Algorithmic Statistics: Normal Objects and Universal Models

Alexey Milovanov[1,2(✉)]

[1] Moscow State University, Moscow, Russian Federation
almas239@gmail.com
[2] National Research University Higher School of Economics, Moscow, Russian Federation
https://www.hse.ru/en/org/persons/176000050

Abstract. In algorithmic statistics quality of a statistical hypothesis (a model) P for a data x is measured by two parameters: Kolmogorov complexity of the hypothesis and the probability $P(x)$. A class of models S_{ij} that are the best at this point of view, were discovered. However these models are too abstract.

To restrict the class of hypotheses for a data, Vereshchaginintroduced a notion of a strong model for it. An object is called normal if it can be explained by using strong models not worse than without this restriction. In this paper we show that there are "many types" of normal strings. Our second result states that there is a normal object x such that all models S_{ij} are not strong for x. Our last result states that every best fit strong model for a normal object is again a normal object.

Keywords: Algorithmic statistics · Minimum description length · Stochastic strings · Total conditional complexity · Sufficient statistic · Denoising

1 Introduction

Let us recall the basic notion of algorithmic information theory and algorithmic statistics (see [4,6,8] for more details).

We consider strings over the binary alphabet $\{0,1\}$. The set of all strings is denoted by $\{0,1\}^*$ and the length of a string x is denoted by $l(x)$. The empty string is denoted by Λ.

1.1 Algorithmic Information Theory

Let D be a partial computable function mapping pairs of strings to strings. *Conditional Kolmogorov complexity* with respect to D is defined as

$$C_D(x|y) = \min\{l(p) \mid D(p,y) = x\}.$$

In this context the function D is called a *description mode* or a *decompressor*. If $D(p,y) = x$ then p is called a *description of x conditional to y* or a *program mapping y to x*.

A.S. Kulikov and G.J. Woeginger (Eds.): CSR 2016, LNCS 9691, pp. 280–293, 2016.
DOI: 10.1007/978-3-319-34171-2_20

A decompressor D is called *universal* if for every other decompressor D' there is a string c such that $D'(p, y) = D(cp, y)$ for all p, y. By Solomonoff—Kolmogorov theorem [2] universal decompressors exist. We pick arbitrary universal decompressor D and call $C_D(x|y)$ *the Kolmogorov complexity* of x conditional to y, and denote it by $C(x|y)$. Then we define the unconditional Kolmogorov complexity $C(x)$ of x as $C(x|\Lambda)$.

Kolmogorov complexity can be naturally extended to other finite objects (pairs of strings, finite sets of strings, etc.). We fix some computable bijection ("encoding") between these objects are binary strings and define the complexity of an object as the complexity of the corresponding binary string. It is easy to see that this definition is invariant (change of the encoding changes the complexity only by $O(1)$ additive term).

In particular, we fix some computable bijection between strings and finite subsets of $\{0,1\}^*$; the string that corresponds to a finite $A \subset \{0,1\}^*$ is denoted by $[A]$. Then we understand $C(A)$ as $C([A])$. Similarly, $C(x|A)$ and $C(A|x)$ are understood as $C(x|[A])$ and $C([A]|x)$, etc.

1.2 Algorithmic Statistics: Basic Notions

Algorithmic statistics studies explanations of observed data that are suitable in the algorithmic sense: an explanation should be simple and capture all the algorithmically discoverable regularities in the data. The data is encoded, say, by a binary string x. In this paper we consider explanations (statistical hypotheses) of the form "x was drawn at random from a finite set A with uniform distribution".

Kolmogorov suggested in a talk [3] in 1974 to measure the quality of an explanation $A \ni x$ by two parameters: Kolmogorov complexity $C(A)$ of A and the log-cardinality $\log |A|$[1] of A. The smaller $C(A)$ is the simpler the explanation is. The log-cardinality measures the *fit* of A—the lower is $|A|$ the more A fits as an explanation for any of its elements. For each complexity level m any model A for x with smallest $\log |A|$ among models of complexity at most m for x is called a *best fit hypothesis for* x. The trade off between $C(A)$ and $\log |A|$ is represented by the *profile* of x.

Definition 1. *The* profile *of a string x is the set P_x consisting of all pairs (m, l) of natural numbers such that there exists a finite set $A \ni x$ with $C(A) \le m$ and $\log_2 |A| \le l$.*

Both parameters $C(A)$ and $\log |A|$ cannot be very small simultaneously unless the string x has very small Kolmogorov complexity. Indeed, $C(A) + \log |A| \gtrsim C(x)$, since x can be specified by A and its index in A. A model (we also use the word "statistic") $A \ni x$ is called *sufficient* if $C(A) + \log |A| \approx C(x)$. The value

$$\delta(x|A) = C(A) + \log |A| - C(x)$$

is called the *optimality deficiency* of A as a model for x. On Fig. 1 parameters of sufficient statistics lie on the segment BD. A sufficient statistic that has the

[1] by log we denote $\log_2$.

