Algorithmic Minimal Sufficient Statistic Revisited

Nikolay Vereshchagin

Moscow State University, Leninskie gory 1,
Moscow 119991, Russia
ver@mccme.ru
http://lpcs.math.msu.su/~ver

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 might have better properties.

1 Introduction

Let x be a binary string. A finite set A containing x is called an (algorithmic) sufficient statistic of 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 of 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 information present in x and i contains only accidental information (noise).

Sufficient statistics may also contain noise. For example, this happens if x is a random string 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 [3]. In this paper we argue that (1) this notion is not well-defined for some x (although for some x the notion is well-defined) and (2) even for those x for which the notion of a minimal sufficient statistic is well-defined not every minimal sufficient statistic qualifies for a "denoised version of x". We propose another definition of a (minimal) sufficient statistic that might have better properties.

K. Ambos-Spies, B. Löwe, and W. Merkle (Eds.): CiE 2009, LNCS 5635, pp. 478–487, 2009.

2 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. Every model A corresponds to the statistical hypothesis "x was obtained by selecting a random element of A". In which case is such a hypothesis plausible? As argued in [3,4,5], 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 $O(\log n)$ term). In the expressions $C(x|A), C(A|x)$ the set A is understood as a finite object. More precisely, we fix any computable bijection $A \mapsto [A]$ between finite sets of binary strings and binary strings and let $C(x|A) = C(x|[A])$, $C(A|x) = C([A]|x)$, $C(A) = C([A])$.

As shown in [3,5] this is equivalent to saying that $C(A) + \log|A| \approx C(x)$. Indeed, assume that A contains x and $C(A) \leq n$. Then, given A, the string x can be specified by its $\log|A|$-bit index in A. Recalling the symmetry of information and omitting additive terms of order $O(\log n)$, we obtain

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

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 in these inequalities coincide. Thus $C(x|A) \approx \log|A|$ and $C(A|x) \approx 0$.

The inequality

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

(which is true up to an additive $O(\log n)$ term) has the following meaning. Consider the two part code (A^*, i) of x, consisting of the minimal program A^* for x and the $\log|A|$-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 $C(x)$, then we call A a *sufficient statistic* of x. To make this notion rigorous we have to specify what we mean by "closeness". In [3] this is specified as follows: fix a constant c and call A a sufficient statistic if

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

More precisely, [3] uses prefix complexity K in place of plain complexity C. For prefix complexity the inequality (2) holds up to a constant error term. If we choose c large enough then sufficient statistics exists, witnessed by $A = \{x\}$. (The paper [1] suggests to set $c = 0$ and to use $C(x|n)$ and $C(A|n)$ in place of $C(x)$ and $C(A)$ in the definition of a sufficient statistic. For such definition sufficient statistics might not exist.)

To avoid the discussion on how small c should be let us call $A \ni x$ a *c-sufficient statistic* if (3) holds. The smaller c is the more sufficient A is. This notion is non-vacuous only for $c = O(\log n)$ as the inequality (2) holds only with logarithmic precision.

3 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 only have the notion of a c-sufficient statistic (where c is either a parameter or a constant). That is, we have actually defined the notion of a minimal c-sufficient statistic. Is this a good notion? We argue that for some strings x it is not whatever the value of c is. There are strings x for which it is impossible to identify MSS in an intuitively appealing way. For those x the complexity of the minimal c-sufficient statistic decreases substantially, as c increases a little.

To present such strings we need to recall a theorem from [7]. Let S_x stand for the *structure set* of x:

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

This set can be identified by either of its two "border line" functions:

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

The function h_x is called the *Kolmogorov structure function* of x; for small i it might take infinite values due to lack of models of small complexity. In contrast, the function g_x is total for all x.

As pointed out by Kolmogorov [4], 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) $g_x(0) = k + O(1)$ (witnessed by $A = \{x\}$). (2) $g_x(n) = O(\log n)$ (witnessed by $A = \{0,1\}^n$). (3) g_x in non-increasing and $g_x(j+l) \ge g_x(j) - l - O(\log l)$ for every $j, l \in \mathbb{N}$.

