Electronic Colloquium on Computational Complexity, Report No. 90 (2010)

Algorithmic Minimal Sufficient Statistics: a New Definition

Nikolay Vereshchagin*

May 14, 2010

Abstract

We express some criticism about the definition of an algorithmic sufficient statistic and, in particular, of an algorithmic minimal sufficient statistic. We propose another definition, which has better properties.

1 Introduction

Let x be a binary string. A finite set A containing x is called an (algorithmic) sufficient statistic for x if the sum of Kolmogorov complexity of A and the log-cardinality of A is close to Kolmogorov complexity $C(x)$ of x:

$$C(A) + \log_2 |A| \approx C(x). \tag{1}$$

Let A^* denote a minimal length description of A and i the index of x in the list of all elements of A, arranged lexicographically. The equality (1) means that the two part description (A^*, i) of x is as concise as the minimal length code of x.

It turns out that A is a sufficient statistic for x iff $C(A|x) \approx 0$ and $C(x|A) \approx \log |A|$. The former equality means that the information in A^* is a part of information in x. The latter equality means that x is a typical member of A: x has no regularities that allow to describe x given A in a shorter way than just by specifying its $\log |A|$-bit index in A. Thus A^* contains all useful

*Moscow State University

ISSN 1433-8092

information present in x and i contains only an accidental information (a noise).

Sufficient statistics may also contain a noise. For example, it happens for x being a random string[1] and $A = \{x\}$. Is it true that for all x there is a sufficient statistic that contains no noise? To answer this question we can try to use the notion of a minimal sufficient statistics defined in [4]. In this paper we argue that this notion is not well defined for some x (although for some x the notion is well defined). Moreover, and even for those x for which the notion of a minimal sufficient statistic is well defined not every minimal sufficient statistic qualifies for "denoised version of x". We propose another definition of a (minimal) sufficient statistic that has better properties.

2 Kolmogorov complexity

We denote by $\{0,1\}^*$ the set of all strings over the binary alphabet $\{0,1\}$, and by $l(x)$ the length of a string x. See the textbook [6] for the definitions of Kolmogorov complexity $C(x)$ of a binary string x and conditional Kolmogorov complexity $C(x|y)$ of a binary string x given another string y, and their properties.

For this paper, the following understanding suffice: $C(x|y)$ is the minimal length of a program that maps y to x:

$$C(x|y) = \min\{l(p) \mid p \in \{0,1\}^*,\ D(p,y) = x\}$$

where $D(p,y)$ denotes the result of the program p for input y (thus D is the interpreter of the programming language). Programs are assumed to be written as binary strings. The programming language (also called a description mode) is chosen in such a way that for any other programming language D' there is the constant c such that for the resulting complexity $C'(x|y)$ it holds $C(x|y) \leqslant C'(x|y) + c$ for all x, y.

By definition $C(x) = C(x|\text{empty string})$. Kolmogorov complexity of a finite set of strings is defined as follows. We fix any computable bijection $A \mapsto [A]$ between finite sets of binary strings and binary strings and let $C(A) = C([A])$. It is not hard to see that if we switch to another computable bijection $A \mapsto [A]$ the value of $C(A)$ changes at most by an additive constant. The expressions $C(x|A), C(A|x)$ are understood as $C(x|[A]), C([A]|x)$, respectively.

[1] A string x is called *random* if $C(x)$ is close to the length of x.

Throughout the paper we will use the notation $\log i$ for $\lceil \log_2 i \rceil$.

We will use in the sequel without reference the following properties of Kolmogorov complexity:

- The number of strings of Kolmogorov complexity less than k is less than 2^k.
- $C(x) \leqslant l(x) + c$, $C(x|y) \leqslant C(x) + c$, for some c and all x, y;
- For every computable function f mapping strings to strings there is c such that $C(f(x)|x) \leqslant c$ and $C(f(x)) \leqslant C(x) + c$ for all x;
- (Conditional version of the previous inequality.) For every computable function f mapping pairs of strings to strings there is c such that $C(f(x, y)|y) \leqslant C(x|y) + c$ for all x, y;
- (Symmetry of information). Fix a computable bijection between pairs of binary strings and binary strings and let $\langle x, y \rangle$ denote the string corresponding to the pair x, y. Then $C(\langle x, y \rangle) \approx C(x) + C(y|x)$. Unfortunately, this equality holds only up to a "logarithmic error term". Specifically, we have

  $$C(\langle x, y \rangle) \leqslant C(x) + C(y|x) + 2 \log \min\{C(x), C(y|x)\} + c$$

  for some c and all x, y, and conversely

  $$C(x) + C(y|x) \leqslant C(\langle x, y \rangle) + 4 \log(C(x) + C(y|x)) + c.$$

- (Conditional version of symmetry of information). For all x, y, z,

  $$C(\langle x, y \rangle | z) = C(x|z) + C(y|\langle x, z \rangle).$$

  Here one inequality is true up to a $2 \log \min\{C(x|z), C(y|\langle x, z \rangle)\} + c$ error term and the other one up to a $4 \log(C(x|z) + C(y|\langle x, z \rangle)) + c$ error term.

Logarithmic error terms. Many useful inequalities in Kolmogorov complexity theory are hardly readable because of ubiquitous logarithmic terms. To make complexity inequalities more transparent we have used in the last property and will use in the sequel expressions like "the inequality $C(x|z) \leqslant C(x|y) + C(y|z)$ holds up to a $O(\log C(x|y))$ error term", which actually means that $C(x|z) \leqslant C(x|y) + C(y|z) + c \log C(x|y) + c$ for some c and all x, y, z.

3 Algorithmic sufficient statistics

Let x be a given string of length n. The goal of algorithmic statistics is to "explain" x. As possible explanations we consider finite sets containing x. We call any finite $A \ni x$ a *model for* x or a *statistic for* x. Every model A corresponds the statistical hypothesis "x was obtained by selecting a random element of A". In which case is such hypothesis plausible? As argued in [5, 4, 7], it is plausible if $C(x|A) \approx \log |A|$ and $C(A|x) \approx 0$ (we prefer to avoid rigorous definitions up to a certain point; approximate equalities should be thought as equalities up to an additive $O(\log n)$ error term).

