Combinatorial foundations of information theory and the calculus of probabilities

Uspekhi Mat. Nauk 38:4 (1983), 27-36 Russian Math. Surveys 38:4 (1983), 29-40

Combinatorial foundations of information theory and the calculus of probabilities[1]

A.N. Kolmogorov

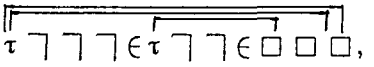
CONTENTS

§1. The growing role of finite mathematics

I wish to begin with some arguments that go beyond the framework of the basic subject of my talk. The formalization of mathematics according to Hilbert is nothing but the theory of operations on schemes of a special form consisting of finitely many signs arranged in some order or another and linked by some connections or other. For instance, according to Bourbaki's conception, the entire set theory investigates exclusively expressions composed of the signs

$$\square,\ \tau,\ \vee,\ \neg,\ =,\ \in,\ \supset$$

and of "letters" connected by "links" $\sqcap$ as, for instance, in the expression



which is the "empty set". Keeping the finite point of view, it would be logical to adopt for an infinite sequence of "letters" some standard notation, or another, for example,

$$\prod_{0},\ \prod_{1},\ \prod_{10},\ \prod_{11},\ \prod_{100},\ \dots$$

[1]The text published here was prepared in 1970 in connection with my talk at the International Congress of Mathematicians in Nice.

Curiously enough, owing to the presence of "links" $\sqcap$, expressions of the formalized Bourbaki mathematics are not "words" extended in one line as, for instance, in A.A. Markov's theory of normal algorithms, but in fact one-dimensional complexes with vertices marked by definite symbols.

But this conception of mathematics as occupied with the reorganization, according to well-defined rules, of specified one-dimensional complexes is only indirectly relevant to the real, intuitively accepted content of mathematics. Bourbaki remarks that in his conception the expression with the meaning "the number 1" contains some tens of thousands of signs, but this does not make the concept of the "number 1" inaccessible to our intuitive understanding.

Pure mathematics develops securely and predominantly as the science of the infinite. And Hilbert, the founder of the conception of completely formalized finite mathematics, undertook his titanic work merely to secure for the mathematicians the right of staying in "Cantor's paradise" of set theory. Apparently, this state of affairs is deeply grounded in the structure of our consciousness, which operates with great facility with intuitive ideas of unbounded sequences, limit passages, continuous and even "smooth" manifolds, and so on.

Until recently, in the mathematical treatment of natural science too, the prevailing way of modelling real phenomena was by means of mathematical models constructed on the mathematics of the infinite and the continuous. For example, in studying the process of molecular heat conductivity, we imagine a continuous medium in which the temperature is subject to the equation

$$\frac{\partial u}{\partial t} = K\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right). \tag{1}$$

Mathematicians usually regard the corresponding difference scheme

$$\Delta_t u = K(\Delta_{xx} u + \Delta_{yy} u + \Delta_{zz} u) \tag{2}$$

only as arising out of the approximate solution of the "exact" equation (1). But the real process of heat conduction is no more similar to its continuous model expressed by (1) than to the discrete model directly expressed by (2).

Quite probably, with the development of the modern computing technique it will be clear that in very many cases it is reasonable to conduct the study of real phenomena avoiding the intermediary stage of stylizing them in the spirit of ideas of mathematics of the infinite and the continuous, and passing directly to discrete models. This applies particularly to the study of systems with a complicated organization capable of processing information. In the most developed such systems the tendency to discrete work was due to reasons that are by now sufficiently clarified. It is a paradox requiring an explanation that while the human brain of a mathematician works essentially according to a discrete principle, nevertheless to a mathematician the intuitive grasp, say, of the properties of geodesics on

smooth surfaces is much more accessible than that of properties of combinatorial schemes capable of approximating them.

Using his brain, as given by the Lord, a mathematician may not be interested in the combinatorial basis of his work. But the artificial intellect of machines must be created by man, and man has to plunge into the indispensable combinatorial mathematics. For the time being it would still be premature to draw final conclusions about the implications for the general architecture of the mathematics of the future.

