REVIEW OF THE INTERNATIONAL
STATISTICAL INSTITUTE
Volume 37 : 3, 1969

ALGORITHMS AND RANDOMNESS[1]

by

P. Martin-Löf
University of Stockholm

1. KOLMOGOROV'S COMPLEXITY MEASURE

Kolmogorov [2] has proposed a definition of randomness for which strong arguments can be given that it is coextensive with our corresponding intuitive concept. It is based on the theory of algorithms developed simultaneously in a variety of equivalent formulations by the logicians in the middle of the thirties. When justifying the definitions to be given we shall lean heavily on the thesis of Church that the precise mathematical notion, that of a partial recursive function, is an adequate formalization of our intuitive concept of an effectively computable function. In the following the terms algorithm and computable function will be used as synonyms for partial recursive function. We shall also have to consider sets which can be generated or, equivalently, effectively enumerated. The corresponding technical notion is that of a recursively enumerable set.

Without restriction of generality all objects which we shall consider will be taken to be binary strings. Other objects can always be given binary codes. In particular, we suppose when convenient that natural numbers are written in the dual system. The length n of a binary string $x = x_1x_2 \ldots x_n$ will be denoted by $l(x)$.

Our aim is to define a measure of the computational complexity of an object y when an object x is already given to us. It will be denoted by $K(y \mid x)$ and is intuitively to be understood as the minimal number of binary units which, for a given x, is required to define y. Equivalently, we might consider $K(y \mid x)$ to be the additional amount of information, measured as usual in bits, contained in y when we have already received the message x. Note that this is a measure of the amount of information contained in an individual object, a notion which is devoid of sense in the usual measure theoretic information theory.

The relation between computational complexity and randomness is simple. Consider a large finite population, for example the set of all binary strings of length 10^4. Four pages of random numbers contain approximately that many digits. Given the population we define the random elements to be those whose complexity is as large as possible. It turns out that almost all elements actually have a complexity which is approximately equal to the maximal value.

Formally we proceed as follows. Let A be an algorithm transforming a pair of binary strings p, x for which it is defined into a binary string $y = A(p, x)$. We shall think of p as a program which, fed into the machine A, causes it to compute y by means of the given data x. The conditional complexity of the object y for given x with respect to A is defined as the shortest length of a program p which computes it,

$$K_A(y \mid x) = \begin{cases} \min\limits_{A(p,x) = y} l(p) \\ +\infty \text{ if there is no } p \text{ such that } A(p, x) = y. \end{cases}$$

This complexity measure depends essentially on the underlying method of programming A. However, Kolmogorov [2] has shown that there exists an algorithm A

[1] Paper read at the European Meeting of Statisticians, London, 1966

which, roughly speaking, is asymptotically at least as efficient as any other competing algorithm B.

A partial recursive function A can be constructed with the property that, for any partial recursive function B, there exists a constant c such that

$$K_A(y \mid x) \leqslant K_B(y \mid x) + c$$

for all x and y.

This is the basic theorem. It is a simple corollary of the fact that there exists a Gödel numbering of the partial recursive functions. Given a computable function f for which

$$\sum_{n=1}^{\infty} 2^{-f(n)} \leqslant 1$$

we can even choose A such that the constant c appearing in the theorem equals $f(b)$ where b denotes the Gödel number of the partial recursive function B. Letting f be such that the sum above converges very slowly, we see that c need only be slightly larger than the number of bits required to define B. For example, one might choose $f(n)$ to be the smallest integer $\geqslant \log n + 2 \log \log n$ for $n \geqslant 3, f(1) = 2$ and $f(2) = 3$. Here and in the following the logarithms are taken to the base 2.

