COMBINATORIAL PROBLEMS MOTIVATED BY COMMA-FREE CODES *

Vladimir I. Levenshtein

Keldysh Institute for Applied Mathematics, RAS, 125047 Moscow, RUSSIA

leven@keldysh.ru

Abstract In the paper some combinatorial problems motivated by comma-free codes are considered. We describe these problems, give the most significant known results and methods used, present some new results and formulate open problems.

Keywords: comma-free codes, comma-free index, comma-free sequence, codes without overlaps, difference systems of sets, synchronization

1. Introduction

The paper "Comma-free codes" [8] by Golomb, Gordon, and Welch opened a new direction in coding theory: investigation of combinatorial problems of *word synchronization* for block codes. Although comma-free codes were introduced one year earlier by Crick, Griffith, and Orgel [3], paper [8] contains essential mathematical results for these codes: like for example an upper bound on the codes' size and a proof that, for any alphabet of size q, this bound is tight when n, the length of the codewords, is odd and sufficiently small. Numerous efforts were overtaken to prove that this bound is tight for each odd length and this task was successfully completed by Eastman in [5]. On the other hand, it was proved in [8] that this bound cannot be attained for even n if q is large compared to n. The effort to strengthen Eastman's result gave rise [12], [22] to an investigation of a very interesting combinatorial problem on the maximum number of pairwise comparable vectors of length n over the alphabet $\{0, 1, *\}$. Section 2 contains the main results for comma-free codes, ideas of their proofs, and open problems.

*The research was supported by The Russian Foundation for Basic Research, grant 01-01-00035, and by the NRC COBASE research grant INT-00023241.

The natural generalization of comma-free codes to codes with given *comma-free index* or *code separation* was caused by considering word synchronization in erroneous channels. The author in [17] showed how it is possible to combine error-correcting and word-synchronization properties. In particular, there exist binary codes C of length n with minimum Hamming distance $2s+1$ and comma-free index $2s+1$ which have asymptotically minimal redundancy

$$n - \log_2 |C| \sim (s+1)\log_2 n \quad \text{for fixed } s \geq 1 \text{ and } n \to \infty.$$

The high encoding and decoding complexity of these codes makes the investigation of the comma-free index of cosets of linear $[n,k]$ codes of essential interest. However, Bassalygo proved in [1] that the redundancy $r = n - k$ of such codes with comma-free index ρ must be at least $\sqrt{\rho n}$. For finding cosets of linear codes whose redundancy is close to this bound, the author introduced and considered in [19] a combinatorial problem on the *difference systems of sets*. The problem is to find a sequence (or a code of size 1) of length n over the alphabet $\{0,1,...,q-1,*\}$ which has a comma-free index ρ and a minimal number of letters distinct of $*$ (here the letter $*$ corresponds to information symbols of a codeword and is disregarded when calculating the Hamming distance). The older results, some new results, and the main open problem in this direction are described in Section 3.

Finally, it is worth noticing that consideration of (block) comma-free codes gives rise to the notion of *bounded synchronization delay* of variable length codes. This notion was independently introduced and investigated by the author in [13], [14], [15] and by Golomb and Gordon in [10]. In this case there exists an integer s such that knowledge of s consecutive letters of any sequence of codewords allows one to uniquely determine at least one place (comma) separating two codewords. Gilbert and Moore in [6] considered the property of *bounded decoding delay* of variable length codes. This property suggests the existence of a *decoding automaton* or a machine with a finite number of states which, starting from a special initial state, decodes any sequence of code words (in the absence of errors). The bounded synchronization delay property of a code is equivalent [14], [15] to the existence of a *self-adjusting* decoding automaton. Such an automaton possesses an additional non-error propagation property: its initialization in *any* state can result in wrong decoding of a bounded number t of initial codewords. In [16], [18], the author introduced and investigated *codes without overlaps* whose characteristic property is the existence of decoding automaton invariant with respect to the initial state (or self-adjusting decoding automata with

$t = 0$). Some results as well as the main open problem for codes without overlaps are given in Section 4.

2. Comma-free codes

Let F_q^n be the set of words (sequences) of length n over the alphabet $F_q = \{0, 1, ..., q-1\}$. For any $x = x_1 \cdots x_n \in F_q^n$, $y = y_1 \cdots y_n \in F_q^n$, and $i = 1, ..., n-1$, we set

$$T_i(x, y) = x_{i+1} \cdots x_n y_1 \cdots y_i \tag{1}$$

and call $T_i(x, y)$ a *splice* of x and y. In particular, $T_i(x, x)$ is a cyclic shift of x. A code $C \subseteq F_q^n$ is called *comma-free*, if any splice of two codewords is not a codeword. Let $M(n, q)$ be the maximum size of a comma-free code $C \subseteq F_q^n$.

Theorem 1 *(Golomb, Gordon, Welch [8], 1958)*

$$M(n, q) \leq B(n, q) = \frac{1}{n} \sum \mu(d) q^{n/d} \tag{2}$$

where the sum is taken over all divisors d of n and $\mu(d)$ is the Möbius function.

