

A THEOREM ON THE CODING OF STATEMENTS IN ARBITRARY LANGUAGES

V. S. Proskurov

1970

SovietRxiv

View the original and related papers at <https://sovietrxiv.org/items/ru-197001.49631>

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

Abstract

Full Text

UDC 519.54

MATHEMATICS

V. S. PROSKUROV

A THEOREM ON THE CODING OF STATEMENTS IN ARBITRARY LANGUAGES

(Presented by Academician I. M. Vinogradov on 26 III 1970)

The problem of coding ⁽¹⁾ meaningful or formal texts is one of the problems of information processing. The principal difficulty is finding an effective one-to-one coding. In ⁽²⁾, various rules are considered for coding the words of a certain dictionary consisting of N words; the chief one among them is the operation of contraction ∇_k of the codes of words containing l_i letters to codes containing k letters ($k \leq l_i$, $1 \leq i \leq N$). In Theorem 2 of ⁽²⁾ it is proved that contraction ∇_k is a non-unique coding with a probability of violating uniqueness tending to zero as $k \rightarrow l = \max_{1 \leq i \leq N} \{l_i\}$. Moreover, the coding becomes unique only for $k = l = \max_{1 \leq i \leq N} \{l_i\}$. The contraction ∇_k does not allow one, from a code of length k , to determine the original word. In extending Theorem 2 to statements, the quantity k will increase correspondingly.

Below we prove a theorem stating that there exists a one-to-one coding that allows any statement to be represented by a code of any prescribed length in advance (for example, $k = 1$), from which the original statement can be recovered.

We introduce the necessary definitions and notation.

An alphabet $\mathfrak{A}$ is a strictly ordered sequence of n pairwise distinct symbols with ordinal numbers $0, 1, 2, \dots, n-1$:

$$\mathfrak{A} = a_0 a_1 a_2 \dots a_{n-1} = (a_i, i = 0, 1, 2, \dots, n-1). \quad (1)$$

We shall call the symbol a_0 a space.

Any collection of letters from the alphabet $\mathfrak{A}$ that contains no spaces is called a word in the alphabet $\mathfrak{A}$:

$$\sigma_l(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_l) = a_{i_1} a_{i_2} \dots a_{i_k} \dots a_{i_l}. \quad (2)$$

In what follows we shall always mean that $1 \leq i_k \leq n-1$ for all $k = 1, 2, \dots, l$, with $l \geq 1$.

The number of letters entering into a given word will be called the length of the word in the alphabet $\mathfrak{A}$ and denoted by l .

By definition, for $l = 0$ we shall set $k = 0$, $i_k = 0$, $\sigma_0(\mathfrak{A}, 0) = a_0$, the empty word in the alphabet $\mathfrak{A}$.

A dictionary $\Sigma_N(\mathfrak{A})$ in the alphabet $\mathfrak{A}$ will mean a subset $\{\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r})\}$ of the set $\Sigma(\mathfrak{A})$ of all words in the alphabet $\mathfrak{A}$, consisting of N words, where $1 \leq r \leq N$:

$$\Sigma_N(\mathfrak{A}) = \{\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r})\} \subset \Sigma(\mathfrak{A}), \quad (3)$$

where $r = 1, 2, \dots, N$.

Similarly, we introduce:

$$\mathfrak{B} = b_0 b_1 b_2 \dots b_{m-1} = (b_i, i = 0, 1, 2, \dots, m-1), \quad (4)$$

b_0 is a space in the alphabet $\mathfrak{B}$,

$$\sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l) = b_{i'_1} b_{i'_2} \dots b_{i'_k} \dots b_{i'_l}. \quad (5)$$

word in the alphabet $\mathfrak{B}$, $1 \leq i \leq m-1$, for all $k = 1, 2, \dots, l$ with $l \geq 1$, where l is the length of the word in the alphabet $\mathfrak{B}$.

By definition, for $l = 0$ we shall take $k = 0$, $i_k = 0$, $\sigma_0(\mathfrak{B}, 0) = b_0$ —the empty word in the alphabet $\mathfrak{B}$.

$$\Sigma_N(\mathfrak{B}) = \{\sigma_{l_r}(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_{l_r})\} \subset \Sigma(\mathfrak{B}), \quad (6)$$

where $r = 1, 2, \dots, N$; $\Sigma_N(\mathfrak{B})$ is a dictionary in the alphabet $\mathfrak{B}$, containing N words; $\Sigma(\mathfrak{B})$ is the set of all possible words in the alphabet $\mathfrak{B}$.