An algorithm whose existence is guaranteed by the theorem will be called asymptotically optimal. It is of course not unique, but given two such algorithms A and B

$$| K_A(y \mid x) - K_B(y \mid x) | \leqslant c$$

for some constant c, so that for large quantities of information the difference becomes negligible. In the following we shall fix once and for all an asymptotically optimal algorithm, speaking simply of the conditional complexity of y given x. Accordingly, we shall drop the index and write

$$K(y \mid x).$$

This is the searched for mathematical definition of the number of bits required to define y when x is already given to us.

Inserting the empty string for x in $K(y \mid x)$ we obtain a quantity to be denoted

$$K(y)$$

which we shall call the complexity of y or the amount of information contained in y.

It is intuitively not astonishing that the measures we have introduced have an asymptotical character. Indeed, it can hardly have any sense to say that 00000 is more or less complex than 01011, whereas a sequence of a thousand zeros should be less complex than a sequence of the same length obtained by coin tossing.

The theorem just stated allows us to estimate the complexity from above. To obtain inequalities in the opposite direction the following lemma is constantly applied.

For a fixed x the number of elements y with

$$K_A(y \mid x) < c$$

is less than 2^c.

$K_A(y \mid x) < c$ if and only if there exists a program p such that $A(p, x) = y$ and $l(p) < c$. It only remains to note that the total number of programs p with $l(p) < c$ equals $2^c - 1$.

Let us now return to the problem of making mathematically precise the notion of

randomness. Consider the population of all binary strings $x_1x_2 \ldots x_n$ of length n. It is completely specified by the number n and we accordingly consider the conditional complexity $K(x_1x_2 \ldots x_n \mid n)$ as a measure of the randomness of the element $x_1x_2 \ldots x_n$ in the specified population. It is a simple consequence of the basic theorem that

$$K(x_1x_2 \ldots x_n \mid n) \leqslant n + c$$

for some constant c. Just compare the asymptotically optimal algorithm with the reproducing algorithm $B(p, x) = p$ for which $K_B(y \mid x) = l(y)$ so that $K_B(x_1x_2 \ldots x_n \mid n) = n$. On the other hand the lemma tells us that the number of sequences of length n for which

$$K(x_1x_2 \ldots x_n \mid n) \geqslant n - c$$

is greater than $(1 - 2^{-c})\, 2^n$. In particular there exists a sequence $x_1x_2 \ldots x_n$ with $K(x_1x_2 \ldots x_n \mid n) \geqslant n$, and for large values of n the overwhelming majority of sequences have a complexity which is approximately equal to n. These are the random elements of the population.

Random sequences cannot be constructed. More precisely, if an algorithm for every n produces a binary string $x_1x_2 \ldots x_n$ of length n, then

$$K(x_1x_2 \ldots x_n \mid n) \leqslant c$$

for some constant c depending on the algorithm. To see this, let B be defined by $B(\ , n) = x_1x_2 \ldots x_n$, the first argument equalling the empty string. Then $K_B(x_1x_2 \ldots x_n \mid n) = 0$ and it only remains to apply the basic theorem.

The non constructability of random sequences is connected with the fact that the complexity measure is not computable. For suppose that $K(y \mid x)$ were a computable function of x and y. Whatever be the natural number n, we could then calculate $K(x_1x_2 \ldots x_n \mid n)$ for all sequences $x_1x_2 \ldots x_n$ of that length, thus finding effectively one for which

$$K(x_1x_2 \ldots x_n \mid n) \geqslant n.$$

This is in contradiction with the preceding paragraph.

Chaitin [1] has, apparently without knowledge of Kolmogorov [2], made the following proposal. Patternless finite binary sequences of a given length are sequences which in order to be computed require programs of approximately the same length as the longest programs required to compute any binary sequences of that given length. His formal development, which is based on a certain type of Turing machines, differs, however, essentially from Kolmogorov's.