As shown in [4, 7], $C(x|A) \approx \log |A|$ and $C(A|x) \approx 0$ if and only if $C(A) + \log |A| \approx C(x)$. Indeed, assume that A contains x and both $C(A), \log |A|$ are at most n (we do not need more complex or larger models to explain x). Then, given A the string x can be specified by its $\log |A|$-bit index in A. Omitting terms of order $O(\log(C(A) + \log |A|)) = O(\log n)$, we obtain

$$C(x) \leqslant C(x) + C(A|x) = C(A) + C(x|A) \leqslant C(A) + \log |A|.$$

The equality here follows from the symmetry of information [6]. Assume now that $C(x|A) \approx \log |A|$ and $C(A|x) \approx 0$. Then all inequalities here become equalities and hence A is a sufficient statistic. Conversely, if $C(x) \approx C(A) + \log |A|$ then the left hand side and the right hand side of the displayed inequality coincide. Thus $C(x|A) \approx \log |A|$ and $C(A|x) \approx 0$.

The inequality

$$C(x) \leqslant C(A) + \log |A| \tag{2}$$

(which is true up to a $O(\log \min\{C(A), \log |A|\})$ term) has the following meaning. Consider the two part code (A^*, i) of x, consisting of the minimal program A^* for x and $\log |A|$-bit index of x in the list of all elements of A arranged lexicographically. The equality means that its total length $C(A) + \log |A|$ cannot exceed $C(x)$. If $C(A) + \log |A|$ is close to $C(x)$, we call A a sufficient statistic for x. To make this notion rigorous we have specify what means "close". In [4] this is specified as follows: fix a natural constant ε and call A a sufficient statistic for x if

$$(C(A) + \log |A|) - C(x) \leqslant \varepsilon. \tag{3}$$

More precisely, [4] uses prefix complexity K in place of plain complexity C and require that the absolute value of the left hand side be at most ε. (For prefix complexity the inequality (2) holds up to a constant error term.) If we

choose ε large enough then sufficient statistics exists, witnessed by $A = \{x\}$. (The paper [1] suggests to set $\varepsilon = 0$ and to use $C(x|n)$ and $C(A|n)$ in place of $K(x)$ and $K(A)$ in the definition of a sufficient statistic. Such sufficient statistics might not exist.)

To avoid the discussion on how small should be ε let us call $A \ni x$ an *ε-sufficient statistic for x* if (3) holds. The smaller ε is the more sufficient A is. In this paper we will be interested in values of ε of order $\Omega(\log n)$.

4 Algorithmic minimal sufficient statistics

Naturally, we are interested in squeezing as much noise from the given string x as possible. What does it mean? Every sufficient statistic A identifies $\log |A|$ bits of noise in x. Thus a sufficient statistic with maximal $\log |A|$ (and hence minimal $C(A)$) identifies the maximal possible amount of noise in x. So we arrive at the notion of a minimal sufficient statistic: a sufficient statistic with minimal $C(A)$ is called a minimal sufficient statistic (MSS).

Is this notion well defined? Recall that actually we have only the notion of a ε-sufficient statistic (where ε is either a parameter, or a constant). That is, we have actually defined the notion of a minimal ε-sufficient statistic. Is this a sound notion? We argue that for some strings x it is not (for every ε). There are strings x for which it is impossible to identify MSS in an intuitively appealing way, since the complexity of the minimal ε-sufficient statistic decreases much, as ε increases a little.

Theorem 1. *Let $k \mapsto \alpha$ be a computable mapping from the naturals to the naturals such that $\alpha \leqslant k$. For every natural k there is a string x of length $2k$ and complexity $k + O(\log k)$ with*

$$g_x(j) = \begin{cases} k - \frac{jk}{k+\alpha} + O(\log k), & \text{if } j \leqslant k + \alpha, \\ O(\log k) & \text{if } k + \alpha \leqslant j. \end{cases}$$

The structure function of x is shown in Fig. 1.

Choose the mapping $k \mapsto \alpha$ so that both k/α and α are large (for example, let $\alpha = \sqrt{k}$). For very small j the graph of g_x is close to the sufficiency line and for $j = k + \alpha$ it is already at a large distance α from it. As j increments by one, the value $g_x(j) + j - C(x)$ increases by at most $k/(k+\alpha) + O(\log n)$, which is negligible. Therefore, it is not clear where the graph of g_x leaves

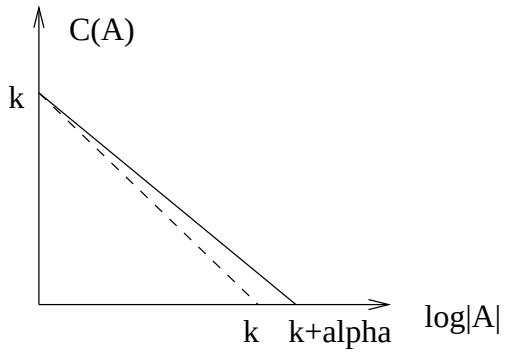

Figure 1: The structure function of a string for which MSS is not well defined

the sufficiency line. The complexity of the minimal ε-sufficient statistic is $k-(\varepsilon+O(\log n))(k/\alpha)$ and decreases fast as a function of ε provided $\varepsilon/\log n$ is large enough.

Theorem 1 is a direct corollary from a result of [9], which will be presented now. Let x be a binary string. Denote by S_x the *structure set* of x:

$$S_x = \{(i,j) \mid \exists A \ni x,\ C(A) \leqslant i,\ \log|A| \leqslant j\}.$$

This set can be identified by either of two its "border line" functions $h_x, g_x : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$:

$$h_x(i) = \min\{\log|A| \mid A \ni x,\ C(A) \leqslant i\}, \quad g_x(j) = \min\{C(A) \mid A \ni x,\ \log|A| \leqslant j\}.$$

The functions h_x, g_x are called the *Kolmogorov structure functions* of x. The structure function h_x may take the infinite value for small i due to lack of models of small complexity. In contrast, the function g_x takes only finite values for all x.

As pointed by Kolmogorov [5], the structure set S_x of every string x of length n and Kolmogorov complexity k has the following three properties (we state the properties in terms of the function g_x):

1. $g_x(0) = k + O(1)$ (there is the unique model, $A = \{x\}$, for x of log-cardinality 0 and $C(\{x\}) = C(x) + O(1)$).

2. $g_x(n) = O(\log n)$ (witnessed by $A = \{0,1\}^n$).

