МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2016
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2016
Proceedings of the Forty Fifth Spring Conference
of the Union of Bulgarian Mathematicians
Pleven, April 6–10, 2016

OPTIMAL BINARY t-DELETION-CORRECTING CODES*

Emil Kolev

In this paper binary t-deletion-correcting codes are considered. By $L_2(n,t)$ we denote the minimum cardinality of t-deletion-correcting binary code of length n. We prove that $L_2(10,2) = 16$, thus solving the first undecided case for $t = 2$. We also describe all optimal inequivalent codes achieving $L_2(9,3) = 11$.

1. Introduction. The deletion-correcting codes have been introduced by Levenshtein in 1965 [1], [2]. The main goal of such codes is to recover a message that has some of its symbols lost during the transmission. In this scenario the receiver gets shorter message and he does not know which of the symbols have been lost. Levenstein found an asymptotically optimal family of 1-deletion correcting codes. For larger values of t though there has been a little or no research.

Any subset of the n-dimensional vector space F_2^n, where $F_2 = \{0,1\}$, is referred to as a binary code. For given positive integers n and t we wish to design a code of length n having largest possible cardinality with the following property:

For any two codewords $\mathbf{x}$ and $\mathbf{y}$ the sets obtained by deleting t symbols from $\mathbf{x}$ and t symbols from $\mathbf{y}$ are disjoint.

If the above is true then the receiver can recover the codeword sent in the case at most t deletions have occurred. A code is called t-deletion-correcting if it corrects any t deletions.

Example 1. Consider the binary code $\mathcal{C} = \{0000, 1101, 0011\}$. For a given codeword we may delete any of its four symbols. As a result we obtain a set of vectors of length 3. Direct verification shows that all three sets obtained from the three codewords are disjoined. Therefore $\mathcal{C}$ is 1-deletion-correcting code.

In general, the described problem is an open problem in coding theory. As in the case of error-correcting codes the efforts are concentrated on finding the largest code size for a fixed number of deletions and a codeword length. In Table 1 [10] some of the known results for $t \geq 2$ are presented.

As it is seen this table not much is known for the exact values of $L_2(n,t)$ for $t \geq 2$. The first undecided case is $L_2(10,2)$. In this paper we prove that $L_2(10,2) = 16$ and we find all inequivalent optimal 2-deletion-correcting codes of length 9.

---

* **2010 Mathematics Subject Classification:** 94B05.
Key words: insertion/deletion codes, Varshamov-Tennengolts codes, multiple insertion/deletion codes.

Table 1

n	$t=2$	$t=3$	$t=4$	$t=5$
4	2	2	–	–
5	2	2	2	–
6	4	2	2	2
7	5	2	2	2
8	7	4	2	2
9	11	5	2	2
10	16–22	6–10	4	2
11	21–44	7–14	5	2
12	31–88	11–22	6–10	4
13	49–176	12–44	6–14	5
14	75–352	16–88	7–22	5–10
15	109–704	24–176	9–44	6–14

2. Preliminaries.

Definition 1. *Levenstein distance* $d_L(\mathbf{x}, \mathbf{y})$ *between two binary vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *is defined as the minimum number of deletions and insertions needed to transform* $\mathbf{x}$ *into* $\mathbf{y}$.

Note that the above definition applies also for vectors $\mathbf{x}$ and $\mathbf{y}$ of different lengths. Deletion distance $\mathrm{dd}(\mathbf{u}, \mathbf{v})$ between two vectors $\mathbf{u}$ and $\mathbf{v}$ of equal length is defined as one-half of the smallest number of deletions and insertions needed to change $\mathbf{u}$ to $\mathbf{v}$ [3]. For example, $\mathrm{dd}(00000, 11111) = 5$ whereas $\mathrm{dd}(00011, 10101) = 2$. It is clear that for vectors $\mathbf{u}$ and $\mathbf{v}$ of equal length we have

$$\mathrm{dd}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} d_L(\mathbf{u}, \mathbf{v}).$$

For a given code $\mathcal{C}$ the deletion distance $\mathrm{dd}(\mathcal{C})$ is defined as

$$\mathrm{dd}(\mathcal{C}) = \min\{\mathrm{dd}(\mathbf{u}, \mathbf{v}) \mid \mathbf{u}, \mathbf{v} \in \mathcal{C}\}.$$

Denote by $L_2(n, t)$ the maximum cardinality of a binary t-deletion-correcting code $\mathcal{C}$ of length n, i.e. for any two distinct codewords $\mathbf{u}$ and $\mathbf{v}$ we have $\mathrm{dd}(\mathbf{u}, \mathbf{v}) > t$ (or, equivalently $d_L(\mathbf{u}, \mathbf{v}) > 2t$). A binary code $\mathcal{C}$ of length n and cardinality $L_2(n, t)$ is called optimal. For more information and useful results the reader is referred to [4], [5], [6], [7], [8], [9].