2. TESTS FOR RANDOMNESS

What is the relation between the complexity measure and the various statistical tests which have been proposed for tables of random numbers? If the definition given by Kolmogorov is adequate, we ought to be able to prove that the sequences of maximal complexity pass all the familiar randomness tests. We shall see that, as a matter of fact, the random sequences can equivalently be defined to be those sequences which pass a certain universal test. Roughly speaking, this universal test is such that if a sequence passes it, then it passes every conceivable test, neglecting a change in the level of significance. The difference $n - K(x_1x_2 \ldots x_n \mid n)$ may be interpreted as minus the logarithm of the critical level with respect to the universal test.

Since we are always interested merely in the order of magnitude of the level of

significance, we shall restrict our attention to levels $\varepsilon = 2^{-m}$, $m = 1, 2, \ldots$. As an example consider the test which rejects if the number of zeros n_0 differs too much from the number of ones n_1,

$$n_0 + n_1 = n, \qquad n_1 = x_1 + x_2 + \ldots + x_n.$$

It is given by the following prescription.

Reject the hypothesis that $x_1 x_2 \ldots x_n$ is random on the level $\varepsilon = 2^{-m}$ provided

$$| n_1 - n_0 | > f(m, n).$$

Here f is determined by the requirement that the number of sequences of length n for which the inequality holds should be $\leqslant 2^{n-m}$ and that it should not be possible to diminish f without violating this condition.

Generally, a randomness test is given by a prescription which, for every level of significance ε, tells us for what sequences $x_1 x_2 \ldots x_n$ the hypothesis should be rejected. Let U_m denote the critical region on the level $\varepsilon = 2^{-m}$. The conditions to be satisfied by the family of critical regions are as usual, firstly, that

$$U_1 \supseteq U_2 \supseteq \ldots \supseteq U_m \supseteq \ldots$$

and, secondly, that the number of sequences of length n contained in U_m is to be

$$\leqslant 2^{n-m}$$

for all m and n. The whole family of critical regions will be denoted by U so that U_m appears as the section of U at m.

Supporting ourselves on Church's thesis we now formalize the fact that the test is given by an effective prescription by assuming the family of critical regions U to be recursively enumerable. Every test which has been proposed for practical use is of this type and on the basis of Church's thesis it seems safe to say that this is the most general definition we can imagine as long as we confine ourselves to tests which can actually be carried out and are not pure set theoretic abstractions. Anyway, in the following we shall understand by a test U a recursively enumerable set whose sections satisfy the two conditions above.

We are now able to prove a theorem which is closely related to the basic one of the previous section. Their interrelation will soon be established.

A test U can be constructed with the property that, for any test V, there exists a constant c such that

$$V_{m+c} \subseteq U_m, m = 1, 2, \ldots.$$

Just as for Kolmogorov's theorem the proof is achieved by proving that the tests can be Gödel numbered. And, again, given a computable function f with

$$\sum_{n=1}^{\infty} 2^{-f(n)} \leqslant 1$$

we can choose U such that the constant c appearing in the theorem equals $f(v)$ where v denotes the Gödel number of V.

A test with the property stated in the theorem will be called universal. For any test U it is convenient to introduce the critical level

$$m_U(x) = \max_{x \in U_m} m,$$

where we have taken U_0 to be the set of all binary strings. With this convention

$$0 \leqslant m_U(x) \leqslant l(x).$$

Note that we have committed ourselves to a slight abuse of language since $m_U(x)$ is actually minus the logarithm of the critical level of x with respect to U. The defining property of a universal test U is that for any test V there exists a constant c such that

$$m_V(x) \leqslant m_U(x) + c.$$

Henceforth we shall fix a universal test and drop the index, speaking of $m(x)$ simply as the critical level of x. It is a measure of the irregularity of x, a low value of $m(x)$ being tantamount to a high degree of randomness.

The following theorem which, together with the previous one, was proved by Martin-Löf [3] shows the relation between the critical level and Kolmogorov's complexity measure.

There exists a constant c such that

$$| l(x) - K(x \mid l(x)) - m(x) | \leqslant c$$

for all binary strings x.

