Reports of the Academy of Sciences of the USSR

1960

Abstract

Full Text

Reports of the Academy of Sciences of the USSR
1960. Volume 131, No. 5

MATHEMATICS

V. I. LEVENSHTEIN

ON ONE CLASS OF SYSTEMATIC CODES

(Presented by Academician M. V. Keldysh, 9 XII 1959)

In the present note a class of systematic codes with detection and correction of errors is considered, obtained by using one of the algorithms for constructing codes proposed by V. I. Siforov ([2]). The power of the codes of the class under consideration lies within the same bounds that are presently known for the maximal power of codes. Some properties of these codes are investigated, and a method is indicated by means of which one can reduce the search in their practical construction.

1. Consider the sets B_n $(n = 0, 1, ...)$ of all n-digit binary sequences* and their union, the set B. To denote the operation of digitwise addition of arbitrary sequences $a = (\alpha_1, \alpha_2, ...)$ and $b = (\beta_1, \beta_2, ...)$ from B, as well as of binary symbols modulo 2, we shall use the sign $\oplus$ $(\sum)$, and we shall omit the multiplication sign:

$$a \oplus b = (\alpha_1 \oplus \beta_1, \alpha_2 \oplus \beta_2, ...), \qquad \alpha a = (\alpha\alpha_1, \alpha\alpha_2, ...).$$

The **value** $|a|$ of the sequence $a = (\alpha_1, \alpha_2, ...)$ will mean the integer whose binary representation is the sequence a, i.e.

$$|a| = \sum_{i=1} \alpha_i 2^{i-1}.$$

The **norm** $\|a\|$ of the sequence a is defined as the number of unit digits of this sequence:

$$\|a\| = \sum_{i=1} \alpha_i.$$

With this definition of the norm, the **Hamming distance** ([1]) between any two sequences a and b from B, i.e. the number of noncoinciding digits of these sequences, can be expressed as follows:

$$\rho(a, b) = \|a \oplus b\|.$$

For arbitrary sequences a, b, and c of the set B the following hold:

Lemma 1. $\rho(a, b) = \rho(a \oplus c, b \oplus c)$.

Lemma 2. The inequalities $|a| < |b \oplus c|$, $|b| < |a \oplus c|$, and $|c| < |a \oplus b|$ are incompatible.

A subset of sequences from B such that the distance between any two of them is not less than a certain number d will be called a d-code; a d-code will be called **systematic** if it forms a group with respect to the operation of digitwise addition of sequences modulo 2. We shall also agree to call a d-code all of whose sequences belong to B_n **complete** in B_n if no sequence from B_n can be adjoined to it so that it remains a d-code, and **maximal** in B_n if there does not exist in B_n a d-code of greater power.

2. Arrange all sequences of the set B_n in a definite order and consider the following algorithm for constructing a set S, possessing—

* In the note it is assumed that appending and deleting zeros on the right does not change sequences, so that they may be regarded as infinite. In particular, one may consider the set B_0 to consist of the single sequence $(0, 0, ...)$, and the set

$$B = \bigcup_{n=0} B_n$$

of all infinite binary sequences with a finite number of unit digits.

···possessing some property Γ. As the first word a_0 of the set S, one takes the first, in order, word of the set B_n possessing property Γ. If the words $a_0, a_1, ..., a_{i-1}$ have already been chosen, then as the word a_i one chooses the first, in order, word of the set B_n (if there still is one), different from those already chosen and such that the words $a_0, a_1, ..., a_{i-1}, a_i$ possess property Γ. The algorithm described above for constructing the set S will be called **trivial**. In particular, the trivial algorithm for constructing a d-code means that as the first word a_0 of the d-code one chooses the first, in order, word of the set B_n; if the words $a_0, a_1, ..., a_{i-1}$ have already been chosen, then as the word a_i one chooses the first, in order, word of the set B_n that is at distance at least d from each of the words $a_0, a_1, ..., a_{i-1}$, if such a word still exists. It is easy to observe that a d-code obtained by the trivial algorithm is maximal in B_n. As V. I. Siforov showed ([2]), the cardinality of such a d-code, generally speaking, depends on the order in which the words of the set B_n are arranged.

