UDC 621. 391. 12

# THREE APPROACHES TO THE QUANTITATIVE DEFINITION OF INFORMATION

A. N. Kolmogorov

Problemy Peredachi Informatsii, Vol. 1, No. 1, pp. 3-11, 1965

There are two common approaches to the quantitative definition of "information": combinatorial and probabilistic. The author briefly describes the major features of these approaches and introduces a new algorithmic approach that uses the theory of recursive functions.

## 1. The Combinatorial Approach

Assume that a variable x is capable of taking values in a finite set X containing N elements. We say that the "entropy" of the variable x is

$$H(x) = \log_2 N.$$

By giving x a definite value

$$x = a$$

we "remove" this entropy and communicate "information"

$$I = \log_2 N.$$

If the variables $x_1, x_2, \ldots, x_k$ are capable of independently taking values in sets respectively containing $N_1, N_2, \ldots, N_k$ members, then

$$H(x_1, x_2, \ldots, x_k) = H(x_1) + H(x_2) + \ldots + H(x_k). \tag{1}$$

Transmission of a quantity of information I requires

$$I' = \begin{cases} I & \text{for integral I} \\ [I] + 1 & \text{for fractional I} \end{cases}$$

binary digits. For example, the number of different "words" consisting of k zeros and ones and one two is

$$2^k (k + 1).$$

Hence, the information content of such a message is

$$I = k + \log_2 (k + 1),$$

i. e., the "coding" of such words in a purely binary system requires*

$$I' \approx k + \log_2 k$$

zeros and ones.

Discussions of information theory do not usually go into this combinatorial approach at any length, but I consider it important to emphasize its logical independence of probabilistic assumptions. Suppose, for example, that we are faced with the problem of coding a message written in an alphabet consisting of s letters, it being known that the frequencies

$$p_r = \frac{s_r}{s} \tag{2}$$

of occurrence of individual letters in a message of length n satisfy the inequality

$$\chi = -\sum_{r=1}^{s} p_r \log_2 p_r \leqslant h. \tag{3}$$

*Here and in what follows $f \approx g$ indicates that the difference $f - g$ is bounded, while $f \sim g$ indicates that the ratio $f : g$ approaches one.

It is easy to see that for large n, the binary logarithm of the number of messages satisfying requirement (2) has the asymptotic estimate

$$H = \log_2 N \sim nh.$$

In transmitting such messages, therefore, it is sufficient to use approximately nh binary digits.

A universal coding method that permits the transmission of any sufficiently long message in an alphabet of s letters with no more than nh binary digits is not necessarily excessively complex; in particular, it is not essential to begin by determining the frequencies $p_r$ for the entire message. In order to make this clear, it is sufficient to note that by splitting the message S into m segments $S_1, S_2, \ldots, S_m$, we obtain the inequality

$$\chi \geqslant \frac{1}{n} [n_1\chi_1 + n_2\chi_2 + \ldots + n_m\chi_m]. \qquad (4)$$

However, I will not go into the details of this special problem here. It is only important for me to show that the mathematical problems associated with a purely combinatorial approach to the measure of information are not limited to trivialities.

It is perfectly natural to take a purely combinatorial approach to the notion of the "entropy of language" if we have in mind an estimate of its "flexibility," an index of the diversity of the possibilities for developing a language with a given dictionary and given rules for the construction of sentences. M. Ratner and N. Svetlova obtained the following estimate for the binary logarithm of the number N of Russian texts of length n, expressed as the "number of symbols including spaces," composed of words in S. I. Ozhegov's Russian dictionary and subject only to the requirement of "grammatical correctness"

$$h = \frac{\log_2 N}{n} = 1.9 \pm 0.1.$$

This is considerably larger than the upper estimate for the "entropy of literary texts" that can be obtained by various methods of "guessing continuations." This discrepancy is quite natural, since literary texts must meet many requirements beyond simple "grammatical correctness."

It is more difficult to estimate the combinatorial entropy of texts subject to definite, more elaborate constraints. It would, for example, be of interest to estimate the entropy of Russian texts that could be regarded as sufficiently accurate (in terms of content) translations of a given foreign-language text. It is only "residual entropy" that makes it possible to translate poetry, where the "entropy cost" of adhering to a given meter and rhyme scheme can be calculated rather accurately. It can be shown that the classical rhyming iambic tetrameter, with certain natural restraints on the frequency of syllables, etc., requires a freedom in handling verbal material characterized by a "residual entropy" of the order of 0. 4 (this estimate is based on the above method of measuring the length of a text in terms of the "number of symbols, including spaces"). On the other hand, if we take into account the fact that the stylistic limitations of a particular genre probably reduce the above estimate of the "total" entropy from 1. 9 to no more than 1. 1-1. 2, the situation becomes remarkable both in the case of translation and in the case of original poetry.

