AUTHOR	Ville Junnila, Tero Laihonen, Tuomo Lehtilä
TITLE	On the List Size in the Levenshtein's Sequence Reconstruction Problem
YEAR	2022
DOI	10.1109/ISIT50566.2022.9834612
VERSION	Author's accepted manuscript
CITATION	V. Junnila, T. Laihonen and T. Lehtilä, "On the List Size in the Levenshtein's Sequence Reconstruction Problem," 2022 IEEE International Symposium on Information Theory (ISIT), 2022, pp. 510-515, doi: 10.1109/ISIT50566.2022.9834612.

On the List Size in the Levenshtein's Sequence Reconstruction Problem

Ville Junnila
Department of Mathematics
and Statistics
University of Turku
Turku FI-20014, Finland
Email: viljun@utu.fi

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

Tuomo Lehtilä
LIRIS
Université Claude Bernard Lyon 1
Lyon, France
Email: tualeh@utu.fi

***Abstract*—In the paper, the Levenshtein's sequence reconstruction problem is considered in the case where at most t substitution errors occur in each of the N channels and the decoder outputs a list of length at most $\mathcal{L}$. Moreover, it is assumed that the transmitted words are chosen from an e-error-correcting code C $(\subseteq \{0,1\}^n)$. Previously, when $t = e + \ell$ and the length n of the transmitted word is large enough, the exact numbers of required channels is determined for $\mathcal{L} = 1, 2$ and $\ell + 1$. Here we determine the number of channels in the cases $\mathcal{L} = 3, 4, \ldots, \ell$. Furthermore, with the aid of covering codes, we also consider the list sizes in the cases where the length n is rather small. Finally, the majority algorithm is discussed for decoding; in particular, we demonstrate that with high probability a decoder based on it, is verifiably successful, i.e., outputs a list (sometimes even of size one) such that it contains the transmitted word.**

***Index Terms*—Levenshtein's Sequence Recontruction, Information Retrieval, Substitution Errors, Majority Algorithm.**

I. Introduction

In this paper, the Levenshtein's *sequence reconstruction problem*, introduced in [1], is studied when the errors are substitution errors. For related sequence reconstruction problems (concerning, for instance, deletion and insertion errors) consult, for example, [1]–[6]. Originally, the motivation for the sequence reconstruction problem came from biology and chemistry where the familiar redundancy method of error correction is unsuitable. The sequence reconstruction problem has returned to the focus, as it was recently pointed out that the problem is highly relevant to information retrieval in advanced storage technologies where the stored information is either a single copy, which is read many times, or it has several copies [4], [7]. This problem (see [4]) is especially applicable to DNA data storage systems (see [8]–[11]) where DNA strands provide numerous erroneous copies of the information and the goal is to recover the information using these copies.

Let us denote the set $\{1, 2, \ldots, n\}$ by $[1, n]$, by $\mathbb{F}$ the finite field of two elements, and by $\mathbb{F}^n$ the *binary Hamming space*. The *support* of the word $\mathbf{x} = x_1 \ldots x_n \in \mathbb{F}^n$ is defined as $\mathrm{supp}(\mathbf{x}) = \{i \mid x_i \neq 0\}$. Let us denote the zero word $\mathbf{0} = 00 \ldots 0 \in \mathbb{F}^n$ and by $\mathbf{e}_i \in \mathbb{F}^n$ a word with $\mathrm{supp}(\mathbf{e}_i) = \{i\}$. The *Hamming weight* $w(\mathbf{x})$ of $\mathbf{x} \in \mathbb{F}^n$ is $|\mathrm{supp}(\mathbf{x})|$. The *Hamming distance* is defined as $d(\mathbf{x}, \mathbf{y}) = w(\mathbf{x} + \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$. Let us denote the radius t *Hamming ball* centered at $\mathbf{x} \in \mathbb{F}^n$ by $B_t(\mathbf{x}) = \{\mathbf{y} \in \mathbb{F}^n \mid d(\mathbf{x}, \mathbf{y}) \leq t\}$ and $|B_t(\mathbf{x})|$ by $V(n, t) = \sum_{i=0}^{t} \binom{n}{i}$. *Code* is a nonempty subset of $\mathbb{F}^n$ and its elements are called *codewords*. The *minimum distance* of a code $C \subseteq \mathbb{F}^n$ is $d_{\min}(C) = \min_{\mathbf{c}_1, \mathbf{c}_2 \in C, \mathbf{c}_1 \neq \mathbf{c}_2} d(\mathbf{c}_1, \mathbf{c}_2)$. Moreover, the code C has the error-correcting capability $e = e(C) = \lfloor (d_{\min}(C) - 1)/2 \rfloor$.

Next we consider the sequence reconstruction problem. For the rest of the paper, let $C \subseteq \mathbb{F}^n$ be any e-error-correcting code. A codeword $\mathbf{x} \in C$ is transmitted through N channels where, in each of them, at most t substitution errors can occur. In the sequence reconstruction problem, our aim is to reconstruct $\mathbf{x}$ based on the N different outputs $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ from the channels (see Fig. 1).

