Algorithmic information theory and martingales*

Laurent Bienvenu, Alexander Shen

Appendix A: Abstracts of Kolmogorov's talks

Some talks at the meetings of Moscow Mathematical Society have short abstracts published in the journal "Успехи математических наук" (Uspekhi matematicheckikh nauk, partially translated as "Russian mathemathical surveys"; these abstracts were not translated). Here we reproduce abstracts of three talks given by A.N. Kolmogorov devoted to algorithmic information theory (translated by Leonid Levin).

I. [vol. 23, no. 2, March-April 1968].

1. A.N. Kolmogorov, "Several theorems about algorithmic entropy and algorithmic amount of information."

Algorithmic approach to the foundations of information theory and probability theory was not developed far in several years from its appearance since some questions raised at the very start remained unanswered. Now the situation has changed somewhat. In particular, it is ascertained that the decomposition of entropy $H(x,y) \sim H(x) + H(y|x)$ and the formula $J(x|y) \sim J(y|x)$ hold in algorithmic concept only with accuracy $O([\log H(x,y)])$ (Levin, Kolmogorov).

Stated earlier cardinal distinction of algorithmic definition of a Bernoulli sequence (a simplest collective) from the definition of Mises-Church is concretized in the form of a theorem: there exist Bernoulli (in the sense of Mises-Church) sequences $x = (x_1, x_2, ...)$ with density of ones $p = \frac{1}{2}$, with initial segments of entropy ("complexity") $H(x^n) = H(x_1, x_2, ..., x_n) = O(\log n)$ (Kolmogorov).

For understanding of the talk an intuitive, not formal, familiarity with the concept of a computable function suffices.

(Moscow Mathematical Society meeting, October 31, 1967)

II. [vol. 27, no. 2, 1972]

1. A.N. Kolmogorov. "Complexity of specifying and complexity of constructing mathematical objects."

1. Organizing machine computations requires dealing with evaluation of (a) complexity of programs, (b) the size of memory used, (c) duration of computation. The talk describes a group of works that consider similar concepts in a more abstract manner.

2. It was noticed in 1964-1965 that the minimal length $K(x)$ of binary representation of a program specifying construction of an object x can be defined invariantly up to an additive constant (Solomonoff, A.N. Kolmogorov). This permitted using the concept of *definition complexity* $K(x)$ of constructive mathematical objects as a base for a new approach to foundations of information theory (A.N. Kolmogorov, Levin) and probability theory (A.N. Kolmogorov, Martin-Löf, Schnorr, Levin).

3. Such characteristics as "required memory volume," or "required duration of work" are harder to free of technical peculiarities of special machine types. But some results may already be extracted from axiomatic "machine-independent" theory of broad class of similar characteristics (Blum, 1967). Let

*[Bienvenu, Shen 2009], [Bienvenu, Shafer, Shen 2009]

$\Pi(p)$ be a characteristic of "construction complexity" of the object $x = A(p)$ by a program p, and $\Lambda(p)$ denotes the length of program p. The formula $K^n\Pi(x) = \inf(\Lambda(p) : x = A(p), \Pi(p) = n)$ defines "n-complexity of definition" of object x (for unsatisfiable condition the inf is considered infinite).

4. Barzdin's Theorem on the complexity $K(M_\alpha)$ of prefixes M_α of an enumerable set of natural numbers (1968) and results of Barzdin, Kanovich, and Petri on corresponding complexities $K^n\Pi(M_\alpha)$, are of general mathematical interest, as they shed some new light on the role of extending previously used formalizations in the development of mathematics. The survey of the state of this circle of problems was given in the form free from cumbersome technical apparatus.

(Moscow Mathematical Society meeting, November 23, 1971)

III. [Vol. 29,. no. 4 (155), 1974]

1. A.N. Kolmogorov. "Complexity of algorithms and objective definition of randomness."

To each constructive object corresponds a function $\Phi_x(k)$ of a natural number k – the log of minimal cardinality of x-containing sets that allow definitions of complexity at most k. If the element x itself allows a simple definition, then the function Φ drops to 1 even for small k. Lacking such a definition, the element is "random" in a negative sense. But it is positively "probabilistically random" only when function Φ, having taken the value Φ_0 at a relatively small $k = k_0$, then changes approximately as $\Phi(k) = \Phi_0 - (k - k_0)$.

(Moscow Mathematical Society meeting, April 16, 1974)

Appendix B. Levin's letters to Kolmogorov

These letters do not have dates but were written after submission of [Zvonkin, Levin 1970] in August 1970 and before Kolmogorov went (in January 1971) to the oceanographic expedition ("Dmitry Mendeleev" ship). Copies provided by L. Levin (and translated by A. Shen).

I.

Dear Andrei Nikolaevich! Few days ago I've obtained a result that I like a lot. May be it could be useful to you if you work on these topics while traveling on the ship.

This result gives a formulation for the foundations of probability theory different from Martin-Löf. I think it is closer to your initial idea about the relation between complexity and randomness and is much clearer from the philosophical point of view (as, e.g., [Yu. T.] Medvedev says).

