Improved Codes for List Decoding in the Levenshtein's channel and Information Retrieval

Tero Laihonen and Tuomo Lehtilä
Department of Mathematics and Statistics, University of Turku, FI-20014 Turku, Finland
Email: terolai@utu.fi and tualeh@utu.fi

***Abstract*—In this paper, we introduce t-revealing codes in the binary Hamming space $\mathbb{F}^n$. Let $C \subseteq \mathbb{F}^n$ be a code and denote by $I_t(C;\mathbf{x})$ the set of codewords of C which are within (Hamming) distance t from a word $\mathbf{x} \in \mathbb{F}^n$. A code C is t-revealing if the majority voting on the coordinates of the words in $I_t(C;\mathbf{x})$ gives unambiguously $\mathbf{x}$. These codes have applications, for instance, to the list decoding problem of the Levenshtein's channel model, where the decoder provides a list based on several different outputs of the channel with the same input, and to the information retrieval problem of the Yaakobi-Bruck model of associative memories. We give t-revealing codes which improve some of the key parameters for these applications compared to earlier code constructions, namely, the length of the output list $\mathcal{L}$ of the decoder and the maximal number of input clues $\hat{m}$ needed for information retrieval.**

I. Introduction

Let us first define mathematically the codes we are interested in and then consider the motivations and applications of them.

Let $\mathbb{F}$ be the binary field and denote by $\mathbb{F}^n$ the Hamming space. As usual, the Hamming distance $d(\mathbf{x},\mathbf{y})$ between two words is the number of coordinate places in which they differ. The all-zero word is denoted by $\mathbf{0} = 00\ldots0$ and the all-one word by $\mathbf{1} = 11\ldots1$. The *support* of a word $\mathbf{x}$ is defined as $\mathrm{supp}(\mathbf{x}) = \{i \mid x_i \neq 0\}$. The Hamming weight $w(\mathbf{x})$ of $\mathbf{x}$ is the cardinality of the support of $\mathbf{x}$. For $\mathbf{x} \in \mathbb{F}^n$ we denote the Hamming ball of radius t and centred at $\mathbf{x}$ by $B_t(\mathbf{x}) = \{\mathbf{y} \in \mathbb{F}^n \mid d(\mathbf{x},\mathbf{y}) \leq t\}$. The *symmetric difference* $A \triangle B$ of two sets A and B is, as usual, $(A\backslash B)\cup(B\backslash A)$. The word $\mathbf{e}_i$ is a word of weight one such that $\mathrm{supp}(\mathbf{e}_i) = \{i\}$. The *complement* of a word $\mathbf{x}$ is the word $\bar{\mathbf{x}} = \mathbf{1} + \mathbf{x}$. A *code* is a subset of $\mathbb{F}^n$ with at least two elements and its elements are called *codewords*. The *minimum distance* of a code C is defined as $d_{\min}(C) = \min_{\substack{\mathbf{c}_1,\mathbf{c}_2 \in C \\ \mathbf{c}_1 \neq \mathbf{c}_2}} d(\mathbf{c}_1,\mathbf{c}_2)$ and the *covering radius* of C as $R(C) = \max_{\mathbf{x}\in\mathbb{F}^n} \min_{\mathbf{c}\in C} d(\mathbf{x},\mathbf{c})$. For $\mathbf{x} = x_1x_2\ldots x_n$, let the function π_i pick the i-th coordinate, that is, $\pi_i(\mathbf{x}) = x_i$. For a subset $A \subseteq \mathbb{F}^n$, we generalize this in the following way by considering the majority voting on the i-th coordinates of the words in A. If there are more 0's (resp. 1's) among the coordinates $\pi_i(\mathbf{a})$, where $\mathbf{a} \in A$, then $\pi_i(A) = 0$ (resp. $\pi_i(A) = 1$). If there is an equal amount of 0's and 1's, the value $\pi_i(A)$ is defined to be the symbol $*$.

Let C be a code and $t \geq 1$ an integer. For any $\mathbf{x} \in \mathbb{F}^n$, we define the set of codewords within distance t from $\mathbf{x}$ as

$$I_t(\mathbf{x}) = I_t(C;\mathbf{x}) = \{\mathbf{c} \in C \mid d(\mathbf{x},\mathbf{c}) \leq t\}.$$

We call this the *I-set* of $\mathbf{x}$. Let $I_t(\mathbf{x})$ be non-empty for a word $\mathbf{x} = x_1x_2\ldots x_n \in \mathbb{F}^n$. We say that the word $\mathbf{x}$ is *accessible*, if $\pi_i(I_t(\mathbf{x})) = x_i$ for all $i = 1,2,\ldots,n$. In other words, using the majority voting on the coordinates of $I_t(\mathbf{x})$ we get $\mathbf{x}$. Otherwise, we say that $\mathbf{x}$ is *non-accessible* (in particular, if I-set of $\mathbf{x}$ is empty).

Next we define a useful function $m_t(\mathbf{x})$ on an accessible word $\mathbf{x}$. Let k be the smallest integer such that if we take *any* subset $U \subseteq I_t(\mathbf{x})$ of size $|U| \geq k$, then $\pi_i(U) = x_i$ for all $i = 1,\ldots n$. In other words, it is enough to take any k codewords from $I_t(\mathbf{x})$ in order to find $\mathbf{x}$ using the majority voting on the coordinates of U. The smallest such k is denoted by $m_t(\mathbf{x}) = m_t(C;\mathbf{x})$. We say that $\mathbf{x}$ is *revealed* from $I_t(\mathbf{x})$ using any $m_t(\mathbf{x})$ (or more) words of $I_t(\mathbf{x})$.

Example 1. Let $C = \{0000, 0100, 1100, 0110, 0111, 1011\}$. For $\mathbf{x} = 0100$ we have $I_1(\mathbf{x}) = \{0000, 0100, 1100, 0110\}$. Clearly, now $\pi_i(I_1(\mathbf{x})) = x_i$ for all $i = 1,2,3,4$, so $\mathbf{x}$ is accessible. It is easy to check that any subset of three codewords of $I_1(\mathbf{x})$ also reveals $\mathbf{x}$ using the majority voting. Hence, $m_1(\mathbf{x}) \leq 3$. Since $U = \{0100, 1100\}$ gives $\pi_1(U) = *$, we get $m_1(\mathbf{x}) = 3$.