It is assumed that $t > e(C)$ (if $t \leq e(C)$, then only one channel is enough to reconstruct $\mathbf{x}$). For $\ell \geq 1$, let us denote

$$t = e(C) + \ell = e + \ell$$

for the rest of paper. The situation where we obtain sometimes a short list of possibilities for $\mathbf{x}$ instead of always recovering $\mathbf{x}$ uniquely, is considered in [12], [13]. Based on the set Y and the code C, the list decoder (see Fig. 1) $\mathcal{D}$ gives an estimation $T_{\mathcal{D}} = T_{\mathcal{D}}(Y) = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|T_{\mathcal{D}}|}\}$ on the sequence $\mathbf{x}$ which we try to reconstruct. We denote by $\mathcal{L}_{\mathcal{D}}$ the maximum cardinality of the list $T_{\mathcal{D}}(Y)$ over all possible sets Y of output words. The decoder is said to be *successful* if $\mathbf{x} \in T_{\mathcal{D}}$. In this paper, we focus on the smallest possible value of $\mathcal{L}_{\mathcal{D}}$ over all successful

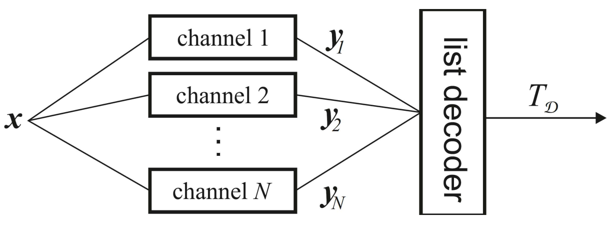

Fig. 1. The Levenshtein's sequence reconstruction.

The first and second author were funded in part by the Academy of Finland grant 338797 and the third author was funded by the Finnish Cultural Foundation.

decoders $\mathcal{D}$, i.e., on $\mathcal{L} = \min_{\mathcal{D} \text{ is successful}}\{\mathcal{L}_{\mathcal{D}}\}$. Let us denote

$$T = T(Y) = C \cap (\bigcap_{\mathbf{y} \in Y} B_t(\mathbf{y})).$$

Consequently,

$$\mathcal{L} = \max\{|T(Y)| \mid Y \text{ is a set of } N \text{ output words}\}.$$

The value $\mathcal{L}$ depends on e, ℓ, n and N. Obviously, one would like to have as small $\mathcal{L}$ as possible. This problem is studied, for example, in [12]–[17]. In this paper, we mainly consider the relation between N and $\mathcal{L}$ for various n after we fix two parameters ℓ and e (while letting C be any e-error-correcting code). The sequence reconstruction problem is also closely related (see [14]) to *information retrieval in associative memory* introduced by Yaakobi and Bruck [12], [13].

The structure of the paper is as follows. In Section II, we recall some of the known results. In particular, it is pointed out that if we have at least (resp. less than) $V(n, \ell-1)+1$ channels, then the list size is constant with respect to n (resp. there are e-error-correcting codes with list size depending on n). In Section III, we give the complete correspondence between the list size and the number of channels when we have more than $V(n, \ell-1)+1$ channels and n is large enough. It is sometimes enough to increase the number of channels only by a constant amount in order to decrease the list size (see Corollary 10). Section IV focuses on improving the bounds on the list size when n is not restricted and we obtain strictly more channels than $V(n, \ell-1)+1$. Section V is devoted to list size when we have *less* than $V(n, \ell-1)+1$ channels. The final section deals with the reconstruction with the aid of a majority algorithm on the coordinates among the output words in Y. Some of the proofs are omitted due to the lack of space.

II. Known results

In this section we present some known results on how the two values N and $\mathcal{L}$ are linked. The basic idea on estimating $\mathcal{L}$ is the following: we analyse the maximum number of output words (N) we can fit in the intersection of $\mathcal{L}$ t-radius balls centered at codewords. As expected, the length $\mathcal{L}$ of the outputted list strongly depends on the number of channels.

Previously, in [1] and [12], the problem has been considered for $\mathcal{L} = 1$ and $\mathcal{L} = 2$, respectively. Moreover, in [18], the exact number of channels N required to have $\mathcal{L}$ constant on n has been presented, see Theorems 3 and 4. Following theorem gives an exact number of channels required to have $\mathcal{L} = 1$.

Theorem 1 ([1]). *We have $\mathcal{L} \leq 1$ if*

$$N \geq \sum_{i=0}^{\ell-1} \binom{n-2e-1}{i} \sum_{k=e+1+i-\ell}^{t-i} \binom{2e+1}{k} + 1.$$

The next result is a a reformulation of a result by Yaakobi and Bruck [12, Algorithm 18] proven in [18].

Theorem 2. *Let $n \geq 2\ell - 1$ and C be an e-error-correcting code in $\mathbb{F}^n$. If $N \geq V(n, \ell-1)+1$, then we have*

$$\mathcal{L} \leq \binom{2\ell}{\ell}.$$

The bound in Theorem 2 can be improved to 2^ℓ which has been shown to be tight in [18].

Theorem 3 ([18]). *Let $n \geq \ell$ and C be an e-error-correcting code in $\mathbb{F}^n$. If $N \geq V(n, \ell-1)+1$, then we have $\mathcal{L} \leq 2^\ell$.*

Besides the 2^ℓ part, also the value $V(n, \ell-1)+1$ for the number of channels is tight, that is, if the value for N is less, then list size $\mathcal{L}$ can be linear with respect to n.

Theorem 4 (Theorem 10, [18]). *If $N \leq V(n, \ell-1)$, then there exists an e-error-correcting code such that $\mathcal{L} \geq \lfloor n/(e+1) \rfloor$.*

Let us denote for the rest of the paper $n(e, \ell, b) = (\ell - 1)^2 \left(b - e + (e+1)\left(b - 3e - 2e^2 + eb + \binom{b-2e-1}{2}\right)\right) + \ell - 2$. Although the bound for $\mathcal{L}$ in Theorem 3 cannot be improved in general, we can improve it, when n is large, to $\ell + 1$.

