Doklady Mathematics, Vol. 68, No. 1, 2003, pp. 132–134. Translated from Doklady Akademii Nauk, Vol. 391, No. 6, 2003, pp. 738–740.

An Improvement of Kolmogorov's Estimates Related to Random Number Generators and a Definition of Randomness in Terms of Complexity

A. L. Semenov and An. A. Muchnik

Presented by Academician Yu.I. Zhuravlev January 4, 2003

Received January 10, 2003

In the 1930s, A.N. Kolmogorov constructed a substantiation of probability theory by means of measure theory. However, not all problems related to the substantiation were solved. In 1963 [1], Kolmogorov started to develop a new approach to the problem, the theory of descriptive complexity. In [1], he obtained upper and lower bounds for the maximal number of admissible place-selection rules for which a random number generator surely exists and stated the problem of obtaining an estimate of exact order. In this paper, we solve this problem of Kolmogorov; namely, we show that his lower bound is of exact order.

In [1], Kolmogorov defined the notion of a place-selection rule for a finite binary sequence **t**. Its informal description is as follows (see [1] for the precise definition). Let us imagine that we have a set of cards; the number of cards equals the length of **t**. The figures from the sequence **t** are written on the card's faces; the backs of all cards are identical. First, the cards lie on their faces in the same order in which the figures are arranged in **t**. The rule decides which card is to be overturned and (before the card is overturned) whether the figure written on the card should be included in the subsequence. In making the current decision, the rule can take into account the figures written on the cards already overturned. The subsequence selected by a rule r from a sequence **t** is denoted by $r[\mathbf{t}]$.

Sometimes, it is useful to consider narrower classes of rules. The monotonic rules (considered for the first time by Church in 1940) always overturn cards in their initial order. The nonadaptive rules specify at once a set of cards, and the subsequence is formed by the figures written on these cards and arranged in the initial order.

Definition 1. Let $\mathcal{R}_L$ be a set of rules on sequences of length L. A sequences **t** of length L is called an (n, ε)-random number generator for $\mathcal{R}_L$ if each rule from $\mathcal{R}_L$ selects a subsequence $r[\mathbf{t}]$ in **t** with the following property:

If the length of the subsequence is not less than n, then the fraction of zeros in this subsequences differs from $\frac{1}{2}$ by a value smaller than ε.

The absolute value of the difference between $\frac{1}{2}$ and the fraction of zeros in the sequence is called the deviation (of the fraction of zeros).

Remark. The results given below can be generalized to the case where 0 and 1 are encountered at frequencies close to p and $1 - p$, respectively, in all long subsequences selected by simple rules.

To give a precise meaning to the notion of not too complex rules, we define the complexity of a finite set to be the binary logarithm of its cardinality. Let us introduce the notation $d(n, \varepsilon) \rightleftharpoons 2n\varepsilon^2\log_2 e$.

Theorem 1 (Kolmogorov, 1963). *Consider arbitrary numbers L (sequence length), $\varepsilon > 0$ (the deviation of frequency from probability), and $n \geq \varepsilon^{-4}$ (length of a selected subsequence). For any set of rules $\mathcal{R}_L$ of complexity less than $d(n, \varepsilon)(1 - \varepsilon)$, there exists an (n, ε)-random number generator.*

In [1], Kolmogorov only gave an outline of the proof of this theorem. The complete proof (which involves some delicacies) is contained in our paper which is being prepared for publication in the journal "Problemy Peredachi Informatsii."

The following theorem is an algorithmic analog of Theorem 1 (the notion of conditional entropy was introduced by Kolmogorov in [2]).

Theorem 1'. *Consider arbitrary numbers L (sequence length), $\varepsilon > 0$ (the deviation of frequency from probability), and $n \geq \varepsilon^{-4}$ (length of a selected subsequence). For the set $\mathcal{R}_L$ consisting of all rules such that their entropies conditional to a known L are*

Kibernetika Scientific Council,
Russian Academy of Sciences,
ul. Vavilova 40, Moscow, 117333 Russia
e-mail: alsemenov@mtu-net.ru,
muchnik@lpcs.math.msu.ru

smaller than $d(n, \varepsilon)(1 - \varepsilon)$, *there exists an* (n, ε)*-random number generator.*

Theorem 2 (Kolmogorov, 1963). *Consider arbitrary numbers* L (*sequence length*), $\varepsilon \in \left(0, \frac{1}{20}\right)$ (*the deviation of frequency from probability*), *and* $n \in \left[\varepsilon^{-3}, \frac{L}{2}\right]$ (*length of a selected subsequence*). *There exists a set* $\mathcal{R}_L$ *of nonadaptive rules of complexity less than* $4n\varepsilon(1 + 5\varepsilon)$ *for which there exists no* (n, ε)*-random number generator.*

Theorems 1 and 2 give lower and upper bounds, respectively, for the maximal number τ such that, for any L and any set of rules of complexity less than τ, there exists at least one (n, ε)-random number generator of length L. Since $d(n, \varepsilon) = 2n\varepsilon^2\log_2 e$ is much smaller than $4n\varepsilon(1 + 5\varepsilon)$ at small ε, Kolmogorov wanted to remove the discrepancy between the exponents of ε in the lower and upper estimates. It turned out that the lower estimate obtained by Kolmogorov is of exact order (even for nonadaptive rules).