The proof of Theorem is based on the fact that $B(n, q)$ is the number of words $x \in F_q^n$ of period n which are not pairwise equivalent with respect to their cyclic shifts. By definition, a comma-free code can contain only one representative of the equivalence class. In [8] it was also proved that (2) is attained for odd $n \leq 15$ and it was assumed that (2) is tight for all odd n.

This last assumption was proved by Eastman [5] using a graceful construction. In fact, his inductive construction gives a comma-free code E^n (n is an odd integer) which belong to the set F_∞^n of words of length n over the alphabet of all nonnegative integers, $\{0, 1, ... \}$. Initially, $E^1 = F_\infty^1$ is a trivial comma-free code. To describe the inductive step we need some definitions. A sequence $z = (z_1, ..., z_{2k+1}) \in F_q^{2k+1}$ of an odd length $2k + 1 \geq 3$ is called a *brick*, if $z_1 > z_2$, $z_{2k} \leq z_{2k+1}$, and $z_{2i} \leq z_{2i+1}$ implies $z_{2i+2} \leq z_{2i+3}$ for any $1 \leq i \leq k-1$ when $k \geq 2$. For any brick $z = (z_1, ..., z_{2k+1}) \in F_q^{2k+1}$ we define the *numerical value* $v(z) = \sum_{i=1}^{2k+1} z_i q^{2k+1-i}$. For any odd $n \geq 3$, a sequence $x \in F_q^n$ is included to E^n if there exists an odd $m \geq 1$ such that x can be represented as a concatenation of m bricks $z^{(1)}, ..., z^{(m)}$ for which $e(x) = (v(z^{(1)}), ..., v(z^{(m)})) \in E^m$. (The uniqueness of the representation, if it exists, follows from the definition of a brick.) As an example, note that $x \in 1001110110100 \in F_2^{13}$ can be represented as a concatenation of three bricks 10011, 101, and 10100, and $e(x) = (19, 5, 20) \in F_\infty^3$. Since $e(x)$ is a brick and since $e(x) \in E^3$ we conclude that $x \in E^{13}$.

Let $E_q^n = E^n \bigcap F_q^n$. As a result we have:

Theorem 2 *(Eastman [5], 1965) For any $q \geq 2$ and any odd n*

$$E_q^n \quad \textit{is a comma-free code and } |E_q^n| = B(n, q). \qquad (3)$$

Another construction of maximum comma-free codes of odd length was presented later by Scholtz [21]. His construction was iterative and applicable to codes of variable length. In the meantime in [8] it was noted that, in general, (2) is not tight for even n and it was proved that $M(2, q) = \lfloor \frac{1}{3} q^2 \rfloor$ whereas $B(2, q) = \frac{q(q-1)}{2}$. Thus, for $n = 2$, (2) is tight only if $q = 2$ or $q = 3$. The same situation holds for $n = 4$. Moreover, Jiggs proved in [12] that $M(4, 4) = 57$ whereas $B(4, 4) = 60$. This case was of a special interest in the connection with genetic applications considered in [3], [9].

One of the results of [8] can be formulated as follows: for any even n, e.g. $n = 2l$, there exists a minimal number $m(l)$ such that $M(n, q) < B(n, q)$ if $q > m(l)$. In particular, it is shown that $m(1) = m(2) = 3$. In [8] it was also shown that $m(l) \leq 3^l + l$, and in [12] it was proved that $m(l) \leq 2^l + l$. The combinatorial problem that helped to prove the previous inequality was refined in [22] allowing for better upper bounds on $m(l)$.

In the current and following sections we use the extended alphabet $\overline{F}_q = \{0, 1, ..., q-1, *\}$ and denote by $\overline{F}_q^n$ the set of words (sequences) of length n over the alphabet $\overline{F}_q$. We say that a binary word $x = x_1 \cdots x_n \in F_2^n$ *covers* $y = y_1 \cdots y_n \in F_2^n$ if $x_i \geq y_i$ for all i, $i = 1, ..., n$, and if there exists a position where this inequality is strict. This definition can be extended to words from $\overline{F}_2^n$ if the condition $x_i \geq y_i$ is only applied to the positions i where $x_i \neq *$ and $y_i \neq *$. We call two words from F_2^n or $\overline{F}_2^n$ *comparable* if one of the words covers the other. It is easily seen that the maximum number of pairwise comparable words of F_2^n equals $n+1$; for instance, $\{000,\ 001,\ 011,\ 111\}$ forms a maximal set for $n = 3$. We denote by $t(n)$ the maximum size of a set of pairwise comparable words of $\overline{F}_2^n$. In particular, $\{000,\ 0{*}1,\ {*}10,\ 10{*},\ 111\}$ forms a maximal set of pairwise comparable words of $\overline{F}_2^3$ and $t(3) = 5$. The following statement in fact shows that $m(l) \leq t(l) + l$.

