The Levenshtein's Sequence Reconstruction Problem and the Length of the List

Ville Junnila, Tero Laihonen, and Tuomo Lehtilä

***Abstract*— In the paper, the Levenshtein's sequence reconstruction problem is considered in the case where the transmitted words are chosen from an e-error-correcting code, at most t substitution errors occur in each of the N channels and the decoder outputs a list of length $\mathcal{L}$. Previously, when $t = e + \ell$ and the transmitted word is long enough, the numbers of required channels were determined for $\mathcal{L} = 1, 2$ and $\ell + 1$. Here we determine the exact number of channels in the cases $\mathcal{L} = 3, 4, \ldots, \ell$. This also provides the size of the largest intersection of $\mathcal{L}$ balls of radius t (with respect to substitutions) centered at the words with mutual Hamming distances at least $2e + 1$. Furthermore, with the aid of covering codes, we also consider the list sizes in the cases where the length n is rather small (improving previously known results). After that we study how much we can decrease the number of required channels when we use list-decoding codes. Finally, the majority algorithm is discussed for decoding in a probabilistic set-up; in particular, we show that the output word of the decoder can be verified to be the transmitted one with high probability.**

***Index Terms*— Information retrieval, Levenshtein's sequence reconstruction, list decoding, majority algorithm, substitution errors.**

I. Introduction

IN THIS paper, the Levenshtein's *sequence reconstruction problem*, which was introduced in [2], is studied when the errors are substitution errors. For many related sequence reconstruction problems (concerning, for instance, deletion and insertion errors) consult, for example, [2], [3], [4], [5], [8], [9]. Originally, the motivation for the sequence reconstruction problem came from biology and chemistry where the familiar redundancy method of error correction is not suitable. The sequence reconstruction problem has returned to the focus, since it was recently discovered that this problem is highly relevant to information retrieval in advanced storage technologies. In these storage systems the stored information is either a single copy, which is read many times, or the stored information has several copies [5], [10]. This problem (see [5]) is especially applicable to DNA data storage systems (see [11], [12], [13], [14]) where DNA strands provide numerous erroneous copies of the information and the goal is to recover the information using these copies.

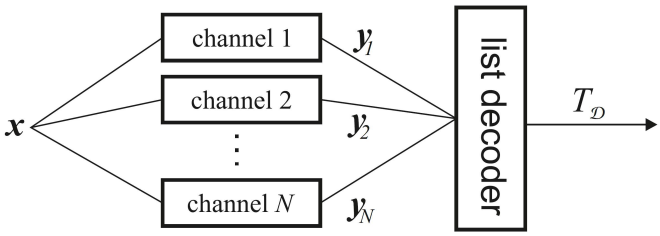

Fig. 1. The Levenshtein's sequence reconstruction.

Let us denote the set $\{1, 2, \ldots, n\}$ by $[1, n]$. Denote by $\mathbb{F}$ the finite field of two elements, and denote the binary Hamming space by $\mathbb{F}^n$. The *support* of the word $\mathbf{x} = x_1 \ldots x_n \in \mathbb{F}^n$ is defined via $\mathrm{supp}(\mathbf{x}) = \{i \mid x_i \neq 0\}$. Let us denote the all-zero word $\mathbf{0} = 00 \ldots 0 \in \mathbb{F}^n$ and by $\mathbf{e}_i \in \mathbb{F}^n$ a word with 1 in the ith coordinate and zeros elsewhere. 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 *Hamming ball* of radius t 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 its cardinality by $V(n, t) = \sum_{i=0}^{t} \binom{n}{i}$. The *Hamming sphere* of radius t centered at $\mathbf{x} \in \mathbb{F}^n$ is $S_t(\mathbf{x}) = \{\mathbf{y} \in \mathbb{F}^n \mid d(\mathbf{x}, \mathbf{y}) = t\}$. A nonempty subset of $\mathbb{F}^n$ is called a *code* and its elements are called *codewords*. The *minimum distance* of a code $C \subseteq \mathbb{F}^n$ is defined as $d_{\min}(C) = \min_{\mathbf{c}_1, \mathbf{c}_2 \in C, \mathbf{c}_1 \neq \mathbf{c}_2} d(\mathbf{c}_1, \mathbf{c}_2)$. Consequently, the code C has the error-correcting capability $e = e(C) = \lfloor (d_{\min}(C) - 1)/2 \rfloor$. Let us denote by $\varepsilon \simeq 2.71828\ldots$ the Napier's constant. With $\Theta(g(n))$ we denote a set of functions such that $f(n) \in \Theta(g(n))$ if there exist positive reals k_1, k_2 and n_0 such that for all $n > n_0$ we have $k_1 g(n) \leq f(n) \leq k_2 g(n)$.