Let $\mathbb{N} = \{0,1,\ldots\}$ be the set of natural numbers. For a word $\mathbf{x} = x_1x_2\ldots x_n$ we define a vector $\mathbf{h}_t(\mathbf{x}) = \mathbf{h}_t(C;\mathbf{x}) = (h_1,h_2,\ldots,h_n) \in \mathbb{N}^n$ where h_i is the number of codewords in $I_t(\mathbf{x})$ such that their i-th coordinate differs from x_i. Hence, $\mathbf{x}$ is accessible, if

$$|I_t(\mathbf{x})| \geq 2 \max_{i=1,\ldots,n} h_i + 1 \tag{1}$$

and, in that case,

$$m_t(\mathbf{x}) = 2 \max_{i=1,\ldots,n} h_i + 1. \tag{2}$$

Definition 2. Let $t \geq 1$ and $n \geq 2$ be integers. A code $C \subseteq \mathbb{F}^n$ is a *coordinatewise revealing code of radius* t (a t-*revealing* code for short) if every word $\mathbf{x} \in \mathbb{F}^n$ is accessible. For such a code, denote the parameter $\hat{\mu}_t(C) = \max_{\mathbf{x}\in\mathbb{F}^n} m_t(C;\mathbf{x})$. Furthermore, let $\hat{\mu}_t(n)$ denote the minimum of $\hat{\mu}_t(C)$ over all t-revealing codes C in $\mathbb{F}^n$.

Example 3. Let $C = \mathbb{F}^3 \setminus \{000, 111\}$. For the word $\mathbf{z} = 000$, we get $\mathbf{h}_1(\mathbf{z}) = (1,1,1)$ and $|I_1(\mathbf{z})| = 3$. Due to (1) and (2) it follows that $m_1(\mathbf{z}) = 3$. For $\mathbf{y} = 001$, the vector $\mathbf{h}_1(\mathbf{y}) = (1,1,0)$ and $|I_1(\mathbf{y})| = 3$. Again $\mathbf{y}$ is accessible and $m_1(\mathbf{y}) = 3$. Similarly, one can check that $m_1(\mathbf{x}) = 3$

for all $\mathbf{x} \in \mathbb{F}^3$. Consequently, C is a 1-revealing code with $\hat{\mu}_1(C) = 3$. Later (in Theorem 8) we will see that $\hat{\mu}_1(3) = 3$.

Lemma 4. *Let C be a t-revealing code, and let $\mathbf{x}$ and $\mathbf{y}$ be any distinct words in $\mathbb{F}^n$.*

(i) *Then we have*

$$|I_t(\mathbf{x}) \cap I_t(\mathbf{y})| \leq \max\{m_t(\mathbf{x}), m_t(\mathbf{y})\} - 1. \quad (3)$$

(ii) *We also have*

$$|I_t(\mathbf{x}) \triangle I_t(\mathbf{y})| \geq 2. \quad (4)$$

Proof. (i) Because C is a t-revealing code, the values $m_t(\mathbf{x})$ and $m_t(\mathbf{y})$ exist. Assume, without loss of generality, that $m_t(\mathbf{y}) \geq m_t(\mathbf{x})$. Suppose to the contrary that $|I_t(\mathbf{x}) \cap I_t(\mathbf{y})| \geq \max\{m_t(\mathbf{x}), m_t(\mathbf{y})\} = m_t(\mathbf{y})$. Consider the codewords in $U = I_t(\mathbf{x}) \cap I_t(\mathbf{y})$. Since C is t-revealing, we know that any subset of $m_t(\mathbf{y})$ or more codewords of $I_t(\mathbf{y})$ — in particular, the set U — reveals $\mathbf{y}$ uniquely. Also these same codewords in U should reveal uniquely $\mathbf{x}$ because $|U| \geq m_t(\mathbf{x})$ and $U \subseteq I_t(\mathbf{x})$. However, this is a contradiction, since $\mathbf{x} \neq \mathbf{y}$. For the case (ii) see [1]. □

Next we consider the applications and the motivations of the codes defined above. The first application is the list decoding problem of Levenshtein's channel model [2], [3], which finds its original motivation in molecular biology and chemistry, where the usual redundancy method is not feasible, and it is also relevant for recent advanced storage technologies [4]. The second application is the information retrieval in associative memories [3], [5], [6], [7]. There are also applications in sensor networks [8], [9].

1) *The list decoding problem for the Levenshtein's channel model:* A codeword $\mathbf{x} \in C$ is transmitted through N channels where at most t errors can occur in each of them as illustrated in Figure 1. It is also assumed that $t > \lfloor (d_{\min}(C) - 1)/2 \rfloor$.

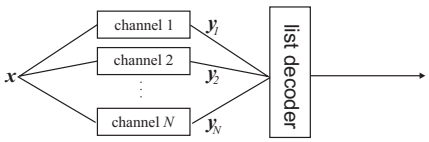

Fig. 1. The channel model.

Based on the N different outputs $\mathbf{y}_1, \ldots, \mathbf{y}_N$ of the channels, the list decoder $\mathcal{D}_{\mathcal{L}}$ gives estimations $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ (where $\ell \leq \mathcal{L}$) on the transmitted word $\mathbf{x}$. In [3], [6], a successful decoder is considered (successful means that the transmitted word $\mathbf{x}$ belongs to the outputted list) and the maximal length of the list $\mathcal{L}$ is considered with respect to the number of channels N. Naturally, we would like to have as short output list as possible while keeping N small and the cardinality of the code as large as possible. In [4], it is shown that if we wish to have a unique output (that is, $\mathcal{L} = 1$), then the number of channels can be inconveniently large — see also Remark 12.