Theorem 3 *(Tang, Golomb, Graham [22], 1987) For even $n = 2l$*

$$M(n, q) < B(n, q) \quad \textit{if } \ q > t(l) + l. \qquad (4)$$

The idea of the proof is based on the fact that if $C \in F_q^n$ is a comma-free code and $|C| = B(n, q)$, then C contains a subset D of $\binom{q}{2}$ representatives from each cyclic class of the type

$$a0^{l-1}b0^{l-1} \text{ where } 0 \leq a < b \leq q-1. \tag{5}$$

The main argument is that, for any $a, b \in F_q,\ a \neq b$, any $r,\ r = 1, ..., l$, and any $s,\ r = 1, ..., l$, the set D cannot contain simultaneously the following four representatives:

$$w_1 = 0^{r-1}y0^{l-1}a0^{l-r}, \quad w_2 = 0^{r-1}b0^{l-1}z0^{l-r},$$

$$w_3 = 0^{s-1}u0^{l-1}b0^{l-s}, \quad w_4 = 0^{s-1}a0^{l-1}v0^{l-r},$$

because the words w_1w_2 and w_3w_4 would contain $0^{l-1}a0^{l-1}b0^{l-1}$ and $0^{l-1}b0^{l-1}a0^{l-1}$ as subwords and hence they would contain all cyclic shifts of (5). In particular when $r = s$ there exists at most one element of F_q which occurs in words belonging to D at both positions r and $l + r$. Since r takes l values in the set E of such elements, we have $|E| \leq l$. For any $a \in F_q \setminus E$ consider a word

$$x^{(a)} = x_1^{(a)}x_2^{(a)} \cdots x_l^{(a)} \in \overline{F}_2^l$$

where $x_r^{(a)} = 0$, if a occurs in words of D at position r; $x_r^{(a)} = 1$, if a occurs in words of D at position $l + r$; and $x_r^{(a)} = *$, if a does not occur in words of D on the positions r and $l + r$ $(r = 1, ..., l)$. By construction, all vectors of the set $\{x^{(a)} : a \in F_q \setminus E\}$ are pairwise comparable. Indeed, for any $a, b \in F_q \setminus E,\ a \neq b$, a cyclic shift of (5) belongs to D and hence there exists $r,\ r = 1, ..., l$, such that either $x_r^{(a)} > x_r^{(b)}$ or $x_r^{(a)} < x_r^{(b)}$; converse inequality on another position s is not possible by the main argument above.

Results on $t(l)$ presently known are described in the paper by van Lint [23] with full proofs. An upper bound

$$t(l) < l^{c \log_2 l} \tag{6}$$

with a constant $c > 1$ was proved in [22]. Using the known recurrence

$$t(l) \leq t(l-1) + t(\lfloor \frac{l-1}{2} \rfloor) \tag{7}$$

van Lint [23] proved (6) with $c = 1$. Note that using (6) in (7) we get

$$t(l) \leq t(l-1) + t(\lfloor \frac{l-1}{2} \rfloor) \leq (l-1)^{c \log_2 l} \left(1 + \frac{2^c}{(l-1)^{2c}}\right).$$

In fact, this shows that $c = 1/2$ is the minimum constant in (6) which can be obtained by induction from (7). Moreover, we have

$$t(l) < l^{1/2\log_2 l} \quad \text{for } l \geq 8. \tag{8}$$

Indeed, using the first values $t(1) = 2,\ t(2) = 3,\ t(3) = 5,\ t(4) = 6,\ t(5) = 8$ and the recurrence (7) one can check that (8) is true for $8 \leq l \leq 16$ (and is not true for smaller l) and then can apply induction beginning from $l = 17$, since $(1 - \frac{1}{l})^{1/2\log_2 l}(1 + \frac{\sqrt{2}}{l-1}) < 1$ already for $l \geq 7$.

The best known lower bound

$$t(h^2+h+1) \geq h(h^2+h+1) + 2 \quad \text{for any } h = 1, 2, ... \tag{9}$$

was found in an unpublished paper by Collins, Shor, and Stembridge (1984). The corresponding set $C \in \overline{F}_2^{h^2+h+1}$ of pairwise comparable words consists of the all-zero and all-one words and of all h^2+h+1 cyclic shifts of each of the words

$$z^{(i)} = 1z_0^{(i)}z_1^{(i)}\cdots z_{h^2+h-1}^{(i)}, \quad i = 0, 1, ..., h-1, \tag{10}$$

which we define as follows. Given $j,\ j = 0, 1, ..., h^2+h-1$, let integers ξ and η, $0 \leq \xi \leq h-1,\ 0 \leq \xi \leq h$, be uniquely defined by $j = \xi(h+1) + \eta$. Then $z_j^{(i)} = 0$, if $\xi \geq i$ and $\xi + \eta \leq h-1$; $z_j^{(i)} = 1$, if $\xi \leq i-1$ and $\xi + \eta \geq h-1$; $z_j^{(i)} = *$, otherwise. In particular, for $h = 3$ (and $l = 13$) we have

$$\begin{array}{lcccccccccccccc}
z^{(0)} & = & 1 & 0 & 0 & 0 & * & 0 & 0 & * & * & 0 & * & * & * \\
z^{(1)} & = & 1 & * & * & 1 & 1 & 0 & 0 & * & * & 0 & * & * & * \\
z^{(2)} & = & 1 & * & * & 1 & 1 & * & 1 & 1 & 1 & 0 & * & * & *
\end{array}$$