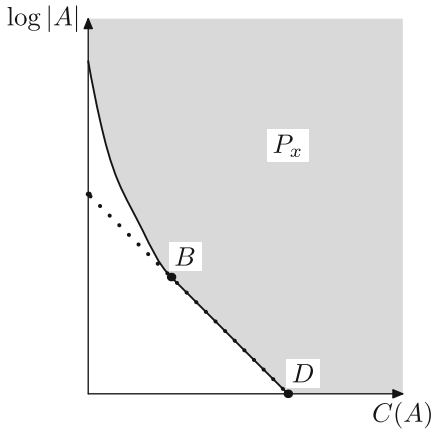

Fig. 1. The profile P_x of a string x.

minimal complexity is called *minimal* (MSS), its parameters are represented by the point B on Fig. 1.

Example 1. Consider a string $x \in \{0,1\}^{2n}$ such that leading n bits of x are zeros, and the remaining bits are random, i.e. $C(x) \approx n$. Consider the model A for x that consists of all strings from $\{0,1\}^{2n}$ that have n leading zeros. Then $C(A) + \log|A| = \log n + O(1) + n \approx C(x)$, hence A is a sufficient statistic for x. As the complexity of A is negligible, A is a minimal sufficient statistic for x.

The string from this example has a sufficient statistic of negligible complexity. Such strings are called *stochastic*. Are there strings that have no sufficient statistics of negligible complexity? The positive to this question was obtained in [7]. Such strings are called *non-stochastic*. Moreover, under some natural constraints for every set P there is a string whose profile is close to P. The constraints are listed in the following theorem:

Theorem 1. *Let x be a string of length n and complexity k. Then P_x has the following properties:*

(1) $(k + O(\log n), 0) \in P_x$.
(2) $(O(\log n), n) \in P_x$.
(3) if $(a, b + c) \in P_x$ *then* $(a + b + O(\log n), c) \in P_x$.
(4) if $(a, b) \in P_x$ *then* $a + b > k - O(\log n)$.

In other words, with logarithmic accuracy, the boundary of P_x contains a point $(0, a)$ with $a \le l(x)$, contains the point $(C(x), 0)$, decreases with the slope at least -1 and lies above the line $C(A) + \log|A| = C(x)$. Conversely, given a curve with these property that has low complexity one can find a string x of length n and complexity about k such that the boundary of P_x is close to that curve:

Theorem 2 [10]. *Assume that we are given k, n and an upward closed set P of pairs of natural numbers such that $(0,n),(k,0) \in P$, $(a,b+c) \in P \Rightarrow (a+c,b) \in P$ and $(a,b) \in P \Rightarrow a+b \geq k$. Then there is a string x of length n and complexity $k+O(\log n)$ whose profile is $C(P)+O(\log n)$-close to P. (We call subsets of N^2 ϵ-*close *if each of them is in the ϵ-neighborhood of the other.) By $C(P)$ we denote the Kolmogorov complexity of the boundary of P, which is a finite object.*

1.3 Universal Models

Assume that A is a sufficient statistic for x. Then A provides a two-part code $y =$ (the shortest description of A, the index of x in A) for x whose total length is close to the complexity of x. The symmetry of information implies that $C(y|x) \approx C(y) + C(x|y) - C(x)$. Obviously, the term $C(x|y)$ here is negligible and $C(y)$ is at most its total length, which by assumption is close to $C(x)$. Thus $C(y|x) \approx 0$, that is, x and y have almost the same information. That is, the two-part code y for x splits the information from x in two parts: the shortest description of A, the index of x in A. The second part of this two-part code is incompressible (random) conditional to the first part (as otherwise, the complexity of x would be smaller than the total length of y). Thus the second part of this two-part code can be considered as accidental information (noise) in the data x. In a sense every sufficient statistic A identifies about $C(x) - C(A)$ bits of accidental information in x. And thus any minimal sufficient statistic for x extracts almost all useful information from x.

However, it turns out that this viewpoint is inconsistent with the existence of universal models, discovered in [1]. Let L_m denote the list of strings of complexity at most m. Let p be an algorithm that enumerates all strings of L_m in some order. Notice that there is such algorithm of complexity $O(\log m)$. Denote by Ω_m the cardinality of L_m. Consider its binary representation, i.e., the sum:

$$\Omega_m = 2^{s_1} + 2^{s_2} + ... + 2^{s_t}, \text{ where } s_1 > s_2 > ... > s_t.$$

According to this decomposition and p, we split L_m into groups: first 2^{s_1} elements, next 2^{s_2} elements, etc. Let us denote by $S^p_{m,s}$ the group of size 2^s from the partition. Notice that $S^p_{m,s}$ is defined only for s that correspond to ones in the binary representation of Ω_m, so $m \geq s$.

If x is a string of complexity at most m, it belongs to some group $S^p_{m,s}$ and this group can be considered as a model for x. We may consider different values of m (starting from $C(x)$). In this way we get different models $S^p_{m,s}$ for the same x. The complexity of $S^p_{m,s}$ is $m-s+O(\log m+C(p))$. Indeed, chop L_m into portions of size 2^s each, then $S^p_{m,s}$ is the last full portion and can be identified by m, s and the number of full portions, which is less than $\Omega_m/2^s < 2^{m-s+1}$. Thus if m is close to $C(x)$ and $C(p)$ is small then $S^p_{m,s}$ is a sufficient statistic for x. More specifically $C(S^p_{m,s}) + \log |S^p_{m,s}| = C(S^p_{m,s}) + s = m + O(\log m + C(p))$.

For every m there is an algorithm p of complexity $O(\log m)$ that enumerates all strings of complexity at most m. We will fix for every m any such algorithm p_m and denote $S^{p_m}_{m,s}$ by $S_{m,s}$.