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 distinct outputs $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ from the channels (see Fig. 1).

Manuscript received 16 November 2022; revised 13 September 2023; accepted 15 September 2023. Date of publication 22 September 2023; date of current version 22 January 2024. This work was supported in part by the Academy of Finland under Grant 338797. The work of Tuomo Lehtilä was supported by the Finnish Cultural Foundation. An earlier version of this paper was presented in part at the 2022 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT50566.2022.9834612]. *(Corresponding author: Ville Junnila.)*

Ville Junnila and Tero Laihonen are with the Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland (e-mail: viljun@utu.fi; terolai@utu.fi).

Tuomo Lehtilä was with Université Claude Bernard, CNRS, LIRIS, UMR 5205, University of Lyon, 69622 Villeurbanne, France, and also with the Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland. He is now with the Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland (e-mail: tualeh@utu.fi).

Communicated by E. Yaakobi, Associate Editor for Coding and Decoding.
Digital Object Identifier 10.1109/TIT.2023.3318354

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 [15] and [16]. 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 $\mathbf{x} \in T_{\mathcal{D}}$ for all input words $\mathbf{x}$ and all possible output sets Y. In this paper, we focus on the smallest possible value of $\mathcal{L}_{\mathcal{D}}$ over all successful decoders $\mathcal{D}$, in other words, on $\mathcal{L} = \min_{\mathcal{D} \text{ is successful}}\{\mathcal{L}_{\mathcal{D}}\}$. Let us denote

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

Consequently,

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

The value of $\mathcal{L}$ depends on e, ℓ, n and N. Obviously, one would like to have as small $\mathcal{L}$ as possible. Observe that we consider the worst case scenario of the output channels regarding $\mathcal{L}$. In such a situation, the channels are sometimes called adversarial; for example, see [17]. The problem of minimizing $\mathcal{L}$ is studied, for example, in [15], [16], [18], [19], [20], and [21]. Also a probabilistic versions of this problem have been studied (often under the name *trace reconstruction*) for example in [22] and [23]. In this paper, we mainly consider the relation between N and $\mathcal{L}$ for various n after we fix the parameters ℓ and e (while letting C be any e-error-correcting code). The sequence reconstruction problem is also closely related (see [18]) to *information retrieval in associative memory* introduced by Yaakobi and Bruck [15],[16].

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 14). 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.

II. Known Results

In this section we present some known results on how the two values of 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 [2] and [15], the problem has been considered for $\mathcal{L} = 1$ and $\mathcal{L} = 2$, respectively. Moreover, in [24], the exact number of channels N required to have $\mathcal{L}$ constant on n has been presented, see Theorems 4 and 5. The following theorem gives an exact number of channels required to have $\mathcal{L} = 1$.

Theorem 1 [2]:
We have $\mathcal{L} \leq 1$ if

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

Theorem 2 [15]:

If $N \geq \sum_{i_1,i_2,i_3,i_4} \binom{n-\lceil \frac{3d}{2} \rceil}{i_1} \binom{\lceil \frac{d}{2} \rceil}{i_2} \binom{\lceil \frac{d}{2} \rceil}{i_3} \binom{\lfloor \frac{d}{2} \rfloor}{i_4} + 1$ for

- $0 \leq i_1 \leq t - \lceil \frac{d}{2} \rceil$,
- $i_1 + \lfloor \frac{d}{2} \rfloor - t \leq i_4 \leq t - \lceil \frac{d}{2} \rceil - i_1$,
- $2\lceil \frac{d}{2} \rceil - t + i_1 \leq i_3 \leq t - (i_1 + i_4)$ and
- $\max\{i_1 - i_3 - i_4 + \lceil \frac{3d}{2} \rceil - t, i_1 + i_3 + i_4 + \lceil \frac{d}{2} \rceil - t\} \leq i_2 \leq t - (i_1 + i_4 + \lceil \frac{d}{2} \rceil - i_3)$,

then $\mathcal{L} \leq 2$ for any code C with minimum distance d.

The following theorem reformulates a result achieved by Yaakobi and Bruck [15, Algorithm 18] proven in [24].

Theorem 3: 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 3 can be improved to 2^ℓ which has been shown to be tight in [24].

Theorem 4 (Theorem 7, [24]): 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 of N is smaller, then list size $\mathcal{L}$ can be linear with respect to n.