A detailed proof of the fact that this elegant construction gives pairwise comparable words was published by van Lint [23]. The main open problem is whether there exists a constant $c,\ c \geq 3/2$, such that

$$t(l) = O(l^c) \quad \text{as } l \to \infty. \tag{11}$$

Since (2) is, in general, not attained for even n, a problem of finding the asymptotic behavior of $M(n, q)$ for a fixed n and $q \to \infty$ was also investigated in [8]. Jiggs proved in [12] that for a fixed $n \geq 5$ and $q \to \infty$

$$\frac{2q^n}{en} \lesssim M(n, q) \lesssim \frac{q^n}{n}.$$

The author assumes that the bound (2) is tight in the asymptotic sense typical for coding theory, that is,

$$M(n,q) \sim \frac{q^n}{n} \text{ when } q \ (q \geq 2) \text{ is fixed and } n \to \infty.$$

However, this problem is still open.

3. Codes with given comma-free index

The notion of comma-free codes gives natural rise to a parameter $\rho(C)$ of a code $C \subseteq F_q^n$ which is called the *comma-free index* or code separation and is defined as follows: $\rho(C)$ is the minimum Hamming distance $d(T_i(x,y),z)$ where $T_i(x,y)$ is a splice defined by (1). The minimum is taken over all $x,y,z \in C$, and $i = 1,\dots,n-1$. We extend this definition to codes $C \subseteq \overline{F}_q^n$ by assuming that in finding the Hamming distance between two words one does not take into account the positions where letters $*$ occur. If a word $z \in \overline{F}_q^n$ forms a code (of size 1) with a comma-free index of at least ρ then we call z a *comma-free sequence of index* ρ or simply a *comma-free sequence* if $\rho = 1$. In particular, every word (10) is a comma-free sequence.

Let $M(n,q,\rho)$ be the maximum cardinality of a code $C \subseteq F_q^n$ with $\rho(C) = \rho$. Since a comma-free code is defined as a code C with $\rho(C) = 1$, we have $M(n,q,1) = M(n,q)$.

Theorem 4 *(Levenshtein [17], 1969) For any fixed* $\rho \geq 1$,

$$M(n,2,\rho) \gtrsim \frac{1}{c(\rho)e}\frac{2^n}{n} \qquad as \qquad n \to \infty \tag{12}$$

where $c(\rho)$ *is a constant (in particular,* $c(1) = c(2) = 1$, $c(3) = 14$, $c(4) = 18$*).*

Note that as a result we have $M(n,2) \gtrsim \frac{2^n}{en}$ over even $n \to \infty$.

Codes $C \subseteq F_2^n$ which were used to prove Theorem 4 have redundancy $n - \log_2 |C|$ which grows only by $\log_2 n$ when $n \to \infty$. However, their encoding and decoding is complex. Therefore, it is natural to consider the same problem for cosets of linear $[n,k]$-codes $C \subseteq F_q^n$. (Any linear code contains the zero vector and, hence, its comma-free index equals zero.) Considering linear codes we tacitly assume that q is a prime power and F_q is the Galois field $GF(q)$. The following result shows that the redundancy $r = n - k$ of a linear $[n,k]$-code whose coset has a given comma-free index should be significantly larger than those mentioned before.

Theorem 5 *(Bassalygo [1], 1966) If q is a prime power and a coset of a linear $[n,k]$-code C has comma-free index ρ, then*

$$r = n - k \geq \sqrt{\rho n}. \qquad \textit{(13)}$$

The proof of this statement is based on the fact that for a coset of a linear $[n,k]$ code over $GF(q)$ there exist k information positions and $r = n-k$ check positions such that any check position is a linear function over $GF(q)$ of the previous information positions. If $r(r-1) < (n-1)\rho$, then one can show that there exists an $i,\ 1 \leq i \leq n-1$, and words $x, y, z \in C$ such that $d(T_i(x,y), z) < \rho$. Therefore, $r(r-1) \geq \rho(n-1)$ which implies (13).

For constructing cosets of linear $[n,k]$-codes $C \subseteq F_q^n$ whose redundancy is close to the bound (13), the author introduced and investigated in [19] the following combinatorial notion. A collection Q of q disjoint subsets Q_i of $N_n = \{0, 1, \ldots, n-1\},\ i = 0, 1, ..., q-1$, is called a *difference system of sets* (DSS) of index ρ if for each number $s,\ s = 1, ..., n-1$, the equation

$$x - y = s \bmod n \qquad (14)$$

has at least ρ solutions for $x \in Q_i,\ y \in Q_j,\ i, j = 0, 1, ..., q-1,\ i \neq j$. (It is worth to underline that x and y should belong to different subsets.)