Martin-Löf considered (for an arbitrary computable measure P) an algorithm that studies a given sequence and finds more and more deviation from P-randomness hypothesis. Such an algorithm should be P-consistent, i.e., find deviations of size m only for sequences in a set that has measure at most 2^{-m}. It is evident that a number m produced by such an algorithm on input string x should be between 0 and $-\log_2 P(x)$. Let us consider the complementary value $(-\log_2 P(x)) - m$ and call it the "complementary test" (the consistency requirement can be easily reformulated for complementary tests).

Theorem. *The logarithm of a priori probability* [*on the binary tree*] $-\log_2 R(x)$ *is a P-consistent complementary test for every measure P and has the usual algorithmic properties.*

Let me remind you that by a priori probability I mean the universal semicomputable measure introduced in our article with Zvonkin. [See [Zvonkin, Levin 1970].] It is shown there that it [its minus logarithm] is numerically close to complexity.

Let us consider a specific computable measure P. Compared to the universal Martin-Löf test f (specific to a given measure P) our test is not optimal up to an additive constant, but is asymptotically optimal. Namely, if the universal Martin-Löf test finds a deviation m, our test finds a deviation at least $m - 2\log_2 m - c$. Therefore, the class of random infinite banry sequences remains the same.

Now look how nice it fits the philosophy. We say that a hypothesis "x appeared randomly according to measure P" can be rejected with certainty m if the measure P is much less consistent with the appearence of

x than a priori probability (this means simply that $P(x) < R(x)/2^m$. This gives a law of probability theory that is violated with probability at most 2^{-m}. Its violation can be established effectively since R is [lower] semicomputable [=enumerable from below]. But if this law holds, all other laws of probability theory [i.e., all Martin-Löf tests] hold, too. The drawback is that it gives a bit smaller value of randomness deficiency (only $m - 2\log_2 m - c$ instead of m), but this is a price for the universality (arbitrary probability distribution). The connection with complexity is provided because $-\log_2 R(x)$ almost coincides with complexity of x. Now this connection does not depend on measure.

It is worth noting that the universal semicomputable measure has many interesting applications besides the above mentioned. You know its application to the analysis of randomized algorithms. Also it is ofter useful in proofs (e.g., in the proof of J.T.Schwartz' hypothesis regarding the complexity of almost all trajectories of dynamic systems). Once I used this measure to construct a definition of intuitionistic validity. All this show that it is a rather natural quantity.

L.

II.

Dear Andrei Nikolaevich!

I would like to show that plain complexity does not work if we want to provide an *exact* definition of randomness, even *for a finite case.* For the uniform distribution on strings of fixed length n the randomness deficiency is defined as n minus complexity. For a non-uniform distribution length is replaced by minus the logarithm of probability.

It turns out that even for a distribution on a finite set the randomness deficiency could be high on a set of large measure.

Example. Let

$$P(x) = \begin{cases} 2^{-(l(x)+100)}, \text{ if } l(x) \le 2^{100}; \\ 0, \text{ if } l(x) > 2^{100}. \end{cases}$$

Then $|\log_2 P(x)| - K(x)$ exceeds 100 *for all* strings x.

A similar example can be constructed for strings of some fixed length (by adding zero prefixes). The violation could be of logarithmic order.

Let me show you how to sharpen the definition of complexity to get an exact result (both for finite and infinite sequences).

Definitions. Let A be a monotone algorithm, i.e., for every x and every y that is a prefix of x, if $A(x)$ is defined, then $A(y)$ is defined too and $A(y)$ is a prefix of $A(x)$. Let us define

$$KM_A(x) = \begin{cases} \min\ l(p)\colon x \text{ is a prefix of } A(p); \\ \infty, \text{ if there is no such } p \end{cases}$$

The complexity with respect to an optimal algorithm is denoted by $KM(x)$.

Let $P(x)$ be a computable distribution on the Cantor space Ω, i.e., $P(x)$ is the measure of the set Γ_x of all infinite extensions of x.

Theorem 1.

$$KM(x) \le |\log_2 P(x)| + O(1);$$

Theorem 2.

$$KM((\omega)_n) = |\log_2 P((\omega)_n)| + O(1)$$

for P-almost all ω; here $(\omega)_n$ stands for n-bit prefix of ω. Moreover, the probability that the randomness deficiency exceeds m for some prefix is bounded by 2^{-m}.

Theorem 3. *The sequences ω such that $KM((\omega)_n) = |\log_2 P((\omega)_n)| + O(1)$, satisfy all laws of probability theory (all Martin-Löf tests).*

Let me use this occasion to tell you the results from my talk in the laboratory [of statistical methods in Moscow State University]: why one can omit non-computable tests (i.e., tests not definable without a strong language).