Theorem 5 (Theorem 10, [24]): 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 4 cannot be improved in general, we can improve it, when n is large, to $\ell + 1$.

Theorem 6 (Theorem 20, [24]): 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 following theorem shows that the previous upper bound $\mathcal{L} \leq \ell + 1$ is tight when $N = V(n, \ell-1) + 1$.

Theorem 7 (Theorem 9, [24]): If $n \geq \ell + \ell e + e$ and the number of channels satisfies $N \leq V(n, \ell-1) + 1$, then there exists an e-error-correcting code $C \subseteq \mathbb{F}_2^n$ such that $\mathcal{L} \geq \ell + 1$.

As we see, N is known precisely only for three values when $\mathcal{L}$ is constant on n. In the following section, we give the missing values of N.

III. List Size With More Channels

In this section, we give exact bounds for the number of channels N (when n is large) which is required for satisfying $\mathcal{L} < h$ for every constant value of h. As noticed above, N was previously known only for three values $h = 2, 3, \ell + 2$. To achieve this, we need to introduce two technical lemmas from [24].

In the following lemma, when n is large, it is shown that if any three codewords in $T(Y)$ differ within some subset of coordinates $\overline{D}$ of constant size b, then 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 $\text{supp}(\mathbf{w}+\mathbf{z})$ gives the set of coordinates in which $\mathbf{w}$ and $\mathbf{z}$ differ.

Lemma 8 (Lemma 18, [24]): Let $b \geq 3t$ be an integer with $t = e + \ell$ 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 $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \in T(Y)$. If now $\overline{D} \subseteq [1, n]$ is a set such that $|\overline{D}| = b$ and

$$\text{supp}(\mathbf{c}_1 + \mathbf{c}_2) \cup \text{supp}(\mathbf{c}_1 + \mathbf{c}_3) \cup \text{supp}(\mathbf{c}_2 + \mathbf{c}_3) \subseteq \overline{D},$$

then for any word $\mathbf{w} \in \mathbb{F}^n$ we have $\text{supp}(\mathbf{w} + \mathbf{c}_1) \setminus \overline{D} = \text{supp}(\mathbf{w} + \mathbf{c}_2) \setminus \overline{D} = \text{supp}(\mathbf{w} + \mathbf{c}_3) \setminus \overline{D}$ and there exists an output word $\mathbf{y} \in Y$ such that

$$|\text{supp}(\mathbf{y} + \mathbf{c}_1) \setminus \overline{D}| \geq \ell - 1.$$

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

Lemma 9 (Lemma 19, [24]): 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 $\mathbf{c}_1, \mathbf{c}_2 \in T(Y)$.

We denote by $N'(n, \ell, e, h)$ the maximum number of t-error channels such that there exists a set of output words $Y \subseteq \mathbb{F}^n$ satisfying $|Y| = N'(n, \ell, e, h)$ and $|T(Y)| \geq h$ for some e-error correcting code C. Observe that $N'(n, \ell, e, h)$ is a non-increasing function on h. By Theorems 5 and 6, $N'(n, \ell, e, h) = V(n, \ell - 1)$ for all $\ell + 2 \leq h \leq \lfloor n/(e + 1) \rfloor$ when n is large enough. Hence, when we use notation $N(n, \ell, e, h)$, we assume that h is the smallest integer among all h' for which $N'(n, \ell, e, h) = N'(n, \ell, e, h')$. When the exact formulation is not necessary for clarity, we denote $N(n, \ell, e, h) = N_h$. Observe that, if $N \geq N_h + 1$, then $\mathcal{L} < h$ for all e-error-correcting codes. In particular, the difference between $N'(n, \ell, e, h)$ and $N(n, \ell, e, h)$ is that $N'(n, \ell, e, h)$ exists for each value of h but may give the same values for different choices of h while $N(n, \ell, e, h)$ does not exist for every choice of h but each value it attains is unique. For example, $N(n, \ell, e, h)$ does not exist (when n is large) for $\ell + 3 \leq h \leq \lfloor n/(e + 1) \rfloor$.

Remark 10: Values $N'(n, \ell, e, h)$ and $N(n, \ell, e, h)$ can also be interpreted as the maximum sizes of intersections of any h Hamming balls of radius $t = e + \ell$ centered at words with pairwise minimum distance of $2e + 1$ in the binary Hamming space of length n. In other words, $N'(n, \ell, e, h) = \max |\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)|$ where the maximum is taken over h words such that $d(\mathbf{c}_i, \mathbf{c}_j) \geq 2e+1$ and $\mathbf{c}_i \in \mathbb{F}^n$ for each $1 \leq i, j \leq h$ with $i \neq j$. For a more detailed introduction to this perspective, the reader is encouraged to see the discussion related to notation $N_t(m, d)$ in [15].