The order in the set B_n (or B) according to which the words are arranged in increasing order of their values will be called **natural**. The trivial construction algorithm in the case when the words of the set B are arranged in the natural order will also be called the **natural** algorithm. Denote by $S_{n,d}$ the code obtained from B_n by the natural algorithm for constructing a d-code, and let $S_d = \bigcup_{n=0} S_{n,d}$.

The main result is the following theorem.

Theorem 1. For any n and d, the codes S_d and $S_{n,d}$ are systematic.

Proof. To prove the theorem it is enough to show that, for the words a_i $(i = 0, 1, ...)$ obtained successively by the natural algorithm for constructing a d-code, the formula

$$a_i = \sum_{j=1}^{\infty} \sigma_{i,j} a_{2^{j-1}}, \qquad \text{when } i = \sum_{j=1}^{\infty} \sigma_{i,j} 2^{j-1}. \tag{1}$$

is valid. We shall prove this formula by induction on the number i. For $i = 0$, and also $i = 2^m$ $(m = 0, 1, ...)$, it is trivial. Consequently, to prove formula (1) it is enough to show that

$$a_{2^m+r} = a_{2^m} \oplus a_r, \qquad 1 \le r \le 2^m - 1, \tag{2}$$

under the assumption that formula (1) is valid for all $i < 2^m + r$. To prove equality (2), suppose the contrary, i.e.

$$a_{2^m+r} \ne a_{2^m} \oplus a_r. \tag{3}$$

Relying on the induction hypothesis and Lemma 1, it is easy to show that

$$\rho(a_{2^m+r} \oplus a_{2^m}, a_j) = \rho(a_{2^m+r}, a_{2^m+j}) \ge d, \qquad 0 \le j < r; \tag{4}$$

$$\rho(a_{2^m+r} \oplus a_r, a_l) = \rho(a_{2^m+r}, a_{l'}) \ge d, \qquad 0 \le l, l' \le 2^m - 1; \tag{5}$$

$$\rho(a_{2^m} \oplus a_r, a_i) \ge d, \qquad i < 2^m + r. \tag{6}$$

From the definition of the natural algorithm and inequalities (3)–(6), the following three inequalities follow:

$$|a_r| < |a_{2^m+r} \oplus a_{2^m}|, \quad |a_{2^m}| < |a_{2^m+r} \oplus a_r|, \quad |a_{2^m+r}| < |a_{2^m} \oplus a_r|,$$

which are incompatible by Lemma 2. The theorem is proved.

From formula (1) there follow two important properties of the codes $S_{n,d}$.

1°. The word a_{2^i} $(i = 1, 2, ...)$ has zeros in the positions with numbers $n(j)$, $1 \le j \le i$, where $n(j)$ denotes the number of the last unit position of the word $a_{2^{j-1}}$.

2°. If $1 \le r \le 2^i - 1$, then $a_{2^i+r} = a_{2^i} \oplus a_r$.

Proceeding from these properties, it is easy to show that the codes $S_{n,d}$ can also be constructed by the following algorithm.

As the first tuple a_0 one takes the tuple $(0, 0, \dots, 0)$. If the tuples $a_0, a_1, \dots, a_{2^i-1}$ have already been chosen, then as the tuple a_{2^i} one takes the smallest, in value, tuple of the set B_n having zeros in the positions with numbers $n(j)$, $1 \le j \le i$, and lying at distance at least d from each of the tuples a_l, $0 \le l \le 2^i - 1$, if such a tuple still exists; and the tuples a_{2^i+r}, $1 \le r \le 2^i - 1$, are defined by the formula $a_{2^i+r} = a_{2^i} \oplus a_r$.

The latter algorithm differs advantageously from the natural algorithm underlying the definition of the codes $S_{n,d}$ in that in it the search is considerably reduced, both by reducing the length of the tuples searched through and by reducing their number.