Theorem 3. *Consider arbitrary numbers* L (*sequence length*), $\varepsilon \in \left(0, \frac{1}{3}\right)$ (*the deviation of frequency from probability*), *and* $n \in \left[2\varepsilon^{-3}\log_2 L, \frac{L}{2}\right]$ (*length of a selected subsequence*). *There exists a set* $\mathcal{R}_L$ *of nonadaptive rules of complexity less than* $d(n, \varepsilon)\frac{1 + \varepsilon}{1 - n/(L - 1)}$ *for which there exists no* (n, ε)*-random number generator.*

The existence of such a set of rules is proved probabilistically, as in Theorem 1. However, this time, we consider probability distribution over the rules and show that the event "for a set $\mathcal{R}_L$ of rules, there exists an (n, ε)-generator" has a probability of less than 1.

We seek the required set of rules among the nonadaptive rules that select subsequences of length precisely n; i.e., a rule is specified by an n-element subset of the set $1, 2, \ldots, L$. The number of such rules is $\binom{L}{n}$; we introduce the uniform probability distribution on the set of these rules.

Take a sequence $\mathbf{t}$ of length L. Let us estimate from below the probability that it is not an (n, ε)-generator for a randomly selected rule r, i.e., that the deviation in $r[\mathbf{t}]$ is not less than ε. Suppose that the number of zeros in $\mathbf{t}$ is not smaller that the number of ones (the opposite case is handled symmetrically).

Consider the situation where the numbers $\frac{L}{2}$ and $n\left(\frac{1}{2} + \varepsilon\right)$ are integer; the general case is easily reduced to this situation. We want to estimate from below the probability of such a deviation for which the fraction of zeros in a sample is larger than the number of ones by at least ε; obviously, this probability is minimal when $\mathbf{t}$ contains equally many zeros and ones. It is sufficient to estimate the probability of the deviation equal to precisely ε. Obviously, this probability is

$$\frac{\dbinom{\frac{L}{2}}{\left(\frac{1}{2} - \varepsilon\right)n}\dbinom{\frac{L}{2}}{\left(\frac{1}{2} + \varepsilon\right)n}}{\dbinom{L}{n}}.$$

Using the upper and lower bounds for the binomial coefficients implied by the Stirling formula and an estimate of the Shannon entropy, we conclude that the sought probability is larger than e^{-K}, where

$$K = \frac{2n\varepsilon^2}{1 - \frac{n}{L}}\left(1 + \frac{4\varepsilon^2}{3}\right) + \frac{1}{2}(1 + \ln n).$$

The probability that, for a fixed sequence, one rule does not select a subsequence with deviation at least ε, turns out to be smaller than $1 - e^{-K}$. Now, let us take independently N random rules (some of these rules may coincide). The probability that a fixed sequence $\mathbf{t}$ is an (n, ε)-generator for the set of these rules is smaller than

$$(1 - e^{-K})^N < e^{-Ne^{-K}} \qquad (*)$$

(we have used the inequality $\left(1 - \frac{1}{x}\right)^x < e^{-1}$, which holds for $x > 1$). Multiplying the value on the right-hand side of inequality (*) by the number of sequences of length L, we obtain a sharp upper bound for the probability of the existence of at least an (n, ε)-generator for the given set of rules; namely,

$$2^L e^{-Ne^{-K}} = e^{L\ln 2 - Ne^{-K}},$$

which does not exceed 1 for $N = \lceil e^K L \ln 2 \rceil < e^K L$. The complexity of the set of rules under consideration is less than

$$\frac{2n\varepsilon^2}{1 - \frac{n}{L}}\left(1 + \frac{4\varepsilon^2}{3}\right)\log_2 e + 2\log_2 L.$$

It remains to show that this value is less than $d(n, \varepsilon)\frac{1 + \varepsilon}{1 - n/L}$ for $\varepsilon < \frac{1}{3}$ and $n \geq 2\varepsilon^{-3}\log_2 L$.

The following theorem is an algorithmic analog of Theorem 3.