3. g_x in non-increasing and the "derivative" of g_x is bounded by 1 in absolute value: $g_x(j+l) \geqslant g_x(j) - l - O(\log \min\{l, g_x(j)\})$ for every $j, l \in \mathbb{N}$.

For the proof of the last property see [7, 9]. Properties (1) and (3) imply that $i + j \geqslant k - O(\log j)$ for every $(i, j) \in S_x$. Sufficient statistics correspond to those $(i, j) \in S_x$ with $i + j \approx k$. The line $i + j = k$ is therefore called *the sufficiency line*.

A result of [9, Remark IV.4] states that for every function g that satisfies (1)–(3) there is a string x of length n and complexity close to k such that g_x is close to g. [2] More specifically, the following holds:

Theorem 2 ([9]). *Let g be any non-increasing function $g : \{0, \ldots, n\} \to \mathbb{N}$ such that $g(0) = k$, $g(n) = 0$ and such that $g(j + l) \geqslant g_x(j) - l$ for every $j, l \in \mathbb{N}$ with $j + l \leqslant n$. Then there is a string x of length n and complexity $k \pm \varepsilon$ such that $|g_x(j) - g(j)| \leqslant \varepsilon$ for all $j \leqslant n$. Here $\varepsilon = O(\log n + C(g))$, where $C(g)$ stands for the Kolmogorov complexity of the graph of g:*

$$C(g) = C(\{\langle j, g(j)\rangle \mid 0 \leqslant j \leqslant n\}).$$

Proof of Theorem 1. Let $n = 2k$ and

$$g(j) = \begin{cases} k - \frac{jk}{k+\alpha}, & \text{if } j \leqslant k + \alpha, \\ 0 & \text{if } k + \alpha \leqslant j \leqslant n. \end{cases}$$

Then n, k, g satisfy all conditions of Theorem 2. Hence there is a string x of length n and complexity $k + O(\log n)$ with $g_x(j) = g(j) + O(\log n)$ (notice that $C(g) = O(\log n)$). □

Thus there are strings for which it is hard to identify the complexity of MSS. There is also another minor point regarding minimal sufficient statistics. Namely, there are strings x for which the complexity of minimal sufficient statistic is well defined but not all MSS qualify as denoised versions of x. Namely, some of them have a weird structure function. What kind of structure set we expect of a denoised string? To answer this question consider the following model example. Let y be a string, m a natural number and z a string of length $l(z) = m$ that is random relative to y. The latter

[2] Actually, [9] provides the description of possible shapes of S_x in terms of the Kolmogorov structure function h_x. We use here g_x instead of h_x, as in terms of g_x the description is easier-to-understand.

means that $C(z|y) \geqslant m - \beta$ for a small β. Consider the string $x = \langle y, z\rangle$. Intuitively, z is a noise in x and thus y is obtained from x by removing m bits of noise. What is the relation between the structure set of x and that of y?

Theorem 3. *Assume that y is an arbitrary string and z is a string of length m with $C(z|y) \geqslant m - \beta$ and let $x = \langle y, z\rangle$. Then*

$$g_x(j) = \begin{cases} C(x) - j, & \text{if } j \leqslant m, \\ g_y(j-m), & \text{if } j \geqslant m. \end{cases} \tag{4}$$

The equalities here hold up to $O(\log C(y) + \log m + \log j + \beta)$ term.

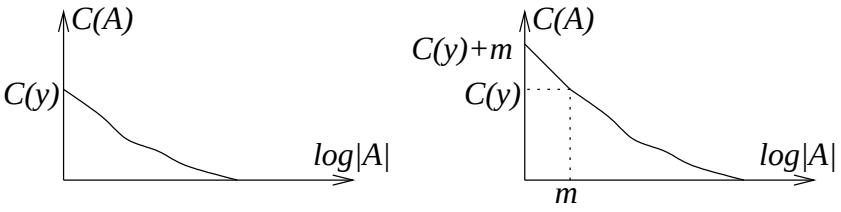

Figure 2: Structure functions of y and x

Proof. In the proof we will ignore terms of order $O(\log C(y) + \log m + \log j + \beta)$.

The equality $g_x(j) = C(x) - j$ (for $j \leqslant m$) is easy. Indeed, we have $g_x(m) \leqslant C(y) = C(x) - C(z|y) = C(x) - m$ witnessed by $A = \{\langle y, z'\rangle \mid l(z') = m\}$. On the other hand, $g_x(0) = C(x)$ (by Property 1 of g_x). Thus $g_x(j)$ should have maximal possible rate of decrease on the segment $[0, m]$ to drop from $C(x)$ to $C(x) - m$.

The inequality $g_x(j) \leqslant g_y(j-m)$ is easy as well. Indeed, let a model A for y witness $g_y(j-m)$ so that $|A| \leqslant 2^{j-m}$ and $C(A) = g_y(j-m)$. Consider the model

$$A' = A \times \{0,1\}^m = \{\langle y', z'\rangle \mid y' \in A,\ |z'| = m\}$$

of cardinality at most 2^j. Its complexity is at most that of $|A|$ (plus $O(\log m)$), which proves $g_x(j) \leqslant g_y(j-m)$.

The non-trivial part of the theorem is the inverse inequality $g_x(j) \geqslant g_y(j-m)$. Let A be a model for x with $|A| \leqslant 2^j$ and $C(A) = g_x(j)$ where

$j \geqslant m$. We need to show that there is a model of y of cardinality at most 2^{j-m} and of the same (or lower) complexity. We will prove it in a non-constructive way using a result from [9].

Consider the projection of A: $\{y' \mid \langle y', z' \rangle \in A\}$. Unfortunately, this set may be as large as A itself. Reduce it as follows. Consider the yth section of A: $A_y = \{z' \mid \langle y, z' \rangle \in A\}$. Define i as the natural number such that $2^i \leqslant |A_y| < 2^{i+1}$. Let A' be the set of those y' whose y'th section has at least 2^i elements. Then by counting arguments we have $|A'| \leqslant 2^{j-i}$.

Assume first that $i \geqslant m$. Then $|A'| \leqslant 2^{j-m}$ and $C(A') \leqslant C(A) + O(\log i)$. Note that $i \leqslant j$ and thus $g_y(j-m) \leqslant g_x(j) + O(\log j)$.