For the proof of the last property see [5,7]. Properties (1) and (3) imply that $i + j \ge k - O(\log n)$ 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 [7, Remark IV.4] states that for every g that satisfies (1)–(3) there is x of length n and complexity close to k such that g_x is close to g.[1] More specifically, the following holds:

Theorem 1 ([7]). *Let g be any non-increasing function $g : \{0, \dots, n\} \to \mathbb{N}$ such that $g(0) = k$, $g(n) = 0$ and such that $g(j+l) \ge g_x(j) - l$ for every $j, l \in \mathbb{N}$ with $j + l \le n$. Then there is a string x of length n and complexity $k \pm \varepsilon$ such that $|g_x(j) - g(j)| \le \varepsilon$ for all $j \le n$. Here $\varepsilon = O(\log n + C(g))$ and $C(g)$ stands for the Kolmogorov complexity of the graph of g: $C(g) = C(\{\langle j, g(j)\rangle \mid 0 \le j \le n\}$.*

[1] Actually, [7] 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.

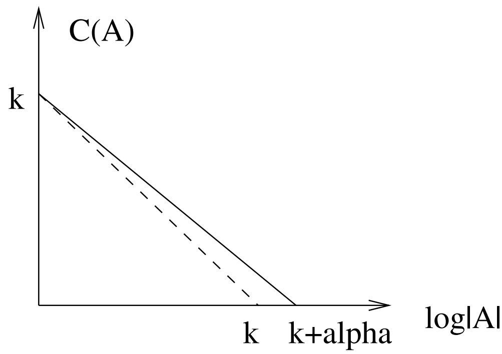

Fig. 1. The structure function of a string for which MSS is not well-defined

We are ready to present strings for which the notion of a MSS is not well-defined. Fix a large n and let $k = n/2$ and $g(j) = \max\{k - jk/(k+\alpha), 0\}$, where $\alpha = \alpha(k) \leq k$ is a computable function of k with natural values. Then n, k, g satisfy all conditions of Theorem 1. Hence there is a string x of length n and complexity $k+O(\log n)$ with $g_x(j) = g(j)+O(\log n)$ (note that $C(g) = O(\log n)$). Its structure function is shown on Fig. 1. Choose α so that α/k is negligible (compared to k) but α is not.

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 increases by one, the value $g_x(j) + j - C(x)$ increases by at most $\alpha/(k + \alpha) + O(\log n)$, which is negligible. Therefore, it is not clear where the graph of g_x leaves the sufficiency line. The complexity of the minimal c-sufficient statistic is $k - (c + O(\log n)) \cdot k/\alpha$ and decreases fast as a function of c.

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 is a string 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 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 means that $C(z|y) \geq m - \beta$ for a small β. Consider the string $x = \langle y, z\rangle$. Intuitively, z is a noise in x. In other words, we can say that 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 2. *Assume that z is a string of length m with $C(z|y) \geq m-\beta$. Then for all $j \geq m$ we have $g_x(j) = g_y(j-m)$ and for all $j \leq m$ we have $g_x(j) = C(y)+m-j = g_y(0) + m - j$. The equalities here hold up to $O(\log m + \log C(y) + \beta)$ term.*

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

The easy part is the equality $g_x(j) = C(y) + m - j$ for $j \leq m$. Indeed, we have $g_x(m) \leq C(y)$ witnessed by $A = \{\langle y, z'\rangle \mid l(z') = m\}$. On the other hand,

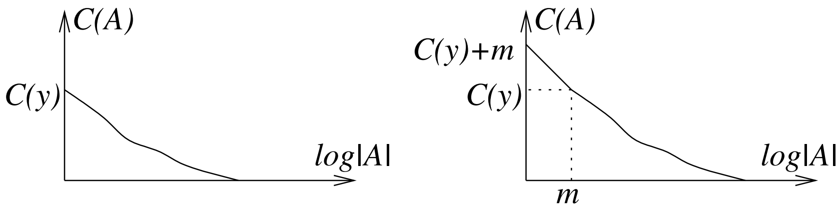

Fig. 2. Structure functions of y and x

$g_x(0) = C(x) = C(y) + C(z|y) = C(y) + m$. Thus $g_x(j)$ should have maximal possible rate of decrease on the segment $[0, m]$ to drop from $C(y) + m$ to $C(y)$.