The models $S_{m,s}$ were introduced in [1]. The models $S^p_{m,s}$ are universal in the following sense:

Theorem 3 [10]. [2] *Let A be any finite set of strings containing a string x of length n. Then for every p there are $s \le m \le n + O(1)$ such that*

(1) $x \in S^p_{m,s}$,
(2) $C(S^p_{m,s}|A) = O(\log n + C(p))$ *(and hence* $C(S^p_{m,s}) \le C(A) + O(\log n + C(p))$*),*
(3) $\delta(x|S^p_{m,s}) \le \delta(x|A) + O(\log n + C(p))$.

It turns out that the model $S^p_{m,s}$ has the same information as the the number Ω_{m-s}:

Lemma 1 [10]. *For every $a \le b$ and for every $s \le m$:*

(1) $C(\Omega_a|\Omega_b) = O(\log b)$.
(2) $C(\Omega_{m-s}|S^p_{m,s}) = O(\log m + C(p))$ *and* $C(S^p_{m,s}|\Omega_{m-s}) = O(\log m + C(p))$.
(3) $C(\Omega_a) = a + O(\log a)$.

By Theorem 3 for every data x there is a minimal sufficient statistic for x of the form $S_{m,s}$. Indeed, let A be any minimal sufficient statistic for x and let $S_{m,s}$ be any model for x that exists by Theorem 3 for this A. Then by item 3 the statistic $S_{m,s}$ is sufficient as well and by item 2 its complexity is also close to minimum. Moreover, since $C(S_{m,s}|A)$ is negligible and $C(S_{m,s}) \approx C(A)$, by symmetry of information $C(A|S_{m,s})$ is negligible as well. Thus A has the same information as $S_{m,s}$, which has the same information as Ω_{m-s} (Lemma 1(2)). Thus if we agree that every minimal sufficient statistic extracts all useful information from the data, we must agree also that information is the same as the information in the number of strings of complexity at most i for some i.

1.4 Total Conditional Complexity and Strong Models

The paper [9] suggests the following explanation to this situation. Although conditional complexities $C(S_{m,s}|A)$ and $C(S_{m,s}|x)$ are small, the short programs that map A and x, respectively, to $S_{m,s}$ work in a huge time. A priori their work time is not bounded by any total computable function of their input. Thus it may happen that practically we are not able to find $S_{m,s}$ (and also Ω_{m-s}) from a MSS A for x or from x itself.

Let us consider now programs whose work time is bounded by a total computable function for the input. We get the notion of *total conditional complexity* $CT(y|x)$, which is the length of the shortest *total* program that maps x to y. Total conditional complexity can be much greater than plain one, see for example [5]. Intuitively, good sufficient statistics A for x must have not only negligible

[2] This theorem was proved in [10, Theorem VIII.4] with accuracy $O(\max\{\log C(y) \mid y \in A\} + C(p))$ instead of $O(\log n)$. Applying [10, Theorem VIII.4] to $A' = \{y \in A \mid l(y) = n\}$ we obtain the theorem in the present form.

conditional complexity $C(A|x)$ (which follows from definition of a sufficient statistic) but also negligible *total* conditional complexity $CT(A|x)$. The paper [9] calls such models A *strong models for* x.

Is it true that for some x there is no **strong** MSS $S_{m,s}$ for x? The positive answer to this question was obtained in [9]: there are strings x for which all minimal sufficient statistics are not strong for x. Such strings are called *strange*. In particular, if $S_{m,s}$ is a MSS for a strange string x then $CT(S_{m,s}|x)$ is large. However, a strange string has no strong MSS at all. An interesting question is whether there are strings x that do have strong MSS but have no strong MSS of the form $S_{m,s}$? This question was left open in [9]. In this paper we answer this question in positive. Moreover, we show that there is a "normal" string x that has no strong MSS of the form $S_{m,s}$ (Theorem 7). A string x is called *normal* if for every complexity level i there is a best fitting model A for x of complexity at most i (whose parameters thus lie on the border of the set P_x) that is strong. In particular, every normal string has a strong MSS.

Our second result answers yet another question asked in [9]. Assume that A is a strong MSS for a normal string x. Is it true that the code $[A]$ of A is a normal string itself? Our Theorem 10 states that this is indeed the case.

Our last result (which comes first in the following exposition) states that there are normal strings with any given profile, under the same restrictions as in Theorem 1 (Theorem 4 in Sect. 2).

2 Normal Strings with a Given Profile

In this section we prove an analogue of Theorem 2 for normal strings. We start with a rigorous definition of strong models and normal strings.

Definition 2. *A set $A \ni x$ is called* ϵ-strong statistic (model) for a string x *if* $CT(A|x) < \epsilon$.

To represent the trade off between size and complexity of ϵ-strong models for x consider the *ϵ-strong profile of x*:

$$P_x^\epsilon = \{(a,b) \mid \exists A \ni x : CT(A|x) \le \epsilon, C(A) \le a, \log|A| \le b\}.$$

It is not hard to see that the set P_x^ϵ satisfies the item (3) from Theorem 1:

for all $x \in \{0,1\}^n$ if $(a, b+c) \in P_x^\epsilon$ then $(a+b+O(\log n), c) \in P_x^{\epsilon+O(\log n)}$.