To formulate the assertion that the natural order is, in a certain sense, the only order under which systematic codes are always obtained, let us introduce the notion of equivalence of orders. Two orders $b_0, b_1, \dots, b_i, \dots$ and $b'_0, b'_1, \dots, b'_i, \dots$ of arranging the tuples of the set B will be called **equivalent** if from $b_j \oplus b_k = b_l$ it follows that $b'_j \oplus b'_k = b'_l$.

Theorem 2. *In order that, for every order of arrangement of the tuples of the set B equivalent to the given one, the trivial algorithm for constructing a d-code, applied to the first 2^n tuples, should yield systematic codes for every n, it is necessary and sufficient that the given order be equivalent to the natural one.*

3°. To estimate the size of the code $S_{n,d}$, denote by $m(n, d)$ the number of generators of this code. Then the size of the code $S_{n,d}$ will be equal to $2^{m(n,d)}$. It can be shown that the quantity $m(n, d)$ satisfies the very same relations that are known ([1,3]) for the number of generators of the maximal systematic code

$$m(n, 2s + 1) = m(n + 1, 2s + 2); \tag{7}$$

$$\left[\log_2 \frac{2^n}{2 + C^1_{n-1} + \dots + C^{2s-1}_{n-1}}\right] \le m(n, 2s + 1) \le \left[\log_2 \frac{2^n}{1 + C^1_n + \dots + C^s_n}\right], \tag{8}$$

Hence, in particular, it follows that the codes $S_{n,3}$ and $S_{n,4}$ are maximal in the class of systematic codes, and

$$m(n, 3) = m(n + 1, 4) = \left[\log_2 \frac{2^n}{n + 1}\right]. \tag{9}$$

Lemma 3. *The code $S_{n,3}$ coincides with the Hamming code ([1]).*

Proof. The Hamming code H_n with correction of one error can be defined in another way as the aggregate of all tuples $a = (\alpha_1, \alpha_2, \dots, \alpha_n)$ for which

$$e = \sum_{i=1}^{n} \alpha_i e_i = (0, 0, \dots), \tag{10}$$

where $e_i = (\sigma_{i,1}, \sigma_{i,2}, ...)$, when $i = \sum_{j=1} \sigma_{i,j} 2^{j-1}$. Let us first prove by induction that if $a_l \in S_{n,3}$, then $a_l \in H_n$. For the tuple $a_0 = (0, 0, ..., 0)$ this is obvious. Suppose that this is true for all tuples a_r, $0 \leq r \leq l-1$, and show that it is then true also for the tuple $a_l = (\alpha_{l,1}, \alpha_{l,2}, ..., \alpha_{l,n})$.

Assume the contrary; in this case

$$e = \sum_{i=1}^{n} \alpha_{l,i} e_i \neq (0, 0, ...).$$

Let the number of the last unit position of the tuple e be t. Then there exists such a p that $\alpha_{l,p} = 1$, $\sigma_{p,t} = 1$, and, consequently, the tuple $e \oplus e_p$ is equal to some tuple e_q, where $0 \leq q < p \leq n$. Consider the tuple $b = (\beta_1, \beta_2, ..., \beta_n)$, where $\beta_p = 0$, $\beta_q = \alpha_{l,q} \oplus 1$ (if $q \neq 0$), and $\beta_j = \alpha_{l,j}$ in

remaining cases. It is clear that $|b| < |a_l|$, and at the same time, because the words $a_0, a_1, ..., a_{l-1}$ and, as is easy to verify, b belong to H_n, the inequalities $\rho(b, a_i) \geq d$, $i = 0, 1, ..., l-1$, hold. We have arrived at a contradiction with the fact that the word a_l of the code $S_{n,3}$ was chosen by the natural algorithm after the words $a_0, a_1, ..., a_{l-1}$. Thus, $S_{n,3} \subset H_n$. On the other hand, since a maximal d-code cannot be a proper part of another d-code, $S_{n,3} = H_n$, and the lemma is proved.