Another easy part is the inequality $g_x(j) \leq g_y(j-m)$. Indeed, for every model A of y with $|A| \leq 2^{j-m}$ consider the model

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

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

The tricky part is the inverse inequality $g_x(j) \geq g_y(j-m)$. Let A be a model for x with $|A| \leq 2^j$ and $C(A) = g_y(j)$. 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 [7].

The first idea is to consider the projection of A: $\{y' \mid \langle y', z'\rangle \in A\}$. However 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 $2^i \leq |A_y| < 2^{i+1}$. Let A' be the set of those y' whose y'th section has at least 2^i elements. Then by counting arguments we have $|A'| \leq 2^{j-i}$. If $i \geq m$, we are done. However, it might be not the case. To lower bound i, we will relate it to the conditional complexity of z given y and A. Indeed, we have $C(z|A, y) \leq i$, as z can be identified by its ordinal number in yth section of A. Hence we know that $\log|A'| \leq j - C(z|A, y)$. Now we will improve A' using a result of [7]:

Lemma 1 (Lemma A.4 in [7]). *For every $A' \ni y$ there is $A'' \ni y$ with $C(A'') \leq C(A') - C(A'|y)$ and $\lfloor \log|A''| \rfloor = \lfloor \log|A'| \rfloor$.*

By this lemma we get the inequality

$$g_y(j - C(z|A, y)) \leq C(A') - C(A'|y).$$

Note that

$$C(A') - C(A'|y) = I(y : A') \leq I(y : A) = C(A) - C(A|y),$$

as $C(A'|A)$ is negligible. Thus we have

$$g_y(j - C(z|A, y)) \leq C(A) - C(A|y).$$

We claim that by the property (3) of the structure set this inequality implies that $g_y(j-m) \leq C(A)$. Indeed, as $C(z|A, y) \leq m$ we have by property (3):

$$g_y(j-m) \leq m - C(z|A, y) + C(A) - C(A|y) \leq m + C(A) - C(z|y) = C(A).$$

In all the above inequalities, we need to be careful about the error term, as they include sets, denoted by A or A', and thus the error term includes $O(\log C(A))$ or $O(\log C(A'))$. All the sets involved are either models of y or of x. W.l.o.g. we may assume that their complexity is at most $C(x)+O(1)$. Indeed, there is no need to consider models of y or x of larger complexity, as the models $\{y\}$ and $\{x\}$ have the least possible cardinality and their complexity is at most $C(x)+O(1)$. Since $C(x) \le C(y)+O(C(z|y)) \le C(y)+O(m)$, the term $O(\log C(A))$ is absorbed by the general error term.

This theorem answers our question: if y is obtained from x by removing m bits of noise then we expect that g_y satisfy Theorem 2. Now we will show that there are strings x as in Theorem 2 for which the notion of the MSS is well-defined but the structure function of some minimal sufficient statistics does not satisfy Theorem 2. The structure set of a finite set A of strings is defined 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 by an additive constant. Thus S_A and g_A are well-defined for finite sets A.

Theorem 3. *For every k there is a string y of length $2k$ and Kolmogorov complexity $C(y) = k$ such that*

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

and hence for any z of length k and conditional complexity $C(z|y) = k$ the structure function of the sting $x = \langle y, z\rangle$ is the following

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

(See Fig. 3.) Moreover, for every such z the string $x = \langle y, z\rangle$ has a model B of complexity $C(B) = k$ and log-cardinality $\log |B| = k$ such that $g_B(j) = k$ for all $j \le 2k$. All equalities here hold up to $O(\log k)$ additive error term.