I trust the reader of a utilitarian bent will forgive me this example, but it should be noted that the broader problem of measuring the information connected with creative human endeavor is of the utmost significance.

At this point, let us turn to a discussion of the extent to which a purely combinatorial approach permits one to estimate the information conveyed by a variable x with respect to a related variable y. The relation between the variables x and y, which respectively take values in the sets X and Y, consists in that not all pairs (x, y) belonging to the Cartesian product X × Y are "possible." The set U of possible pairs determines the set $Y_a$ of y such that for a given $a \in X$

$$(a, y) \in U.$$

It is natural to define the conditional entropy by the equation

$$H(y/a) = \log_2 N\ (Y_a) \qquad (5)$$

(where $N(Y_x)$ is the number of members of $Y_x$) and the information conveyed by x with respect to y by the formula

$$I\ (x : y) = H(y) - H(y/x). \qquad (6)$$

For the case shown in the table, for example, we have

| $x \backslash y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | + | + | + | + |
| 2 | + | − | + | − |
| 3 | − | + | − | − |

$$I(x=1:y)=0,$$
$$I(x=2:y)=1,$$
$$I(x=3:y)=2.$$

Clearly, $H(y/x)$ and $I(x:y)$ are functions of x (whereas y takes the form of a "bound variable").

It is not difficult to introduce in a purely combinatorial conception the notion of the "quantity of information necessary to designate an object x with given requirements imposed on the accuracy of the designation." (Apropos of this see the extensive literature on the "ε-entropy" of sets in metric spaces.)

It is obvious that

$$H(x/x)=0, \qquad I(x:x)=H(x). \tag{7}$$

## 2. The Probabilistic Approach

The possible advantages of further developing information theory on the basis of definitions (5) and (6) have been overshadowed by the fact that if we make the variables x and y "random variables" with given joint probability distributions, we can obtain a considerably richer system of concepts and relationships. Paralleling the quantities introduced in §1, here we have

$$H_W(x) = -\sum_x p(x)\log_2 p(x), \tag{8}$$

$$H_W(y/x) = -\sum_y p(y/x)\log_2 p(y/x), \tag{9}$$

$$I_W(x:y) = H_W(y) - H_W(y/x). \tag{10}$$

As before, $H_W(y/x)$ and $I_W(x:y)$ are functions of x, and we have the inequalities

$$H_W(x) \leqslant H(x),\ H_W(y/x) \leqslant H(y/x), \tag{11}$$

where the equality holds when the corresponding distributions (on both X and $Y_X$) are uniform. The quantities $I_W(x:y)$ and $I(x:y)$ are not related by an inequality of a particular direction. As in §1,

$$H_W(x/x)=0,\ I_W(x:x)=H_W(x). \tag{12}$$

The difference lies in the fact that we can form the mathematical expectations

$$MH_W(y/x),\ MI_W(x:y),$$

while the quantity

$$I_W(x,y) = MI_W(x:y) = MI_W(y:x) \tag{13}$$

symmetrically characterizes the "closeness of the relation" between x and y.

However, it should be noted that the probabilistic approach gives rise to a paradox: In the combinatorial approach, $I(x:y)$ is always non-negative, which is natural in a naive conception of information content, but $I_W(x:y)$ may be negative. Now only the averaged quantity $I_W(x, y)$ is a true measure of the information content.

The probabilistic approach is natural in the theory of information transmission over communications channels carrying "bulk" information consisting of a large number of unrelated or weakly related messages obeying definite probabilistic laws. In this type of problem there is a harmless and (in applied work) deep-rooted tendency to mix up probabilities and frequencies within a sufficiently long time sequence (which is rigorously justified if it is assumed that "mixing" is sufficiently rapid). In practice, for example, it can be assumed that the problem of finding the "entropy" of a flow of congratulatory telegrams and the channel "capacity" required for timely and undistorted transmission is validly represented by a probabilistic treatment even with the usual substitution of empirical frequencies for probabilities. If something goes wrong here, the problem lies in the vagueness of our ideas of the relationship between mathematical probability theory and real random events in general.

But what real meaning is there, for example, in asking how much information is contained in "War and Peace"? Is it reasonable to include this novel in the set of "possible novels," or even to postulate some probability distribution for this set? Or, on the other hand, must we assume that the individual scenes in this book form a random sequence with "stochastic relations" that damp out quite rapidly over a distance of several pages?

Actually, we are just as much in the dark over the fashionable question of the "quantity of hereditary information" necessary, say, for the reproduction of particular form of roach. Still, within the limits of the probabilistic approach, two variants are possible. In the first variant, we must consider the set of "possible forms" with a probability distribution of uncertain origin on this set*. In the second variant, the characteristics of the form are assumed to be a set of weakly dependent random variables. The real nature of the mechanism of mutation provides arguments favoring the second variant, but these arguments are undermined if we assume that natural selection causes a system of consistent characteristics to appear.