In this paper, we will focus on the case when there are *only two* channels, that is, $N = 2$, and we try to find large codes giving a short output list from the decoder. For a general bound for the size of the code see (6). Suppose that C is a t-revealing code and $N = 2$. Next we see that we obtain a successful decoder with $\mathcal{L} \leq \hat{\mu}_t(C) - 1$. Two different words $\mathbf{y}_1$ and $\mathbf{y}_2$ are received from the channels and the decoder outputs all the codewords $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ of C such that $d(\mathbf{y}_j, \mathbf{x}_i) \leq t$ for all $j = 1, 2$ and $i = 1, \ldots, \ell$. In other words, the list consists of the codewords in $I_t(\mathbf{y}_1) \cap I_t(\mathbf{y}_2)$. By (3), the length of this list is at most $\hat{\mu}_t(C) - 1$. The decoder is clearly successful, since $\mathbf{x} \in I_t(\mathbf{y}_1) \cap I_t(\mathbf{y}_2)$ due to the fact that at most t errors occurred in the channels.

2) *Information retrieval in an associative memory:* In the model of Yaakobi and Bruck [3], an associative memory is given as a (simple and undirected) graph $G = (V, E)$. A vertex in the graph corresponds to a stored information unit and if two information units are associated, then there is an edge between them. Moreover, two vertices are called *t-associated*, if the graphical distance (that is, the number of edges) between them is at most t. An unknown information unit $x \in V$ is retrieved from the associative memory using *input clues* (provided by an information seeker) which are t-associated to x and also belong to a *reference set* $C \subseteq V$. The reference set should be such that given enough input clues, the sought information unit x can be unambiguously found. Naturally, we want the maximum number $\hat{m}$ of input clues, which are needed to retrieve any information unit from the memory, to be as small as possible. In this paper (like in [3], [6], [7]), we concentrate on the binary hypercube $\mathbb{F}^n$. Here two words (i.e., information units) $\mathbf{a}$ and $\mathbf{b}$ are t-associated if and only if $d(\mathbf{a}, \mathbf{b}) \leq t$. According to the model above, we wish to find a sought information unit $\mathbf{x}$ with the aid of input clues coming from the code C (the reference set) which are t-associated to the unknown word $\mathbf{x}$. In other words, the input clues come from the set $I_t(\mathbf{x})$. If the reference set C is a t-revealing code, then we can uniquely and efficiently (due to the majority voting) find the information unit by receiving at most $\hat{\mu}_t(C)$ input clues. Therefore, the maximum number of needed input clues satisfies $\hat{m} \leq \hat{\mu}_t(C)$. Here it is natural to have as small code as possible for the reference set.

Earlier in [3], [7], [6] the length $\mathcal{L}$ of the output of the list decoder and the maximum number of input clues $\hat{m}$ in an associative memory was considered using codes $C \subseteq \mathbb{F}^n$ which are based on limiting the size of the intersections $I_t(\mathbf{x}) \cap I_t(\mathbf{y})$ while the codes have the property that $I_t(\mathbf{x}) \setminus I_t(\mathbf{y}) \neq \emptyset$ for all $\mathbf{x} \neq \mathbf{y}$ (see, for instance, Theorem 9 in [6]). In this paper, we use the idea of majority voting on coordinates in designing the codes and not the intersections. But as we saw in (3), we can still estimate the intersections (needed, for example, in the list decoding problem as explained above).

We will see that the new class of t-revealing codes provides better results for the length $\mathcal{L}$ and for the number of input clues $\hat{m}$ than the earlier code constructions.

II. Linear codes and optimal results

We can often benefit from codes being linear, that is, $C = \{\mathbf{x} \in \mathbb{F}^n \mid H\mathbf{x}^T = \mathbf{0}\}$ where $H = (\mathbf{h}^{(1)} \mid \cdots \mid \mathbf{h}^{(n)})$ is

the check matrix. The *syndrome* of a word $\mathbf{y}$ is $s(\mathbf{y}) = H\mathbf{y}^T$ and a word of minimum weight in a coset $\mathbf{x}+C$ is the *coset leader*. Let us denote the cardinality of the size of the ball of radius t in $\mathbb{F}^n$ by $V(n,t)$.

For any subset $A \subseteq \mathbb{F}^n$ and a word $\mathbf{b} \in \mathbb{F}^n$ we define $d(\mathbf{b}, A) = \min\{d(\mathbf{b},\mathbf{a}) \mid \mathbf{a} \in A\}$ and $\mathbf{b}+A = \{\mathbf{b}+\mathbf{a} \mid \mathbf{a} \in A\}$. Next we will consider useful results regarding linear codes and the codes of type $\mathbf{x}+C$.

Theorem 5. (i) *Let $C \subseteq \mathbb{F}^n$ be code and $\mathbf{x} \in \mathbb{F}^n$. We have $\mathbf{h}_t(\mathbf{x}+C;\mathbf{y}) = \mathbf{h}_t(C;\mathbf{x}+\mathbf{y})$ and $|I_t(\mathbf{x}+C;\mathbf{y})| = |I_t(C;\mathbf{x}+\mathbf{y})|$ for all $\mathbf{y} \in \mathbb{F}^n$. If the word $\mathbf{x}+\mathbf{y}$ is accessible with respect to the code C, then $\mathbf{y}$ is accessible with respect to $\mathbf{x}+C$ and, moreover, $m_t(\mathbf{x}+C;\mathbf{y}) = m_t(C;\mathbf{y}+\mathbf{x})$.*

(ii) *Let C be a linear t-revealing code. Then $s(\mathbf{x}) = s(\mathbf{y})$ implies that $m_t(\mathbf{x}) = m_t(\mathbf{y})$. In particular, all the words in a coset have the same minimum number of revealing codewords as the coset leader.*

Proof. This can be obtained (see for details [1]) using the observation $I_t(\mathbf{x}+C;\mathbf{y}) = I_t(C;\mathbf{x}+\mathbf{y})+\mathbf{x}$. □

Shortening a code is a useful operation as we shall see.