The structure set of $x = \langle y, z\rangle$ clearly leaves the sufficiency line at the point $j = k$. Thus k is intuitively the complexity of minimal sufficient statistic and both models $A = y \times \{0,1\}^k$ 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. In contrast, the shape of the structure set of B is intuitively incompatible with the hypothesis that B, as a finite object, is a denoised x.

4 Desired Properties of Sufficient Statistics and a New Definition

We have seen that there is a string x that has two very different minimal sufficient statistics A and B. Recall the probabilistic notion of sufficient statistic [2]. In the

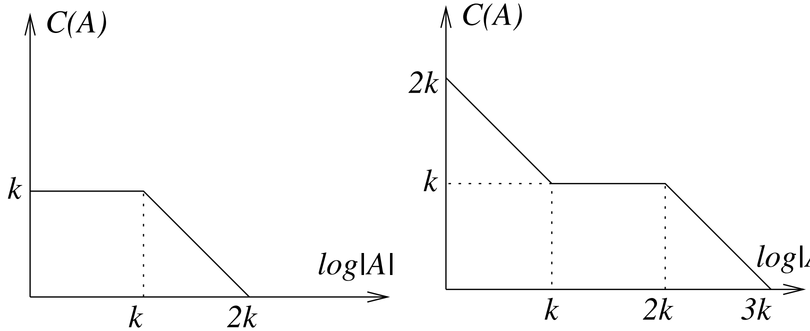

Fig. 3. Structure functions of y and x

probabilistic setting, we are given a parameter set Θ and for each $\theta \in \Theta$ we are given a probability distribution on a set X. For every probability distribution on Θ we thus obtain a probability distribution on $\Theta \times X$. A function $f : X \to Y$ (where Y is any set) is called a sufficient statistic, if for every probability distribution on Θ, the random variables x and θ are independent relative to $f(x)$. That is, for all $a \in X$, $c \in \Theta$,

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

In other words, $x \to f(x) \to \theta$ is a Markov chain (for every probability distribution on Θ). We say that a sufficient statistic f is *less* than a sufficient statistic g if for some function h with probability 1 it holds $f(x) \equiv h(g(x))$. An easy observation is that there is always a sufficient statistic f that is less than any other sufficient statistic: $f(a)$ is equal to the function $c \mapsto \mathrm{Prob}[\theta = c|x = a]$. Such sufficient statistics are called minimal. Any two minimal sufficient statistics have the same distribution and by definition every minimal sufficient statistic is a function of every sufficient statistic. Is it possible to define a notion of an algorithmic sufficient statistic that has similar properties? More specifically, we wish it to have the following properties.

(1) If A is an (algorithmic) sufficient statistic of x and $\log|A| = m$ then the structure function of $y = A$ satisfies the equality of Theorem 2. In particular, structure functions of every MSS A, B of x coincide.

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

As the example of Theorem 3 demonstrates, the property (1) does not hold for the definitions of Sections 2 and 3, and we do not know whether (2) holds. Now we propose an approach towards a definition that (hopefully) satisfies both (1) and (2). The main idea of the definition is as follows. As observed in [6], in order to have the same structure sets, the strings x, y should be equivalent in the following strong sense: there should exist short *total* programs p, q with $D(p, x) = y$ and $D(q, y) = x$ (where D is an optimal description mode 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 conditional description mode D have the following property. For any other description mode D' there is a constant c such that $CT_D(x|y) \leq CT_{D'}(x|y) + c$ 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*. The following lemma is straightforward.

Lemma 2. *For all x, y we have $|g_x(j) - g_y(j)| \leq 2\max\{CT(x|y), CT(y|x)\} + O(1)$. (If x, y are strongly equivalent then their structure sets are close.)*

Call A a *strongly sufficient statistic* of x if $CT(A|x) \approx 0$ and $C(x|A) \approx \log|A|$. More specifically, call a model A of x an *α, β-strongly sufficient statistic* of x if $CT(A|x) \leq \alpha$ and $C(x|A) \geq \log|A| - \beta$. The following theorem states that strongly sufficient statistics satisfy the property (1). It is a direct corollary of Theorem 2 and Lemma 2.