For a collection Q of disjoint subsets Q_i of $N_n = \{0, 1, \ldots, n-1\},\ i = 0, 1, ..., q-1$ (in particular, for a DSS) consider a sequence

$$z(Q) = z_0 z_1 \cdots z_{n-1} \in \overline{F}_q^n \qquad (15)$$

where $z_i = *$, if $i \notin \bigcup_{i=0}^{q-1} Q_i$, and $z_i = j$, if $i \in Q_j$ $(i = 0, 1, ..., n-1)$. It is easily seen that such a collection Q is a DSS of index ρ if and only if $z(Q)$ is a comma-free sequence of index ρ. As an example, note that the collection Q of subsets $Q_0 = \{0, 9\}$, $Q_1 = \{1, 18\}$, $Q_2 = \{3, 14\}$ of N_{25} form a DSS of index 1. The corresponding sequence

$$z(Q) = 01*2*****0****2***1****** \in \overline{F}_3^{25} \qquad (16)$$

is a comma-free sequence.

For a DSS Q of index ρ, consider a code $C(Q) \subseteq F_q^n$ of redundancy $|\bigcup_{i=0}^{q-1} Q_i|$ whose information positions are in the places, where $z(Q)$ has $*$, and the remaining positions equal 0. From the definition of a DSS of index ρ, the shift of $C(Q) \subseteq F_q^n$ on the vector obtained from $z(Q)$ by replacement all $*$ for zeros gives a comma-free code of index ρ. For this reason we have $\sum_{i=0}^{q-1} |Q_i|$ being the *redundancy* of a DSS. We denote by

$r_q(n, \rho)$ the minimum redundancy of all DSS of index ρ with parameters n and q. Such a DSS with redundancy equal to $r_q(n, \rho)$ is referred to as *optimal.*

A DSS of index ρ is called *perfect* if for every number s, $1 \leq s \leq n-1$, equation (14) has exactly ρ required solutions. A DSS is called *regular* if all subsets Q_i are of the same size. We use the notation DSS-(n, m, q, ρ) for a regular DSS of index ρ with q subsets of size m of the set N_n; its redundancy equals $r = qm$. In particular, the sequence (16) generates a regular perfect DSS-$(25, 2, 3, 1)$. Any *cyclic difference set* (v, k, λ) (see [2]) is a perfect regular DSS-$(v, 1, q, \rho)$ with $q = k$ and $\rho = \lambda$. Thus a DSS can be seen as a generalization of cyclic difference sets.

Theorem 6 *(Levenshtein [19], 1971) For any DSS with parameters n, q, and ρ,*

$$r_q(n, \rho) \geq \sqrt{\frac{q\rho(n-1)}{q-1}} \qquad (17)$$

with equality if and only if the DSS is perfect and regular.

It follows that any perfect regular DSS is optimal. Note also that the condition $n = (q-1)qm^2/\rho + 1$ is necessary for attainability of (17); in particular, ρ must be a divisor of $(q-1)qm^2$.

Theorem 7 *(Levenshtein [19], 1971)*

$$r_2(n, 1) = \lceil \sqrt{2(n-1)} \rceil, \quad r_2(n, 2) = \lceil 2\sqrt{n-1} \rceil. \qquad (18)$$

In particular, for any $n \geq 2$ and $\tau_0 = \lceil \sqrt{\frac{n-1}{2}} \rceil$, $\tau_1 = \lceil \frac{n-1}{2\tau_0} \rceil$, the sets

$$Q_0 = \{\tau_0 + 1, 2\tau_0 + 1, ..., \tau_1\tau_0 + 1\} \quad \text{and} \quad Q_1 = \{1, 2, ..., \tau_0\}$$

form an optimal DSS with $q = 2$ and $\rho = 1$ which is perfect and regular if $n = 2m^2 + 1$. For $n = 13$ we have $\tau_1 = 3$, $\tau_0 = 2$ and get the following example of a comma-free sequence

$$1 \ 1 \ 0 \ * \ 0 \ * \ 0 \ * \ * \ * \ * \ * \ *$$

It is interesting to note that even though every word in (10) of length $n = h^2 + h + 1$ is a comma-free sequence, non of them is optimal for $h \geq 3$.

The case $q \geq 3$ remains unsolved despite the example of a perfect regular DSS-$(25, 2, 3, 1)$ shown above. The main open problem is to prove the following assumption:

$$r_q(n, 1) = O(\sqrt{n}) \quad \text{for a fixed } q \geq 2 \text{ as } n \to \infty. \qquad (19)$$

Note that $r_2(n,1) \sim \sqrt{2n}$ holds by (18).