Theorem 6. *Let $C \subseteq \mathbb{F}^n$ be a t-revealing code. Then the (p-times) shortened code C_p is also t-revealing and $\hat{\mu}_t(C) \geq \hat{\mu}_t(C_p)$ provided that for all $\mathbf{x} \in \mathbb{F}^{n-p}$ we have*

$$|I_t(C;\mathbf{x}0^p)| - \sum_{i=n-p+1}^{n} h_i \geq m_t(C;\mathbf{x}0^p) \tag{5}$$

where $\mathbf{h}_t(C;\mathbf{x}0^p) = (h_1,\ldots,h_n)$.

Next we give some constructions to revealing codes.

Theorem 7. *There exist codes giving*

(i) $\hat{\mu}_1(n) \leq 3$ *for all* $n \geq 3$,

(ii) $\hat{\mu}_2(n) \leq 3$ *for all* $n = 2^r - 1 - p$ *where* $r \geq 3$ *and* $0 \leq p \leq 2^{r-1}-3$.

Proof. (i) Consider first the radius $t = 1$. We will show that the linear code C with $r \times n$ check matrix H such that it contains every non-zero column (of $\mathbb{F}^r$) at least 3 times and there are no zero-columns in H is 1-revealing. Since every word of $\mathbb{F}^r$ appears as a column of H, the covering radius of C equals one and, therefore, the weight of any coset leader is at most one. In addition, $d_{\min}(C) = 2$. By Theorem 5(ii), it is enough to consider coset leaders when we want to calculate the values $m_t(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{F}^n$. Suppose first that the weight of the coset leader $\mathbf{x}$ equals zero, so in other words $\mathbf{x} = \mathbf{0}$. Since $d_{\min}(C) = 2$, we know that $I_1(\mathbf{0}) = \{\mathbf{0}\}$. Trivially, $\mathbf{h}_1(\mathbf{0}) = (0,0,\ldots,0)$, so $\max h_i = 0$ and thus, by (1) and (2), we get $m_1(\mathbf{0}) = 1$. Assume then that the coset leader $\mathbf{x}$ has weight one. Let the syndrome $s(\mathbf{x}) = \mathbf{s}$ (where $\mathbf{s} \neq \mathbf{0}$). Now the $I_1(\mathbf{x}) = \{\mathbf{x}+\mathbf{e}_i \mid i \in \mathcal{I}\}$ where $\mathcal{I}$ consists of all of those indices j for which the column $\mathbf{h}^{(j)} = \mathbf{s}$. Since H contains as a column each word of $\mathbb{F}^r$ at least three times, we get $|I_1(\mathbf{x})| \geq 3$. Now the vector $\mathbf{h}_1(\mathbf{x}) = (h_1,\ldots,h_n)$ is such that $h_i = 1$ for $i \in \mathcal{I}$ and $h_i = 0$ if $i \notin \mathcal{I}$. Therefore, by (1) and (2), we obtain $m_1(\mathbf{x}) = 3$. This yields that $\hat{\mu}_1(C) = 3$ and $\hat{\mu}_1(n) \leq 3$. The case (ii) comes similarly considering the Hamming code $\mathcal{H}_r$ of length 2^r-1 together with Theorem 6. □

The previous constructions in (i) and (ii) are *optimal* according to the next result.

Theorem 8. *For $t \geq 1$ and $n \geq 3$ we have $\hat{\mu}_t(n) \geq 3$.*

Proof. Let C be a t-revealing code in $\mathbb{F}^n$, $n \geq 3$. We show that $\hat{\mu}_t(C) \geq 3$ from which the claim follows. If there exists $\mathbf{c} \in C$ such that $I_t(\mathbf{c})$ contains at least two codewords, say $\mathbf{c}$ and $\mathbf{c}'$, then they both belong to the set $I_t(\mathbf{c}) \cap I_t(\mathbf{c}')$ and hence, by (3), we know that $m_t(\mathbf{c}) \geq 3$ or $m_t(\mathbf{c}') \geq 3$. Assume therefore, that for all $\mathbf{c} \in C$ we have $I_t(\mathbf{c}) = \{\mathbf{c}\}$. Choose any $\mathbf{x} \in B_1(\mathbf{c})$ with $\mathbf{x} \neq \mathbf{c}$. The words $\mathbf{c}$ and $\mathbf{x}$ differ in exactly one coordinate, say $c_i \neq x_i$. Now $\mathbf{h}_t(\mathbf{x}) = (h_1,\ldots,h_n)$ has $h_i \geq 1$ and hence $\max_{j=1,2\ldots,n} h_j \geq 1$. By (2), we obtain $m_t(\mathbf{x}) \geq 3$. This yields the assertion $\hat{\mu}_t(C) \geq 3$. □

The result $\hat{\mu}_2(n) = 3$ in Theorem 7(ii) together with Theorem 6 gives an infinite family of codes with the bound $\mathcal{L} = 2$ for the length of the decoder list and the bound $\hat{m} = 3$ for the maximal number of input clues in information retrieval. This improves on the earlier known constructions (see, e.g., [6], [7]), which provided the bounds $\mathcal{L} = 4$ and $\hat{m} = 5$, respectively.

Theorem 9. *(i) If a code $C \subseteq \mathbb{F}^n$ is such that the intersection of I-sets of any distinct words $\mathbf{x}$ and $\mathbf{y}$ satisfies $|I_t(\mathbf{x}) \cap I_t(\mathbf{y})| \leq \mathcal{L}'$, then there we have the upper bound*

$$|C| \leq \mathcal{L}' \frac{2^n}{V(n,t) - \binom{n-1}{t}}. \tag{6}$$

If C is a t-revealing code, then this bound holds for $\mathcal{L}' = \hat{\mu}_t(C) - 1$.

