ON ALPHABETIC CODING

1961

Abstract

Full Text

CYBERNETICS AND CONTROL THEORY

A. A. MARKOV

ON ALPHABETIC CODING

(Presented by Academician A. I. Berg, 11 VIII 1960)

1. In paper ([1]) a method was described which makes it possible effectively to solve the question of mutual uniqueness in the general case of alphabetic coding without memory. Here results will be set forth concerning certain properties of one-to-one alphabetic coding, solving most of the problems formulated for the case of binary coding of variable length at the end of the paper by Gilbert and Moore ([2]) in the general case.

2. Let $\mathfrak{U} : \{u_1, u_2, \dots, u_m\}$ be a coding system of words ([1]) in the alphabet

$$\mathfrak{B} = \{b_1, \dots, b_r\}, \quad \sum_{i=1}^{m} l(u_i) = n \quad \text{and} \quad \mathcal{L} = \max_{1 \le i \le m} \{l(u_i)\} - 1^*.$$

A word in the alphabet $\mathfrak{B}$ is called a complete message if it is some sequence of words from $\mathfrak{U}$. To each word u_i from $\mathfrak{U}$ there is put into correspondence a letter a_i of the alphabet $\mathfrak{A} = \{a_1, a_2, \dots, a_m\}$, and by decoding is meant the restoration of the preimage of a given message, which is the image of some word in $\mathfrak{A}$ under coding by the system $\mathfrak{U}$. Considering the coding process in time, we shall assume that the decoding machine receives the words successively, letter by letter. We shall now be interested in the question: can the decoding device have finite memory? This memory is characterized by the number $T_{(\mathfrak{U})}$—the decoding delay—defined by the fact that for any t, knowing the first $t + T_{(\mathfrak{U})}$ letters of a message, we can with certainty decipher the first t letters. If, for the given coding system of words $\mathfrak{U}$, there exists a finite $T_{(\mathfrak{U})}$, then the coding by this system is said to have the property of finite delay ([2], [4]).

3. The finite-state source ([3]), constructed in the proof of Theorem 1 in ([1]), in the case of one-to-one coding by this theorem contains no closed circuits passing through vertex 1. We shall now be interested only in a part of this source, namely that based on a graph which is a subgraph of the graph underlying the original source, containing vertex 1 and all vertices to which from 1 there is at least one path. The connections and the words assigned to the edges of the graph remain the same as in the original source, except that unnecessary vertices are removed (those to which there is no path from vertex 1), together with the connections belonging to them. In addition, each terminal vertex i, i.e. a vertex from which no edge issues, is

connected with some additional vertex μ by an edge to which is assigned the word in $\mathfrak{B}$ from $\mathfrak{M}$ corresponding to vertex i.

We shall denote the source obtained by $G(\mathfrak{U})^{**}$, and the graph underlying it by $G(\mathfrak{U})$.

Theorem 1. *In order that the coding by the system $\mathfrak{U}$ have the property of finite delay, it is necessary and sufficient that the graph $G(\mathfrak{U})$ contain no closed circuits.*

---

* Everywhere in this article, without special reservation, we shall consider the coding to be one-to-one.

** The construction is indeed a finite-state source, since the vertices 1 and μ, as initial and terminal, may be identified. The finite-state language generated by $G(\mathfrak{U})$ will be denoted by $L\{G(\mathfrak{U})\}$.

If the coding by the system $\mathfrak{U}$ has the property of finite delay, then

$$T_{(\mathfrak{U})} = \max_{a \in L\{\bar{G}(\mathfrak{U})\}} \{l(a)\}.$$

In view of the fact that the number of vertices of the graph $\bar{G}(\mathfrak{U})$ does not exceed $n - m + 1$, and the lengths of the words assigned to the edges do not exceed $\mathcal{L}$, we obtain:

Theorem 2. If the coding by the system $\mathfrak{U}$ has the property of finite delay, then

$$T_{(\mathfrak{U})} \leq \mathcal{L}(n - m + 1).$$

Let Ω be the class of messages whose beginning cannot be deciphered before the entire message as a whole has been received.

Theorem 3. $\Omega = L\{\bar{G}(\mathfrak{U})\}$.

4. Let the system $\mathfrak{U}$ carry out a coding that does not have the property of finite delay. In paper ([2]) the question was posed: is the set of infinite messages that do not have the property of finite delay always finite? The question was posed imprecisely: it is clear that if there exists at least one message with the property of infinite delay, then there will already be an infinite (countable) number of such messages. The question, however, may be posed as the existence or nonexistence of some finite basis, to which we assign infinite messages such that for no t does there exist a finite decoding delay $T_{(\mathfrak{U})}$, and we denote the class of such messages by Ω^{∞}. We now define the class $L^{\infty}\{G(\mathfrak{U})\}$ of infinite words in the alphabet $\mathfrak{B}$, compatible with the source $G(\mathfrak{U})$, as the set of sequences $b_{i_1} b_{i_2} \ldots b_{i_l} \ldots$ such that for any $k < \infty$, $b_{i_1} \ldots b_{i_k}$ is the beginning of some word in $L\{G(\mathfrak{U})\}$.

Theorem 4. $\Omega^{\infty} = L^{\infty}\{G(\mathfrak{U})\}$.

Theorem 5. In order that a system of words $\mathfrak{U}$ admit no more than a finite number of basis messages with the property of infinite delay, it is necessary and sufficient that the graph $\bar{G}(\mathfrak{U})$ contain no connected pairs of circuits generating different words in $\mathfrak{B}$ when passing from one to the other.

In the case where the condition of Theorem 5 is not fulfilled, the set of messages in $\mathfrak{B}$ having the property of infinite delay is uncountable if and only if there is a two-way connection between the circuits.

5. A coding is called exhaustive if every infinite message is the code of some (infinite) word in $\mathfrak{A}$. In paper ([2]) it is shown that every exhaustive binary coding has the prefix property (i.e., none of the words of the system is the beginning of another) and satisfies the condition

$$\sum_{i=1}^{m} 2^{-l(u_i)} = 1.$$

Theorem 6. There exist codes that do not have the prefix property and are not inverses of such codes, satisfying the condition

$$\sum_{i=1}^{m} r^{-l(u_i)} = 1,$$

where r is the base of the alphabet $\mathfrak{B}$.

Such, for example, is the binary code C: $u_1 = 1$; $u_2 = 01$; $u_3 = 100$; $u_4 = 0100$; $u_5 = 0000$. We have: C^*: $u_1^* = 1$; $u_2^* = 10$; $u_3^* = 001$; $u_4^* = 0010$; $u_5^* = 0000$ (the code obtained by reversing C). It is easy to see that neither C nor C^* has the prefix property and that they satisfy the required condition. C and C^* do not have the property of finite delay. The question of whether there exist codes with the conditions of Theorem 6 such that C and C^* have the property of finite delay remains open.

Research Physico-Technical Institute
at Gorky State University
named after N. I. Lobachevsky

Received
8 VIII 1960

CITED LITERATURE

1. Al. A. Markov, DAN, 132, No. 3 (1960).

2. E. N. Gilbert, E. F. Moore, Bell Syst. Techn. J., 38, No. 4, 933 (1959).

3. N. Chomsky, G. A. Miller, Inf. and Control, 1, No. 2, 91 (1958).

4. M. P. Schützenberger, Trans. IRE, IT-2, No. 3, 47 (1956).

Note: Figure translations are in progress. See original paper for figures.

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.