It follows from the definition that $P_x^\epsilon \subset P_x$ for all x, ϵ. Informally a string is called normal if for a negligible ϵ we have $P_x \approx P_x^\epsilon$.

Definition 3. *a string x is called* (ϵ,δ)-normal *if $(a,b) \in P_x$ implies $(a+\delta, b+\delta) \in P_x^\epsilon$ for all a, b.*

The smaller ϵ, δ are the stronger is the property of (ϵ,δ)-normality. The main result of this section shows that for some $\epsilon, \delta = o(n)$ for every set P satisfying the assumptions of Theorem 1 there is an ϵ,δ-normal string of length n with $P_x \approx P$:

Theorem 4. *Assume that we are given an upward closed set P of pairs of natural numbers satisfying assumptions of Theorem 2. Then there is an $(O(\log n), O(\sqrt{n \log n}))$-normal string x of length n and complexity $k + O(\log n)$ whose profile P_x is $C(P) + O(\sqrt{n \log n})$-close to P.*

To prove this theorem we do an excursus to Algorithmic statistics with models of restricted type.

Models of Restricted Type. It turns out that Theorems 1 and 2 remain valid (with smaller accuracy) even if we restrict (following [11]) the class of models under consideration to models from a class $\mathcal{A}$ provided the class $\mathcal{A}$ has the following properties.

(1) The family $\mathcal{A}$ is enumerable. This means that there exists an algorithm that prints elements of $\mathcal{A}$ as lists of strings, with some separators (saying where one element of $\mathcal{A}$ ends and another one begins).

(2) For every n the class $\mathcal{A}$ contains the set $\{0,1\}^n$.

(3) There exists some polynomial p with the following property: for every $A \in \mathcal{A}$, for every natural n and for every natural $c < |A|$ the set of all n-bit strings in A can be covered by at most $p(n) \cdot |A|/c$ sets of cardinality at most c from $\mathcal{A}$.

Any family of finite sets sets of strings that satisfies these three conditions is called *acceptable*.

Let us define the *profile of x with respect to $\mathcal{A}$*:

$$P_x^{\mathcal{A}} = \{(a,b) \mid \exists A \ni x : A \in \mathcal{A}, C(A) \le a, \log |A| \le b\}.$$

Obviously $P_x^{\mathcal{A}} \subseteq P_x$. Let us fix any acceptable class $\mathcal{A}$ of models.

Theorem 5 [11]. *Let x be a string of length n and complexity k. Then $P_x^{\mathcal{A}}$ has the following properties:*

(1) $(k + O(\log n), 0) \in P_x^{\mathcal{A}}$.
(2) $(O(\log n), n) \in P_x^{\mathcal{A}}$.
(3) if $(a, b+c) \in P_x^{\mathcal{A}}$ *then* $(a + b + O(\log n), c) \in P_x^{\mathcal{A}}$.
(4) if $(a,b) \in P_x^{\mathcal{A}}$ *then* $a + b > k - O(\log n)$.

Theorem 6 [11]. *Assume that we are given k, n and an upward closed set P of pairs of natural numbers such that $(0,n), (k,0) \in P$, $(a, b+c) \in P \Rightarrow (a+c, b) \in P$ and $(a,b) \in P \Rightarrow a + b \ge k$. Then there is a string x of length n and complexity $k + O(\log n)$ such that* both sets $P_x^{\mathcal{A}}$ *and P_x are $C(P) + O(\sqrt{n \log n})$-close to P.*

Remark 1. Originally, the conclusion of Theorem 6 stated only that the set $P_x^{\mathcal{A}}$ is close to the given set P. However, as observed in [8], the proof from [11] shows also that P_x is close to P.

Proof (Proof of Theorem 4). We will derive this theorem from Theorem 6. To this end consider the following family $\mathcal{B}$ of sets. A set B is in this family if it has the form

$$B = \{uv \mid v \in \{0,1\}^m\},$$

where u is an arbitrary binary string and m is an arbitrary natural number. Obviously, the family $\mathcal{B}$ is acceptable, that is, it satisfies the properties (1)–(3) above.

Note that for every x and for every $A \ni x$ from $\mathcal{B}$ the total complexity of A given x is $O(\log n)$. So $P_x^{\mathcal{B}} \subseteq P_x^{O(\log n)}$. By Theorem 6 there is a string x such that P_x and $P_x^{\mathcal{B}}$ are $C(P) + O(\sqrt{n \log n})$-close to P. Since $P_x^{\mathcal{B}} \subseteq P_x^{O(\log n)} \subseteq P_x$ we conclude that x is $(O(\log n), O(\sqrt{n \log n}))$-normal.

Instead of using Theorem 4 one can, in special cases, show this result directly even within a better accuracy range.

For instance, this happens for the smallest set P, satisfying the assumptions of Theorem 6, namely for the set

$$P = \{(m, l) \mid m \geq k, \text{ or } m + l \geq n\}.$$

Strings with such profile are called "antistochastic".