(ii) If C is t-revealing, we have a lower bound

$$|C| \geq \frac{3 \cdot 2^n}{V(n,t)+2}. \tag{7}$$

Proof. (i) For the upper bound, choose a set $S = B_t(\mathbf{0}) \cap B_t(\mathbf{e}_1)$. One obtains $\sum_{\mathbf{x} \in \mathbb{F}^n} |(\mathbf{x}+S) \cap C| = |S||C|$. Since $\mathbf{x}+S = B_t(\mathbf{x}) \cap B_t(\mathbf{x}+\mathbf{e}_1)$, and thus, $(\mathbf{x}+S) \cap C = I_t(\mathbf{x}) \cap I_t(\mathbf{x}+\mathbf{e}_1)$, we get by the assumption that $|(\mathbf{x}+S) \cap C| \leq \mathcal{L}'$. This implies that $2^n \mathcal{L}' \geq |S||C|$. For the claim (6) it suffices to notice that $|S| = V(n,t) - \binom{n-1}{t}$. By virtue of (3) we obtain the claim with $\mathcal{L}' = \hat{\mu}_t(C) - 1$ for a t-revealing code. For (ii) see [1]. □

Notice that the lower bound (7) can be attained (the small codes is what we prefer for the information retrieval). For example, the infinite family of codes in the proof of Theorem 7(i) for the lengths $n = 3(2^r-1)$ achieve the bound where $r \geq 1$. Indeed, each non-zero column of H appears exactly three times giving $|I_1(\mathbf{x})| = 3$ for non-codewords and $|I_1(\mathbf{x})| = 1$ for the codewords.

For $t = 2$ the above upper bound (6) gives for $\mathcal{L}' = 2$ that $|C| \leq 2^n/n$. The codes in Theorem 7(ii) give $\hat{\mu}_2(C) = 3$, so these codes satisfy $\mathcal{L}' = 2$. The ratio between the cardinality

of codes $\mathcal{H}_r$ in Theorem 7(ii) and the bound (6) approaches to 1 when n tends to infinity. Large codes is what we prefer for the Levenshtein's channel problem.

III. Optimal results for the radius $t = 3$

In this section, we consider the case of radius $t = 3$. Let $C_1 \subseteq \mathbb{F}^n$ and $C_2 \subseteq \mathbb{F}^n$ be codes (not necessarily revealing). We will utilize the following additive properties valid for all $t \geq 1$ and $\mathbf{x} \in \mathbb{F}^n$: if $C_1 \cap C_2 = \emptyset$, then

$$\mathbf{h}_t(C_1 \cup C_2; \mathbf{x}) = \mathbf{h}_t(C_1; \mathbf{x}) + \mathbf{h}_t(C_2; \mathbf{x})$$

and

$$|I_t(C_1 \cup C_2; \mathbf{x})| = |I_t(C_1; \mathbf{x})| + |I_t(C_2; \mathbf{x})|.$$

In Theorem 7(ii), we gave codes with minimum distance three and the radius was two. Recall that for the Levenshtein's channel problem, we have $t > \lfloor (d_{\min}(C) - 1)/2 \rfloor$. In the next theorem, we consider codes in the case where the minimum distance is three and the radius equals three also. These codes provide $\hat{\mu}_3(n) \leq 5$, which is shown to be optimal in Theorem 11. Moreover, the cardinality of the codes is large as pointed out in Remark 13.

Theorem 10. *We have $\hat{\mu}_3(n) \leq 5$ for $n = 2^{2r} - 1 - p$ where $r \geq 2$ and $0 \leq p \leq n/3 - 5$.*

Proof. Let the radius $t = 3$. Denote by $\mathcal{P}_r$ the punctured Preparata code [10, p. 51] of length $n = 2^{2r} - 1$ where $r \geq 2$. It is well-known that $d_{\min}(\mathcal{P}_r) = 5$ and $R(\mathcal{P}_r) = 3$. Let us first determine $m_3(\mathbf{x})$ for those words $\mathbf{x} \in \mathbb{F}^n$ that are accessible (not all are). Since $R(\mathcal{P}_r) = 3$, we know that $d(\mathbf{x}, \mathcal{P}_r) \leq 3$.

Let first $2 \leq d(\mathbf{x}, \mathcal{P}_r) \leq 3$. Since $\mathcal{P}_r$ is a nearly perfect code [10, p. 313], we have $|I_3(\mathcal{P}_r; \mathbf{x})| = n/3$. Let us consider $\mathbf{h}_3(\mathbf{x}) = (h_1, \ldots, h_n)$. We will see that $h_i \leq 1$ for all $i = 1, \ldots, n$. Indeed, suppose to the contrary that $h_i \geq 2$ for some i. Consequently, there are (at least) two codewords $\mathbf{c}$ and $\mathbf{c}'$ in $I_3(\mathcal{P}_r; \mathbf{x})$ such that they differ from $\mathbf{x}$ in the coordinate i. But now $d(\mathbf{c}, \mathbf{c}') \leq 4$ and this is a contradiction with $d_{\min}(\mathcal{P}_r) = 5$. Moreover, since $|I_3(\mathcal{P}_r); \mathbf{x})| = n/3$, in the vector $\mathbf{h}_3(\mathbf{x})$ all entries h_i are equal to 1 or exactly one is 0 and the others are 1. Therefore, by (1) and (2), we get $m_3(\mathbf{x}) = 3$.

Let then $0 \leq d(\mathbf{x}, \mathcal{P}_r) \leq 1$. If $\mathbf{x} \in \mathcal{P}_r$ we obtain $\mathbf{h}_3(\mathbf{x}) = (0, \ldots, 0)$ and $|I_3(\mathbf{x})| = 1$ due to the fact that the minimum distance is five. Thus, $m_3(\mathbf{x}) = 1$. If $d(\mathbf{x}, \mathcal{P}_r) = 1$, then $\mathbf{x}$ is *not* accessible (and $m_3(\mathbf{x})$ does not exist), since $I_3(\mathcal{P}_r; \mathbf{x}) = \{\mathbf{c}\}$ where $\mathbf{x} \neq \mathbf{c}$ and $\mathbf{h}_3(\mathbf{x})$ contains zeros except 1 in the position where $\mathbf{x}$ and $\mathbf{c}$ differ.

