Logical Basis for Information Theory and Probability Theory

ANDREI N. KOLMOGOROV

Abstract—**A new logical basis for information theory as well as probability theory is proposed, based on computing complexity.**

Section I

WE SHALL be concerned with the main basic concepts of information theory, beginning with the traditional concept of the conditional entropy of x when the value of y is known, $H(x \mid y)$, which can be interpreted as the quantity of information required for computing ("programming") the value x when the value y is already known. By using ϕ to denote a particular given known value, we get the unconditional entropy

$$H(x \mid \phi) = H(x).$$

Information given by y concerning the value of x can, as is well known, be expressed:

$$I(x \mid y) = H(x) - H(x \mid y).$$

It is evident that

$$I(x \mid x) = H(x).$$

The ordinary definition of entropy uses probability concepts, and thus does not pertain to individual values, but to random values, i.e., to probability distributions within a given group of values. In order to stress this difference, we will denote random values by Greek letters. By limiting the case to discrete distributions, let us recall the standard formula:

$$H(\xi \mid \eta) = -\sum_{x,y} p(\xi = x, \eta = y) \cdot \log_2 p(\xi = x \mid \eta = y) \cdots . \quad (1)$$

As we know, "probability" expressions can be realistically interpreted in statistical terms. In other words, our definition (1) can be used practically only in the application to the broad statistical populations of value pairs

$$(x_1, y_1), (x_2, y_2) \cdots , (x_n, y_n) \cdots .$$

By far, not all applications of information theory fit rationally into such an interpretation of its basic concepts. I believe that the need for attaching definite meaning to the expressions $H(x \mid y)$ and $I(x \mid y)$, in the case of individual values x and y that are not viewed as a result of random tests with a definite law of distribution, was realized long ago by many who dealt with information theory.

As far as I know, the first paper published on the idea of revising information theory so as to satisfy the above conditions was the article by Solomonov [1]. I came to similar conclusions, before becoming aware of Solomonov's work, in 1963–1964, and published my first article on the subject [2] in early 1965. A young Swedish mathematician, Martin-Löf, who worked in Moscow during 1964–1965, began developing this concept. His lectures [3] which he gave in Erlangen in 1966 represent a better introduction to the subject of my paper.

The meaning of the new definition is very simple. Entropy $H(x \mid y)$ is the minimal length of the recorded sequence of zeros and ones of a "program" P that permits construction of the value of x, the value of y being known,

$$H(x \mid y) = \min_{A(P,y)=x} l(P). \quad (2)$$

This concept is supported by the general theory of "computable" (partially recursive) functions, i.e., by the theory of algorithms in general. We will return again to the interpretation of the notation $A(P, y) = x$.

Although Martin-Löf and I realized the importance of the new concept, the development was hindered because the simplest formulas that can be produced as a result of simple algebraic transposition of (1) could not be derived from the new definitions. One such formula which *cannot* be derived is

$$I(x \mid y) = I(y \mid x). \quad (3)$$

However, on further examination, its content is not trivial; it does raise doubts as to the unconditional application of this formula as an information "analysis" formula. Nor can we derive

$$H(x, y) = H(x) + H(y \mid x). \quad (4)$$

Formulas (3) and (4) are so familiar that it is not immediately evident that within the meaning of the new concept they are simply inaccurate and can be derived only in the form of approximate equalities

$$|I(x \mid y) - I(y \mid x)| = O(\log I(x, y)) \quad (3')$$

$$H(x, y) = H(x) + H(y \mid x) + O(\log I(x, y)). \quad (4')$$

Manuscript received December 13, 1967. This work is based on an invited lecture given at the International Symposium on Information Theory, San Remo, Italy, September, 1967. Translation courtesy of AFOSR, USAF. Edited by A. V. Balakrishnan.

The author is with the Academy of Sciences, The Institute of Mathematics and Mechanics, Moscow State University, and the Division of Probability Theory and Mathematical Statistics, Steklov Mathematical Institute, all in the U.S.S.R.