We shall call $K = (K'(\mathfrak{A}, \mathfrak{B}), K''(\mathfrak{B}, \mathfrak{A}))$ a one-to-one coding of words (or simply a coding) in the alphabet $\mathfrak{A}$ by words in the alphabet $\mathfrak{B}$ and conversely, if

$$K'(\mathfrak{A}, \mathfrak{B})(\sigma_{l_1}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_1})) = \sigma_{l_2}(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_{l_2}), \quad (7)$$

$$K''(\mathfrak{B}, \mathfrak{A})(\sigma_{l_2}(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_{l_2})) = \sigma_{l_1}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_1}), \quad (8)$$

where $K'(\mathfrak{A}, \mathfrak{B})$ is a single-valued transformation (coding) of a word in the alphabet $\mathfrak{A}$ into a word in the alphabet $\mathfrak{B}$ (the case $\mathfrak{A} = \mathfrak{B}$ is allowed) for any word from $\Sigma(\mathfrak{A})$; $K''(\mathfrak{B}, \mathfrak{A})$ is a single-valued transformation (coding) of a word in the alphabet $\mathfrak{B}$ into a word in the alphabet $\mathfrak{A}$ (the case $\mathfrak{B} = \mathfrak{A}$ is allowed) for any word from $\Sigma(\mathfrak{B})$.

According to the definition, the coding K establishes a one-to-one correspondence between the words (2) from $\Sigma(\mathfrak{A})$ and (6) from $\Sigma(\mathfrak{B})$

$$\sigma_l(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_l) \xleftrightarrow{K} \sigma_{l'}(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_{l'}). \quad (9)$$

Consequently, knowing $\Sigma_N(\mathfrak{A})$, with the aid of the coding $K'(\mathfrak{A}, \mathfrak{B})$ one can obtain $\Sigma_N(\mathfrak{B})$, and conversely—knowing $\Sigma_N(\mathfrak{B})$, with the aid of the coding $K''(\mathfrak{B}, \mathfrak{A})$ one can obtain $\Sigma_N(\mathfrak{A})$.

Theorem 1. *For any dictionary $\Sigma_N(\mathfrak{A}) \subset \Sigma(\mathfrak{A})$ there exist a one-to-one coding K and an alphabet $\mathfrak{B}$ such that, for each*

$$\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}) \in \Sigma_N(\mathfrak{A}), \quad 1 \leq r \leq N,$$

the following relations hold:

$$K'(\mathfrak{A}, \mathfrak{B})(\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r})) = \sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l),$$

$$K''(\mathfrak{B}, \mathfrak{A})(\sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l)) = \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}),$$

where l may be any preassigned positive integer satisfying $l > l_r$ or $l \leq l_r, l_{r-1}, l_{r-2}, \dots, 1$.

Proof. Let the alphabet (1) be given. Take a positional numeral system with base n and digits

$$0, 1, 2, \dots, n-1 = (k, \quad k = 0, 1, 2, \dots, n-1). \quad (10)$$

Establish a one-to-one correspondence between the symbols of the alphabet $\mathfrak{A}$ and the digits of the numeral system (10) so that the symbol $a_k \in \mathfrak{A}$, occupying the k -th place in the sequence (1), is assigned the digit k from the sequence (10), occupying the k -th place, and conversely. Denote the rule establishing this correspondence by F^n :

$$a_k \xleftrightarrow{F^n} k, \quad 0 \leq k \leq n-1. \quad (11)$$

Take the word $\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}) \in \Sigma_N(\mathfrak{A})$. By virtue of the rule F^n (11), to the word $\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r})$ there corresponds one and only one positive integer

$$i_1 i_2 \dots i_k \dots i_{l_r}(n) = i_1 n^{l_r-1} + i_2 n^{l_r-2} + \dots + i_{l_r} = \sum_{k=1}^{l_r} i_k n^{l_r-k}$$

in the positional numeral system with base n , and conversely:

$$F^n(\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r})) = i_1 i_2 \dots i_k \dots i_{l_r}(n), \quad (12)$$

$$F^n(i_1 i_2 \dots i_k \dots i_{l_r}(n)) = \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}). \quad (13)$$

The number of digits in the notation of a number will be called the length of the number.

The following assertion holds:

Lemma. Let $l' = \max_{r \leq N} \{l_r\}$ and let l be positive integers. For any positive integer not exceeding $i_1, i_2, \dots, i_k, \dots, i_{l'}(n)$ in the positional numeral system with base n , there exists a number equal to it, of length not exceeding l , in the positional numeral system with base

$$m = [n^{l'/l}]',$$

where

$$[n^{l'/l}]' = \begin{cases} n^{l'/l}, & \text{if } n^{l'/l} \text{ is an integer,} \\ [n^{l'/l}] + 1, & \text{otherwise,} \end{cases}$$

$[n^{l'/l}]$ is the integer part of the number $n^{l'/l}$.

From the validity of the lemma it follows that any number of length not exceeding l' in the positional numeral system with base n can be represented by an equal number of length not exceeding l in the positional numeral system with base m .