Theorem 5 (Theorem 20, [18]). *Let $n \geq n(e, \ell, b)$, $b = \max\{3t, 4e+4\}$, $|Y| = N \geq V(n, \ell-1)+1$ and C be an e-error-correcting code. Then we have*

$$\mathcal{L} \leq \ell + 1.$$

Moreover, the bound $\ell + 1$ is tight.

Theorem 6 (Theorem 9, [18]). *There exists an e-error-correcting code $C \subseteq \mathbb{F}_2^n$ such that $\mathcal{L} \geq \ell + 1$ if $n \geq \ell + \ell e + e$ and the number of channels satisfies $N \leq V(n, \ell-1)+1$.*

Finally, in [12, Theorem 6], the authors have given exact number of channels required to have $\mathcal{L} \leq 2$. All in all, these three values for N are all we know when $\mathcal{L}$ is constant on n. In the following section, we give the missing values for N.

III. List size with more channels

In this section, we give exact bounds for the number of channels $N = N_h + 1$ (when n is large) which is required that $\mathcal{L} < h$ for every constant value h. Previously, N_h was known only for three values $h = 2, 3, \ell+2$, *i.e.*, $\mathcal{L} = 1, 2, \ell+1$. In Theorem 9, we give a solution for $\mathcal{L} = 3, \dots, \ell$. To achieve this, we need to introduce two technical lemmas from [18].

In the following lemma, when n is large, it is shown that any three codewords in $T(Y)$ differ within some subset of coordinates $\overline{D}$ of size constant size b and there exists an output word $\mathbf{y}$ which differs from these codewords in at least $\ell - 1$ coordinate positions outside of $\overline{D}$. Notice that supp$(\mathbf{w} + \mathbf{z})$ gives the set of coordinates in which $\mathbf{w}$ and $\mathbf{z}$ differ.

Lemma 7 (Lemma 18, [18]). *Let $b \geq 3t$ be an integer and C_1 be an e-error-correcting code. Assume that $n \geq n(e, \ell, b)$, $|Y| = N \geq V(n, \ell-1)+1$, $|T(Y)| \geq 3$ and $\mathbf{c}_0, \mathbf{c}_1, \mathbf{c}_2 \in T(Y)$. If now $\overline{D} \subseteq [1, n]$ is a set such that $|\overline{D}| = b$ and*

$$\mathit{supp}(\mathbf{c}_0 + \mathbf{c}_1) \cup \mathit{supp}(\mathbf{c}_0 + \mathbf{c}_2) \cup \mathit{supp}(\mathbf{c}_1 + \mathbf{c}_2) \subseteq \overline{D},$$

then for any word $\mathbf{w} \in \mathbb{F}^n$ we have $\mathit{supp}(\mathbf{w} + \mathbf{c}_0) \setminus \overline{D} = \mathit{supp}(\mathbf{w} + \mathbf{c}_1) \setminus \overline{D} = \mathit{supp}(\mathbf{w} + \mathbf{c}_2) \setminus \overline{D}$ and there exists an output word $\mathbf{y} \in Y$ such that

$$|\mathit{supp}(\mathbf{y} + \mathbf{c}_0) \setminus \overline{D}| \geq \ell - 1.$$

The following lemma shows that the distance between any codewords in $T(Y)$ is either $2e+1$ or $2e+2$.

Lemma 8 (Lemma 19, [18]). *Let $n \geq n(e,\ell,3t)$, $|Y| = N \geq V(n,\ell-1)+1$, C be an e-error-correcting code and $|T(Y)| \geq 3$. Then we have $d(\mathbf{c}_1,\mathbf{c}_2) \leq 2e+2$ for any two $\mathbf{c}_1,\mathbf{c}_2 \in T(Y)$.*

Let us denote by $N(n,\ell,e,h) = N_h$ (when the exact formulation is not necessary for clarity) the maximum number of t-error channels such that there exists a set of output words $Y \subseteq \mathbb{F}^n$ satisfying $|Y| = N_h$ and $|T(Y)| \geq h$ for some e-error-correcting code C. By Theorems 4 and 5 $N_{\ell+2} = N_{\lfloor n/(e+1)\rfloor} = V(n,\ell-1)$. Observe that in general, if $N \geq N_h+1$, then $\mathcal{L} < h$ for all e-error-correcting codes.

We require the following two technical notations. Let

$$W_w = \left\{(i_1,\dots,i_{\mathcal{L}}) \in \mathbb{N}^{\mathcal{L}} \mid \frac{w+1-\ell}{2} \leq i_j \leq e+1 \text{ and } w \geq \sum_{j=1}^{\mathcal{L}} i_j\right\}$$

and

$$W'_w = \left\{(i_1,\dots,i_{\mathcal{L}}) \in \mathbb{N}^{\mathcal{L}} \mid \frac{w-\ell}{2} \leq i_1 \leq e \text{ and for}\right.$$
$$\left. j \geq 2 : \frac{w+1-\ell}{2} \leq i_1 \leq e+1 \text{ and } w \geq \sum_{j=1}^{\mathcal{L}} i_j\right\}.$$

In the following theorem, we give the maximum number of channels N_h which gives list size $\mathcal{L} \geq h$.