Cyclic difference sets $(v,k,1)$ (or DSS-$(v,1,k,1)$) are known to exist for parameters $k = t+1$ and $v = t^2+t+1$ where t is a prime power. There exists a very possible conjecture that these are the only valid parameters (see [4]). This implies that $r_{t+1}(t^2+t+1,1) \leq t+1$ if t is a prime power. Note that from (17) it follows that

$$r_q(n,1) \gtrsim \sqrt{n} \quad \text{as } q \text{ and } n \to \infty. \tag{20}$$

Thus, for the known cyclic difference sets, the asymptotic bound (20) is tight, however, for q grows as $\sqrt{n}$ when $n \to \infty$. The following result allows one to construct asymptotic optimal comma-free sequences with a slower growth of q as a function in n.

Theorem 8 *If there exists a DSS-$(v,1,q,\rho)$, $2 \leq q < v$, then, for any $h = 2,3,\dots$, there exists a regular DSS-(n,m,q,ρ) with $n = v^h$ and $m = \frac{q^h-1}{q-1}$.*

To describe the construction denote $D = \{a_0, a_1, ..., a_{q-1}\}$ a q-subset of $N_v = \{0,1,...,v-1\}$ which forms a DSS-$(v,1,q,\rho)$. We can assume that $0 \notin D$. For any $i = 0,1,...,q-1$ and $t = 1,...,h$, let

$$Q_{i,t} = \{\sum_{j=t}^{h} x_j v^{j-1} \ : \ x_t = a_i, \ x_j \in D \text{ when } j = t+1,...,h\}.$$

The required DSS-$(v^h, \frac{q^h-1}{q-1}, q, \rho)$ consists of the sets

$$Q_i = \bigcup_{t=1}^{h} Q_{i,t}, \quad i = 0,1,...,q-1. \tag{21}$$

Using the existence of DSS-$(t^2+t+1,1,t+1,1)$ which happen to be the known cyclic difference sets we get the following statements.

Corollary 1 *For any prime power t and integer h there exists a regular DSS-$(n,m,t+1,1)$ with $n = (t^2+t+1)^h$ and $m = \frac{(t+1)^h-1}{t}$.*

Corollary 2 *If $q = t+1$ and t runs prime powers, then for the subsequence of $n = (t^2+t+1)^t$*

$$r_q(n,1) \lesssim \sqrt{en} \quad \textit{and} \quad q \sim \frac{\ln n}{2\ln(\ln n)}. \tag{22}$$

Corollary 2 follows from Corollary 1 since for $h = t$ we have $n \sim t^{2t}e$ and $mq \sim t^t e$ as $t \to \infty$. For the sequence $q = t+1$ there exists also a subsequence of n which implies $r_q(n,1) \sim \sqrt{n}$ for a slightly faster growth of q as a function in n.

4. Codes without overlaps

A finite (or countable) code $C \subset \bigcup_{n=1} F_q^n$ is called a *code without overlaps* [16], [18] if two conditions are satisfied. At first a prefix of a codeword, which is not empty and does not coincide with this codeword, is not a suffix of a codeword and secondly, for variable-length codes, a codeword does not contain another codeword as a subword. Any code $C \subset F_q^n$ without overlaps is a comma-free code. Moreover this code has the strongest non-error propagation property which is that an error in a codeword or in a state of a suitable decoding automaton has no influence on decoding of subsequent codewords.

Let $L(n, q)$ denote the maximum cardinality of a code $C \subset F_q^n$ without overlaps. For any $H \subset F_q^m$ the code $C_H \subset \bigcup_{n=1} F_q^n$ with the set H of *synchronizing suffices* is defined as the set of all words c such that for any $h \in H$ the word hc contains elements of H only in the first and last m positions. Note that some words of C_H can have length smaller than m. Codes with one synchronizing suffix ($|H| = 1$) were investigated by Gilbert in [7]. In the case of suffix 0^m these are codes without overlaps which we denote by $G_q^{0,m}$. Let $G_q^{0,m}(n) = G_q^{0,m} \bigcap F_q^n$ and $G(n, q) = \max |G_q^{0,m}(n)|$ where the maximum is taken over all $m \geq 1$. This results in $L(n, q) \geq G(n, q)$.

Theorem 9 *(Gilbert [7], 1960, Levenshtein [16], 1964). For a fixed* $q \geq 2$,

$$G(n, q) \gtrsim q^{\frac{-q}{q-1}} \ln q \frac{q^n}{n} \quad \text{as } n \to \infty \qquad (23)$$

$$\text{and} \quad G(n, q) \gtrsim \frac{q-1}{eq} \frac{q^n}{n} \text{ over the subsequence } n = \frac{q^i - 1}{q-1}, \; i = 1, 2, \ldots \; . \qquad (24)$$

Let $F_2^{n,w}$ be the subset of all words of F_2^n which have w ones. For any relatively prime n and w, a maximum code $C \subset F_2^{n,w}$ without overlaps was constructed by Markov and Noskov in [20]. Their construction is based on the remarkable fact that every cyclic class of words of $F_2^{n,w}$ contains a unique representative $x \in F_2^{n,w}$ such that for any of its prefices $y \in F_2^{n',w'}$, $1 \leq n' < n$, there holds $wn' > w'n$. This maximum code consists of all $\binom{n}{w}/n$ such representatives and is unique up to writing code words in the reverse order.