## 3. An Algorithmic Approach

Actually, it is most fruitful to discuss the quantity of information "conveyed by an object" (x) "about an object" (y). It is not an accident that in the probabilistic approach this has led to a generalization to the case of continuous variables, for which the entropy is infinite but, in a large number of cases,

$$I_W(x, y) = \iint P_{xy}(dx\,dy) \log_2 \frac{P_{xy}(dx\,dy)}{P_x(dx)\,P_y(dy)}$$

is finite. The real objects that we study are very (infinitely) complex, but the relationships between two separate objects diminish as the schemes used to describe them become simpler. While a map yields a considerable amount of information about a region of the earth's surface, the microstructure of the paper and the ink on the paper have no relation to the microstructure of the area shown on the map.

In practice, we are most frequently interested in the quantity of information "conveyed by an individual object x about an individual object y." It is true, as we have already noted, that such an individual quantitative estimate of information is meaningful only when the quantity of information is sufficiently large. It is, for example, meaningless to ask about the quantity of information conveyed by the sequence

$$0\ 1\ 1\ 0$$

about the sequence

$$1\ 1\ 0\ 0.$$

But if we take a perfectly specific table of random numbers of the sort commonly used in statistical practice, and for each of its digits we write the unit's digit of the units of its square according to the scheme

$$\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 0 & 1 & 4 & 9 & 6 & 5 & 6 & 9 & 4 & 1, \end{matrix}$$

the new table will contain approximately

$$\left(\log_2 10 - \frac{8}{10}\right) n$$

bits of information about the initial sequence (where n is the number of digits in the tables).

Accordingly, below we propose to define

$$I_A(x : y)$$

so that some indeterminacy remains. Different equivalent variants of this definition will lead to values equivalent only in the sense that $I_{A_1} \approx I_{A_2}$, i. e.,

$$|I_{A_1} - I_{A_2}| \leqslant C_{A_1A_2},$$

where the constant $C_{A_1A_2}$ depends on the two basic ways of defining the universal methods of programming $A_1$ and $A_2$.

Consider an "indexed domain of objects," i. e., a countable set

$$X = \{x\},$$

binatorial calculation of the number of possible forms extant (or once extant) on the earth would
ner limit (something like < 100 bits).

with a finite sequence n(x) of zeros and ones, beginning with a one, associated with each element as its index. Denote the length of the sequence n(x) by $l(x)$, and assume that:

1) the correspondence between X and the set D of binary sequences of the form described above is one-to-one;

2) $D \subset X$, the function n(x) on D is generally recursive [1], and for $x \in D$

$$l(n(x)) \leqslant l(x) + C,$$

where C is a constant;

3) together with x and y, the set X contains the ordered pair (x, y), whose index is a generally recursive function of the indices of x and y and

$$l(x, y) \leqslant C_x + l(y),$$

where $C_x$ depends only on x.

Not all of these requirements are essential, but they do simplify the discussion. The end result of the construction is invariant under transition to a new indexing n'(x) that has the same properties as the old system, and can be generally recursively expressed in terms of it; moreover, X retains its properties when embedded in a larger system X' (provided that, for the members of the initial system, the index n' in the expanded system can be generally recursively expressed in terms of the initial index n). The new "complexity" K and quantity of information remain equivalent under these transformations in the sense of $\approx$.

As the "relative complexity" of an object y with a given x, we will take the minimal length $l(p)$ of the "program" p for obtaining y from x. The definition thus formulated depends on the "programming method," which is nothing other than the function

$$\varphi(p, x) = y,$$

that associates on object y with a program p and an object x.

In accordance with the views now universally accepted in modern mathematical logic, we must assume that the function $\varphi$ is partially recursive. For any such function we have

$$K_\varphi(y/x) = \begin{cases} \min\limits_{\varphi(p, x)=y} l(p) \\ \infty, \text{ if there is no p such that } \varphi(p, x) = y. \end{cases}$$

In this case a function

$$v = \varphi(u)$$

of $u \in X$ with range $v \in X$ is said to be partially recursive if it generates a partially recursive function of the index transformation

$$n(v) = \Psi[n(u)].$$

In order to understand the definition, it is important to note that, in general, partially recursive functions are not defined everywhere, and there is no fixed method for determining whether application of the program p to an object k will lead to a result or not. As a result, the function $K_\varphi(y/x)$ cannot be effectively calculated (generally recursive) even if it is known to be finite for all x and y.