Theorem 9. *Let $n \geq n(e,\ell,b)$, $b \geq \max\{3t, 4e+4\}$, $\ell \geq 3$, $3 \leq h \leq \ell+1$, $|D| = \mathcal{L}(e+1)$ and $|D'| = \mathcal{L}(e+1)-1$. Then*

$$N(n,\ell,e,h) = N_h = V(n,\ell-1)+$$
$$\max\left\{\sum_{w\geq\ell}\sum_{(i_1,\dots,i_{\mathcal{L}})\in W_w}\binom{n-|D|}{w-\sum_{j=1}^{\mathcal{L}} i_j}\prod_{j=1}^{\mathcal{L}}\binom{e+1}{i_j},\right.$$
$$\left.\sum_{w\geq\ell}\sum_{(i_1,\dots,i_{\mathcal{L}})\in W'_w}\binom{n-|D'|}{w-\sum_{j=1}^{\mathcal{L}} i_j}\binom{e}{i_1}\prod_{j=2}^{\mathcal{L}}\binom{e+1}{i_j}\right\}.$$

Proof. Let us have $N = N(n,\ell,e,h)$, $n \geq n(e,\ell,b)$, $b \geq \max\{3t, 4e+4\}$, $\ell \geq 3$ and $3 \leq h \leq \ell+1$. Moreover, let C be such an e-error-correcting code, that it maximizes $\mathcal{L}$ (we have $\mathcal{L} \geq h$) when $N = N(n,\ell,e,h)$ and let Y be a set of outputs such that $|T(Y)| = \mathcal{L}$ and let us denote $T(Y) = \{\mathbf{c}_1,\dots,\mathbf{c}_{\mathcal{L}}\}$.

Since C is an e-error-correcting code and by Lemma 8, we have $d(\mathbf{c}_i,\mathbf{c}_j) \in \{2e+1, 2e+2\}$ for each $i \neq j$. Since $h \geq 3$, each pairwise distance cannot be $2e+1$. Let us assume w.l.o.g. that $d(\mathbf{c}_1,\mathbf{c}_2) = 2e+2$ and let us then translate the Hamming space so that $\mathbf{c}_1 = \mathbf{0}$. Now $w(\mathbf{c}_2) = 2e+2$ and $w(\mathbf{c}_3) \in \{2e+1, 2e+2\}$. Moreover, $|\mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_3)| = e+1$. Let $\overline{D}$ be any subset of $[1,n]$ satisfying $\mathrm{supp}(\mathbf{c}_1+\mathbf{c}_2)\cup\mathrm{supp}(\mathbf{c}_1+\mathbf{c}_3)\cup\mathrm{supp}(\mathbf{c}_2+\mathbf{c}_3) \subseteq \overline{D}$ and $|\overline{D}| = b$. Observe that $\mathrm{supp}(\mathbf{c}_1)\setminus\overline{D} = \mathrm{supp}(\mathbf{c}_2)\setminus\overline{D} = \mathrm{supp}(\mathbf{c}_3)\setminus\overline{D} = \emptyset$.

By Lemma 7, there exists an output word $\mathbf{y} \in Y$ such that $|\mathrm{supp}(\mathbf{c}_1+\mathbf{y})\setminus\overline{D}| \geq \ell-1$. Since $d(\mathbf{y},\mathbf{c}_2) \leq t$, we have $|\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_2)| \geq e+1$. Moreover, since $d(\mathbf{y},\mathbf{c}_1) \leq t$, we have $w(\mathbf{y}) \leq t$ and hence, $|\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_2)| = e+1$. Thus, $\mathrm{supp}(\mathbf{y}) = (\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_2))\cup(\mathrm{supp}(\mathbf{y})\setminus\overline{D})$ and $\mathrm{supp}(\mathbf{y})\cap(\mathrm{supp}(\mathbf{c}_3)\setminus\mathrm{supp}(\mathbf{c}_2)) = \emptyset$. Hence, $\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_3) \subseteq \mathrm{supp}(\mathbf{c}_2)$ and moreover, $\mathrm{supp}(\mathbf{y})\cap(\mathrm{supp}(\mathbf{c}_2)\setminus\mathrm{supp}(\mathbf{c}_3)) = \emptyset$ as otherwise $d(\mathbf{y},\mathbf{c}_3) \geq 1+1+e+(l-1) = t+1 > t$ (a contradiction). Together these give that $\mathrm{supp}(\mathbf{y})\cap\overline{D} = \mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_3)$. Notice that for each $i \in [4,\mathcal{L}]$ we may choose $\overline{D}$ in such a way that also $\mathrm{supp}(\mathbf{c}_i) \subseteq \overline{D}$ since $|\overline{D}| = b \geq 4e+4$. Thus, there exists an output word $\mathbf{y}' \in Y$ such that $|\mathrm{supp}(\mathbf{c}_1+\mathbf{y}')\setminus\overline{D}| \geq \ell-1$. Therefore, as above, $\mathrm{supp}(\mathbf{y}')\cap\overline{D} = \mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_i)$ and $\mathrm{supp}(\mathbf{y}')\cap\overline{D} = \mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_3)$ implying $\mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_i) = \mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_3)$. Finally, translate the Hamming space so that the word $\mathbf{z}$ with $\mathrm{supp}(\mathbf{z}) = \mathrm{supp}(\mathbf{c}_2)\cap\mathrm{supp}(\mathbf{c}_3)$ becomes $\mathbf{z} = \mathbf{0}$. Then we have $w(\mathbf{c}_i) \in \{e, e+1\}$ and $\mathrm{supp}(\mathbf{c}_i)\cap\mathrm{supp}(\mathbf{c}_j) = \emptyset$ for each $i \neq j$ since $d(\mathbf{c}_i,\mathbf{c}_j) \in \{2e+1, 2e+2\}$. Moreover, at most one of $\mathbf{c}_i$ can have weight e by the minimum distance of C.