Definition 4. *A string x of length n and complexity k is called* ϵ-antistochastic *if for all $(m, l) \in P_x$ either $m > k - \epsilon$, or $m + l > n - \epsilon$ (Fig. 2).*

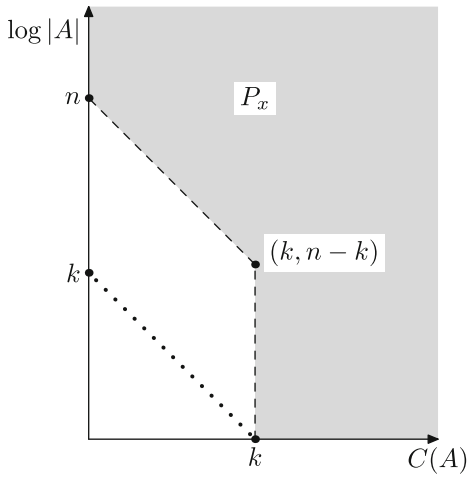

Fig. 2. The profile of an ϵ-antistochastic string x for a small ϵ.

We will need later the fact that for every n there is an $O(\log n)$-antistochastic string x of length n and that such strings are normal:

Lemma 2. *For all n and all $k \leq n$ there is an $O(\log n)$-antistochastic string x of length n and complexity $k + O(\log n)$. Any such string x is $(O(\log n), O(\log n))$-normal.*

Proof. Let x be the lexicographic first string of length n that is not covered by any set A of cardinality 2^{n-k} and complexity less than k. By a direct counting such a string exists. The string x can be computed from k, n and the number of halting programs of length less than k hence $C(x) \le k + O(\log n)$. To prove that x is normal it is enough to show that for every $i \le k$ there is a $O(\log n)$-strong statistics A_i for x with $C(A_i) \le i + O(\log n)$ and $\log |A_i| = n - i$.

Let $A_k = \{x\}$ and for $i < k$ let A_i be the set of all strings of length n whose the first i bits are the same as those of x. By the construction $C(A_i) \le i + O(\log n)$ and $\log |A_i| = n - i$.

3 Normal Strings Without Universal MSS

Our main result of this section is Theorem 7 which states that there is a normal string x such that no set $S_{m,l}$ is a strong MSS for x.

Theorem 7. *For all large enough k there exist an $(O(\log k), O(\log k))$-normal string x of complexity $3k + O(\log k)$ and length $4k$ such that:*

(1) The profile P_x of x is $O(\log k)$-close to the gray set on Fig. 3.
(2) The string x has a strong MSS. More specifically, there is an $O(\log k)$-strong model A for x with complexity $k + O(\log k)$ and log-cardinality $2k$.
(3) For all simple q and all m, l the set $S^q_{m,l}$ cannot be a strong sufficient statistic for x. More specifically, for every ϵ-strong ϵ-sufficient model $S^q_{m,l}$ for x of complexity at most $k + \delta$ we have $O(\epsilon + \delta + C(q)) \ge k - O(\log k)$

(The third condition means that there are constants r and t such that $r(\epsilon + \delta + C(q)) \ge k - t \log k$ for all large enough k).

In the proof of this theorem we will need a rigorous definition of MSS and a related result from [9].

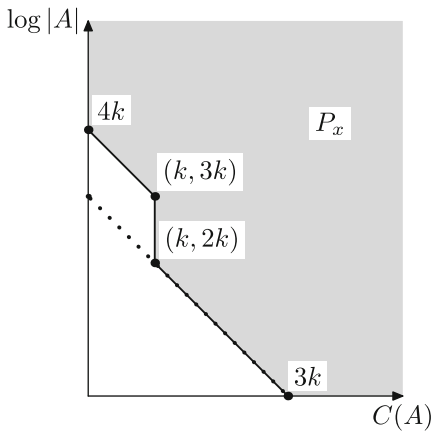

Fig. 3. The profile P_x of a string x from Theorem 7.

Definition 5. *A set A is called a (δ, ϵ, D)-minimal sufficient statistic (MSS) for x if A is an ϵ-sufficient statistic for x and there is no model B for x with $C(B) < C(A) - \delta$ and $C(B) + \log |B| - C(x) < \epsilon + D \log C(x)$.*

The next theorem states that for every strong MSS B and for every sufficient statistic A for x the total conditional complexity $CT(B|A)$ is negligible.

Theorem 8 ([9], Theorem 13). *For some constant D if B is ϵ-strong (δ, ϵ, D)-minimal sufficient statistic for x and A is an ϵ-sufficient statistic for x then $CT(B|A) = O(\epsilon + \delta + \log C(x))$.*

Let us fix a constant D satisfying Theorem 8 and call a model (δ, ϵ)-MSS if it is (δ, ϵ, D)-MSS. Such models have the following property.

Theorem 9 ([9], Theorem 14). *Let x be a string of length n and A be an ϵ-strong ϵ-sufficient statistic for x. Then for all $b \geq \log |A|$ we have*

$$(a, b) \in P_x \Leftrightarrow (a + O(\epsilon + \log n), b - \log |A| + O(\epsilon + \log n)) \in P_{[A]}$$

and for $b \leq \log |A|$ we have $(a, b) \in P_x \Leftrightarrow a + b \geq C(x) - O(\log n)$.

Proof (The proof of Theorem 7). Define x as the concatenation of strings y and z, where y is an $O(\log k)$-antistochastic string of complexity k and length $2k$ (existing by Lemma 2) and z is a string of length $2k$ such that $C(z|y) = 2k - O(\log k)$ (and hence $C(x) = 3k + O(\log k)$). Consider the following set $A = \{yz' \mid l(z') = 2k\}$. From the shape of P_x it is clear that A is an $(O(\log k), O(\log k))$-MSS for x. Also it is clear that A is an $O(\log k)$-strong model for x. So, by Theorem 9 the profile of x is $O(\log k)$-close to the gray set on Fig. 3. From normality of y (Lemma 2) it is not difficult to see that x is $(O(\log k), O(\log k))$-normal.