Fundamental theorem. There exists a partially recursive function A(p, x) such that for any other partially recursive function $\varphi$(p, x) we have the inequality

$$K_A(y/x) \leqslant K_\varphi(y/x) + C_\varphi,$$

where the constant $C_\varphi$ does not depend on x or y.

The proof is based on the existence of a universal partially recursive function

$$\Phi(n, u),$$

which has the property that by fixing an appropriate index n, we can use the formula

$$\varphi(u) = \Phi(n, u)$$

to obtain any other partially recursive function. The function A(p, x) we require is given by the formula*

$$A((n, q), x) = \Phi(n, (q, x)).$$

Indeed, if

$$y = \varphi(p, x) = \Phi(n(p, x)),$$

then

$$A((n, p), x) = y,$$
$$l(n, p) \leqslant l(p) + C_n.$$

We will call functions A(p, x) that satisfy the requirements of the fundamental theorem (and the programming methods defined by them) asymptotically optimal. It is clear that the corresponding "complexity" $K_A(y/x)$ is finite for all x and y. For two such functions $A_1$ and $A_2$

$$|K_{A_1}(y/x) - K_{A_2}(y/x)| \leqslant C_{A_1A_2},$$

where $C_{A_1A_2}$ does not depend on x and y, i. e., $K_{A_1}(y/x) \approx K_{A_2}(y/x)$.

Finally,

$$K_A(y) = K_A(y/1)$$

can be taken for the "complexity of y" and we can define the "quantity of information conveyed by x about y" by the formula

$$I_A(x : y) = K_A(y) - K_A(y/x).$$

It is easy to show** that this quantity is always essentially positive,

$$I_A(x : y) \gtrsim 0,$$

which means that $I_A(x : y)$ is no less than some negative constant C that depends only on the characteristics of the selected programming method. As we have already noted, the theorem was designed for application to a quantity of information so large that, in comparison, |C| is negligibly small.

Note, finally, that $K_A(x/x) \approx 0$, $I_A(x : x) \approx K_A(x)$.

Of course, one can avoid the indeterminacies associated with the constant $C_\varphi$, etc., by considering particular domains of the objects X, indexing, and the function A, but it is doubtful that this can be done without explicit arbitrariness. One must, however, suppose that the different "reasonable" variants presented here will lead to "complexity estimates" that will converge on hundreds of bits instead of tens of thousands. Hence, such quantities as the "complexity" of the text of "War and Peace" can be assumed to be defined with what amounts to uniqueness. Experiments on guessing continuations of literary texts make it possible to obtain an upper estimate for the conditional complexity in the presence of a given consumption of "a priori information" (about language, style, textual content) available to the guesser. In tests conducted at the Moscow State University Department of Probability Theory, such upper estimates fluctuated between 0. 9 and 1. 4. The estimates of the order of 0. 9-1. 1 obtained by N. G. Rychkov have led less successful guessers to suggest that he telepathically communicated with the authors of the texts.

I believe that the approach proposed here yields, in principle, a correct definition of the "quantity of hereditary information," although it would be difficult to obtain a reliable estimate of this quantity.

## 4. Conclusion

The concept discussed in §3 have one important disadvantage: They do not allow for the "difficulty" of preparing a program p for passing from an object x to an object y. By introducing appropriate definitions, it is possible to prove rigorously formulated mathematical propositions that can be legitimately interpreted as an indication of the existence of cases in which an object permitting a very simple program, i. e., with a very small complexity K(x), can be restored by short programs only as the result of computations of a thoroughly unreal duration. Sometime in the future, I intend to study the relationship between the necessary complexity of a program

*$\Phi(n, u)$ is defined only when $n \in D$, and A(p, x) is defined only when p is of the form (n, q), $n \in D$.

**By choosing a "comparison function" of the form $\varphi(p, x) = A(p, 1)$, we obtain $K_A(y/x) \leq K_\varphi(y/x) + C_\varphi = K_A(y) + C_\varphi$.

$$K^t(x)$$

and its permissible difficulty t. The complexity K(x) that was obtained in §3, is, in this case, the minimum of $K^t(x)$ on the removal of constraints on t.

It is beyond the scope of this article to consider the use of the constructions of §3 in providing a new basis for probability theory. Roughly speaking, the situation is as follows: If a finite set M containing a very large number of members N admits determination by means of a program of length negligibly small in comparison with $\log_2 N$, then almost all members of M have complexity K(x) close to $\log_2 N$. The elements $x \in M$ of this complexity are also treated as "random" members of the set M. An incomplete discussion of this idea may be found in [2].

## REFERENCES

1. V. A. Uspenskii, Lectures on Computable Functions [in Russian], Fizmatgiz, Moscow, 1960.

2. A. N. Kolmogorov, "On tables of random numbers," Sankhya. The Indian Journal of Statistics, Series A, 25, 4, 369-376, 1963.

9 January 1965