Let us then count the number of words in $\bigcap_{i=1}^{\mathcal{L}} B_t(\mathbf{c}_i)$. Clearly, each word $\mathbf{y}$ with $w(\mathbf{y}) \leq \ell-1$ belongs to the intersection contributing $V(n,\ell-1)$ words to it. Assume then that $w(\mathbf{y}) = w \geq \ell$. As $d(\mathbf{y},\mathbf{c}_j) \leq t$ for all $j \in [1,\mathcal{L}]$, we have $w(\mathbf{y})+w(\mathbf{c}_j)-2|\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_j)| \leq t$. Denote $i_j = |\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_j)|$. Assume first that $w(\mathbf{c}_j) = e+1$ for all j. Then $\mathbf{y} \in B_t(\mathbf{c}_j)$ if and only if we have $w+e+1-2i_j \leq t$. Hence, $e+1 \geq i_j \geq (w+1-\ell)/2$. Moreover, $\sum_{j=1}^{\mathcal{L}} i_j \leq w$ since $w(\mathbf{y}) = w$ and $\mathrm{supp}(\mathbf{c}_{j_1})\cap\mathrm{supp}(\mathbf{c}_{j_2}) = \emptyset$ for each $j_1 \neq j_2$. In other words, $\mathbf{y} \in \bigcap_{i=1}^{\mathcal{L}} B_t(\mathbf{c}_i)$ if and only if $(i_1,\dots,i_{\mathcal{L}}) \in W_w$. In the case where $w(\mathbf{c}_k) = e$ for some k, say $k=1$, we have $e \geq i_1 \geq (w-\ell)/2$. Thus, $\mathbf{y} \in \bigcap_{i=1}^{\mathcal{L}} B_t(\mathbf{c}_i)$ if and only if $(i_1,\dots,i_{\mathcal{L}}) \in W'_w$. Together these give the claim. □

Observe that by the proof the bounds given in Theorem 9 are tight. If we increase N by one, then $\mathcal{L}$ decreases by at least one since we cannot place the output word within the intersection of t-balls centered at codewords in $T(Y)$. Notice that geometrically the output sets giving maximal list size are more complicated than, for example, in Theorem 4 (where a ball of volume $V(n,\ell-1)$ is essential). Another observation is that although the sums do not include an upper bound for w, there is one. Namely the definition, for W_w, gives that $w \leq 2e+\ell+1$ and for W'_w that $w \leq 2e+\ell$.

Theorem 9 allows improving the bound $\mathcal{L} \leq \ell+1$ of Theorem 5 just by increasing N by constant number $(e+1)^{\ell+1}$.

Corollary 10. *Let $n \geq n(e,\ell,b)$, $b \geq \max\{3t, 4e+4\}$ and $\ell \geq 3$. If $N \geq V(n,\ell-1)+(e+1)^{\ell+1}+1$, then $\mathcal{L} \leq \ell$.*

Proof (sketch). First calculate the value N_h of Theorem 9 with $h = \ell+1$. For this purpose, first consider the set W_w with $w \geq \ell$. We get that each $i_j \geq 1$ and $\ell+1 \leq w \leq \ell+1$ as $\mathcal{L} \geq h = \ell+1$. Indeed, if $w \geq \ell+2$, then $w \geq \sum_{j=1}^{\mathcal{L}} i_j \geq \sum_{j=1}^{\ell+1} i_j \geq (\ell+1)(w+1-\ell)/2 = (\ell-1)(w+1-\ell)/2+w-(\ell-1) > w$ (a contradiction). Therefore, as $w = \ell+1$ and $i_j \geq 1$, we have $\mathcal{L} = \ell+1$ implying $W_w = \{(1,1,...,1)\}$. Thus, the sum corresponding to W_w in Theorem 9 gives $V(n,\ell-1)+(e+1)^{\ell+1}$. Analogously, for W'_w, we obtain that $\ell \leq w \leq \ell+1$, $W'_\ell = \{(0,1,1,\dots,1)\}$ and $W'_{\ell+1} = \{(1,1,1,\dots,1)\}$. Hence, the corresponding sum is equal to $V(n,\ell-1)+e(e+$

$1)^\ell + (e+1)^\ell = V(n, \ell-1) + (e+1)^{\ell+1}$. Thus, in conclusion, if $N \geq N_h + 1 = V(n, \ell-1) + (e+1)^{\ell+1} + 1$, then $\mathcal{L} \leq \ell$. □

IV. New Bounds with the aid of Covering Codes

Notice that although we have the bound $\mathcal{L} \leq \ell + 1$ when n is rather large (see Theorems 5), for smaller lengths of the codes our best bound is still $\mathcal{L} \leq 2^\ell$ (see Theorem 3) when the number of channels satisfies $N \geq V(n, \ell-1) + 1$. Although this bound is attained in some cases (see [18]) and thus cannot be improved in general, we can try to get a smaller list size $\mathcal{L}$ when we increase the number of channels. Indeed, recall that in [12, Theorem 6] the authors give a (fairly large) number of channels (which depends also on e whereas $N \geq V(n, \ell - 1) + 1$ does not) such that $\mathcal{L} \leq 2$. In this section, we utilize covering codes when we increase the number of channels. A code $C \subseteq \mathbb{F}^n$ is an *R-covering code* if for every word $\mathbf{x} \in \mathbb{F}^n$ there exists a codeword $\mathbf{c} \in C$ such that $d(\mathbf{x}, \mathbf{c}) \leq R$. For an excellent source on results concerning covering codes, see [19]. Let us denote by $k[n, R]$ the smallest possible dimension of a *linear* R-covering code of length n.