Theorem 4. *Assume that y is an α, β-strongly sufficient statistic of x and $\log|y| = m$. Then for all $j \geq m$ we have $g_x(j) = g_y(j - m)$ and for all $j \leq m$ we have $g_x(j) = C(y) + m - j$. The equalities here hold up to a $O(\log C(y) + \log m + \alpha + \beta)$ term.*

Let us turn now to the second desired property of algorithmic sufficient statistics. We do not know whether (2) holds in the case when both A, B are strongly sufficient statistics. Actually, for strongly sufficient statistics it is more natural to require that the property (2) hold in a stronger form: (2') Assume that A is a MSS and both A, B are strongly sufficient statistics of x. Then $CT(A|B) \approx 0$. Or, in an even stronger form: (2") Assume that A is a minimal strongly sufficient statistic (MSSS) of x and B is a strongly sufficient statistic of x. Then $CT(A|B) \approx 0$.

An interesting related question: (3) Is there always a strongly sufficient statistic that is a MSS?

Of course, we should require that properties (2), (2') and (2") hold only for those x for which the notion of MSS or MSSS is well-defined. Let us state the properties in a formal way. To this end we introduce the notation $\Delta_x(A) = CT(A|x) + \log|A| - C(x|A)$, which measures "the deficiency of strong sufficiency" of a model A of x. In the case $x \notin A$ we let $\Delta_x(A) = \infty$. To avoid cumbersome notations we reduce generality and focus on strings x whose structure set is as in Theorem 3. In this case the properties (2') and (3) read as follows: (2') For all models A, B of x,

$$CT(A|B) = O(|C(A) - k| + \Delta\mathrm{T}_x(A) + \Delta\mathrm{T}_x(B) + \log k).$$

(3) Is there always a model A of x such that $CT(A|x) = O(\log k)$, $\log|A| = k + O(\log k)$ and $C(x|A) = k + O(\log k)$.

It is not clear how to formulate property (2") even in the case of strings x satisfying Theorem 3 (the knowledge of g_x does not help).

We are only able to prove (2') in the case when both A, B are MSS. By a result of [7], in this case $C(A|B) \approx 0$ (see Theorem 5 below). Thus our result strengthens this result of [7] in the case when both A, B are strongly sufficient statistics (actually we need only that A is strong).

Let us present the mentioned result of [7]. Recalling that the notion of MSS is not well-defined, the reader should not expect a simple formulation. Let $d(u, v)$ stand for $\max\{C(u|v), C(v|u)\}$ (a sort of algorithmic distance between u and v).

Theorem 5 (Theorem V.4(iii) from [7]). *Let N^i stand for the number of strings of complexity at most i.* [2] *For all $A \ni x$ and i, either $d(N^i, A) \leq C(A) - i$, or there is $T \ni x$ such that $\log |T| + C(T) \leq \log |A| + C(A)$ and $C(T) \leq i - d(N^i, A)$, where all inequalities hold up to $O(\log(|A| + C(A)))$ additive term.*

Theorem 6. *There is a function $\gamma = O(\log n)$ of n such that the following holds. Assume that we are given a string x of length n and natural numbers $i \leq n$ and $\varepsilon < \delta \leq n$ such that the complexity of every $\varepsilon + \gamma$-sufficient statistic of x is greater than $i - \delta$. Then for every ε-sufficient statistics A, B of x of complexity at most $i + \varepsilon$, we have $CT(A|B) \leq 2 \cdot CT(A|x) + \varepsilon + 2\delta + \gamma$.*

Let us see what this statement yields for the string $x = \langle y, z \rangle$ from Theorem 3. Let $i = k$ and $\varepsilon = 100 \log k$, say. Then the assumption of Theorem 6 holds for $\delta = O(\log k)$ and thus $CT(A|B) \leq 2 \cdot CT(A|x) + O(\log k)$ for all $100 \log k$-sufficient B, A of complexity at most $k + 100 \log k$.