Assume now that $i < m$. To be sure that all the following inequalities hold up to a $O(\log C(y) + \log m + \log j + \beta)$ error term let us upper bound $C(x), C(A), C(A'), \log |A|, \log |A'|$. We have:

- $C(x) \leqslant C(y) + O(C(z|y)) \leqslant C(y) + O(m)$,
- $C(A) \leqslant C(x) + O(1) \leqslant C(y) + O(m)$,
- $C(A') \leqslant C(A) + O(\log i) \leqslant C(A) + O(\log j)$,
- $\log |A'| \leqslant \log |A| \leqslant j$,

Thus in the sequel we may neglect terms of order $\log(C(x) + C(A) + C(A') + \log |A| + \log |A'|)$.

First we will improve A' using a result of [9]:

Lemma 4 (Lemma A.4 in [9]). *For every $A' \ni y$ there is $A'' \ni y$ with $C(A'') \leqslant C(A') - C(A'|y) + O(\log(C(A') + \log |A'|))$ and $\log |A''| = \log |A'|$.*

By this lemma we get the inequality

$$g_y(j-i) \leqslant C(A'') \leqslant C(A') - C(A'|y).$$

We claim that

$$C(A') - C(A'|y) \leqslant C(A) - C(A|y). \tag{5}$$

Indeed, $C(A'|A)$ is negligible hence $C(y|A') \geqslant C(y|A)$. This implies that

$$C(y) - C(y|A') \leqslant C(y) - C(y|A).$$

By the symmetry of information this inequality is equivalent to (5). Thus we have

$$g_y(j-i) \leqslant C(A) - C(A|y).$$

By the Property 3 (page 7) of the structure set this inequality implies that

$$g_y(j-m) \leqslant C(A) - C(A|y) + (m-i).$$

We need to prove that the right hand side of this inequality is at most $C(A)$, that is, $i \geqslant m - C(A|y)$. To lower bound i, we will relate it to the conditional complexity of z given y and A. Indeed, we have $C(z|A,y) \leqslant i$, as z can be identified by its ordinal number in yth section of A. On the other hand,

$$C(z|A,y) \geqslant C(z|y) - C(A|y) \geqslant m - \beta - C(A|y). \quad \square$$

This theorem answers our question: if y is obtained from x by removing m bits of noise then we expect that g_y, g_x satisfy Theorem 3.

Let us define the structure set of a finite set A of strings as that of $[A]$. It is not hard to see that if we switch to another computable bijection $A \mapsto [A]$ the value of $g_{[A]}(j)$ changes at most by an additive constant. Thus S_A and g_A are well defined for finite sets A.

Is it true that for every x the structure function g_A of every ε-sufficient statistic A for x satisfies (4) (where we let $y = A$ and $m = \log|A|$) with an error term $O(\varepsilon + \log n)$? The answer is twofold: the first equality in (4) is true (with precision $O(\varepsilon + \log n)$) for all x and all ε-sufficient statistics A for x and the second equality in (4) is false for some x, A.

The first equality in (4) (for $y = A$ and $m = \log|A|$) is proved as follows: $g_x(0) = C(x) + O(1)$ (by Property 1 of g_x) and $g_x(\log|A|) \leqslant C(A) = C(x) - \log|A| + O(\varepsilon)$. Thus g_x must decrease with maximal speed to fall down from $C(x)$ to about $C(x) - \log|A|$ on the segment $[0, \log|A|]$ (Property 3). Thus $g_x(j) = C(x) - j + O(\varepsilon + \log n)$ for all $j \in [0, \log|A|]$.

Now we will show that there is a string x as in Theorem 3 and a $O(\log n)$-sufficient statistic B for x that does not satisfy the equality $g_B(j) \approx g_x(j + \log|B|)$.

Theorem 5. *For every k there is a string y of length $2k$ and Kolmogorov complexity $C(y) = k$ such that*

$$g_y(j) = \begin{cases} k & \text{if } j \leqslant k, \\ 2k - j & \text{if } k \leqslant j \leqslant 2k \end{cases}$$

and hence for any z of length k and conditional complexity $C(z|y) = k$ the structure function of the sting $x = yz$ (concatenation of y and z) satisfies

$$g_x(j) = \begin{cases} 2k - j & \text{if } j \leqslant k, \\ k & \text{if } k \leqslant j \leqslant 2k, \\ 3k - j & \text{if } 2k \leqslant j \leqslant 3k. \end{cases}$$

(See Fig. 3.) Moreover, for every such z the string $x = yz$ has a model $B \subset \{0,1\}^{3k}$ of complexity $C(B) = k$ and log-cardinality $\log|B| = k$ such that $g_B(j) = k$ for all $j \in [k, 2k)$. All equalities here hold up to $O(\log k)$ additive error term.

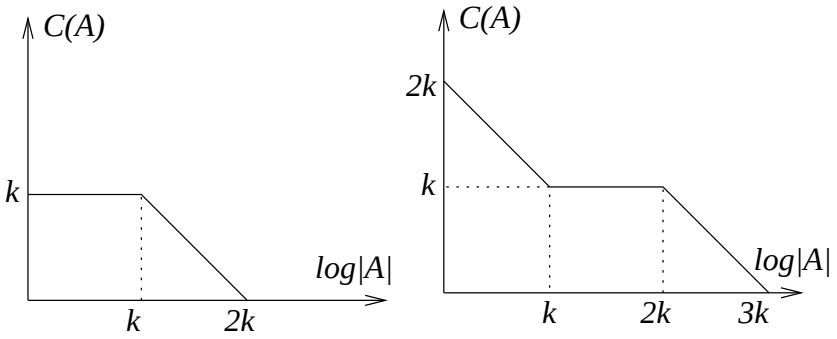

Figure 3: Structure functions of y and x

The structure function of $x = yz$ clearly leaves the sufficiency line at the point $j = k$. Thus k is intuitively the complexity of minimal sufficient statistic. The function $g_B(j)$ is far from $g_x(j + k)$ on the segment $[k, 2k)$.