Section II

Let us analyze the long sequence

$$x = (x_1, x_2, \cdots x_l); \qquad l = l(x)$$

consisting of zeros and ones. It is easy to understand that there are sequences with entropy $H(x)$ not smaller than their length.

$$H(x) \geq l(x).$$

Such sequences cannot be determined by the use of a program that is shorter than their length. In order to program them we must write them out. They cannot be defined in any less simple manner. It is natural to recall that the absence of periodicity is, according to common sense, a symptom of randomness. We start with the premise that "tables of random numbers" used in mathematical statistics and probability theory are constructed in this manner.

To proceed further, let us examine how we see the sequence of zeros and ones, as a result of independent tests with a probability p of getting a one during each test. If l is large, the number of ones is approximatelv equal to lp, i.e., the frequency

$$k/l \approx p$$

gives us some idea about the periodicity present in the sequence x. On the basis of this recurrence, the unconditional entropy of x can be estimated by the inequality

$$H(x \mid l, k) \leq \log_2 C_l^k + O(1)$$

(Addition of the term $O(1)$ in the formula will be discussed below). If the entropy $H(x \mid l, k)$ is close to this upper limit, the most economical method of programming x is by showing l, k, and the number of sequences x among all C_l^k sequences with given l and k. We view, approximately, in this manner "Bernoulli sequences" where separate signs are "independent" and appear with a certain probability p.

Thus we see that a concept analogous to the Bernoulli sequence can be formulated by using the language of the above mentioned algorithmic information theory.

Martin-Löf's work [4] is devoted especially to these Bernoulli sequences in the new light. Of course, the concept of Bernoulli finite sequences has to be "relative." In finite sequences there is no sharp division between "recurring" and "random." According to Martin-Löf, sequences are of m-Bernoulli type if

$$H[x \mid l(x), k(x)] \geq C_l^k - m.$$

Clear distinction between Bernoulli and non-Bernoulli sequences is possible only after limit transition to infinite sequences of zeros and ones,

$$x = (x_1, x_2, \cdots, x_n \cdots).$$

In such a sequence let us denote by x^l its initial segment of length l,

$$x^l = (x_1, \cdots, x_l).$$

It is tempting to define x as Bernoulli of type m if all x^l are m-Bernoulli type, i.e., if always

$$H(x^l \mid l, k_l) \geq C_l^{k_l} - m.$$

But such infinite sequences of zeros and ones, as shown by Martin-Löf, do not exist. The reason is easy to explain. According to traditional probability theory, we know that in true random sequences there are continuous zero sequences and continuous one sequences. It is clear that the description of such segments of infinite sequences can be substantially simplified in comparision with the standard description.

The most natural definition of infinite Bernoulli sequences is the following: x is considered m-Bernoulli type if m is such that all x^l are *initial segments* of the finite m-Bernoulli sequences. Martin-Löf gives another, possibly narrower definition.

Another concept, the one-half-Bernoulli sequence, which Martin-Löf used initially, was developed independently by Chaitin [5].

It is possible that some readers have already noticed that by referring to infinite sequences we are dealing with a task already set by Mises in his concept of "collectives." As is known, Mises' concept was formulated by Church [6] using the approach of computable-function theory. Substantial addition to Mises' concept (the development of the "permissible choice system" concept) was given in my work [7]. Strictly formal presentation of this expanded theory can be found in Loveland's article [9].

However, the Bernoulli sequences class, according to Church and Loveland, is too large. Their segments can be relatively "recurring." There are Bernoulli sequences that fit both concepts, with segments having only a logarithmically increasing entropy,

$$H(x^l) = O(\log l).$$

On the other hand, according to Martin-Löf (and according to previous definition) Bernoulli sequences have segments the complexity of which is almost maximal within the meaning of inequality,

$$H(x^l) \geq l - O((\log l)^{1+\epsilon}), \qquad \epsilon > 0 \quad \text{arbitrary}.$$

Bernoulli sequences, according to Martin-Löf, possess all constructive qualities which, according to the modern probability theory, are proved (for any probability p) "with probability equal to 1." Nothing analogous can be stated about Bernoulli sequences according to Church or Loveland.

Conclusions

The preceding rather superficial discourse should prove two general theses.

1) Basic information theory concepts must and can be founded without recourse to the probability theory, and in such a manner that "entropy" and "mutual information" concepts are applicable to individual values.

2) Thus introduced, information theory concepts can form the basis of the term *random*, which naturally suggests that random is the absence of periodicity.