Proof. Fix models A, B as in Theorem 6. We claim that if $\gamma = c \log n$ and c is a large enough constant, then the assumption of Theorem 6 implies $d(B, A) \leq 2\delta + O(\log n)$. Indeed, we have $K(A) + \log |A| = O(n)$. Therefore all the inequalities of Theorem 5 hold with $O(\log n)$ precision. Thus for some constant c, by Theorem 5 we have $d(N^i, A) \leq \varepsilon + c \log n$ (in the first case) or we have a T with $C(T) + \log |T| \leq i + \varepsilon + c \log n$ and $d(N^i, A) \leq i - C(T) + c \log n$ (in the second case). Let $\gamma = c \log n$. The assumption of Theorem 6 then implies that in the second case $C(T) > i - \delta$ and hence $d(N^i, A) < \delta + c \log n$. Thus anyway we have $d(N^i, A) \leq \delta + c \log n$. The same arguments apply to B and therefore $d(A, B) \leq 2\delta + O(\log n)$.

In the course of the proof, we will neglect terms of order $O(\log n)$. They will be absorbed by γ in the final upper bound of $CT(A|B)$ (we may increase γ).

Let p be a total program witnessing $CT(A|x)$. We will prove that there are many $x' \in B$ with $x' \in p(x') = A$ (otherwise $C(x|B)$ would be smaller than assumed). We will then consider all A' such that there are many $x' \in B$ with $x' \in p(x') = A'$. We will then identify A given B in few bits by its ordinal number among all such A's.

Let $D = \{x' \in B \mid x' \in p(x') = A\}$. Obviously, D is a model of x with

$$C(D|B) \leq C(A|B) + l(p) \leq 2\delta + l(p).$$

Therefore

$$C(x|B) \leq C(D|B) + \log |D| \leq \log |D| + 2\delta + l(p).$$

[2] Actually, the authors of [7] 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.

On the other hand, $C(x|B) \geq \log|B| - \varepsilon$, hence $\log|D| \geq \log|B| - \varepsilon - 2\delta - l(p)$.
Consider now all A' such that

$$\log|\{x' \in B \mid x' \in p(x') = A'\}| \geq \log|B| - \varepsilon - 2\delta - l(p).$$

These A' are pairwise disjoint and each of them has at least $|B|/2^{\varepsilon+2\delta+l(p)}$ elements of B. Thus there are at most $2^{\varepsilon+2\delta+l(p)}$ 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 $\varepsilon + 2\delta + 2l(p)$ bits that maps B to A and $CT(A|B) \leq \varepsilon + 2\delta + 2l(p)$.

Another interesting related question is whether the following holds: (4) *Merging strongly sufficient statistics:* If A, B are strongly sufficient statistics for x then x has a strongly sufficient statistic D with $\log|D| \approx \log|A| + \log|B| - \log|A \cap B|$.

It is not hard to see that (4) implies (2"). Indeed, as merging A and B cannot result in a strongly sufficient statistic larger than A we have $\log|B| \approx \log|A \cap B|$. Thus to prove that $CT(A|B)$ is negligible, we can argue as in the last part of the proof of Theorem 6.

References

1. Antunes, L., Fortnow, L.: Sophistication revisited. In: Baeten, J.C.M., Lenstra, J.K., Parrow, J., Woeginger, G.J. (eds.) ICALP 2003. LNCS, vol. 2719, pp. 267–277. Springer, Heidelberg (2003)
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
3. Gács, P., Tromp, J., Vitányi, P.M.B.: Algorithmic statistics. IEEE Trans. Inform. Th. 47(6), 2443–2463 (2001)
4. Kolmogorov, A.N.: Talk at the Information Theory Symposium in Tallinn, Estonia (1974)
5. Shen, A.K.: Discussion on Kolmogorov complexity and statistical analysis. The Computer Journal 42(4), 340–342 (1999)
6. Shen, A.K.: Personal communication (2002)
7. Vereshchagin, N.K., Vitányi, P.M.B.: Kolmogorov's structure functions and model selection. IEEE Trans. Information Theory 50(12), 3265–3290 (2004)