As we saw, there are three types of words in $\mathbb{F}^n$ with respect to the code $\mathcal{P}_r$. Those words which have $m_3(\mathbf{x}) = 3$ and $|I_3(\mathbf{x})| = n/3$ we call *type 3* words. The (code)words with $m_t(\mathbf{x}) = 1$ and $|I_3(\mathbf{x})| = 1$ are called *type 1* words. The rest of the words (the non-accessible ones) are of *type 0*.

In order to find a 3-revealing code we take advantage of the additive properties mentioned above and consider the code $C = \mathcal{P}_r \cup (\mathbf{g} + \mathcal{P}_r)$ where $\mathbf{g}$ is a word of weight three such that $d(\mathbf{g}, \mathcal{P}_r) = 3$ (for such words, see [11, p. 475]). Due to the fact that $d_{\min}(\mathcal{P}_r) = 5$ we have $\mathcal{P}_r \cap (\mathbf{g} + \mathcal{P}_r) = \emptyset$, so we can use the additive properties. By [11, p. 475], we know that $d_{\min}(C) = 3$.

Next we estimate $m_3(C; \mathbf{y})$ for $\mathbf{y} \in \mathbb{F}^n$ by considering the different types of the words. Using Theorem 5(i) we know that the words in $\mathbb{F}^n$ have the same three types with respect the code $\mathbf{g} + \mathcal{P}_r$ as they had in the code $\mathcal{P}_r$.

If a word $\mathbf{y}$ is of type 3 in $\mathcal{P}_r$ and also of type 3 in $\mathbf{g} + \mathcal{P}_r$, then by the additive properties we get $|I_3(C; \mathbf{y})| = |I_3(\mathcal{P}_r; \mathbf{y})| + |I_3(\mathbf{g} + \mathcal{P}_r; \mathbf{y})| = 2n/3$ and $\mathbf{h}_3(C; \mathbf{y}) = \mathbf{h}_3(\mathcal{P}_r; \mathbf{y}) + \mathbf{h}_3(\mathbf{g} + \mathcal{P}_r; \mathbf{y}) = (h_1, \ldots, h_n)$, where the maximal h_i is equal to 2. Consequently, $m_3(C; \mathbf{y}) = 5$. The case where $\mathbf{y}$ is of type 3 in $\mathcal{P}_r$ and of type 0 or 1 in $\mathbf{g} + \mathcal{P}_r$ goes similarly.

Now the only possibility left to be studied is when $\mathbf{y}$ is of type 0 or 1 in both of the subcodes of C. This means that there would be codewords $\mathbf{c} \in \mathcal{P}_r$ and $\mathbf{g} + \mathbf{c}' \in \mathbf{g} + \mathcal{P}_r$ such that $d(\mathbf{y}, \mathbf{c}) \leq 1$ and $d(\mathbf{y}, \mathbf{g} + \mathbf{c}') \leq 1$. But then we get $d(\mathbf{c}, \mathbf{g} + \mathbf{c}') \leq 2$, which contradicts the fact that $d_{\min}(C) = 3$. Therefore, there does not exist such a possibility for the word $\mathbf{y}$. Consequently, C is 3-revealing with the parameter $\hat{\mu}_3(C) \leq 5$. Hence $\hat{\mu}_3(n) \leq 5$ for $n = 2^{2r} - 1$, $r \geq 2$. In order to get the result for the lengths $n - p$, where $0 < p \leq n/3 - 5$, we use Theorem 6. □

The result $\hat{\mu}_3(n) \leq 5$ found in the previous theorem is actually *optimal* for $t = 3$ as will be seen next.

Theorem 11. *For $t \geq 3$ and $n \geq 5$ we have $\hat{\mu}_t(n) \geq 5$.*

Remark 12. For the radius $t = 3$, the construction in Theorem 10 gives an infinity family of codes with $\mathcal{L} = 4$ for the length of the list decoder and $\hat{m} = 5$ for the information retrieval. In earlier constructions, the best results (see, e.g., [6]) for $t = 3$ are $\mathcal{L} = 6$ and $\hat{m} = 7$. Recall that these results on the list decoding are for the case when we use only two channels, $N = 2$. If we would like to find the transmitted word uniquely [2] (that is, $\mathcal{L} = 1$) we would need as many as $N = 6n - 9$ channels to do that when the minimum distance is three as for the codes in Theorem 10.

Remark 13. The upper bound of Theorem 9 for the maximal size of intersection $\mathcal{L}' = 4$ and for lengths $n = 2^{2r} - 1$ equals

$$|C| \leq \frac{2^{4^r+1}}{16^r - 3 \cdot 4^r + 4}. \tag{8}$$

The codes of length $n = 2^{2r} - 1$ in Theorem 10 give $\mathcal{L}' = \hat{\mu}_t(C) - 1 = 4$ and the ratio between the cardinality of these codes and the bound (8) approaches to 1 as r tends to infinity. Therefore, these codes are good also in this respect for the Levenshtein's list decoding problem.

IV. More constructions

In this section, we will study how to get $(n-t-1)$-revealing codes from t-revealing ones. In addition, we discuss the use of direct sum $D = C \oplus \mathbb{F}$ when C is t-revealing.

Theorem 14. *Let $C \subseteq \mathbb{F}^n$ be such a t-revealing code that each coordinate has 0 in exactly half of the codewords in any given coordinate. Then C is also $(n - t - 1)$-revealing with*

$$m_{n-t-1}(\mathbf{x}) = |C| - 2|I_t(\bar{\mathbf{x}})| + m_t(\bar{\mathbf{x}})$$

for all $\mathbf{x} \in \mathbb{F}^n$. *In particular,* $\hat{\mu}_{n-4}(n) \leq 2^{n-4r+1} - 2(n/3 + 1) + 5$ *for all* $n = 2^{2r} - 1$, $r \geq 2$.