§2. Information theory

Discrete forms of storing and processing information are fundamental. They are at the base of the very measure of the "quantity of information" expressed in "bits", numbers of binary symbols. According to what has been said before, the discrete part of information theory is to some extent destined to play a leading organizing part in the development of combinatorial finite mathematics. From the general considerations that have been briefly developed it is not clear why information theory should be based so essentially on probability theory, as the majority of text-books would have it. It is my task to show that this dependence on previously created probability theory is not, in fact, inevitable. However, I shall confine myself to two examples.

The real substance of the entropy formula

$$H = -\sum p_i \log p_i \tag{1}$$

(here, and everywhere, the logarithms are binary) is as follows: if we carry out a large number n of independent experiments with a probability distribution

$$(p_1, p_2, \ldots, p_s)$$

of s possible outcomes of each experiment, then to record the result of a whole series of n experiments we need approximately

$$nH$$

binary digits. But this result holds under incomparably weaker and purely combinatorial assumptions. To record the result of our experiments it suffices to state that each of the results appeared

$$m_1, m_2, \ldots, m_s$$

times, respectively, and only afterwards to indicate the ordinal number of that of the

$$C(m_1, m_2, \ldots, m_s) = \frac{n!}{m_1!\, m_2! \ldots m_s!}$$

arrangements that took place.

For this no more than

$$s \log n + \log C(m_1, m_2, \ldots, m_n)$$

binary digits are needed, and for a large n this is approximately

$$n\left(-\sum \frac{m_i}{n} \log \frac{m_i}{n}\right) \sim nH. \tag{2}$$

By the law of large numbers, in the case of independent experiments with the probability distribution specified above, $m_i/n \sim p_i$. But our assumptions in deducing (2) were far weaker.

A second example with the entropy of a Markov chain is completely analogous. Here, too, the assumption that it is required to record the information about the realization of the Markov process is grossly superfluous.

§3. The definition of "complexity"

If any object is "simply" constructed, then for its description a small quantity of information is sufficient; but if it is "complicated", then its description must contain much information. According to certain arguments (see below §7), it is convenient to call the quantity thus introduced the "complexity".

We regard as the standard way of conveying information binary sequences beginning with 1,

$$1, 10, 11, 100, 101, 110, 111, 1000, 1001, \ldots,$$

which are the binary expressions of the natural numbers. We denote by $l(n)$ the length of the sequence n.

Suppose that we are dealing with some domain D of objects in which there is already some standard numbering of objects by numbers $n(x)$. However, indicating the number $n(x)$ is by no means always the most economical way of identifying an object x. For example, the binary notation for the number

$$9^{9^{9}}$$

is immensely long, but we have defined it fairly simply. We have to carry out a comparative study of the various means of specifying objects in D. It suffices to restrict ourselves to ways of identification that establish a correspondence between any number p written in binary notation and some number

$$n = S(p).$$

Thus, the way of specifying an object in D becomes nothing but a function S of a natural argument taking natural values. Somewhat later we shall turn to the case when this function is computable. Such methods can be called "effective". But for the time being we preserve full generality. For each object in D it is natural to consider among the numbers p leading to it the

one of smallest length $l(p)$. This smallest length is the "complexity" of the object x with the "specifying method S":

$$K_S(x) = \min l(p),$$
$$S(p) = n(x).$$

In the language of the mathematics of computation, p can be called a "programme", and S a "method of programming". Then we can say that p is the minimal length of a programme by which the object x can be obtained following the programming method S.

If there are several distinct methods

$$S_1, S_2, \ldots, S_r,$$

of specifying elements of D, then it is easy to construct a new method S that gives us any object $x \in D$ whose complexity $K_S(x)$ exceeds only, for example, by $\log r$ the original minimum of the complexities

$$K_{S_1}(x), K_{S_2}(x), \ldots, K_{S_r}(x).$$

The construction of such a method is very simple. We have to reserve sufficiently many initial digits of the sequence p to fix the method S_i that should be followed, by using as a programme the remaining digits of p.