For a set $C \subset \bigcup_{n=1} F_q^n$, denote by $\hat{C}$ the set of all words from $\bigcup_{n=0} F_q^n$ which contain no word of C. Consider the generating functions

$$f(C,z) = \sum_{n=1} |C(n)| z^n \quad \text{and} \quad g(C,z) = \sum_{n=0} |\hat{C}(n)| z^n$$

where $C(n) = C \cap F_q^n$ and $\hat{C}(n) = \hat{C} \cap F_q^n$. The author proved in [18] that for any code $C \subset \bigcup_{n=1} F_q^n$ without overlaps

$$g(C,z)(1 - qz + f(C,z)) = 1.$$

The following statement is a consequence of this equality.

Theorem 10 *(Levenshtein [18], 1970)*

$$L(n,q) \le \left(1 - \frac{1}{n}\right)^{n-1} \frac{q^n}{n} < \frac{1}{e} \frac{q^n}{n-1}. \tag{25}$$

The main open problems are to strengthen bounds (23)–(25) and prove or contradict the conjecture that $L(n,q) = G(n,q)$.

To investigate this conjecture we present a new class of codes without overlaps which are generalizations of the Gilbert codes $G_q^{0,m}$. Denote by $G_q^{k,m}$ the code with the set H of synchronizing suffices where H consists of $(q-1)^k \binom{m-1}{k}$ words $h \in F_q^m$ which have exactly $m-k$ zeros with one of the zeros in the last position. For instance, $H = \{10000, 01000, 00100, 00010\}$ when $q = 2,\ k = 1,\ m = 5$.

Theorem 11 *For any q, k, m, n $(0 \le k \le m-1,\ q \ge 2,\ m \ge 2,\ n \ge 1)$, the code $G_q^{k,m}(n) = G_q^{k,m} \bigcap F_q^n$ is a code without overlaps.*

The code $G_q^{k,m}$ contains q^k words of length $k+1$ which have a zero only in the last position. These words form a code without overlaps. The remaining words have a length of at least $k+2$ and begin with a prefix of $k+1$ nonzero letters. From the definition of a code with the set H of synchronizing suffices we can say that a prefix $a \in F_q^l$ of $x \in G_q^{k,m}(n)$, $n \ge k+2$, differs from a suffix $b \in F_q^l$ of $y \in G_q^{k,m}(n)$ in the case $m < l < n$. In the case $k+1 \le l \le m$, a contains at least $k+1$ nonzero letters whereas b can contain at most k of them. In the case $1 \le l \le k$, a has no zeros whereas b has zero on the last position.

Let $G_k(n,q) = \max G_q^{k,m}(n)$ where the maximum is taken over all $m \ge k+1$. It is natural to compare $G(n,q) = G_0(n,q)$ with $G_k(n,q)$, $k \ge 1$. Words of $G_q^{k,m}(n)$ have a larger number of possible suffices, but a stronger restriction is imposed to their choice. The fact that for $n \geq k+2$ all words of $G_q^{k,m}(n)$ begin with a prefix of $k+1$ nonzero letters allows us to apply Theorem 2.1 (after correcting some misprints in its formulation) of the paper by Guibas and Odlyzko [11] to find a generating function

$$f^-(G_q^{k,m}, z) = \sum_{n=k+2} |G_q^{k,m}(n)| z^{-n}.$$

We give the result for $q = 2$ and $k = 1$.

Theorem 12 $f^-(G_2^{1,m}, z) = \sum_{i=1}^{m-1} x_i$ *where* $x_0, x_1, ..., x_{m-1}$ *is a solution of the following system of* m *linear equations:*

$$\begin{cases} (z-2)x_0 + \sum_{j=1}^{m-1} x_j & = & z^{-2} \\ x_0 - \sum_{j=1}^{m-1} a_{i,j} x_j & = & -\delta_{i,1} z^{-2}, \quad i = 1, 2, ..., m-1, \end{cases}$$

where

$$a_{i,j} = \frac{z^{\min(i-1,m-j)} - 1}{z-1} + b_{i,j}, \quad 1 \leq i, j \leq m-1,$$

$$b_{i,j} = 0, \quad \textit{if} \;\; i > j, \quad \textit{and} \;\; b_{i,j} = z^{m+i-j-1}, \;\; \textit{if} \;\; i \leq j.$$

As an example,

$$z^2 f^-(G_2^{1,5}, z) = \frac{z^6 + 3z^5 - z^4 - z^3 + z^2 - z - 1}{z^{10} - z^9 - z^8 - z^6 - 2z^5 + z^3 + 1}.$$

Calculations show that we still have $G(n,2) > G_1(n,2)$ at least for $n \leq 100$.