Proof. The first claim follows from $I_{n-t-1}(\mathbf{x}) = C \setminus I_t(\bar{\mathbf{x}})$. For $t = n - 4$ we use the codes from Theorem 10. □

If we have a t-revealing code of length n, then we can easily modify it into a t-revealing code of length $n+1$ when following conditions are met.

Theorem 15. *Let* $D = C \oplus \mathbb{F}$ *where* $C \subseteq \mathbb{F}^n$ *is* t*-revealing. Let further* $\mathbf{x} \in \mathbb{F}^n$ *and denote* $\mathbf{h}_{t-1}(C; \mathbf{x}) = (h'_1, \ldots, h'_n)$ *and* $\mathbf{h}_t(C; \mathbf{x}) = (h_1, \ldots, h_n)$. *The word* $\mathbf{x}0$ *is accessible with respect to* D *if and only if*

$$|I_t(C; \mathbf{x})| + |I_{t-1}(C; \mathbf{x})| \geq 2 \max_{i=1,\ldots,n} (h_i + h'_i) + 1$$

and $I_t(C; \mathbf{x}) \setminus I_{t-1}(C; \mathbf{x}) \neq \emptyset$.

Proof. If $I_t(C; \mathbf{x}) \setminus I_{t-1}(C; \mathbf{x}) = \emptyset$, then $\pi_{n+1}(\mathbf{x}0) = *$, since $|I_{t-1}(C; \mathbf{x})0| = |I_{t-1}(C; \mathbf{x})1|$. We have $|I_t(D; \mathbf{x}0)| = |I_t(C; \mathbf{x})| + |I_{t-1}(C; \mathbf{x})|$, $\mathbf{h}_t(D; \mathbf{x}0) = (h''_1, \ldots, h''_{n+1}) = (h_1 + h'_1, \ldots, h_n + h'_n, |I_{t-1}(C; \mathbf{x})|)$ and $|I_t(D; \mathbf{x}0)| \geq 2|I_{t-1}(C; \mathbf{x})| + 1$ if $I_t(C; \mathbf{x}) \setminus I_{t-1}(C; \mathbf{x}) \neq \emptyset$. We get the other condition from the inequality $|I_t(D; \mathbf{x}0)| \geq 2 \max_{i=1,\ldots,n+1} h''_i + 1$. □

We can simplify the conditions of Theorem 15 into following form.

Corollary 16. *If* $C \subseteq \mathbb{F}^n$ *is both* t*-revealing and* $(t-1)$*-revealing with* $I_t(C; \mathbf{x}) \setminus I_{t-1}(C; \mathbf{x}) \neq \emptyset$ *for each* $\mathbf{x} \in \mathbb{F}^n$, *then* $D = C \oplus \mathbb{F}$ *is* t*-revealing.*

We can loosen these conditions a bit for 2-revealing codes.

Theorem 17. *The code* $D = C \oplus \mathbb{F}$ *is* 2*-revealing if* $C \subseteq \mathbb{F}^n$ *is a* 2*-revealing code,* $I_2(C; \mathbf{x}) \setminus I_1(C; \mathbf{x}) \neq \emptyset$ *for each* $\mathbf{x} \in \mathbb{F}^n$ *and* $|I_1(C; \mathbf{x}')| \neq 1$ *for each non-codeword* $\mathbf{x}' \in \mathbb{F}^n$.

Proof. Since $I_2(C; \mathbf{x}) \setminus I_1(C; \mathbf{x}) \neq \emptyset$ for each $\mathbf{x} \in \mathbb{F}^n$, the last coordinate is voted correctly. Let $\mathbf{h}_1(C; \mathbf{x}') = (h'_1, \ldots, h'_n)$ and $\mathbf{h}_2(C; \mathbf{x}') = (h_1, \ldots, h_n)$. Because C is a 2-revealing code, $|I_2(C; \mathbf{x}')| \geq 2 \max_{i=1,\ldots,n} h_i + 1$. If $|I_1(C; \mathbf{x}')| \geq 2$, then $|I_2(C; \mathbf{x}')| + |I_1(C; \mathbf{x}')| \geq 2 \max_{i=1,\ldots,n} h_i + 3 \geq 2 \max_{i=1,\ldots,n} (h_i + h'_i) + 1$ for each non-codeword $\mathbf{x}'$. We have $h'_i \leq 1$ since inside a 1-radius ball at most one coordinate can change. If $|I_1(C; \mathbf{x}')| = 0$, then $\mathbf{h}_1(\mathbf{x}') = \mathbf{0}$. If $\mathbf{c} \in C$ and $|I_1(C; \mathbf{c})| = 1$, then $\mathbf{h}_1(\mathbf{x}') = \mathbf{0}$ and if $|I_1(C; \mathbf{c})| \geq 2$, then situation is similar as above. □

Lemma 18. *The code* $C = \mathcal{H}_r \cup \{\mathcal{H}_r + \mathbf{e}_1\}$ *is a* 2*-revealing code,* $I_2(C; \mathbf{x}) \setminus I_1(C; \mathbf{x}) \neq \emptyset$ *and* $|I_1(C; \mathbf{x})| = 2$ *for each* $\mathbf{x} \in \mathbb{F}^{2^r - 1}$.

Proof. $\mathcal{H}_r$ and $\mathcal{H}_r + \mathbf{e}_1$ are separate because $d_{\min}(\mathcal{H}_r) = 3$. Their covering radius is 1 so $|I_1(C; \mathbf{x})| = 2$. If $\mathbf{x} \notin \mathcal{H}_r$, then $\mathbf{c}, \mathbf{c}' \in I_2(\mathcal{H}_r; \mathbf{x})$. Now $d(\mathbf{c}, \mathbf{c}') = 3$, so $I_2(C; \mathbf{x}) \setminus I_1(C; \mathbf{x}) \neq \emptyset$ for each $\mathbf{x} \in \mathbb{F}^{2^r - 1}$. The same is true for $\mathbf{x} \notin \mathcal{H}_r + \mathbf{e}_1$. Because $\mathcal{H}_r$ and $\mathcal{H}_r + \mathbf{e}_1$ are separate and 2-revealing, their union is also 2-revealing since $\pi_i(I(\mathcal{H}_r; \mathbf{x})) = \pi_i(I(\mathcal{H}_r + \mathbf{e}_1; \mathbf{x})) = \pi_i(I(C; \mathbf{x}))$ for each i and each $\mathbf{x} \in \mathbb{F}^{2^r - 1}$. □

Therefore, if $C = \mathcal{H}_r \cup \{\mathcal{H}_r + \mathbf{e}_1\}$, then $D = C \oplus \mathbb{F}$ is a 2-revealing code. Consequently, we get the following result for lengths not covered in Theorem 7.