Let us now return to the specific test considered in the beginning of this section which rejects if the relative frequency differs too much from 1/2. By comparing it with the universal test we obtain the inequality

$$| n_1 - n_0 | \leqslant f(m(x_1 x_2 \ldots x_n) + c, n)$$

which shows how to estimate the difference between the frequencies of zeros and ones by means of the critical level or, equivalently, the complexity. According to the theorem of de Moivre and Laplace the right hand member is of the order of magnitude $\sqrt{n}$ provided $m(x_1 x_2 \ldots x_n)$ is bounded.

In order to show that our investigations are not of a purely theoretical character we have to say something about the size of the constants which appear as soon as a specific test is compared with the universal one by means of the basic theorem. They depend on the skill with which we construct our Gödel numbering. Using Gödel's original technique with prime factor representations and so on, they would apparently be of an astronomical size. Working in Post's canonical systems the situation is more hopeful. To write down in a straightforward manner the test which rejects if the frequencies of zeros and ones differ too much, we need not more than 10^3 bits. This is still quite a lot but note that merely three lines in an ordinary book contain approximately that amount of information. Since Post's canonical systems are not specially adapted for the definition of tests, it is likely that the constants can be further pressed down.

We have seen that the complexity measure is non computable. Likewise, it is important to understand that the universal test which we have constructed is a recursively enumerable set which is not recursive. This means that, given a binary string and a level of significance, we cannot in general decide whether we should accept or reject the hypothesis. We merely know that if it is false on the given level, which means that the sequence is non random, then we shall get to know this sooner or later by systematically going through all tests. In the opposite case the process will extend indefinitely without our reaching any decision.

The definition of a test which we have given by means of recursive function theory and the fact that we have taken into account the Gödel numbers of the tests provides us with an at least conceptually satisfactory solution of a basic problem of statistics. In order that it be permissible to apply a statistical test it is currently required that it

must have been decided upon without knowledge of the outcome of the experiment. This important non mathematical clause is related to and just as obscure as von Mises' "Auswahl ohne Benutzung der Merkmalunterschiede der auszuwählenden Elemente" which forms part of his definition of a "Kollektiv" and for which he was severely criticized. It is introduced to avoid the following paradox. Given any outcome of a statistical experiment there exists a test which rejects on the lowest possible level of significance, namely the test whose critical region consists of that outcome only. From the usual set theoretical point of view there is no essential difference between such a critical region and those which we intuitively accept as reasonable. The principle cited above is an effort to exclude unreasonable tests such as the mentioned one on the basis that they depend on the outcome of the experiment. The worst thing with the clause is not that it is a non mathematical one but that it is unacceptable to a practically working statistician. When testing for example a table of random numbers, we must scrutinize it with all our ingenuity to find out whether it has some very improbable properties.

The construction of the universal test above amounts to the following. Suppose we choose to work on the level of significance $\varepsilon = 2^{-m}$. We should carry out all possible tests, whether or not they have been suggested by the outcome of the experiment. Only, when using a specific test V, we are to work on the smaller level of significance

$$\varepsilon\, 2^{-f(v)} = 2^{-(m+f(v))},$$

where v is the Gödel number of the test V. Here f is a computable function such that

$$\sum_{n=1}^{\infty} 2^{-f(n)} \leqslant 1.$$

If f is suitably chosen, $f(v)$ equals approximately the number of bits required to define V. We see now that we cannot reject a random sequence by means of the unreasonable test defined above, since its Gödel number is too large. Indeed, if the random sequence to be tested is of length n, we need more than n bits to define the test, and the modified level of significance will be $< 2^{-n}$ so that nothing can be rejected.

3. THE ENTROPY IN EHRENFEST'S MODEL OF DIFFUSION

The complexity measure of Kolmogorov provides us with a new definition of the statistical mechanical entropy. To be specific, we shall consider the Ehrenfest model of diffusion. A large number of balls numbered from 1 to n are distributed in two containers. The microstate of the system is given by the sequence $x_1 x_2 \ldots x_n$, where $x_m = 0$ or 1 is the number of the container in which the mth ball is situated. In all there are 2^n different microstates.