Theorem 3'. *Consider arbitrary integer* L (*sequence length*), *rational* $\varepsilon \in \left(0, \frac{1}{3}\right)$ (*the deviation of frequency from probability*), *and integer* $n \in \left[2\varepsilon^{-3}\log_2 L, \frac{L}{2}\right]$ (*length of a selected subsequence*). *For the set* $\mathcal{R}_L(n, \varepsilon)$ *of all nonadaptive rules whose conditional entropy at given* L, n, *and* ε *is smaller than*

$$d(n, \varepsilon)\frac{1+\varepsilon}{1 - n/(L-1)} + C,$$

there exists no (n, ε)-*random number generator.* (*Here, C is a constant depending only on the choice of an optimal programming language.*)

By Theorem 3, for some set of nonadaptive rules of complexity less than $d(n, \varepsilon)\frac{1+\varepsilon}{1 - n/(L-1)}$, there exists no (n, ε)-generator of length L. Let us show that, for given L, n, and ε, a set of rules with this property can be constructed algorithmically.

Indeed, for any set $\mathcal{R}_L$ of nonadaptive rules and any sequence $\mathbf{t}$ of length L, we can determine effectively whether $\mathbf{t}$ is an (n, ε)-generator for $\mathcal{R}_L$ (for this purpose, we must apply each rule from $\mathcal{R}_L$ to $\mathbf{t}$ and calculate the deviation). Searching through all sequences of length L, we can determine whether there exist (n, ε)-generators for $\mathcal{R}_L$. Searching through all sets of nonadaptive rules of a given size, we can find the required set (if there are several sets with the required property, we take the first in the list).

The conditional (relative to L, n, and ε) entropy of each rule from the found set does not exceed the complexity of the set plus the length of the program implementing the procedure described in the preceding paragraph. The addition of the remaining rules with an entropy smaller than $d(n, \varepsilon)\frac{1+\varepsilon}{1 - n/(L-1)} + C$ preserves the absence of an (n, ε)-generator.

ACKNOWLEDGMENTS

The authors are grateful to N.K. Vereshchagin for interesting discussions of Kolmogorov's theory. The authors thank A.V. Chernov for great help in preparation of the paper for publication. The main content of the paper was reported at a Kolmogorov seminar at Moscow State University in the spring of 2002. We are grateful to the participants of the seminar for their attention.

This work was financially supported by the Russian Foundation for Basic Research (project nos. 01-01-00505, 02-01-10904, and 02-01-22001) and the Grant Council of the President of the Russian Federation for the support of leading scientific schools.

POSTSCRIPT

In [1], Kolmogorov defined the notion of a table of (n, ε, p)-random numbers for an $\mathcal{R}_L$, which differs from an (n, ε)-random number generator mentioned in Definition 1 in that the fractions of ones in the long samples obtained by rules from $\mathcal{R}_L$ is close to p rather than to 1/2. By $l(n, \varepsilon)$, Kolmogorov denoted the maximal number l such that, for any p, any L, and any set of rules having a complexity less than l, there exists at least one table of (n, ε, p)-random numbers of length L. The results obtained in [1] imply that, at sufficiently small ε $\left(\text{say at } \varepsilon < \frac{1}{20}\right)$ and $n \geq \varepsilon^{-4}$,

$$(2\log_2 e)n\varepsilon^2(1-\varepsilon) < l(n, \varepsilon) < 4n\varepsilon(1+5\varepsilon).$$

Kolmogorov stated the problem of removing the discrepancy between the power of ε in the lower and upper bounds for $l(n, \varepsilon)$. Our Theorem 3 implies the inequality

$$l(n, \varepsilon) < (2\log_2 e)n\varepsilon^2(1+2\varepsilon),$$

which holds at sufficiently small ε and $n \geq \varepsilon^{-4}$. (To prove this, it suffices to set $L = \lfloor 2^{n\varepsilon^3/2} \rfloor$ in Theorem 3.) Thus, both the upper and lower bounds for $l(n, \varepsilon)$ have the form $(2\log_2 e)n(\varepsilon^2 + o(\varepsilon^2))$.

REFERENCES

1. Kolmogorov, A.N., *Ind. J. Stat. Ser. A*, 1963, vol. 25, part 4, pp. 369–376. Reprinted in *Theor. Comput. Sci.*, 1998, vol. 207, pp. 387–395. Russian translation with author's addition: *Semiotika Informatika*, 1982, vol. 18, pp. 3–13; reprinted in collection Kolmogorov, A.N., *Teoriya informatsii i teoriya algoritmov* (The Theory of Information and the Theory of Algorithms), Moscow: Nauka, 1987, pp. 204–213.
2. Kolmogorov, A.N., *Prob. Peredachi Inform.*, 1965, vol. 1, no. 1, pp. 3–11.