Let $S^q_{m,l}$ be an ϵ-strong ϵ-sufficient model for x of complexity at most $k + \delta$. We claim that $S^q_{m,l}$ is an $(\epsilon, \delta + O(\log k))$-MSS for x.

By Theorem 8 we get $CT(S^q_{m,l}|A) = O(\epsilon + \delta + \log k)$ and thus $CT(s_0|y) = O(\epsilon + \delta + \log k)$, where s_0 is the lexicographic least element in $S^q_{m,l}$. Denote by p a total program of length $O(\epsilon + \delta + \log k)$ that transforms y to s_0. Consider the following set $B := \{p(y') \mid l(y') = 2k\}$. We claim that if ϵ and δ are not very big, then the complexity of any element from B is not greater than m. Indeed, if $\epsilon + \delta \leq dk$ for a small constant d, then $l(p) < k - O(\log k)$ and hence every element from B has complexity at most $C(B) + \log |B| + O(\log k) \leq 3k - O(\log k) \leq m$. The last inequality holds because $S^q_{m,l}$ is a model for x and hence $m \geq C(x) = 3k + O(\log k)$. (Otherwise, if $\epsilon + \delta > dk$ then the conclusion of the theorem is straightforward.)

Let us run the program q until it prints all elements from B. Since $s_0 \in B$, there are at most 2^l elements of complexity m that we have been printed yet. So, we can find the list of all strings of complexity at most m from B, q and some extra l bits. Since this list has complexity at least $m - O(\log m)$ (as from this list and m we can compute a string of complexity more than m), we get $O(C(B) + C(q)) + l \geq m - O(\log m)$.

Recall that the $C(S^q_{m,l}) + \log |S^q_{m,l}|$ is equal to $m + O(\log m + C(q))$ and is at most $C(x) + \epsilon$ (since $S^q_{m,l}$ is the strong statistic for x). Hence $m \leq 4k$ unless $\epsilon > k + O(\log k + C(q))$. Therefore the term $O(\log m)$ in the last inequality can be re-written as $O(\log k)$.

Recall that the complexity of $S^q_{m,l}$ is $m - l + O(\log m + C(q))$. From the shape of P_x it follows that $C(S^q_{m,l}) \geq k - O(\log k)$ or $C(S^q_{m,l}) + \log |S^q_{m,l}| \geq C(x) + k - O(\log k)$. In the latter case $\epsilon \geq k - O(\log k)$ and we are done. In the former case $m - l \geq k - O(\log k + C(q))$ hence $O(C(B) + C(q)) \geq k - O(\log k + C(q))$ and so $O(\epsilon + \delta + C(q)) \geq k - O(\log k)$.

4 Hereditary of Normality

In this section we prove that every strong MSS for a normal string is itself normal. Recall that a string x is called (ϵ, δ)-normal if for every model B for x there is a model A for x with $CT(A|x) \leq \epsilon$ and $C(A) \leq C(B) + \delta$, $\log |A| \leq \log |B| + \delta$.

Theorem 10. *There is a constant D such that the following holds. Assume that A is an ϵ-strong (δ, ϵ, D)-MSS for an (ϵ, ϵ)-normal string x of length n. Assume that $\epsilon \leq \sqrt{n}/2$. Then the code $[A]$ of A is $O((\epsilon + \delta + \log n) \cdot \sqrt{n})$-normal.*

The rest of this section is the proof of this theorem. We start with the following lemma, which is a simple corollary of Theorem 3 and Lemma 1.

Lemma 3. *For all large enough D the following holds: if A is a (δ, ϵ, D)-MSS for $x \in \{0,1\}^n$ then $C(\Omega_{C(A)}|A) = O(\delta + \log n)$.*

We fix a constant D satisfying Lemma 3 and call a model (δ, ϵ)-MSS if it (δ, ϵ, D)-MSS. This D is the constant satisfying Theorem 10

A family of sets $\mathcal{A}$ is called *partition* if for every $A_1, A_2 \in \mathcal{A}$ we have $A_1 \cap A_2 \neq \varnothing \Rightarrow A_1 = A_2$. Note that for a finite partition we can define its complexity. The next lemma states that every strong statistic A can be transformed into a strong statistic A_1 such that A_1 belongs to some partition of similar complexity.

Lemma 4. *Let A be an ϵ-strong statistic for $x \in \{0,1\}^n$. Then there is a set A_1 and a partition $\mathcal{A}$ of complexity at most $\epsilon + O(\log n)$ such that:*

(1) A_1 is $\epsilon + O(\log n)$-strong statistic for x.
(2) $CT(A|A_1) < \epsilon + O(\log n)$ and $CT(A_1|A) < \epsilon + O(\log n)$.
(3) $|A_1| \leq |A|$.
(4) $A_1 \in \mathcal{A}$.

Proof. Assume that A is an ϵ-strong statistic for x. Then there is a total program p such that $p(x) = A$ and $l(p) \leq \epsilon$.