We shall consider quantities like

$$n_1 = x_1 + x_2 + \ldots + x_n$$

and refer to the value of $n_1 = 0, 1, \ldots, n$ as the macrostate of the system. Suppose now that we have observed the system to be in a certain macrostate. The Boltzmann entropy is then defined as the logarithm of the number of microstates which realize the given macrostate. Thus, if nothing is known about the system except its size n, the entropy equals

$$\log 2^n = n,$$

whereas for a given value of n_1 we obtain

$$\log \binom{n}{n_1}.$$

Intuitively the entropy is to be a measure of the disorder of the system. The Boltzmann entropy, say in the second version, has the strange property that a configuration like

00000000000000000000000000000001111111111111111111

has the same entropy as a chaotic one

10011000110101111000000001011010000010000100001110

since they correspond to the same macrostate. Think of the number of balls n as being of the order of magnitude 10^{23} instead of 50.

It is proposed to replace the Boltzmann entropy by the conditional complexity of the configuration $x_1x_2 \ldots x_n$ given the macrostate of the system. Thus in the first case we obtain

$$K(x_1x_2 \ldots x_n \mid n)$$

and in the second

$$K(x_1x_2 \ldots x_n \mid n_0, n_1), \qquad n_0 + n_1 = n.$$

The entropy is now a function of the unknown microstate of the system and not merely of the macrostate.

The importance of Boltzmann's expression is that it gives the maximum value of the entropy. Indeed, Kolmogorov's theorem tells us that

$$K(x_1x_2 \ldots x_n \mid n) \leqslant n + c$$

and

$$K(x_1\, x_2 \ldots x_n \mid n_0, n_1) \leqslant \log \binom{n}{n_1} + c,$$

where c is some constant, say of the order of magnitude 10^2 and hence truly negligible in comparison with 10^{23}.

A statistical mechanical system is said to be in equilibrium provided the entropy given the observed macrostate is maximal. Usually the macrostate corresponds to a given value of the energy. In the Ehrenfest model the equilibrium conditions are

$$K(x_1x_2 \ldots x_n \mid n) \approx n$$

and

$$K(x_1\, x_2 \ldots x_n \mid n_0, n_1) \approx \log \binom{n}{n_1}$$

depending on whether we merely know the size n of the system or have in addition observed the quantity n_1.

The fundamental principle of statistical mechanics can now be expressed as follows. Since nothing is known about the mechanical system except its macrostate given by the values of a small number of physical quantities, we so to say express our lack of knowledge by assuming the microstate, which we cannot observe, to be as difficult as possible to define. From this assumption of maximal entropy the theorems of statistical mechanics are deduced. A typical example of how this is done was given in the previous section where we proved that

$$| n_1 - n_0 | \leqslant f(n - K(x_1x_2 \ldots x_n \mid n) + c, n).$$

Here the assumption of maximal entropy

$$K(x_1x_2 \ldots x_n \mid n) \approx n$$

leads to the conclusion that $\mid n_1 - n_0 \mid$ is of the order of magnitude $\sqrt{n}$.

REFERENCES

[1] Chaitin, G. J. (1966). On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery*, 13, 547–569, New York.

[2] Kolmogorov, A. N. (1965). Three approaches to the definition of the concept "quantity of information". *Problemy Peredači Informacii*, 1, 3–11, Moscow.

[3] Martin-Löf, P. (1966). The definition of random sequences. *Information and Control*, 9, 602–619, New York.

RESUME

La mesure de complexité de Kolmogoroff est employée, premièrement, pour préciser la notion d'une séquence fortuite finie et, deuxièmement, pour donner une nouvelle définition de l'entropie d'un système mécanique statistique.