We say that a method S "absorbs a method S' with a precision up to l" if always

$$K_S(x) \leqslant K_{S'}(x) + l.$$

We have shown above how to construct a method S that is stronger with a precision up to l than any of the methods $S_1, S_2, \ldots, S_r$, where approximately $l \sim \log r$.

Two methods S_1 and S_2 are called "l-equivalent" if each of them l-absorbs the other. This whole construction would hardly be productive if the hierarchy of methods with respect to absorption were quite odd. Comparatively recently it was noticed that under some fairly natural conditions this is not so. I follow my paper [1], but roughly the same ideas can be found in [3], [4], and [5]; however, in [3] they appear in a somewhat veiled form.

Theorem. *Among the computable functions $S(p)$ there exist optimal ones, that is, such that for any other computable function $S'(p)$,*

$$K_S(x) \leqslant K_{S'}(x) + l(S, S').$$

Clearly, all optimal methods of specifying objects in D are equivalent:

$$|K_{S_1}(x) - K_{S_2}(x)| \leqslant l(S_1, S_2).$$

Thus, from an asymptotic point of view, the complexity $K(x)$ of an element x, when we restrict ourselves to effective methods of specifying, does not depend on accidental peculiarities of the chosen optimal method.

Of course, the purely practical interest of this result depends on how great the divergences in complexity are for various sufficiently elastic, but at the same time convenient and natural, methods of programming.

§4. Regularity and randomness

The idea that "randomness" consists in a lack of "regularity" is thoroughly traditional. But apparently only now has it become possible to found directly on this simple idea precise formulations of conditions for the applicability of results of the mathematical probability theory to real phenomena.

Any results of observations can be registered in the form of a finite, though sometimes very long, entry. Therefore, when we speak of a lack of regularity in observational results, we have in mind merely the absence of a *fairly simple* regularity. For example, the sequence of 1000 digits

$$1274031274031274031 \ldots,$$

changing with a period of six digits is certainly to be regarded as a "regular" and not as "random". The sequence of the first thousand decimal digits of the fractional part of the number π

$$1415 \ldots,$$

is known to have many properties of "random sequences". But knowing the rule of its formation, we also refuse to accept it as "random". But if we are given a polynomial of degree 999 whose values for $x = 1, 2, 3, \ldots, 1000$ yield a sequence of integers $p(x)$ between 0 and 9, obtained as a result of honest random experiments like roulette play, then the presence of such a polynomial does not prevent us from continuing to regard the sequence as "random".

If by one method or another we have come to the conclusion that the sequence of results of given experiments does not admit a complete description in a form acceptable to us from the point of view of the complexity of its form, then we say that this sequence is only partially regular, or is partially "random". But this is still not the "randomness" that is needed to apply deductions of probability theory. In applying probability theory we do not confine ourselves to negating regularity, but from the hypothesis of randomness of the observed phenomena we draw definite positive conclusions.

We see presently that practical deductions of probability theory can be justified as consequences of hypotheses about the *limiting* complexity, under given restrictions, of the phenomena in question.

§5. The stability of frequencies

Following von Mises, the acknowledgement of the hypothesis on the stability of frequencies is often put at the basis of applications of probability theory. In a form close to the practice this conept was also accepted in my well-known booklet on the basis of probability theory published in 1933.

Suppose that the result of a sequence of a large number N of experiments is recorded in the form of a sequence of 0's and 1's.

$$11010010111001011 \ldots$$

We say that the appearance of a 1 is random with probability p if the proportion of 1's is

$$\frac{M}{N} \sim p \tag{1}$$

and this frequency cannot be altered substantially by a selection from our sequence of a reasonably long subsequence according to a fairly simple rule and so that the inclusion of some element of the original sequence in the subsequence proceeds without using the value of this element (for the most careful finite formulation of this principle of von Mises, see [17]).