Let us next present the well-known Sauer-Shelah lemma.

Theorem 11 ([20], [21])**.** *If $Y \subseteq \mathbb{F}^n$ is a set containing at least $V(n, k-1) + 1$ words, then there exists a set S of k coordinates such that for any word $\mathbf{w} \in \mathbb{F}^n$ with supp$(\mathbf{w}) \subseteq S$ there exists a word $\mathbf{s} \in Y$ satisfying supp$(\mathbf{w}) =$ supp$(\mathbf{s}) \cap S$.*

Observe that each Hamming ball of radius e contains at most one codeword of C. Thus, if the intersection of the balls of radius t centered at the output words of Y can be covered by k balls of radius e, then we have $|T(Y)| \leq k$. This approach is formulated in the following lemma.

Lemma 12 ([18])**.** *Let $C \subseteq \mathbb{F}^n$ be an e-error-correcting code. If for any set of output words $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ we have*

$$T(Y) \subseteq \bigcup_{i=1}^{k} B_e(\beta_i)$$

for some words $\beta_i \in \mathbb{F}^n$ ($i = 1, \dots, k$), then $\mathcal{L} \leq k$.

Notice that Lemma 12 also gives a decoding algorithm. Indeed, if the words β_i are known, then there is at most one codeword in each $B_e(\beta_i)$, we can use the decoding algorithm of C on β_i and the codeword can be added to the list T.

Theorem 13. *Let C be an e-error-correcting code. If the number of channels satisfies $N \geq V(n, \ell + 2R - 1) - 2^{\ell+2R-k[\ell+2R,R]} + 2$, then*

$$\mathcal{L} \leq 2^{k[\ell+2R,R]}.$$

Proof. Let $\mathbf{x}$ be the input word. We have $|Y| \geq (V(n, \ell+2R-1)+1) - (2^{\ell+2R-k[\ell+2R,R]} - 1)$. Next we show that with this number of outputs we can guarantee that there exists a set S of $\ell + 2R$ coordinates such that within these coordinates of S a subset $Y' \subseteq Y$ contains a linear R-covering code of length $\ell + 2R$. Due to Theorem 11, we know that if we had more output words, namely, $|Y| \geq V(n, \ell + 2R - 1) + 1$, then we would have a set S of coordinates such that a subset $Y'' \subseteq Y$ contains all the $2^{\ell+2R}$ words of length $\ell + 2R$ among these coordinates of S. Let D be a linear R-covering code in $\mathbb{F}^{\ell+2R}$ with $\dim(D) = k[\ell + 2R, R]$. Notice that any coset $\mathbf{u} + D$, $\mathbf{u} \in \mathbb{F}^{\ell+2R}$, of the linear code D is also an R-covering code, and there are $2^{\ell+2R-\dim(D)}$ distinct cosets. Therefore, the set Y'' can miss any $2^{\ell+2R-\dim(D)} - 1$ words of $\mathbb{F}^{\ell+2R}$ and still the remaining subset contains at least one R-covering code of length $\ell+2R$. Consequently, it follows that Y' contains an R-covering code of size $2^{k[\ell+2R,R]}$ because Y' can be obtained from Y'' by removing some $2^{\ell+2R-\dim(D)} - 1$ words.

Now let $\mathbf{s} \in \mathbb{F}^n$ be a word such that $\mathrm{supp}(\mathbf{s}) = S$ and $Y_1 = \{\mathbf{y}_1, \dots, \mathbf{y}_{2^{k[\ell+2R,R]}}\} \subseteq Y'$ the subset of output words corresponding to the R-covering code. Denote $\beta_i = \mathbf{s} + \mathbf{y}_i$ for $i = 1, \dots, 2^{k[\ell+2R,R]}$. Since the words in set Y_1 form, among the coordinates corresponding to S, an R-covering code of length $\ell + 2R$, we know that there exists $\mathbf{y}_j$, $j \in \{1, \dots, 2^{k[\ell+2R,R]}\}$, such that the words $\mathbf{y}_j$ and $\mathbf{x} + \mathbf{s}$ differ in at most R places among the coordinates of S. Consequently, as $d(\mathbf{x}, \mathbf{y}_j) \leq t$, the words $\mathbf{x}$ and $\beta_j = \mathbf{y}_j + \mathbf{s}$ have distance at most $t - (\ell + R) + R = e$ from one another. Therefore, by Lemma 12, we get that $\mathcal{L} \leq 2^{k[\ell+2R,R]}$. □

Note that if $\ell = 5$ and $N \geq V(n, 4)+1$, then, by Theorem 3, we have $\mathcal{L} \leq 2^5 = 32$. If we have $N \geq V(n, 6) - 6$, then (using as the linear 1-covering code D the Hamming code of length 7), we obtain by the previous result, that $\mathcal{L} \leq 16$.

V. List size with less channels

By the following theorem it is clear that if we have *less* than $V(n, \ell - 1) + 1$ channels, then the list size cannot in general be constant for e-error-correcting codes of length n.

Theorem 14 ([18])**.** *Let $V(n, \ell-b-1)+1 \leq N \leq V(n, \ell-b)$ where $0 \leq b \leq \ell - 1$. Moreover, let $C \subseteq \mathbb{F}^n$ be such an e-error-correcting code that $\mathcal{L}$ is maximal. Then we have*

$$\mathcal{L} = \Theta(n^b).$$

Consequently, let us concentrate on certain e-error-correcting codes, namely, those with at most M codewords within any ball of radius $e + a$, for some $a > 0$.