For a binary vector $\mathbf{u}$ of length n denote by $D_t(\mathbf{u})$ the set of all words of length $n - t$ obtained from $\mathbf{u}$ by deleting t entries in $\mathbf{u}$. In other words, $D_t(\mathbf{u})$ contains all subsequences of $\mathbf{u}$ of length $n - t$.

The size of $D_t(\mathbf{u})$ depends on $\mathbf{u}$. For example, $|D_t(\mathbf{0^n})| = 1$ for any t and $|D_1(\mathbf{x})|$ equals the number of runs in $\mathbf{x}$, that is the number of blocks of consecutive equal symbols.

We introduce a notion of equivalence for deletion-correcting codes. For error-correcting codes the usual definition of equivalence includes coordinate permutation and permutation of the symbols in each coordinate. For deletion-correcting codes these two actions do not preserve deletion-correcting capabilities. That is why we adopt different notion for equivalence.

Definition 2. *Two deletion-correcting codes* $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *are equivalent if one of the following is true:*

1. $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathcal{C}_1$ *if and only if* $\mathbf{x} = (\overline{u_1}, \overline{u_2}, \ldots, \overline{u_n}) \in \mathcal{C}_2$;
2. $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in \mathcal{C}_1$ *if and only if* $\mathbf{x} = (u_n, u_{n-1}, \ldots, u_1) \in \mathcal{C}_2$.

Here, for $x \in \{0, 1\}$ *by* $\overline{x} \in \{0, 1\}$ *we mean* $\overline{x} \neq x$.

Consider two vectors $\mathbf{u}$ and $\mathbf{v}$. We say that a vector $\mathbf{u}$ is t-dominant over a vector $\mathbf{v}$ (alternatively, $\mathbf{v}$ is subordinate of $\mathbf{u}$) if $D_t(\mathbf{v}) \subset D_t(\mathbf{u})$. If $\mathcal{C}$ is t-deletion-correcting code then any dominant codeword may be replaced by its subordinate. Hence, there exists an optimal code having the vectors 0^n and 1^n as codewords. A code is called basic if 0^n and 1^n are codewords. For certain n and t we consider the following research problems:

1. Find $L_2(n, t)$.
2. Find the number of inequivalent basic optimal codes.

3. Optimal 2-deletion-correcting binary codes of length 9 and 10. It is known that $16 \leq L_2(10, 2) \leq 20$ and $L_2(9, 2) = 11$. In this section we prove that $L_2(10, 2) = 16$

Table 2. Optimal binary basic codes of length 9 and 11 codewords.

1.	0, 7, 44, 63, 98, 248, 329, 438, 448, 455, 511
2.	0, 7, 44, 95, 98, 248, 329, 438, 448, 455, 511
3.	0, 7, 44, 98, 159, 248, 329, 438, 448, 455, 511
4.	0, 7, 44, 98, 219, 248, 287, 329, 448, 462, 511
5.	0, 7, 44, 98, 231, 248, 287, 329, 438, 448, 511
6.	0, 7, 52, 63, 171, 224, 290, 380, 413, 483, 511
7.	0, 7, 59, 104, 140, 222, 341, 386, 455, 484, 511
8.	0, 7, 59, 104, 140, 222, 341, 386, 455, 496, 511
9.	0, 7, 59, 104, 140, 249, 293, 438, 448, 455, 511
10.	0, 7, 59, 104, 140, 249, 293, 438, 455, 480, 511
11.	0, 7, 61, 88, 201, 252, 266, 347, 455, 464, 511
12.	0, 7, 61, 88, 201, 252, 266, 347, 455, 480, 511
13.	0, 7, 61, 104, 140, 215, 293, 380, 448, 483, 511
14.	0, 7, 61, 104, 140, 231, 293, 438, 448, 497, 511
15.	0, 7, 62, 85, 112, 140, 231, 386, 438, 497, 511
16.	0, 7, 62, 112, 140, 243, 362, 386, 407, 504, 511
17.	0, 7, 85, 112, 140, 231, 287, 386, 438, 497, 511
18.	0, 7, 110, 112, 140, 243, 341, 386, 399, 500, 511
19.	0, 7, 110, 112, 140, 243, 341, 386, 399, 504, 511
20.	0, 14, 49, 63, 182, 224, 292, 413, 419, 504, 511
21.	0, 14, 49, 63, 182, 224, 292, 413, 467, 504, 511
22.	0, 14, 49, 63, 218, 224, 292, 371, 395, 504, 511
23.	0, 14, 49, 63, 218, 224, 292, 371, 407, 504, 511
24.	0, 14, 63, 112, 145, 182, 388, 413, 467, 504, 511
25.	0, 14, 63, 112, 145, 218, 371, 388, 407, 504, 511
26.	0, 15, 49, 171, 224, 252, 266, 317, 434, 455, 511
27.	0, 15, 52, 171, 224, 252, 289, 317, 434, 455, 511
28.	0, 15, 81, 125, 156, 224, 266, 347, 455, 504, 511
29.	0, 25, 62, 112, 170, 243, 323, 392, 413, 504, 511

and we find all inequivalent optimal 2-deletion-correcting codes of length 9.