Proof. We first construct y. Call a string y of length $2k$ *simple* if it has a model T with $|T| \leqslant 2^k$, $C(T) < k$. Otherwise call y *complex*. The total number of simple strings is strictly less than 2^{2k}. Indeed, there are less than 2^k sets of complexity less than k. Therefore the total number of elements in all T with $|T| \leqslant 2^k$, $C(T) < k$ is less than $2^k \times 2^k$. Let y be the lexicographical first complex string of length $2k$. Let

$$B = \{yz' \mid z' \in \{0,1\}^k\} \cup \{x' \in \{0,1\}^{3k} \mid C(x') < k\}.$$

We claim that $C(y) \leqslant k + O(1)$. Indeed, y can be found given k and the set of all simple strings. The set of simple strings can be found given the set of all halting programs of length less than k, which in turn can be identified by the k-bit number N^k of halting programs of length less than k (we run all programs of length less than k until N^k of them halt). The same argument shows that $C(B) \leqslant k + O(1)$.

It remains to find g_y and g_B. By construction, $g_y(k) \geqslant k$ and $g_y(0) = C(y) \leqslant k$. Thus $g_y(j) \equiv k$ on the segment $[0, k]$. As $g_y(2k) = 0$ (witnessed by

$\{0,1\}^{2k}$), $g_y(j)$ should have maximal possible rate of decrease on the segment $[0,k]$ to drop from k to 0.

It remains to show that $g_B(j) \equiv k$ on the segment $[k,2k)$. The inequality $g_B(j) \leqslant k$ is witnessed by $\{[B]\}$. On the other hand, assume that M is a family of at most 2^{2k-1} finite sets and $[B] \in M$. We have to show that $C(M) \geqslant k - O(\log k)$. To this end let $M' = \{[B'] \in M \mid |B'| < 2^{k+1}\}$. As $|B| < 2^{k+1}$, the family M' is a model of $[B]$ as well. Its complexity is at most $C(M) + O(\log k)$. Given k and M' we can find a string u length $3k$ and complexity at least k: pick the lexicographical first string outside the union of all sets B' with $[B'] \in M'$. There is such a string, as the total number of elements in all such B' is less than $2^{2k-1}2^{k+1} = 2^{3k}$. As that union contains all strings of length $3k$ and complexity less than k we have $C(u) \geqslant k$. Therefore, $k \leqslant C(u) \leqslant C(M) + O(\log k)$. Thus $g_B(2k-1) \geqslant k - O(\log k)$ hence $g_B(j) \geqslant k - O(\log k)$ for all $j < 2k$. □

5 Desired properties of sufficient statistics and a new definition

Recall the probabilistic notion of sufficient statistic [3]. In the probabilistic setting, we are given a parameter set Θ and for each $\theta \in \Theta$ we are given a probability distribution over a set X (for simplicity we assume that both X, Θ are finite or countable). For every probability distribution over Θ we thus obtain a probability distribution over $\Theta \times X$. A function $f : X \to Y$ (where Y is any set) is called a *sufficient statistic*, if for every probability distribution over Θ, the random variables x and θ are independent conditional to $f(x)$. That is, for all $a \in X$, $c \in \Theta$,

$$\mathrm{Prob}[\theta = c | x = a] = \mathrm{Prob}[\theta = c | f(x) = f(a)].$$

Saying differently, $x \to f(x) \to \theta$ is a Markov chain (for every probability distribution over Θ).

In the definition of a sufficient statistics it only matters in which equivalence classes it partitions X (x' is equivalent to x'' if $f(x') = f(x'')$). We say that a sufficient statistic f is less than a sufficient statistic g if the partition defined by g is a sub-partition of that defined by f. It turns out that there is always a sufficient statistic f that is less than any other sufficient statistic. Such sufficient statistics are called minimal. Any two minimal sufficient

statistics partition the set X into the same equivalence classes thus basically there is only one minimal sufficient statistic. Is it possible to define a notion of an algorithmic sufficient statistic that has similar properties? More specifically, we wish it have the following properties.

(1) If A is an (algorithmic) sufficient statistic for x and $\log|A| = m$ then the structure functions of $y = A$ and x satisfy Theorem 3. In particular, structure functions of every MSS A, B for x coincide.

(2) Assume that A is a MSS and B is a sufficient statistic for x. Then $C(A|B) \approx 0$.

As the example of Theorem 5 demonstrates, the property (1) does not hold for the usual definition of sufficient statistics. Indeed, consider the model $A \rightleftharpoons \{y\} \times \{0,1\}^k$ for the string x form that theorem. Both A and B are minimal sufficient statistics. The model A, as finite object, is identical to y and hence the structure function of A coincides with that of y. However g_B is quite different from g_A.

It is unknown whether the property (2) holds or not. We propose here a new definition of an algorithmic sufficient statistic that satisfies both (1) and (2). The main idea of the new definition is the following. As observed in [8], to have the same structure sets strings x, y should be equivalent in the following strong sense: there should be short *total* programs p, q with $D(p, x) = y$ and $D(q, y) = x$ (where D is the programming language in the definition of conditional Kolmogorov complexity). A program p is called *total* if $D(p, z)$ converges for *all* z.

Let $CT_D(x|y)$ stand for the minimal length of p such that p is total and $D(p, y) = x$. For the sequel we need that the programming language D have the following property. For any other programming language D' there is a constant c such that $CT_D(x|y) \leqslant CT_{D'}(x|y) + c$ for all x, y. (The existence of such a D is straightforward.) Fixing such D we get the defintion of the *total* Kolmogorov complexity $CT(x|y)$. If both $CT(x|y), CT(y|x)$ are small then we will say that x, y are *strongly equivalent*.

Lemma 6. *For all x, y and all natural j we have*

$$|g_x(j) - g_y(j)| \leqslant 2\max\{CT(x|y), CT(y|x)\} + O(1).$$

(If x, y are strongly equivalent then their structure functions are close to each other.)

Proof. We will prove the inequality $g_x(j) \leqslant g_y(j) + 2CT(x|y) + O(1)$. The other inequality is proved in a similar way. Let p witness $CT(x|y)$ and let A

witness $g_y(j)$. The set $B = \{D(p, y') \mid y' \in A\}$ contains x and has at most $|A| \leqslant 2^j$ elements. Its complexity is at most $C(A) + 2l(p) + O(1)$. □

We call A a strongly sufficient statistic for x if $CT(A|x) \approx 0$ and $C(A) + \log|A| \approx C(x)$ (recall that the latter inequality implies only that $C(A|x) \approx 0$). More specifically, call a model A for x an *α-strong* statistic for x if $CT(A|x) \leqslant \alpha$. *Strongly sufficient statistics* for x are those statistics that are α-strong and ε-sufficient for small α, ε. It turns out that strongly sufficient statistics satisfy properties (1) and (2). That is, (1) if A is a strongly sufficient statistic for x and $\log|A| = m$ then the structure functions of $y = A$ and x satisfy the equalities of Theorem 3 and (2) $C(A|B) \approx 0$ for every strongly sufficient statistic B and every minimal sufficient statistic A for x (A may be not strong). The exact statements are given in the following Theorems 7 and 9.