Theorem 15. *Let $N \geq V(n, \ell-a-1)+1$ where $0 \leq a \leq \ell-1$. Let C be an e-error-correcting code such that $|B_{e+a}(\mathbf{u}) \cap C| \leq M$ for every $\mathbf{u} \in \mathbb{F}^n$. Consequently,*

$$\mathcal{L} \leq 2^{\ell-a} M.$$

The previous result is useful when our e-error-correcting code is a code for traditional list-decoding, see [22]. For number of channels being less than $V(n, \ell - 1) + 1$, it also gives, for every e-error-correcting code with suitable a, small exponent for n compared to Theorem 14 (see Corollary 16(ii) below), or even constant bounds (see Corollary 16(i)). Let us denote (see [22, Theorem 3.2])

$$r(n, e, M) = \frac{n}{2} \left(1 - \sqrt{1 - \frac{M-1}{M} \frac{2(2+1)}{n}} \right)$$

and the Johnson bound

$$r(n,e) = \frac{n}{2}\left(1-\sqrt{1-\frac{2(2e+1)}{n}}\right).$$

Corollary 16. *Let $M \geq 1$ and $2e+1 < n/2$. We have*

(i) Let $N \geq V(n, \ell - r(n,e,M) + e - 1) + 1$ where $0 \leq r(n,e,M) - e \leq \ell - 1$. Consequently,

$$\mathcal{L} \leq 2^{t-r(n,e,M)}M.$$

(ii) Let $N \geq V(n, \ell - r(n,e) + e - 1) + 1$ where $0 \leq r(n,e) - e \leq \ell - 1$. Consequently,

$$\mathcal{L} \leq 2^{t-r(n,e)}n.$$

The following result considers the case when we have less than $V(n,\ell-1)+1$ channels and the set of output words have certain restrictions on the distances between the output words.

Theorem 17. *Let C be an e-error-correcting code, $s \geq 1$, $N \geq \binom{n}{\ell-s} + 2V(n, \ell-s-1) + 1$ and $d(\mathbf{y}, \mathbf{y}') \geq 2s+1$ for any distinct $\mathbf{y}, \mathbf{y}' \in Y$. Then we have $\mathcal{L} \leq \binom{2\ell}{\ell}$.*

VI. Decoding with majority algorithm

In this section, we focus on decoding the transmitted word $\mathbf{x} \in C$ based on the set Y of the output words using a majority algorithm. First we describe the (well-known) majority algorithm using similar terminology and notation as in [12]. The coordinates of the output words $\mathbf{y}_j \in Y$ are denoted by $\mathbf{y}_j = (y_{j,1}, y_{j,2}, \ldots, y_{j,n})$. For simplicity we assume that N is odd. Furthermore, the number of zeros and ones in the ith coordinates of the output words are respectively denoted by

$$m_{i,0} = |\{j \in \{1,2,\ldots,N\} \mid y_{j,i} = 0\}|$$

and $m_{i,1} = N - m_{i,0}$. Based on Y, the *majority algorithm* outputs the word $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{F}^n$, where

$$c_i = \begin{cases} 0 & \text{if } m_{i,0} > m_{i,1} \\ 1 & \text{if } m_{i,0} < m_{i,1} \end{cases}.$$

In other words, for each coordinate of $\mathbf{c}$, we choose zero or one based on which one occurs more frequently. In [12, Example 1], it is shown that the majority algorithm does not always output the correct transmitted word $\mathbf{x}$ even though we take the e-error-correction capability of C into account. In [6], a modification of the majority algorithm is presented for decoding and it is shown that if the number of channels satisfies the bound of Theorem 1, then the output word of the algorithm belongs to $B_e(\mathbf{x})$ and can be uniquely decoded to $\mathbf{x}$. In what follows, we demonstrate that *with high probability* the word $\mathbf{c}$ is *verifiably* within distance e from $\mathbf{x}$ with significantly smaller number of channels (than in [6]).

For this purpose, notice first that the total number of errors occurring in the ith coordinates of $\mathbf{y}_i$ is at least $m_i = \min\{m_{i,0}, m_{i,1}\}$. On the other hand, there happens at most t errors in each channel and, hence, the total number of errors in the channels is at most tN. Thus, we obtain that

$$\sum_{i=1}^{n} m_i \leq tN. \tag{1}$$

Furthermore, if $\mathbf{x} = \mathbf{c}$, then the number of errors is exactly $\sum_{i=1}^{n} m_i$. In addition, if $\mathbf{x} \neq \mathbf{c}$, then for each coordinate i in which the words differ, $\max\{m_{i,0}, m_{i,1}\} = N - m_i$ is contributed to the sum of errors (instead of m_i). The following theorem is based on the idea that even the modified sum (in the left hand side of (2)) has to satisfy Inequality (1).

