Electronic Colloquium on Computational Complexity, Report No. 16 (2000)

Information Distance and Conditional Complexities

Mikhail V. Vyugin[*]

Abstract

C.H. Bennett, P. Gács, M. Li, P.M.B. Vitányi, and W.H. Zurek have defined information distance between two strings x, y as

$$d(x, y) = \max\{K(x|y), K(y|x)\}$$

where $K(x|y)$ is the conditional Kolmogorov complexity. It is easy to see that for any string x and any integer n there is a string y such that $d(x, y) = n + O(1)$.

In this paper we prove the following (stronger) result: for any n and for any string x such that $K(x) \geqslant 3n + O(1)$ there exists a string y such that both $K(x|y)$ and $K(y|x)$ are equal to $n + O(1)$.

1 Introduction

Conditional Kolmogorov complexity $K(x|y)$ of a binary string x relative to binary string y was defined in [2] as the length of the shortest program that computes x given y as an input. (For main properties of conditional complexity see [3] where the notation $C(x|y)$ is used and $K(x|y)$ is reserved for prefix complexity.)

Assume that a string x and integer n are given. It is easy to find a string y such that $K(y|x) = n+O(1)$ (most strings of length n have this property).

It is also possible to find y such that $K(x|y) = n + O(1)$ (assuming that $n \leqslant K(x)$). (Indeed, let y be a prefix of x of length k. When k increases, $K(x|y)$ decreases continuously and takes all values between $K(x)$ and 0.)

We prove that one can find y that satisfies *both* requirements $K(x|y) = n + O(1)$ and $K(y|x) = n + O(1)$ at the same time (see theorem 1 for an exact statement).

[*] Dept. of Mathematical Logic and Theory of Algorithms, Moscow State University, Vorobjevy Gory, Moscow 119899, Russia. E-mail: misha@vyugin.mccme.ru.

ISSN 1433-8092

2 Main theorem

Theorem 1. *There exists a constant c such that for any n the following holds: for any string x such that $K(x) \geqslant 3n + c$ there exists a string y such that both $K(x|y)$ and $K(y|x)$ are between n and $n + c$:*

$$n \leqslant K(x|y), K(y|x) \leqslant n + c.$$

Proof. For a fixed n we prove the theorem using the following game. There are two players called Man and Nature. Let $X = Y = (0, 1)^*$. Consider X as a set of left-side vertices of an infinite bipartite graph and Y as a set of right-side vertices. The graph is constructed during the game: Man and Nature add edges to the graph. Man adds undirected edges while Nature adds directed edges.

Initially all pairs of type (z, z) are connected by undirected edges which are counted as Man's edges.

Man (at his turn) may connect some left-side vertex x to some right-side vertex y by an undirected edge. He may also declare some left-side vertices as "simple" (mark them).

Nature may connect some left-side vertex x to some right-side vertex y by a directed edge going from x to y or vice versa.

Players must obey the following restrictions:

- (1) Maximum number of (undirected) Man's edges adjacent to any (leftor right-side) vertex is $2^{n+1} - 1$. (This restriction will be used to establish upper bounds on both conditional complexities.)

- (2) Maximum number of (directed) Nature's edges outgoing from any (left- or right-side) vertex is $2^n - 1$. (This restriction will be used to establish lower bounds on both conditional complexities.)

- (3) Maximum number of marked ("simple") vertices is 2^{3n+1}. (This will be used to prove that all marked strings have complexity less than $3n + c$.)

Players make their moves in turn; the game never terminates. Man's goal is to guarantee that any left-side vertex x either will be marked or has a "free" undirected edge after sufficiently large number of moves. We say that undirected edge connecting vertices x and y is "free" if there is no directed edge going from x to y or from y to x. (Otherwise the edge is called "covered".)

Note that the game is determined by parameter n.

Lemma 1 below shows that Man has a winning strategy. Before proving it we explain how this strategy is used to prove Theorem 1.

Consider the following "natural" strategy for Nature. It generates pairs (x, y) such that $K(x|y) < n$; for each pair (x, y) with this property Nature creates two directed edges going from y to x in both directions. It is easy to see that this strategy obeys to the restrictions given above. This strategy is computable (given n). Note that actually Nature ignores Man's moves.

By Lemma 1 below, there exists a computable winning strategy for Man. This strategy wins against any strategy of Nature, including the "natural" strategy. Assuming that Nature follows the natural strategy and Man follows the computable winning strategy described in Lemma 1 we get a computable process creating Man's and Nature's edges.

Let us look at vertex x on the left-hand side (here x is the string given in the statement of the theorem). Man wins, therefore either x will be marked or x has a free edge (after some number of moves). In the first case (as we prove) $K(x) < 3n + c$ for some constant c. In the second case consider a free edge that is adjacent to x after sufficiently large number of moves. (Since the number of outgoing edges is limited, free edge stabilizes.) This edge connects x to some string y. We will prove that this y satisfies the requirements of the theorem.

Lower bounds for $K(x|y)$ and $K(y|x)$ are guaranteed by Nature's strategy, because free edges are not covered by Nature's edges.

Now we prove that $K(y|x)$ and $K(x|y)$ does not exceed $n + c$ if y is connected to x by a free edge. Consider the games (using described strategies) for all n in parallel on the same bipartite graph; all edges will be marked by the corresponding values of n. Note that this process is computable.