Theorem 7 (on the structure function of a strong sufficient statistic). *Assume that A is an ε-strong ε-sufficient statistic for x. Then for all natural j we have $g_A(j) = g_x(j + \log|A|) + O(\varepsilon + \log(C(A) + \log|A| + j))$*

Proof. Let z stand for the index of x with respect to the lexicographical order on A. We claim that both $CT(\langle A, z\rangle|x)$ and $CT(x|\langle A, z\rangle)$ are of order $O(\varepsilon)$. Indeed, there is a total program of constant length that maps $\langle A, z\rangle$ to x. On the other hand, given x we can find A by applying a total ε-bit program and then find z.

By Lemma 6 $g_{\langle A,z\rangle}(j) = g_x(j) + O(\varepsilon)$. As A is ε-sufficient for x, conditions of Theorem 3 are fulfilled for $y = A$, $m = \log|A|$ and $\beta = O(\varepsilon + \log C(A) + \log m)$. Thus $g_A(j) = g_{\langle A,z\rangle}(j + m) = g_x(j + m)$. The error term here is $O(\varepsilon + \log C(A) + \log m + \log j)$. □

To state the second property of strongly sufficient statistics we need to clarify the notion of a minimal sufficient statistic. Technically, it is convenient to separate minimality from sufficiency. Namely, let γ, δ be natural numbers. Call $A \ni x$ a *δ, γ-minimal statistic for x* if

$$g_x(\log|A| + \gamma + \delta) > C(A) - \delta. \tag{6}$$

The meaning of this definition is as follows: it states that it is impossible to decrease the complexity of A by δ bits at the expense of increasing the log-cardinality of A by the same amount plus a little bit more (γ). The notion of a minimal statistic has the following useful property that states that it is

impossible to decrease the complexity of a minimal γ, δ-statistic by δ at the expense of increasing the sum $C(A) + \log |A|$ by at most $\gamma - O(\log C(A))$. This property is not straightforward: indeed, we might decrease $C(A)$ by much more than δ and simultaneously increase $\log |A|$ by the same amount. However, in this case we can exchange the decrease of $C(A)$ for the decrease of $\log |A|$ (Property 3 of the structure set).

Lemma 8. *For some constant c and all γ, δ if A is a δ, γ-minimal statistic for x then x has no model T of complexity at most $C(A) - \delta$ with*

$$C(T) + \log |T| \leqslant C(A) + \log |A| + (\gamma - c \log C(A)).$$

Proof. Assume that there is such T. Then $g_x(\log |T|) \leqslant C(A) - \delta$. The inequality (6) implies that $\log |A| + \gamma + \delta < \log |T|$ (recall that g_x is non-increasing). Therefore we can use Property 3 and conclude that for some constant c' we have

$$g_x(\log |A| + \gamma + \delta) \leqslant g_x(\log |T|) + \log |T| - (\log |A| + \gamma + \delta) + c' \log C(A).$$

Let $c = c'$. Then the right hand side here is at most

$$\begin{aligned} &C(T) + \log |T| - (\log |A| + \gamma + \delta) + c' \log C(A) \\ &\leqslant C(A) + \log |A| + \gamma - c \log C(A) - (\log |A| + \gamma + \delta) + c' \log C(A) = C(A) - \delta, \end{aligned}$$

a contradiction. □

We are able to present the theorem stating that the strong statistics satisfy the second desired property.

Theorem 9 (Chelnokov [2]). *There is a constant c such that for all natural n and $\varepsilon, \delta \leqslant n$ and all strings x of length of n the following is true. If A is a $\delta, c(\varepsilon + \log n)$-minimal ε-sufficient statistic for x and B is an ε-strong ε-sufficient statistic for x with $|B| \leqslant |A|$, then $C(A|B) \leqslant c(\varepsilon + \delta + \log n)$.*

In the proof of this theorem we will need a result of [9] stating that all δ, γ-minimal models of the same complexity are algorithmically equivalent (even if they are models for non-related strings), Theorem 10 below and its straightforward Corollary 11. Let $d(u, v)$ stand for $\max\{C(u|v), C(v|u)\}$ (a sort of algorithmic distance between u and v).

Theorem 10 (Theorem V.4(iii) from [9]). *Let N^i stand for the number of strings of complexity at most i.* [3] *For all $A \ni x$ and i, either $d(N^i, A) \leqslant C(A) - i$, or there is $T \ni x$ such that $\log|T| + C(T) \leqslant \log|A| + C(A)$ and $C(T) \leqslant i - d(N^i, A)$, where all inequalities hold up to $O(\log(C(A) + |A|))$ additive term.*

Corollary 11. *There is a constant c such that for every string x and for every i and δ the following holds. If A is a $\delta, c\log(C(A) + \log|A|)$-minimal model for x then $d(N^i, A) \leqslant |C(A) - i| + \delta + c\log(C(A) + \log|A|)$.*

Proof. Let c' be the constant in the error term in Theorem 10 and c'' be the constant from Lemma 8. Set $c = c' + c''$ and let A, i be as in the corollary. Let $\gamma = c\log(C(A)+\log|A|)$ so that A is a δ, γ-minimal model for x. By way of contradiction assume that $d(N^i, A) > |C(A) - i| + \delta + \gamma$. Then certainly $d(N^i, A) > C(A) - i$ and thus the second option in Theorem 10 holds. That is, there is $T \ni x$ such that

$$\log|T| + C(T) \leqslant \log|A| + C(A) + \gamma' \leqslant \log|A| + C(A) + \gamma - c''\log C(A)$$

and $C(T) \leqslant i - d(N^i, A) + \gamma'$ where $\gamma' = c'\log(C(A) + \log|A|)$ is the error term in Theorem 10. By our assumption the right hand side of the last inequality is less than $i - (|C(A) - i| + \delta + \gamma) + \gamma' \leqslant C(A) - \delta$. By Lemma 8 this implies that A is not δ, γ-minimal. □

This corollary reveals an interesting phenomenon: let x, y be arbitrary strings and let A be a $\delta, c\log(C(A) + \log|A|)$-minimal model for x and B a $\delta, c\log(C(B) + \log|B|)$-minimal model for y. If the complexity A is close to that of B, then A and B are algorithmically equivalent, as they both are equivalent to $N^{\log|A|}$. More specifically $d(A, B) = O(|C(A) - C(B)| + \delta + \log(C(A) + \log|A|) + \log(C(B) + \log|B|))$.