Table 1

Words
11111
11000111
1010010011
01010010101
011001000011
11010010000101
110101001001001
10110010010010001
11100110100100001
111001000000000011
1001001000000000101
111000001001000001001

Table 2

Words
11111
11000111

Words
1010010011
01010010101
011001000011
11010010000101
110101001001001
1011001010010001
11100110100100001
100101000000000011
1011001000000000101
01000100100000001001
1100000000010000010001

The supposition that all codes $S_{n,d}$ are maximal in the class of systematic codes turned out to be incorrect. A more detailed investigation showed that the codes $S_{n,5}$, for example for $n \leq 21$, are indeed maximal in the class of systematic codes, with the possible exception of $S_{18,5}$, whereas the code $S_{22,5}$ is no longer such. Table 1 gives the generators of a maximal systematic code $S_{22,5}$; Table 2 gives the generators of a maximal systematic code for $n = 22$ and $d = 5$.

Table 3

Words
1111
110011
1010101
0101011
01101001
11010101
11011011
10110111
11101111
111010001
100101001
111000111
1001100001
1011010001
0100101001
1100000101
1100010011
1111101011
0110011111
11011000001
01110100001

Words
00011010001
10001001001

4. In the practical construction of the codes $S_{n,d}$, it is useful to take into account properties 1^0 and 2^0 and to perform the search only over the generators of these codes and only over those digits of the generators whose numbers are not of the form $n(j)$. This can be achieved by means of the following device.

Denote by $C_{n,d}$ the set of words having the property that, for any pairwise distinct words $c^1, c^2, \dots, c^{d-1}$ from this set, the inequalities are satisfied (cf. (3))

$$\|c^1\| \geq d-1, \quad \|c^1 \oplus c^2\| \geq d-2, \dots, \|c^1 \oplus c^2 \oplus \dots \oplus c^{d-1}\| \geq 1,$$

and obtained from B_n by the natural algorithm. Let

$$C_d = \bigcup_{n=0}^{\infty} C_{n,d}.$$

From the first m words $c_i = (\gamma_{i,1}, \gamma_{i,2}, \dots, \gamma_{i,k(i)})$, $1 \leq i \leq m$ ($k(i)$ denotes the number of the last unit digit of the word c_i) of the set C_d, form the words

$$c_i^* = (\gamma_{i,1}^*, \gamma_{i,2}^*, \dots, \gamma_{i,k(i)+i}^*), \quad 1 \leq i \leq m,$$

putting

$$\gamma_{i,j}^* = \begin{cases} 0, & j = k(l) + l, \quad 1 \leq l < i; \\ \gamma_{i,j-l}, & k(l) + l < j < k(l+1) + l + 1, \quad 0 \leq l < i \ (k(0) = 0); \\ 1, & j = k(i) + i. \end{cases} \tag{11}$$

The group, with respect to the operation of componentwise addition modulo 2, generated by the words $c_1^*, c_2^*, \dots, c_m^*$, will be denoted by $C_{m,d}^*$.

Lemma 4. $S_{n,d} = C_{m,d}^*$, where $k(m) + m \leq n < k(m+1) + m + 1$, with

$$a_{2^i-1} = c_i^*, \quad i = 1, 2, \dots, m.$$

On the basis of Lemma 4, the problem of constructing the codes $S_{n,d}$ can be reduced to the problem of finding several first words of the set C_d. On a computer the first 23 words of the set C_5 were found (Table 3); with their help, by formula (11), one can determine the generators of all codes $S_{n,5}$ for $n \leq 34$ and thereby construct these codes.

Mathematical Institute named after V. A. Steklov
Academy of Sciences of the USSR

Received
8 XII 1959

References

1. R. W. Hamming, Bell Syst. Techn. J., **29**, No. 2, 147 (1950).

2. V. I. Siforov, *Radiotekhnika i elektronika*, **1**, No. 2, 131 (1956).

3. R. R. Varshamov, DAN, **117**, No. 5, 739 (1957).

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