Vitányi (1994) Statistical properties of finite sequences with high Kolmogorov complexity, *Math. Systems Theory* **27**, 365–376.

[27] M. Li, P. Vitányi (1997) An Introduction to Kolmogorov Complexity and Its Applications, New York: Springer–Verlag.

[28] P. Martin-Löf (1966) The definition of random sequences, *Inform. and Contr.* **9 (6)**, 602–619.

[29] An. A. Muchnik (1998) On common information, *Theor. Comp. Sci.* **207**, 319–328.

[30] J. Rissanen (1983) A universal prior for integers and estimation by minimum description length, *Ann. Statist.* **11**, 416–431.

[31] J. Rissanen (1985) Minimum description length principle, in Encyclopedia of Statistical Sciences" (S.Kotz and N.L.Johnson, Eds.) Vol. **5**, 523–527, New York: Wiley

[32] H. Rogers (1967) Theory of recursive functions and effective computability, New York: McGraw Hill.

[33] Mark J. Schervish (1995) Theory of statistics, New York: Springer.

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

[35] R.J. Solomonoff (1964) A formal theory of inductive inference I, II, *Inform. Control* **7**, 1–22, 224–254.

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

[37] V.A. Uspensky, A.L. Semenov, A.Kh. Shen (1990) Can an individual sequence of zeros and ones be random?, *Russian Math. Surveys* **45 (1)**, 121–189.

[38] V.A. Uspensky, A.L. Semenov (1993) Algorithms: Main Ideas and Applications, Kluwer Academic Publishers, also in Lect. Notes Comput. Sci., vol 122, A.P.Ershov and D.E.Knuth, Eds., Springer-Verlag, 1981, pp.100–234.

[39] V.A. Uspensky, A. Shen (1996) Relations between varieties of Kolmogorov complexities, *Math. Systems Theory* **29**, 271–292.

[40] V.G. Vovk (1986) On the concept of the Bernoulli property, *Russian Math. Surveys* **41**, 247–248.

[41] V.G. Vovk (1993) A logic of probability, with application to the foundations of statistics, *J. Royal. Statist. Soc. B* **55**, N2, 317–351.

[42] V.G. Vovk (1995) Minimum description length estimators under the optimal coding scheme, in "Computational Learning Theory"(P. Vitanyi, Ed.) *Lect. Notes Comp. Sci.* **904**, 237–251.

[43] V.G. Vovk (1997) Learning about parameter of the Bernoulli model, *J. Comp. Syst. Sci.* **55**, N1, 96–104.

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

[45] V.V. V'yugin (1981) Algorithmic entropy (complexity) of finite objects, and its application to defining randomness and quantity of information, *Semiotika and Informatica* **16**, 14–43. In Russian. Translated into English in *Selecta Mathematica* formerly *Sovietica* **13(4)**, 1994, 357–389.

[46] C.S.Wallace, P.R.Freeman (1987) Estimation and inference by compact coding, *J. R. Statist. Soc.* **49:3**, 240–265.

[47] A.K. Zvonkin and L.A. Levin (1970) The complexity of finite objects and the algorithmic concepts of information and randomness, *Russ. Math. Surv.* **25**, 83–124.