For this we need do improve the definition of complexity once more. The plain complexity $K(x)$ has the following property:

Remark. Let A_i be an effectively given sequence of algorithms such that

$$K_{A_{i+1}}(x) \le K_{A_i(x)}$$

for all i and x. Then there exists an algorithm A_0 such that

$$K_{A_0}(x) = 1 + \min_i K_{A_i}(x).$$

Unfortunately, it seems that $KM(x)$ does not have this property. This can be corrected easily. Let A_i be an effective sequence of monotone algorithms with finite domain (provided as tables) such that

$$KM_{A_{i+1}}(x) \le KM_{A_i(x)}$$

for all i and x. Let us define then

$$\overline{KM}_{A_i}(x) = \min_i KM_{A_i}(x).$$

Among all sequences A_i there exists an optimal one, and the complexity with respect to this optimal sequence is denoted by $\overline{KM}(x)$. This complexity coincides with the logarithm of an universal semicomputable semimeasure [=a priori probability on the binary tree].

Theorem 4. *$\overline{KM}(x)$ is a minimal semicomputable [from above] function that makes Theorem 2 true.*

Therefore no further improvements of $\overline{KM}$ are possible.

Now consider the language [=set] of all functions computable with a fixed noncomputable sequence [oracle] α. Assume that α is complicated enough, so this set contains the characteristic function of a universal enumerable set [$\mathbf{0}'$].

We can define then a relativized ["языковую" in the Russian original] complexity $\overline{KM}_\alpha(x)$ replacing algorithms by algorithms with oracle α, i.e., functions from this language.

Definition. A sequence ω is called *normal* if

$$\overline{KM}((\omega)_n) = \overline{KM}_\alpha((\omega)_n) + O(1).$$

For a finite sequence ω_n we define the "normality deficiency" as

$$\overline{KM}(\omega_n) - \overline{KM}_\alpha(\omega_n).$$

Theorem 5. *A sequence obtained by an algorithm from a normal sequence is normal itself.*

Theorem 6. *Let P be a probability distribution that is defined (in a natural encoding) by a normal sequence. Then P-almost every sequence is normal.*

This theorem exhibits a law of probability theory that says that a random process cannot produce a non-normal sequence unless the probability distribution itself is not normal. This is a much more general law than standard laws of probability theory since it does not depend on the distribution. Moreover, Theorem 5 shows that this law is not restricted to probability theory and can be considered as a universal law of nature:

Thesis. Every sequence that appears in reality (finite or infinite) has normality deficiency that does not exceed the complexity of the description (in a natural language) of how it is physically produced, or its location etc.

It turns out that this normality law (that can be regarded as not confined in probability theory) and the law corresponding to the universal computable test together imply any law of probability theory (not necessary computable) that can be described in the language. Namely,the following result holds:

Theorem 7. *Let P be a computable probability distribution. If a sequence ω is normal and passes the universal computable P-test, then ω passes any test defined in our language (i.e., every test computable with oracle α).*

Note that for every set of measure 0 there exists a test (not necessary computable) that rejects all its elements.

Let us give one more iunteresting result that shows that all normal sequences have similar structure.

Theorem 8. *Every normal sequence can be obtained by an algorithm from a sequence that is random with respect to the uniform distribution.*

III.

(This letter has no salutation. Levin recalls that he often gave notes like this to Kolmogorov, who rarely had much time to hear lengthy explanations and preferred something written in any case.)

We use a sequence α that provides a "dense" coding of a universal [recursively] enumerable set. For example, let α be the binary representation of [here the text "the sum of the a priori probabilities of all natural numbers" is crossed out and replaced by the following:] the real number

$$\sum_{p\in A}\frac{1}{p\cdot\log^2 p}$$

where A is the domain of the optimal algorithm.

A binary string p is a "good" code for x if the optimal algorithm converts the pair $(p, K(x))$ into a list of strings that contains x and the logarithm of the cardinality of this list does not exceed $K(x)+3\log K(x)-l(p)$. (The existence of such a code means that x is "random" when $n \geq l(p)$.)

We say that a binary string p is a canonical code for x if every prefix of p either is a "good" code for x or is a prefix of α, and $l(p) = K(x) + 2\log K(x)$.

Theorem 1. *Every x (with finitely many exceptions) has a canonical code p, and p and x can be effectively transformed into each other if $K(x)$ is given.*

Therefore, the "non-randomness" in x can appear only due to some very special information (a prefix of α) contained in x. I cannot imagine how such an x can be observed in (extracted from) the real world since α is not computrable. And the task "to study the prefixes of a specific sequence α" seems to be very special.

References

[Bienvenu, Shen 2009] Laurent Bienvenu, Alexander Shen. Algorithmic information theory and martingales. 6/15/09. http://arxiv.org/abs/0906.2614

[Bienvenu, Shafer, Shen 2009] Laurent Bienvenu, Glenn Shafer and Alexander Shen. On the history of martingales in the study of randomness. Electronic Journal for History of Probability and Statistics. 5(1), June 2009 http://www.jehps.net/juin2009/BienvenuShaferShen.pdf http://www.emis.ams.org/journals/JEHPS/juin2009/BienvenuShaferShen.pdf

[Zvonkin, Levin 1970] Zvonkin, A.K.; Levin, L.A. Slozhnost' konechnykh ob'ektov i obosnovanie ponyatij informatsii i sluchajnosti s pomoshch'yu teorii algoritmov (The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms). Uspehi Mat. Nauk 25 (1970), no. 6(156), 85–127; translated in Russ. Math. Surv. 25:6, 83-124 (1970).