But it turns out that this requirement can be replaced by another one that can be stated much simpler. The complexity of a sequence of 0's and 1's satisfying (1) cannot be substantially larger than

$$nH(p) = n(-p \log p - (1 - p) \log(1 - p)).$$

It can be proved that *the stability of frequencies in the sense of von Mises is automatically ensured if the complexity of our sequence is sufficiently close to the upper bound indicated above.*

I cannot make here this result quantitatively more precise (see [17], although the definition of complexity in [1] is not yet there) nor can I discuss from this point of view more complex problems of probability theory. But the principle is general. For example, assuming that a sequence of 0's and 1's represents a Markov chain with the matrix of transition probabilities

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

then, in essence, we give approximate values of the frequencies of 1's after 1's, 1's after 0's, 0's after 1's, and 0's after 0's. The maximal complexity of such a sequence of length n can be computed. If the complexity of a specific sequence with given transition frequencies is close to this maximum, then automatically all the predictions of the probabilistic theory of Markov chains apply to it.

§6. Infinite random sequences

So far the programme just outlined has not been carried out, but I have no doubt that it can be done. In fact, its execution must connect the mathematical probability theory with its applications more completely than a construction of the type of von Mises. Here I have in mind that there is no need whatsoever to change the established construction of the mathematical probability theory on the basis on the general theory of measure.

I am not inclined to attribute the significance of necessary foundations of probability theory to the investigations I am now going to survey. But they are most interesting in themselves.

For a mathematician it is an attractive problem to determine what infinite sequences of 0's and 1's should be called "random". I confine myself to the simplest case of sequences with frequencies of 0's and 1's equal to 1/2. In close connection with what was said before, one would wish to require from the sequence

$$x = (x_1, x_2, \ldots, x_n, \ldots),$$

that its finite segments

$$x^n = (x_1, x_2, \ldots, x_n)$$

have a complexity

$$K(x^n) \geqslant n - C,$$

where C is some constant (different for different x). But Martin-Löf has proved the following theorem:

Martin-Löf's first theorem. *If $f(n)$ is a computable function such that*

$$\sum 2^{-f(n)} = \infty, \tag{1}$$

then for any binary sequence

$$x = (x_1, x_2, \ldots, x_n, \ldots)$$

there are infinitely many values of n for which

$$K(x^n) < n - f(n).$$

The condition of the theorem is satisfied, for example, by the function $f(n) = l(n)$. But if the series

$$\sum 2^{-f(n)} \tag{2}$$

"converges constructively" (for details, see [8]), then almost all sequences x (in the sense of binary measure) have the property

$$K(x^n) \geqslant n - f(n), \tag{3}$$

from some n onwards. It would be illogical to take the property (3) as definition of a random sequence. Martin-Löf's definition is more profound. I cannot quote it here in full. Random binary sequences in the sense of Martin-Löf have all the "effectively verifiable" (see again [8]) properties that from the point of view of the usual modern probability theory are satisfied "with probability 1" in the case of independent experiments in which $x_n = 1$ with probability 1/2. For such random sequences Martin-Löf has proved a second theorem:

Martin-Löf's second theorem. *Random sequences of* 0*'s and* 1*'s satisfy* (3) *from some n onwards, provided that the function f is such that the series* (2) *converges constructively.*

I quoted these subtle, but fairly special results of Martin-Löf to show that here we have a field for very interesting mathematical research (in this connection, see other papers by Martin-Löf and Schnorr, for example, [9]).

§7. Relative complexity and quantity of information

The complexity of specifying any object can be facilitated when any other object is already specified. This fact reflects the following definition of the relative complexity of an object x, given an object y:

$$K_S(x \mid y) = \min_{S(n(y),\, p)=n(x)} l(p).$$

Here the method S of relative determinations is a function of two arguments, the number of the object y and the number p of the programme for computing the number $n(x)$ when y is given. Concerning relative complexities, everything that was said in §3 can be repeated.