Proposition 1. *Up to equivalence there exist* 29 *basic* 2*-deletion-correcting binary optimal codes of length* 9*.*

Proof. Recall that $L_2(9,2) = 11$ and let $\mathcal{C}$ be a basic 2-deletion-correcting binary optimal code of length 9. Since $\mathcal{C}$ is basic, we have that $\mathbf{0^9}, \mathbf{1^9} \in \mathcal{C}$. To find the remaining 7 codewords we perform exhaustive computer search. The final step is to check all codes found for equivalence. The results are given in Table 2 (the codewords are the binary representations of the given integers). □

Proposition 2. *It is true that* $L_2(10,2) = 16$*.*

Proof. Assume there exists 2-deletion correcting code $\mathcal{C}$ of length 10 and 17 codewords. Assume also that the code $\mathcal{C}$ is basic meaning that $\mathbf{0^{10}}, \mathbf{1^{10}} \in \mathcal{C}$.

Split the codewords of $\mathcal{C}$ into two sets according to their last coordinate and then delete this last coordinate. As a result we obtain a 2-deletion-correcting codes $\mathcal{C}_0$ and $\mathcal{C}_1$ of length 9. Without lost of generality assume $|\mathcal{C}_0| > |\mathcal{C}_1|$. Thus, $|\mathcal{C}_0| \geq 9$ and since $L_2(9,2) = 11$ we conclude that $|\mathcal{C}_0| = 9, 10$ or 11. By exhaustive computer search we first find all codes of length 9 and cardinality 9, 10 or 11. Then we add an extra symbol 0 at the end of each of the obtained vectors. The final step is to search for the elements of $\mathcal{C}_1$. The search gave no result implying that 2-deletion correcting code $\mathcal{C}$ of length 10 and 17 codewords does not exist. □

REFERENCES

[1] V. I. Levenshtein. Binary codes capable of correcting, deletions, insertions and reversals. *Doklady Akad. Nauk SSSR*, **163** (1965), 845–848.

[2] V. I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones, *Problemy Peredchi Informacii*, **1** (1965), 12–25.

[3] N. J. A. Sloane. On single-deletion-coirrecting codes. In: Codes and designs. Proceedings of a conference honoring Professor Dijen K. Ray-Chaudhuri on the occasion of his 65th birthday, The Ohio State University, Columbus, OH, USA, May 18-21, 2000. Berlin: de Gruyter. Ohio State Univ. Math. Res. Inst. Publ., **10** (2002), 273–291.

[4] V. I. Levenshtein. On perfect codes in the deletion-insertion metric. *Diskretnaya Matematika*, **3** (1991), 3–20.

[5] V. I. Levenshtein. Efficient reconstruction of sequences. textitIEEE Trans Inf. Theory, **47** (2001), 2–22.

[6] T. G. Swart, H. C. Fereira. A note on double insertion/deletion correcting codes. *IEEE Trans. Inf. Theory*, **49**(2003), 269–272.

[7] L. Tolhuizen. Upper bounds on the size of insertion/deletion codes, Proc 8th Int. Workshop on ACCT, Tsarskoe Selo, Russia, September 2002, 242–246.

[8] A. S. J. Helberg, H. C. Fereira. On multiple insertion/deletion correcting codes. *IEEE Trans. Inf. Theory*, **48** (2002), 305–308.

[9] I. Landjev, Kr. Haralambiev. On multiple deletion codes. Serdica J. Computing, **1**, No 1 (2006), 13–26.

[10] K. Landgev. Deletion-correcting codes, Sofia University, MSci Thesis, 2015.

Emil Milanov Kolev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: emil@math.bas.bg

ОПТИМАЛНИ КОДОВЕ, КОРИГИРАЩИ t ИЗТРИВАНИЯ

Емил Колев

В статията се разглеждат двоични кодове, коригиращи изтривания. С $L_2(n,t)$ се означава минималната мощност на двоичен код с дължина n, поправящ t изтривания. В статията се доказва, че $L_2(10,2) = 16$, с което се решава първият открит случай за $t = 2$. Намерени са всички нееквивалентни оптимални кодове, за които се достига границата $L_2(9,3) = 11$.