Proof of Theorem 9. Let $n, \varepsilon, \delta, x, A, B$ be as in the theorem (the constant c is to be defined later). On the top level, the proof goes as follows. As B is a sufficient statistic for x, Theorem 7 implies that $g_B(j) \approx g_x(j + \log|B|)$. Therefore the complexity of every minimal sufficient statistic for B is approximately the same as that for A (recall that A is a minimal sufficient statistic for x). Let D be a minimal sufficient statistic for B. As both D and

[3] Actually, the authors of [9] use prefix complexity in place of the plain complexity. It is easy to verify that Theorem V.4(iii) holds for plain complexity as well.

A are minimal sufficient statistics of the same complexity, by Corollary 11 they are algorithmically equivalent. Therefore $C(A|D) \approx 0$. As D is a sufficient statistic for B, we additionally have $C(D|B) \approx 0$. This implies $C(A|B) \leqslant C(A|D) + C(D|B) \approx 0$.

Now we repeat these arguments formally and estimate the resulting error terms.

Proving that $g_B(j) \equiv g_x(j + \log|B|)$: This equality follows from Theorem 7 for B and x. How precise is this equality? In the error term of Theorem 7 $\log C(B)$ and $\log m$ are of order $O(\log n)$ as $C(B) + \log|B| = O(n)$, and the latter holds as B is ε-sufficient and $\varepsilon \leqslant n$. Thus the equality $g_B(j) = g_x(j + \log|B|)$ hold up to an error term $O(\varepsilon + \log j + \log n)$. It implies that the log-cardinality of minimal sufficient statistics for B should be $\log|B|$ less than of those of x, that is, about $\log|A| - \log|B|$.

Constructing D: Let D be a model of B of minimal complexity among models of log-cardinality at most $\log|A| - \log|B|$, so that

$$C(D) = g_B(\log|A| - \log|B|), \quad \log|D| \leqslant \log|A| - \log|B|.$$

Proving that $C(D) \approx C(A)$: The equality $g_B(j) = g_x(j + \log|B|)$ holds for $j = \log|A| - \log|B|$ and therefore

$$C(D) = g_x(\log|A|),$$

up to a $O(\varepsilon + \log n)$ error term (as $j \leqslant \log|A| \leqslant C(x) + \varepsilon = O(n)$). We claim that $g_x(\log|A|) = C(A)$. Indeed, we have

$$C(x) \leqslant g_x(\log|A|) + \log|A| \leqslant C(A) + \log|A| \leqslant C(x).$$

Here the first inequality holds (up to an error term $\log C(A)$), as the graph of g_x lies above sufficiency line and the last inequality is true (up to an error term ε), since A is ε-sufficient. Thus all inequalities here are equalities up to an error term $O(\varepsilon + \log n)$.

Proving that D is a $O(\varepsilon + \log n)$-sufficient statistic for B: Indeed, ignoring terms of order $O(\varepsilon + \log n)$, we have

$$C(D) + \log|D| \leqslant C(A) + \log|A| - \log|B| = C(x) - \log|B| = C(B).$$

Here, the first equality holds, since A is an ε-sufficient statistic for x and the second equality holds, since B is an ε-sufficient statistic for x. As D is an $O(\varepsilon + \log n)$-sufficient statistic for B, $C(D|B)$ is negligible. More specifically,

$C(D|B)$ is of order $O(\varepsilon + \log n + \log(C(D) + \log|D|)) = O(\varepsilon + \log n)$ (the last equality holds, as $C(D) + \log|D| = C(B) = O(n)$). It remains to show that $C(A|D) \approx 0$, which is proved in two steps.

Proving that D is a minimal statistic for B: Recall that A is a δ, γ-minimal model for x where $\gamma = c(\varepsilon + \log n)$. As $g_B(j) = g_x(j + \log|B|)$, we have

$$g_B(\log|D| + \gamma + \delta) \geqslant g_B(\log|A| - \log|B| + \gamma + \delta) = g_x(\log|A| + \gamma + \delta) > C(A) - \delta = C(D) - \delta.$$

The first equality is true by Theorem 7 (with an error term $O(\varepsilon + \log n)$), the second inequality is true as written by δ, γ-minimality of A. The second equality was proved above. Recalling the error term we get

$$g_B(\log|D| + \gamma + \delta) > C(D) - \delta - c''(\varepsilon + \log n)$$

for some constant c''. This means that D is a δ', γ'-minimal model for B where $\delta' = \delta + c''(\varepsilon + \log n)$ and $\gamma' = \gamma - c''(\varepsilon + \log n)$.

Proving that $C(A|D) \approx 0$: We need that γ' be at least $c'(\log(C(D) + \log|D|))$, where c' is the constant in Corollary 11. As $C(D) + \log|D| = O(n)$, to this end we can set $c = c' + c''$. Then by Corollary 11 we have $d(N^i, D) \leqslant \delta + O(\varepsilon + \log n)$ where $i = \log|A|$. Recall that $C(A) + \log|A| = O(n)$ and hence $\gamma \geqslant c' \log(C(A) + \log|A|)$. By Corollary 11 we have $d(N^i, A) = \delta + O(\log n)$. Thus $C(A|D) \leqslant d(N^i, D) + d(N^i, A) = O(\delta + \varepsilon + \log n)$. □

So strongly sufficient statistics have better properties than sufficient statistics. However, many questions are still open. The first one is whether $CT(A|B) \approx 0$ for every strong MSS A for x and every strong sufficient statistic B for x. Formally:

Question 1. *Is it true that there is a constant c such that for all natural n and $\varepsilon, \delta \leqslant n$ and all strings x of length of n the following is true. If A is an ε-strong $\delta, c(\varepsilon + \log n)$-minimal ε-sufficient statistic for x and B is an ε-strong ε-sufficient statistic for x with $|B| \leqslant |A|$, then $CT(A|B) \leqslant c(\varepsilon + \delta + \log n)$?*

An interesting related question is the following: is it true that there is always a strongly sufficient statistic that is a MSS? Formally:

Question 2. *Is it true that for every constant c there is a constant c' such that for all natural n and $\varepsilon, \delta \leqslant n$ and all strings x of length of n the following is true. If there is a $\delta/c', c'c(\varepsilon + \log n)$-minimal ε/c'-sufficient statistic A for x then there is ε-strong $\delta, c(\varepsilon + \log n)$-minimal ε-sufficient statistic B for x?*

Another interesting related question is the following.