For each n, each vertex z and each undirected edge e adjacent to z we define *ordinal number* of e with respect to z as the ordinal number of e in the set of all Man's edges adjacent to z in order of edges' enumeration in n-th game. We write this number as binary string of length $n + 1$ (padding it with zeros if necessary). Let x and y be connected by an undirected edge in the n-th game. We can use the ordinal number of the edge (x, y) in n-th game to reconstruct y given x. Note that n is determined by the ordinal number (which is a string of length $n + 1$) and therefore can be omitted. Therefore, $K(y|x) \leqslant n + c$ for some constant c. The bound for $K(x|y)$ may be obtained in the same way.

We need to prove also that if x was marked in n-th game then $K(x) \leqslant 3n + c$ for some constant c. It can be done by a similar argument which we omit.

□

Lemma 1. *Game described in the proof of Theorem 1 has a winning strategy for Man. This strategy is computable given n.*

Proof. Consider any (left- or right-side) vertex z. Let us define *weight* of z as pair (k, l) where k is the number of Man's edges outgoing from z covered by Nature's edges going from left to right, and l is the number of Man's edges covered by Nature's edges going from right to left.

Man waits until Nature covers his edge. (If this never happens, he wins.) Let us consider the situation after the edge is covered.

All left-side vertices are divided into three classes. First class consists of all marked vertices. There are no free edges adjacent to them. The second class consists of exactly one vertex x which is not marked and has no free adjacent edges. At its last move Nature covered an edge adjacent to x. The third class is the set of all other left-side vertices. Each of them is adjacent to one and only one free edge. Any free edge connects vertices with equal weights. Each right-side vertex has at most one free adjacent edge.

Assume that Nature has covered Man's edge (x, y) (x is a left-side vertex, y is a right-side vertex). By induction hypothesis weights of x and y were equal and remain equal after Nature's move. Man has to mark x or connect it to some vertex $\tilde{y}$. Let Man connect x to such $\tilde{y}$ that $\tilde{y}$ has the same weight as x and x was not connected to $\tilde{y}$ earlier. If there is no such $\tilde{y}$ then Man marks x. After that Man again waits until Nature covers his edge.

The strategy is described. It remains to prove that the number of marked vertices is bounded by 2^{3n+1}.

Let $(k, l) \neq (0, 0)$ be any fixed weight and x be a vertex of this weight. First, note that $k, l < 2^n$. Therefore, the number of covered edges adjacent to vertex x is less than $2^{n+1} - 1$. Therefore the total number of Man's edges adjacent to one vertex is less than 2^{n+1}.

Free right-side vertex is a vertex that has no adjacent free Man's edges. The number of left-side vertices of weight (k, l) is equal to the number of right-side vertices of the same weight. Therefore after Man's move the number of marked left-side vertices of weight $(k.l)$ equals to the number of free right-side vertices of the same weight.

Suppose there are 2^{n+1} marked vertices of weight (k, l) and Man has to connect x of weight (k, l) to $\tilde{y}$ of the same weight. Then such $\tilde{y}$ exists among 2^{n+1} free right-side vertices of weight (k, l) (x was not connected with all of them earlier). Therefore, the number of marked vertices of any fixed weight does not exceed 2^{n+1}. The number of different weights is 2^{2n}. Therefore the total number of marked vertices does not exceed 2^{3n+1}.

□

Remark 1. Modifying the proof of Theorem 1, we can replace the condition $K(x) \geqslant 3n + c$ by a weaker condition $K(x) \geqslant 2n + c$.

Indeed, it is enough to change the definition of weight (see the proof of Lemma 1). Define weight as the difference between elements of a pair. Then the number of different weights is bounded by 2^{n+1} instead of 2^{2n}, and the proof goes as before.

3 Related results

Remark 2. Our results do not cover the case when $n \leqslant K(x) \leqslant 2n$. The only thing we can do for this case is to apply a simple general construction which gives y such that $K(y|x) = n + O(\log K(x))$ and $K(x|y) = n + O(\log K(x))$.

Here is it: take a shortest program that produces x (and has length $K(x)$) and replace n last bits of this program by a random (and independent of x) string of length n.

Remark 3. There is another, more general question. Let x be a string and let m, n be two integers. We want to find a string y such that

$$K(x|y) \approx n, \quad K(y|x) \approx m.$$

Here the following is known:

(1) If $n \leqslant m$, one can find y such that $K(x|y) = n + O(\log m)$, $K(y|x) = m + O(\log m)$. This is an easy consequence of Theorem 1 and Remark 2: take y such that $K(x|y) \approx K(y|x) \approx n$ (up to $O(\log n)$) and add $m - n$ random bits to y.

(2) Andrei A. Muchnik has proved that this result is not valid for the case $n > m$ (personal communication).

Acknowledgements. N.Vereshchagin and A.Shen explained me the main results and methods of Kolmogorov complexity. The problem solved in this paper was posed by Andrei A. Muchnik. He helped me to simplify the proof and allowed to mention his unpublished results. Author is deeply grateful to them for very useful discussions.

References

[1] C.H. Bennett, P. Gács, M. Li, P.M.B. Vitányi, and W.H. Zurek. “Information Distance”, IEEE Trans. on Information Theory **44** (1998), No. 4, 1407–1423.

[2] A.N. Kolmogorov. “Three approaches to the quantitative definition of information.” *Problems of Information Transmission*, 1(1):1–7, 1965.

[3] M. Li, P. Vitányi. An Introduction to Kolmogorov Complexity and its Applications. Springer Verlag, 1997.