Presentation of the first part of the paper (Section I) was somewhat simplified. Only in Section II is the unavoidable relativity in differentiation between random and nonrandom in the application to finite objects emphasized. An analogous situation exists in the principles of information theory. Essentially, it is applicable to large quantities of information, when the initial information (contained in the method on which the theory is based) is infinitesimal. Our basic formula (1) implies a "universal programming method" A, which exists because there are programming methods A possessing the quality

$$H_A(x \mid y) \leq H_{A'}(x \mid y) + C_{A'}.$$

They allow the programming of anything with a program length that exceeds the length of any other programming method by not greater than a constant and is dependent only on this second programming method and not on values of x and y. Credit for noting this relatively simple condition evidently belongs to Solomonov and me.

Therefore, *all* algorithmic information theory assumptions in their general formulation only are valid with "reliability" to members of $O(1)$ type. In the application to formulas (3) and (4) the appearance of logarithmic-order terms was unexpected.

It is important to understand that by using probability theory, we resort to considerably rougher generalization. A realistic interpretation of probability results is always statistical, and error estimates (occurring in the application of probability results to finite objects) are considerably rougher than in the information theory exposition being developed by us.

Bibliography

[1] R. J. Solomonov, *Information and Control*, vol. 7, pp. 1–22, 1964.
[2] A. Kolmogorov, "Three approaches for defining the concept of information quantity," *Information Transmission*, vol. 1, pp. 3–11, 1965.
[3] P. Martin-Löf, "Algorithms and random sequences," University of Erlangen, Germany, 1966.
[4] ——, "The definition of random sequences," *Information and Control*, vol. 9, pp. 602–619, 1966.
[5] S. J. Chaitin, "On the length of programs for computing finite binary sequences," *J. ACM*, vol. 13, no. 4, 1966.
[6] A. Church, "On the concept of a random sequence," *Bull. Am. Math. Soc.*, vol. 46, pp. 254–260, 1940.
[7] A. Kolmogorov, "On tables of random numbers," *Sankhya* vol. 25, pp. 369–376, 1963.
[8] D. Loveland, "A new interpretation of the von Mises concept of random sequence," *Z. Math. Logik und Grundlagen der Math.*, vol. 12, pp. 279–294, 1966.
[9] ——, "Hierarchy classification of recursively random sequences," *Trans. Am. Math. Soc.*, vol. 125, pp. 497–510, 1966.

Nonlinear Prediction of a Class of Random Processes

A. H. HADDAD, MEMBER, IEEE

***Abstract*—This paper is concerned with the minimum mean-squared error (MMSE) nonlinear prediction of a class of random processes. A class of random processes is defined by the property that its MMSE zero-memory predictor is represented by a finite sum of separable terms. Sufficient conditions for the existence of such processes are also considered. The nonlinear predictor is restricted to be composed of a linear filter in parallel with a zero-memory nonlinearity (ZNL) preceded by a variable delay. The optimum predictor is shown to be the solution of linear integral equations with the same kernel as for the optimum linear predictor. The first step of the derivation also yields a simpler scheme which only requires the addition of a ZNL to the optimum linear predictor. The improvements in the MMSE of the two nonlinear systems over the linear case are compared and illustrated by a numerical example.**

Manuscript received July 31, 1967; revised March 4, 1968. This work was supported by the Joint Services Electronics Program under Army Contract DAAB-07-67-C-0199.

The author is with the Coordinated Science Laboratory and the Department of Electrical Engineering, University of Illinois, Urbana, Ill. 61801

Introduction

THE PROBLEM of nonlinear filtering, prediction, and interpolation of random signals is of significant importance. It is known that the use of nonlinear filters for non-Gaussian processes results in an improved performance. However, the derivation and implementation of optimum nonlinear filters generally involve mathematical and practical difficulties. Several aspects of nonlinear filtering and prediction have been discussed in the literature [1]–[4]. One approach to the filtering problem considers classes of random processes from the point of view of filtering. Such classes may be defined and discussed from two different aspects. The first aspect is concerned with the derivation of classes of processes for which the optimum filters or predictors have specified forms. An example of such a class is considered by Wolff *et al.* [5]. The second aspect is concerned with classes of processes for which nonlinear filtering problems are simplified. The