We will use the same construction as in Remark 1 in [9]. For every set B denote by B' the following set: $\{x' \in B \mid p(x') = B,\ x' \in \{0,1\}^n\}$. Notice that $CT(A'|A)$, $CT(A|A')$ and $CT(A'|x)$ are less than $l(p) + O(\log n) = \epsilon + O(\log n)$ and $|A'| \leq |A|$.

For any $x_1, x_2 \in \{0,1\}^n$ with $p(x_1) \neq p(x_2)$ we have $p(x_1)' \cap p(x_2)' = \varnothing$. Hence $\mathcal{A} := \{p(x)' | x \in \{0,1\}^n\}$ is a partition of complexity at most $\epsilon + O(\log n)$.

By Theorem 3 and Lemma 1 for every $A \ni x$ there is a $B \ni x$ such that B is informational equivalent to $\Omega_{C(B)}$ and parameters of B are not worse than those of A. We will need a similar result for normal strings and for strong models.

Lemma 5. *Let x be an (ϵ, α)-normal string with length n such that $\epsilon \le n$, $\alpha < \sqrt{n}/2$. Let A be an ϵ-strong statistic for x. Then there is a set H such that:*

(1) H is an ϵ-strong statistic for x.
(2) $\delta(x|H) \le \delta(x|A) + O((\alpha + \log n) \cdot \sqrt{n})$ and $C(H) \le C(A)$.
(3) $C(H|\Omega_{C(H)}) = O(\sqrt{n})$.

Proof (Sketch of proof). Consider the sequence $A_1, B_1, A_2, B_2, \ldots$ of statistics for x defined as follows. Let $A_1 := A$ and let B_i be an improvement of A_i such that B_i is informational equivalent to $\Omega_{C(B_i)}$, which exists by Theorem 3. Let A_{i+1} be a strong statistic for x that has a similar parameters as B_i, which exists because x is normal. (See Fig. 4.)

Denote by N the minimal integer such that $C(A_N) - C(B_N) \le \sqrt{n}$. For $i < N$ the complexity of B_i is more than $\sqrt{n}$ less that of A_i. On the other hand, the complexity of A_{i+1} is at most $\alpha < \sqrt{n}/2$ larger than that of B_i. Hence $N = O(\sqrt{n})$. Let $H := A_N$. By definition A_N (and H) is strong. From $N = O(\sqrt{n})$ it follows that the second condition is satisfied. From $C(A_N) - C(B_N) \le \sqrt{n}$ and definition of B_N it is follows that the third condition is satisfied too (use symmetry of information).

Proof (Sketch of proof of Theorem 10). Assume that A is a ϵ-strong (δ, ϵ, D)-minimal statistic for x, where D satisfies Lemma 3. By Lemma 3 A is informational equivalent to $\Omega_{C(A)}$. We need to prove that the profile of $[A]$ is close to the strong profile of $[A]$.

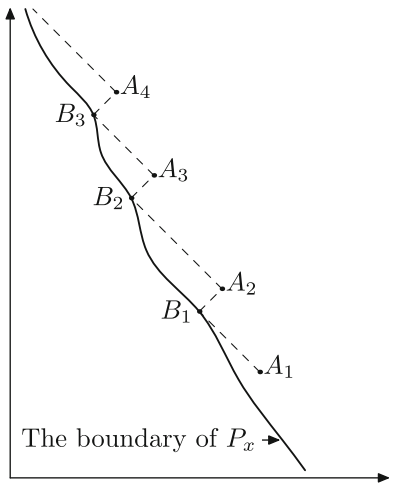

Fig. 4. Parameters of statistics A_i and B_i

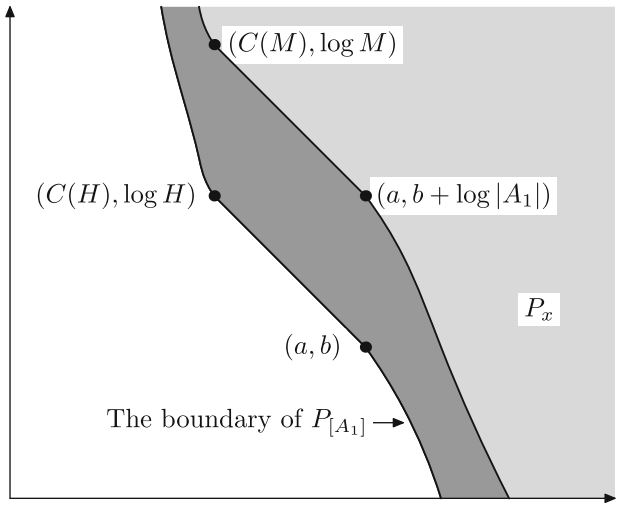

Fig. 5. P_x is located $\log |A_1|$ higher than $P_{[A_1]}$

Let $\mathcal{A}$ be a simple partition and A_1 a model from $\mathcal{A}$ which exists by Lemma 4 applied to A, x. As the total conditional complexities $CT(A_1|A)$ and $CT(A|A_1)$ are small, the profiles of A and A_1 are close to each other. This also applies to strong profiles. Therefore it suffices to show that (the code of) A_1 is normal.

Let $(a, b) \in P_{[A_1]}$. The parameters (complexity and log-cardinality) of A_1 are not larger than those of A and hence A_1 is a sufficient statistic for x. By Theorem 9 we have $(a, b + \log |A_1|) \in P_x$ (see Fig. 5).