If the relative complexity $K(x \mid y)$ is much smaller than the unconditional complexity $K(x)$, then it is natural to interpret it as an indication that the object y contains some "information" about x. It is, therefore, natural to regard the difference

$$\mathcal{I}_S(x \mid y) = K_S(x) - K_S(x \mid y)$$

as a quantitative measure of the information about x contained in y.

As a value of the second argument of the function $S(n, p)$ we admit the number 0, and we put

$$S(n, 0) = n$$

(the zero programme from n produces n). Then

$$K_S(x \mid x) = 0, \qquad \mathcal{I}_S(x \mid x) = K_S(x).$$

Thus, the complexity $K_S(x)$ can be called the information contained in an object about itself.

As regards applications, our definition of the quantity of information has the advantage that it refers to individual objects and not to objects treated as members of a set of objects with a probability distribution given on it. The probabilistic definition can be convincingly applied to the information contained, for example, in a stream of congratulatory telegrams. But it would not be too clear how to apply it, for example, to an estimate of the quantity of information contained in a novel or in the translation of a novel into another language relative to the original. I think that the new definition is capable of introducing in similar applications of the theory at least a clarity of principle.

The question arises whether the new definition allows us to prove a number of basic propositions of information theory that recommend themselves sufficiently. It is clear beforehand that they should hold merely to within additive constants corresponding to the indeterminacy in §3. One could not expect, for example, that the equality

$$\mathcal{J}(x \mid y) = \mathcal{J}(y \mid x) \tag{1}$$

holds exactly, but a priori it would seem that the difference between the left- and the right-hand side should be bounded. In fact, Levin and I have established only a weaker inequality of the type

$$|\mathcal{J}(x \mid y) - \mathcal{J}(y \mid x)| = \mathcal{O}(\log K(x, y)) \tag{2}$$

(see [16]). We have established that the difference can, in fact, be of this order.

But in applications amenable to the probabilistic approach, (2) replaces (1) completely. For the strict equality (1) of the probabilistic information theory allows us to draw real conclusions only in application to a large number of pairs (x_i, y_i), that is, essentially about the information in

$$(x_1, x_2, \ldots, x_r)$$

relative to

$$(y_1, y_2, \ldots, y_r)$$

and vice versa. And deductions of this kind can be made also from (2), where in this case the expression on the right-hand side is negligibly small.

§8. Barzdin's theorem

A new series of concepts turns out to be interesting even beyond the limits of probability theory and applied information theory. To give an example of it, I state a theorem by Barzdin'. It concerns infinite binary sequences

$$x = (x_1, x_2, \ldots, x_n, \ldots),$$

in which the set of numbers n with $x_n = 1$ is countable. If the complementary set of numbers n with $x_n = 0$ were also countable, then the function $f(n) = x_n$ would be computable, and the relative complexity $K(x^n \mid n)$ would be bounded. But in the general case (when the set of 1's is countable), $K(x^n \mid n)$ can grow unboundedly.

Barzdin's theorem [15]. *For any binary sequence with a countable set M of 1's and*

$$K(x^n \mid n) \leqslant \log n + C_M$$

there are sequences such that for any n

$$K(x^n \mid n) \geqslant \log n.$$

This theorem appears to me as having an interest of principle from the point of view of investigations on the foundations of mathematics. To be definite, we consider the following problem: we label all Diophantine equations by the natural numbers. Matiyasevich has proved recently that there is no general algorithm to answer the question whether the equation D_n is soluble in integers. But one can ask about the existence of an algorithm that enables us to answer the question of the existence or non-existence of solutions of the first n Diophantine equations with the help of some supplementary information under one order of growth or another of the quantity of this information as n increases. Barzdin's theorem shows that this growth can be very slow:

$$\log n + C.$$

§9. Conclusion

The talk was necessarily extremely incomplete. A detailed bibliography of relevant papers can be found in [16]. I repeat some conclusions:

1. Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character.