Let R_m^n be the rule for translating numbers from the positional numeral system with base n into numbers of the positional numeral system with base m , and let R_n^m be the inverse rule ⁽³⁾. Then

$$R_m^n(i_1 i_2 \dots i_k \dots i_{l_r}(n)) = i'_1 i'_2 \dots i'_k \dots i'_l(m), \quad (14)$$

$$R_n^m(i'_1 i'_2 \dots i'_k \dots i'_l(m)) = i_1 i_2 \dots i_k \dots i_{l_r}(n), \quad (15)$$

where $i'_1, i'_2, \dots, i'_k, \dots, i'_l$ are the digits of the positional numeral system with base m , and the digits are

$$0, 1, 2, \dots, m-1 = (k, \ k = 0, 1, 2, \dots, m-1). \quad (16)$$

Let F^m be a rule establishing a one-to-one correspondence between the symbols in the alphabet (4) and the digits of the positional numeral system with base m (16), analogous to rule (11):

$$b_k \xleftrightarrow{F^m} k, \quad 0 \leq k \leq m-1. \quad (17)$$

Consequently,

$$F^m(i'_1 i'_2 \dots i'_k \dots i'_l(m)) = \sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l), \quad (18)$$

$$F^m(\sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l)) = i'_1 i'_2 \dots i'_k \dots i'_l(m). \quad (19)$$

Thus, by virtue of (12), (14), (18), we have

$$\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}) \xleftrightarrow{F^n} i_1 i_2 \dots i_k \dots i_{l_r}(n) \stackrel{R_n^n}{=}$$

$$\stackrel{R_m^n}{=} i'_1 i'_2 \dots i'_k \dots i'_l(m) \xleftrightarrow{F^m} \sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l),$$

where

$$m = [n^{l'/l}]', \quad l = \max_{1 \leq r \leq N} \{l_r\}.$$

If, as $K'(\mathfrak{A}, \mathfrak{B})$, we use successively F^n (11), R_m^n , F^m (17), then

$$K'(\mathfrak{A}, \mathfrak{B})(\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r})) = \sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l).$$

Similarly, by virtue of (13), (15), (19), we have

$$\sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l) \xleftrightarrow{R_m^n} i_1 i_2 \dots i_k \dots i_l(m) \stackrel{R_n^n}{=} i_1 i_2 \dots i_k \dots i_{l_r}(n) \xleftrightarrow{F^n}$$

$$\xleftrightarrow{F^n} \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}),$$

where

$$m = [n''/l'], \quad l = \max_{1 \leq r \leq N} \{l_r\}.$$

If, as $K''(\mathfrak{B}, \mathfrak{A})$, one uses successively F^m (17), R_n^m , F^n (11), then

$$K''(\mathfrak{B}, \mathfrak{A})(\sigma_l(\mathfrak{B}, i'_1, i'_2, \dots, i'_k, \dots, i'_l)) = \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_r}).$$

Thus the theorem is completely proved.

We shall call a statement in the alphabet $\mathfrak{A}$ a finite collection of words (2) in the alphabet $\mathfrak{A}$ (1), joined into a single word by means of the blank spaces a_0 :

$$\begin{aligned} &\sigma_{l_1}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_1}) a_0 \sigma_{l_2}(\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_2}) a_0 \dots a_0 \sigma_{l_s} \times \\ &\times (\mathfrak{A}, i_1, i_2, \dots, i_k, \dots, i_{l_s}). \end{aligned}$$

The length of a statement will mean the number of symbols in it, including the spaces between words,

$$l = \sum_{k=1}^s (l_k + 1) - 1.$$

The set of all possible statements from the words of the dictionary $\Sigma_N(\mathfrak{A})$ (3) will be denoted by $D(\Sigma_N(\mathfrak{A}))$.

Analogously we introduce a statement in the alphabet $\mathfrak{B}$ (4), and $D(\Sigma_N(\mathfrak{B}))$ —the set of all possible statements from the words of the dictionary $\Sigma_N(\mathfrak{B})$ (6).

Obviously, the rules F^n (11), R_n^m , F^m (17), R_n^m can also be extended to statements. Then the theorem proved will also be valid for arbitrary statements from $D(\Sigma_N(\mathfrak{A}))$.

Department for the Introduction of Economic-Mathematical Methods
in the Planning of the National Economy
Gosplan of the USSR
Moscow

Received
5 III 1970

REFERENCES

1. A. I. Mal' tsev, *Algorithms and Recursive Functions*, Nauka, 1965.
2. L. N. Korolev, DAN, 113, No. 4 (1957).
3. A. I. Kitov, N. A. Krinitskii, *Electronic Digital Machines and Programming*, Moscow, 1961.

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

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