Theorem 18. *Let C be an e-error-correcting code, m'_i be the integers m_i ordered in such a way that $m'_1 \geq m'_2 \geq \cdots \geq m'_n$ and $\mathbf{c}$ be the output word of the majority algorithm. We have $d(\mathbf{x}, \mathbf{c}) \leq k$ if k is a positive integer such that*

$$\sum_{i=1}^{k+1}(N - m'_i) + \sum_{i=k+2}^{n} m'_i > tN. \tag{2}$$

Observe that (2) allows us to estimate the accuracy of $\mathbf{c}$. In particular, if $k \leq e$, then $d(\mathbf{x}, \mathbf{c}) \leq k \leq e$ and the word $\mathbf{c}$ can be decoded to $\mathbf{x}$ as C is an e-error-correcting code. Furthermore, if $k > e$, then $\mathbf{x} \in C \cap B_k(\mathbf{c})$ and the decoding algorithm outputs a list of words containing $\mathbf{x}$. Moreover, the size of the list is at most $\max_{\mathbf{u} \in \mathbb{F}^n} |C \cap B_k(\mathbf{u})|$, which is closely related to the traditional list decoding (see [22]). In conclusion, the theorem gives us a condition guaranteeing that the transmitted word can be decoded with certain accuracy. In what follows, we further study the probability that for a set Y of outputs there exists k such that $k \leq e$.

For the rest of the section, we assume that each word of $B_t(\mathbf{x})$ is outputted from a channel with equal probability. Here we actually allow — unlike elsewhere in the paper — some of the output words $\mathbf{y}_i$ to be equal. Analysing analytically the probability that in Theorem 18 there exists k such that $k \leq e$ seems rather demanding problem. Hence, in this presentation, we only approximate it using Monte Carlo simulations. In Table I, the probability is approximated using 100000 samples for $n = 24$, $t = 7$, $e = 2, 3, 4$ and varying number of channels N. From the table, we can notice that as the number of channels increases it becomes very likely that the majority algorithm together with the the e-error-correction capability of C correctly gives the transmitted word $\mathbf{x}$. Thus, although the majority algorithm does not always work (as was stated in [12]), it works with high probability when the number of channels is large enough. However, the required number of channels is very modest in comparison to [6] where tens of thousands channels are needed in the cases of Table I. In some applications, we could also request new outputs (from the channels) until the integer k is small enough to obtain the transmitted word $\mathbf{x}$ with desired accuracy.

TABLE I
THE MONTE CARLO APPROXIMATIONS WITH 100000 SAMPLES OF THE PROBABILITY FOR e SATISFYING THE CONDITION OF THEOREM 18 WHEN $n = 24, t = 7, e = 2, 3, 4$ AND $N = 11, 21, 31, 41$.

$N \backslash e$	2	3	4
11	0.068	0.260	0.587
21	0.369	0.790	0.972
31	0.701	0.971	0.999
41	0.887	0.997	0.999

REFERENCES

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

[2] V. Levenshtein, E. Konstantinova, E. Konstantinov, and S. Molodtsov, "Reconstruction of a graph from 2-vicinities of its vertices," *Discrete Appl. Math.*, vol. 156, pp. 1399–1406, 2008.

[3] R. Gabrys and E. Yaakobi, "Sequence reconstruction over the deletion channel," *IEEE Trans. Inform. Theory*, vol. 64, no. 4, pp. 2924–2931, 2018.

[4] M. Horovitz and E. Yaakobi, "Reconstruction of sequences over non-identical channels," *IEEE Trans. Inform. Theory*, vol. 65, no. 2, pp. 1267–1286, 2018.

[5] M. Abu-Sini and E. Yaakobi, "On list decoding of insertions and deletions under the reconstruction model," in *Proceedings of 2021 IEEE International Symposium on Information Theory*, 2021, pp. 1706–1711.

[6] ——, "On Levenshtein's reconstruction problem under insertions, deletions, and substitutions," *IEEE Trans. Inform. Theory*, vol. 67, no. 11, pp. 7132–7158, 2021.

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

[8] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 2, pp. 637–649, 2016.

[9] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[10] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angew. Chem. Int. Edit.*, vol. 54, no. 8, pp. 2552–2555, 2015.

[11] S. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.

[12] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," *IEEE Trans. Inform. Theory*, vol. 65, no. 4, pp. 2155–2165, 2018.

[13] ——, "On the uncertainty of information retrieval in associative memories," in *Proceedings of 2012 IEEE International Symposium on Information Theory*, 2012, pp. 106–110.

[14] V. Junnila and T. Laihonen, "Information retrieval with varying number of input clues," *IEEE Trans. Inform. Theory*, vol. 62, no. 2, pp. 625–638, 2016.

[15] ——, "Codes for information retrieval with small uncertainty," *IEEE Trans. Inform. Theory*, vol. 60, no. 2, pp. 976–985, 2014.

[16] T. Laihonen and T. Lehtilä, "Improved codes for list decoding in the Levenshtein's channel and information retrieval," in *Proceedings of 2017 IEEE International Symposium on Information Theory*, 2017, pp. 2643–2647.

[17] T. Laihonen, "On t-revealing codes in binary Hamming spaces," *Information and Computation*, vol. 268, 2019.

[18] V. Junnila, T. Laihonen, and T. Lehtilä, "On Levenshtein's channel and list size in information retrieval," *IEEE Trans. Inform. Theory*, vol. 67, no. 6, pp. 3322–3341, 2020.

[19] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering codes*, ser. North-Holland Mathematical Library. Amsterdam: North-Holland Publishing Co., 1997, vol. 54.

[20] N. Sauer, "On the density of families of sets," *J. Comb. Theory A*, vol. 13, no. 1, pp. 145–147, 1972.

[21] S. Shelah, "A combinatorial problem; stability and order for models and theories in infinitary languages," *Pac. J. Math.*, vol. 41, no. 1, pp. 247–261, 1972.

[22] V. Guruswami, *List decoding of error-correcting codes*. ProQuest LLC, Ann Arbor, MI, 2001, thesis (Ph.D.)–Massachusetts Institute of Technology. [Online]. Available: http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:0803408