References

[1] L.A. Bassalygo, “On the separation of comma-free codes”, *Probl. Peredachi Inform.*, vol. 2, no. 4, pp. 78–79, 1966 (in Russian).

[2] C. J. Colbourn and J.F. Dinitz, eds., “The CRC Handbook of Combinatorial Designs”, CRC Press, Boca Raton, 1996.

[3] H.C. Crick, J.S. Griffith, and L.E. Orgel, “Codes without commas”, *Proc. Nat. Acad.Sci.,*, vol. 43, pp. 416–421, 1957.

[4] J.F. Dinitz, D.R. Stinson, eds., “Contemporary Design Theory: A Collection of Surveys ”, Wiley-Interscience Publication, New York, 1992.

[5] W.L. Eastman, “On the construction of comma-free codes”, *IEEE Trans. on Inform. Theory*, vol. IT-11, pp. 263–266, 1965.

[6] E.N. Gilbert and E.F. Moore, “Variable-length binary encoding”, *Bell Syst. Techn. J.*, vol. 38, no. 4, pp. 933–967, 1959.

[7] E.N. Gilbert, “Synchronization of binary messages”, *IRE Trans. on Inform. Theory*, vol. IT-6, pp. 470–477, 1960.

[8] S.W. Golomb, B. Gordon, L.R. Welch, “Comma-free codes”, *Canad. J. Math.*, vol. 10, no. 2, pp. 202–209, 1958.

[9] S.W. Golomb, L.R. Welch, M. Delbrück“ Construction and properties of comma-free codes”, *Biol. Medd. Dan. Vid. Selsk.*, vol. 23, no. 9, pp. 202–209, 1958.

[10] S.W. Golomb, B. Gordon, “Codes with bounded synchronization delay”, *Information and Control*, vol. 8, no. 4, pp. 355–372, 1965.

[11] L.J. Guibas, A.M. Odlyzko, “String overlaps, pattern matching, and nontransitive games”, *J. of Combin. Th., A*, vol. 30, pp. 183–208, 1981.

[12] B.H. Jiggs, “Recent results in comma-free codes”, *Canad. J. Math.,*, vol. 15, pp. 178–187, 1963.

[13] V.I. Levenshtein, “Certain properties of code systems”, Dokl. Acad. Nauk, vol. 140, no. 6 (1961), pp. 1274-1277. English translation in *Soviet Physics - Doklady*, vol. 6, no. 10 (1962), pp. 858–860.

[14] V.I. Levenshtein, “Self-adaptive automata for message decoding”, Dokl. Acad. Nauk, vol. 141, no. 6 (1961), pp. 1320-1323. English translation in *Soviet Physics - Doklady*, vol. 6, no. 12 (1962), pp. 1042–1045.

[15] V.I. Levenshtein, “Certain properties of code systems and self-adjusting automata for message decoding”, in *Probl. Cybern.* **11**, Nauka, Moscow, pp. 63–121, 1964 (in Russian)

[16] V.I. Levenshtein, “Decoding automata which are invariant with respect to their initial state”, in *Probl. Cybern.* **12**, Nauka, Moscow, pp. 125–136, 1964 (in Russian).

[17] V.I. Levenshtein, “Bounds for codes ensuring error correction and synchronization”, *Problemy Peredachi Informatsii*, vol. 5, no. 2, pp. 3-13, 1969 (in Russian). English translation in *Probl. Inform. Transm.*, vol. 5, pp. 1–10, 1969.

[18] V.I. Levenshtein, “Maximum number of words in codes without overlaps”, *Problemy Peredachi Informatsii*, vol. 6, no. 4, pp. 88–90, 1970 (in Russian). English translation in *Probl. Inform. Transm.*, vol. 6, pp. 355–357, 1970.

[19] V.I. Levenshtein, “One method of constructing quasilinear codes providing synchronization in the presence of errors”, *Problemy Peredachi Informatsii*, vol. 7, no. 3, pp. 30–40, 1971 (in Russian). English translation in *Probl. Inform. Transm.*, vol. 7, no. 3, pp.215–222, 1971.

[20] Al.A. Markov, V.V. Noskov, “Construction and properties of binary constant-weight codes without overlaps”, *Discrete Analysis*, vol. 18, pp. 49–65, 1971 (in Russian).

[21] R.A. Scholtz, “Maximal and variable word-length comma-free codes”, *IEEE Trans. on Inform. Theory*, vol. IT-15, pp. 300–306, 1969.

[22] B. Tang, S.W. Golomb, R.L. Graham, “A new result on comma-free codes of even word-length”, *Canad. J. Math.*, vol. 39, no. 3, pp. 513–526, 1987.

[23] J.H. van Lint, “$\{0, 1, *\}$ distance problems in combinatorics”, *Surveys in Combinatorics* (1985), London Mathematical Society Lecture Note Series 103, pp. 113–135.