Theorem 19. *We have* $\hat{\mu}_2(2^r) \leq 7$ *where* $r \geq 3$.

Proof. Let $C = \mathcal{H}_r \cup \{\mathcal{H}_r + \mathbf{e}_1\}$ and $D = C \oplus \mathbb{F}$. Let $\mathbf{d} = \mathbf{c}0$, $\mathbf{c} \in C$ (we can assume $\mathbf{c} \in \mathcal{H}_r$), $\mathbf{h}_2(D; \mathbf{d}) = (h_1, \ldots, h_{2^r})$, $\mathbf{h}_2(C; \mathbf{c}) = (h'_1, \ldots, h'_{2^r - 1})$ and $\mathbf{h}_2(\mathcal{H}_r + \mathbf{e}_1; \mathbf{c}) = (h''_1, \ldots, h''_{2^r - 1})$. Now we have I-sets $I_1(D; \mathbf{d}) = \{\mathbf{c}0, \mathbf{c}1, (\mathbf{c} + \mathbf{e}_1)0\}$ and $I_2(D; \mathbf{d}) \setminus I_1(D; \mathbf{d}) = \{(\mathbf{c} + \mathbf{e}_1)1, I_2(C; \mathbf{c})0 \setminus I_1(C; \mathbf{c})0\}$. Furthermore $h_{2^r} = 2$, $h_1 = 2h''_1 = 2$ and $h_i = h'_i = h''_i = 1$ for each $2 \leq i \leq 2^r - 1$, so we have $m_2(D; \mathbf{d}) = 5$.

Let $\mathbf{x}0 \notin D$ and $\mathbf{h}(D; \mathbf{x}0) = (h_1, \ldots, h_{2^r})$. Now $I_1(D; \mathbf{x}0) = I_1(C; \mathbf{x})0$ and $I_2(D; \mathbf{x}0) \setminus I_1(D; \mathbf{x}0) = I_1(C; \mathbf{x})1 \cup (I_2(C; \mathbf{x}) \setminus I_1(C; \mathbf{x}))0$. Since $|I_1(C; \mathbf{x})| = 2$, we have $h_{2^r} = 2$. Because $\mathbf{h}_2(\mathcal{H}_r + \mathbf{e}_1; \mathbf{x}) = \mathbf{h}_2(\mathcal{H}_r; \mathbf{x} + \mathbf{e}_1) = \mathbf{1}$ for a non-codeword $\mathbf{x}$, we have $h_i \leq 2 + 1$, for each $1 \leq i \leq 2^r$, since the error in 1-neighbourhood is doubled and the same code $\mathcal{H}_r$ or $\mathcal{H}_r + \mathbf{e}_1$ cannot also have error in the same coordinate at distance two from $\mathbf{x}$. This gives us upper bound $\hat{\mu}_2(2^r) \leq 7$. □

Finally we have following condition for lengthening 1-revealing codes.

Theorem 20. *The code* $D = C \oplus \mathbb{F}$ *is a* 1*-revealing code if each codeword in* $C \subseteq \mathbb{F}^n$ *is neighbouring another codeword and* C *is* 1*-revealing.*

Proof. Let $\mathbf{x} \in \mathbb{F}^n$. If $\mathbf{x} \notin C$, then $I_1(D; \mathbf{x}0) = I_1(C; \mathbf{x})0$ and because C is 1-revealing and the last coordinate is always 0 in $I_1(D; \mathbf{x}0)$, $\mathbf{x}0$ is accessible in $\mathbb{F}^{n+1}$. If $\mathbf{x} \in C$, then $I_1(D; \mathbf{x}0) = I_1(C; \mathbf{x})0 \cup \{\mathbf{x}1\}$. We have $|I_1(C; \mathbf{x})0| > 1$ because there is a neighbouring codeword, so the last coordinate is voted correctly. Other coordinates are known since C is 1-revealing. We can deduce $\mathbf{x}1$ similarly. □

REFERENCES

[1] T. Laihonen, "On t-revealing codes in binary Hamming spaces," submitted for publication.

[2] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 2–22, 2001.

[3] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," in *Proceedings of ISIT 2012*, pp. 106–110, 2012.

[4] E. Yaakobi, J. Bruck, and P. Siegel, "Constructions and decoding of cyclic codes over b-symbol read channels," *IEEE Trans. Inform. Theory*, vol. 62, no. 4, pp. 1541–1551, 2016.

[5] E. Yaakobi, M. Schwartz, M. Langberg, and J. Bruck, "Sequence reconstruction for grassmann graphs and permutations," in *Proceedings of ISIT 2013*, pp. 874–878, 2013.

[6] V. Junnila and T. Laihonen, "Information retrieval with varying number of input clues," *IEEE Trans. Inform. Theory*, vol. 62, 2016.

[7] V. Junnila and T. Laihonen, "Codes for information retrieval with small uncertainty," *IEEE Trans. Inform. Theory*, vol. 60, 2014.

[8] N. Fazlollahi, D. Starobinski, and A. Trachtenberg, "Connected identifying codes," *IEEE Trans. Inform. Theory*, vol. 58, 2012.

[9] G. Cohen, I. Honkala, A. Lobstein, and G. Zémor, "New bounds for codes identifying vertices in graphs," *Electron. J. Combin.*, vol. 6, pp. Research Paper 19, 14 pp., 1999.

[10] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering codes, North-Holland,1997.*

[11] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes, North-Holland, 1977.*