Question 3. *Merging strongly sufficient statistics: Is it true that for some c and all ε for all $c\varepsilon$-strong $c\varepsilon$-sufficient statistics A, B for x there is a ε-strong ε-sufficient statistic D for x with $\log|D| \geqslant \log|A| + \log|B| - \log|A \cap B| - c(\varepsilon + \log n)$?*

The next theorem answers Question 1 in the case when *both* A, B are minimal. For C in place of CT this was known: all minimal sufficient statistics for x are algorithmically equivalent. This easily follows from Corollary 11 (see the proof of the next theorem).

Theorem 12. *There is a constant c such that for all natural n and $\varepsilon, \delta < n$ the following holds for every string x of length n. For every $\delta, \varepsilon + c\log n$-minimal ε-sufficient models A, B for x such that A is ε-strong we have $CT(A|B) \leqslant c(\varepsilon + \delta + \log n)$.*

Proof. Our plan is as follows. We first show that $C(A) \approx C(B)$ and, moreover, $d(A, B) \approx 0$. To this end we only need that A, B be minimal sufficient statistics for x. Then we show that A has the following feature: A has many elements $x' \in B$ such that A can be retrieved from any such x' using the program p witnessing $CT(A|x) \leqslant \varepsilon$. Finally, we show that there are few A' that have this feature and there is a short program that given B and p finds a list of all such A'. Given B, the set A can be identified by p and its index in that list.

Let us start the formal argument. First notice that $C(A) + \log|A|$ is of order $O(n)$, which implies that error terms of order $\log(C(A) + \log|A|)$ below may be estimated as $O(\log n)$. The same applies to B.

Proving that $C(A)$ and $C(B)$ are at most δ apart: Assume that c is large. By way of contradiction assume that $C(A) < C(B) - \delta$, say. As B is minimal, by Lemma 8 we have

$$C(A)+\log|A| > C(B)+\log|B|+\varepsilon+c\log n-O(\log n) \geqslant C(x)+\varepsilon+c\log n-O(\log n).$$

On the other hand, as A is ε-sufficient, the left hand side here is at most $C(x) + \varepsilon$, which is a contradiction if c is large enough.

Proving that $d(A, B)$ is negligible: If c is large enough, then A is $\delta, c'\log(C(A) + \log|A|)$-minimal model for x, where c' is the constant from Corollary 11. The same holds for B. Applying Corollary 11 for A and B and $i = \log|A|$ we conclude that both $d(N^i, A)$ and $d(N^i, B)$ are of order $O(\delta + \log n)$ hence $d(A, B) = O(\delta + \log n)$.

The feature of A using which we find A given B: Let p be a total program witnessing $CT(A|x) \leqslant \varepsilon$. Let us show that there are many $x' \in B$

with $x' \in D(p, x') = A$ (otherwise B would be not sufficient). We will then identify A given B in few bits by its ordinal number among all A' that have this property.

Let $D = \{x' \in B \mid x' \in D(p, x') = A\}$. Obviously, D includes x and

$$C(D|B) \leqslant C(A|B) + \varepsilon = O(\delta + \varepsilon)$$

(ignoring terms of order $O(\log n)$). Given B and p the string x can be identified by its index in D, therefore

$$C(x|B) \leqslant C(D|B) + \log |D| \leqslant \log |D| + O(\delta + \varepsilon).$$

On the other hand, $C(x|B) \geqslant \log |B| - \varepsilon$, as B is ε-sufficient. Hence $\log |D| \geqslant \log |B| - \varepsilon - O(\delta + \varepsilon)$. Recall that we ignored terms of order $O(\log n)$. Actually, we have shown that for some constant c we have $\log |D| \geqslant \log |B| - c(\varepsilon + \delta + \log n)$.

Proving that the number of A' that have this feature is small: Consider now all A' such that

$$\log |\{x' \in B \mid x' \in D(p, x') = A'\}| \geqslant \log |B| - c(\varepsilon + \delta + \log n).$$

These A' are pairwise disjoint (as A' can be retrieved from any its element by applying program p). Each of them has at least $|B|/2^{c(\varepsilon+\delta+\log n)}$ elements of B. Thus there are at most $2^{c(\varepsilon+\delta+\log n)}$ different such A's. Given B and p, ε, δ we are able to find the list of all A's. The program that maps B to the list of A's is obviously total. Therefore there is a total program of $O(\varepsilon + \delta + \log n)$ bits that maps B to A and $CT(A|B) = O(\varepsilon + \delta + \log n)$. □

Acknowledgement

The author is grateful to Georgij Chelnokov for permission to publish Theorem 9 and for reading the preliminary version of this paper.

References

[1] L. Antunes and L. Fortnow. Sophistication revisited. Theory of Computing Systems, 45(1):150-161, June 2009.

[2] G. Chelnokov, Personal communication, 2009.

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

[4] P. Gács, J. Tromp, P.M.B. Vitányi. Algorithmic statistics, *IEEE Trans. Inform. Th.*, 47:6(2001), 2443–2463.

[5] A.N. Kolmogorov, Talk at the Information Theory Symposium in Tallinn, Estonia, 1974.

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

[7] A.Kh. Shen, Discussion on Kolmogorov complexity and statistical analysis, *The Computer Journal*, 42:4(1999), 340–342.

[8] A.Kh. Shen, Personal communication, 2002.

[9] N.K. Vereshchagin and P.M.B. Vitányi, Kolmogorov's structure functions and model selection, *IEEE Trans. Information Theory*, 50:12 (2004) 3265-3290.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de