2. The applications of probability theory can be put on a uniform basis. It is always a matter of consequences of hypotheses about the impossibility of reducing in one way or another the complexity of the description of the objects in question. Naturally, this approach to the matter does not prevent the development of probability theory as a branch of mathematics being a special case of the general measure theory.

3. The concepts of information theory as applied to infinite sequences give rise to very interesting investigations, which, without being indispensable as a basis of probability theory, can acquire a certain value in the investigation of the algorithmic side of mathematics as a whole.

References

[1] A.N. Kolmogorov, Three approaches to the definition of the concept "quantity of information", Problemy peredachi Informatsii, **1** (1965), 3-11. MR 32 # 2273. = Problems Inform. Transmission **1**:1 (1965), 1-7.

[2] ———, The logical basis for information theory and probability theory, IEEE Trans. Information Theory IT-**14** (1968), 662-664. MR **39** # 3900. = Problemy peredachi informatsii **5**:3 (1969), 3-7.

[3] R.J. Solomonoff, A formal theory of inductive inference. I. Information and control **7** (1964), 1-22. MR **30** # 2963.

[4] G.J. Chaitin, On the length of programs for computing finite binary sequences, J. Assoc. Comput. Mach. **13** (1966), 547-569. MR **35** # 1412.

[5] D.W. Loveland, A new interpretation of von Mises' concept of a random sequence, Z. Math. Logik Grundlagen Math. **12** (1966), 279-294. MR **34** # 5124.

[6] A. Church, On the concept of a random sequence, Bull. Amer. Math. Soc. **46** (1940), 130-135. MR **1**-149.
[7] P. Martin-Löf, Algorithms and random sequences, Information and Control **9** (1966), 602-619. MR **36** # 228.
[8] ———, Algorithms and random sequences, University of Erlangen, 1966.
[9] C.P. Schnorr, Eine Bemerkung zum Begriff der zufälligen Folge Z. Wahrsch. und verw. Geb. **14** (1969-1970), 27-35. MR **41** # 9315.
[10] B.A. Trakhtenbrot, *Slozhnost' algoritmov i vychislenii* (The complexity of algorithms and computations), Novosibirsk 1967.
[11] A.N. Kolmogorov and V.A. Uspenskii, On the definition of an algorithm, Uspekhi Mat. Nauk **13**:4 (1958), 3-28. MR **20** # 5735.
[12] Ya.M. Barzdin', Problems of universality in the theory of growing automata, Dokl. Akad. Nauk SSSR **157** (1964), 542-545. MR **29** # 4644.
= Soviet Physics Dokl. **9** (1967), 535-537.
[13] Yu.P. Ofman, A universal automaton, Trudy Moskov. Mat. Obshch. **14** (1965), 186-199. MR **33** # 5408.
= Trans. Moscow Math. Soc. **14** (1965), 200-215.
[14] ———, The modelling of a self-constructing system on a universal automaton, Problemy Peredachi Informatsii, **211** (1966), 68-73.
= Problems Inform. Transmission 2:1 (1966), 53-56.
[15] Ya.M. Barzdin', Complexity of programmes to determine whether natural numbers not greater than n belong to a recursively enumerable set, Dokl. Akad. Nauk SSSR, **182** (1968), 149-152.
= Soviet Math. Dokl. **9** (1968), 1251-1254.
[16] A.K. Zvonkin and L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, Uspekhi Mat. Nauk **25**:6 (1970), 85-127. MR **46** # 7004.
= Russian Mathematical Surveys **25**:6 (1970), 83-124.
[17] A.N. Kolmogorov, On tables of random numbers, Sankhya **25** (1963), 369-376. MR **31** # 274.
= Semiotika i Informatika, VINITI (All-Union Scientific and Technical Information Institute) **19**, no. 18, 3-13.

Translated by S.K. Zaremba Received by the Editors 20 August 1982