As x is normal, the pair $(a, b + \log |A_1|)$ belongs to the strong profile of x as well. By Lemma 5 there is a **strong** model M for x that has low complexity conditional to $\Omega_{C(M)}$ and whose parameters (complexity, optimality deficiency) are not worse than those of A_1.

We claim that $C(M|A_1)$ is small. As A is informational equivalent to $\Omega_{C(A)}$, so is A_1. From $\Omega_{C(A)}$ we can compute $\Omega_{C(M)}$ (Lemma 1) and then compute M (as $C(M|\Omega_{C(M)}) \approx 0$). This implies that $C(M|A_1) \approx 0$.

However we will need a stronger inequality $CT(M|A_1) \approx 0$. To find such M, we apply Lemma 4 to M, x and change it to a model M_1 with the same parameters that belongs to a simple partition $\mathcal{M}$. Item (2) of Lemma 4 guarantees that M_1 is also simple given A_1 and that M_1 is a strong model for x. Since $C(M|A_1) \approx 0$, we have $C(M_1|A_1) \approx 0$ as well.

As A_1 lies on the border line of P_x and $C(M_1|A_1) \approx 0$, the intersection $A_1 \cap M_1$ cannot be much less than A_1, that is, $\log |A_1 \cap M_1| \approx \log |A_1|$ (otherwise the model $A_1 \cap M_1$ for x would have much smaller cardinality and almost the same complexity as A_1). The model M_1 can be computed by a total program from A_1 and its index among all $M' \in \mathcal{M}$ with $\log |A_1 \cap M'| \approx \log |A_1|$. As $\mathcal{M}$ is a partition, there are few such sets M'. Hence $CT(M_1|A_1) \approx 0$.

Finally, let $H = \{A' \in \mathcal{A} \mid \log|A' \cap M_1| = \log|A_1 \cap M_1|\}$. The model H for A_1 is strong because the partition $\mathcal{A}$ is simple and $CT(M_1|A_1) \approx 0$. The model H can be computed from M_1, $\mathcal{A}$ and $\log|A_1 \cap M_1|$. As $\mathcal{A}$ is simple, we conclude that $C(H) \lesssim C(M_1)$. Finally $\log|H| \leq \log|M_1| - \log|A_1|$, because $\mathcal{A}$ is a partition and thus it has few sets that have $\log|A_1 \cap M_1| \approx \log|A_1|$ common elements with M_1.

Thus the complexity of H is not larger than that of M_1 and the sum of complexity and cardinality of H is at most $a + b - \log|A_1|$. As the strong profile of x has the third property from Theorem 1, we can conclude that it includes the point (a, b).

Acknowledgments. The author is grateful to professor N. K. Vereshchagin for statements of questions, remarks and useful discussions.

This work is supported by RFBR grant 16-01-00362 and partially supported by RaCAF ANR-15-CE40-0016-01 grant. The study has been funded by the Russian Academic Excellence Project '5-100'.

References

1. Gács, P., Tromp, J., Vitányi, P.M.B.: Algorithmic statistics. IEEE Trans. Inform. Theory **47**(6), 2443–2463 (2001)
2. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Probl. Inf. Trans. **1**(1), 1–7 (1965)
3. Kolmogorov, A.N.: The complexity of algorithms, the objective definition of randomness. Usp. Matematicheskich Nauk **29**(4(178)), 155 (1974). Summary of the talk presented April 16, at Moscow Mathematical Society
4. Li, M., Vitányi, P.: An Introduction to Kolmogorov complexity and its applications, 3rd edn., xxiii+790 p. Springer, New York (2008). (1st edn. 1993; 2nd edn. 1997), ISBN 978-0-387-49820-1
5. Shen, A.: Game arguments in computability theory and algorithmic information theory. In: Cooper, S.B., Dawar, A., Löwe, B. (eds.) CiE 2012. LNCS, vol. 7318, pp. 655–666. Springer, Heidelberg (2012)
6. Shen, A.: Around kolmogorov complexity: basic notions and results. In: Vovk, V., Papadoupoulos, H., Gammerman, A. (eds.) Measures of Complexity: Festschrift for Alexey Chervonenkis. Springer, Heidelberg (2015)
7. Shen, A.: The concept of (α, β)-stochasticity in the Kolmogorov sense, and its properties. Sov. Math. Dokl. **271**(1), 295–299 (1983)
8. Shen, A., Uspensky, V., Vereshchagin, N.: Kolmogorov complexity and algorithmic randomness. MCCME (2013) (Russian). English translation: http://www.lirmm.fr/~ashen/kolmbook-eng.pdf
9. Vereshchagin, N.: Algorithmic minimal sufficient statistics: A new approach. Theory Comput. Syst. **56**(2), 291–436 (2015)
10. Vereshchagin, N., Vitányi, P.: Kolmogorov's structure functions with an application to the foundations of model selection. IEEE Trans. Inf. Theory **50**(12), 3265–3290 (2004). Preliminary version: Proceedings of the 47th IEEE Symposium on Foundations of Computer Science, pp. 751–760 (2002)
11. Vitányi, P., Vereshchagin, N.: On algorithmic rate-distortion function. In: Proceedings of 2006 IEEE International Symposium on Information Theory, Seattle, Washington, 9–14 July 2006