We need the following two technical notations in the next theorem. Let us have $\ell, e, h, w \in \mathbb{N}$, then

$$W_w = \{(i_1, \ldots, i_h) \mid \text{for each } j : i_j \in \mathbb{N}, e + 1 \geq i_j \geq \frac{w + 1 - \ell}{2} \text{ and } w \geq \sum_{j=1}^{h} i_j\}$$

and

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

In the following theorem, we give the maximum number of channels $N'(n, \ell, e, h)$ which gives list size $\mathcal{L} \geq h$ for some e-error-correcting code and if $N > N'(n, \ell, e, h)$, then $\mathcal{L} < h$ for all e-error-correcting codes. Later in Theorem 12, we show that the two sums in the maximum of Equation (1) are equal. Then, in Theorem 13, we show that the following theorem actually holds for N_h as well. Theorem 11 is further analysed in Corollaries 14, 15 and 16.

Theorem 11: Let $n \geq n(e, \ell, b)$, $b \geq \max\{3t, 4e + 4\}$, $\ell \geq 2$, $3 \leq h \leq \ell + 1$. Then

$$N'(n, \ell, e, h) = V(n, \ell - 1) + \max\left\{\sum_{w \geq \ell} \sum_{(i_1, \ldots, i_h) \in W_w} \binom{n - h(e+1)}{w - \sum_{j=1}^{h} i_j} \prod_{j=1}^{h} \binom{e+1}{i_j}, \quad (1)\right.$$
$$\left. \sum_{w \geq \ell} \sum_{(i_1, \ldots, i_h) \in W'_w} \binom{n + 1 - h(e+1)}{w - \sum_{j=1}^{h} i_j} \binom{e}{i_1} \prod_{j=2}^{h} \binom{e+1}{i_j}\right\}.$$

Proof: Let us have $N = N'(n, \ell, e, h)$, $n \geq n(e, \ell, b)$, $b \geq \max\{3t, 4e + 4\}$, $\ell \geq 2$ and $3 \leq h \leq \ell + 1$. Moreover, let C be an e-error-correcting code maximizing $|\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)|$ for some h codewords $\mathbf{c}_i$ and that we have $\mathcal{L} = h$ when $N = N'(n, \ell, e, h)$. Furthermore, let Y be a set of outputs such that $|T(Y)| = \mathcal{L} = h$ and let us denote $T(Y) = \{\mathbf{c}_1, \ldots, \mathbf{c}_h\}$. Observe that we have $|Y| = N'(n, \ell, e, h) = |\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)|$.

By Theorem 5, we have $N \geq V(n, \ell-1)+1$. By Lemma 9 and due to the fact that C is an e-error-correcting code,

we have $d(\mathbf{c}_i, \mathbf{c}_j) \in \{2e + 1, 2e + 2\}$ for each $i \neq j$. Since we are considering binary Hamming space and $h \geq 3$, all pairwise distances cannot be $2e+1$. Let us assume without loss of generality 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,

$$|\text{supp}(\mathbf{c}_2) \cap \text{supp}(\mathbf{c}_3)| = e + 1, \quad (2)$$

since $e+1 = \lceil d(\mathbf{c}_2, \mathbf{c}_3)/2 \rceil = \lceil (w(\mathbf{c}_2)+w(\mathbf{c}_3)-2|\text{supp}(\mathbf{c}_2) \cap \text{supp}(\mathbf{c}_3)|)/2 \rceil = 2e + 2 - |\text{supp}(\mathbf{c}_2) \cap \text{supp}(\mathbf{c}_3)|$. Let $\overline{D}$ be any subset of $[1, n]$ satisfying $\text{supp}(\mathbf{c}_1 + \mathbf{c}_2) \cup \text{supp}(\mathbf{c}_1 + \mathbf{c}_3) \cup \text{supp}(\mathbf{c}_2 + \mathbf{c}_3) = \text{supp}(\mathbf{c}_2) \cup \text{supp}(\mathbf{c}_3) \subseteq \overline{D}$ and $|\overline{D}| = b$.

We can make this choice for $\overline{D}$ since $|\mathrm{supp}(\mathbf{c}_2)\cup\mathrm{supp}(\mathbf{c}_3)| \leq 3e+3$. 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 8, there exists an output word $\mathbf{y} \in Y$ such that

$$|\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| = |\mathrm{supp}(\mathbf{c}_1 + \mathbf{y}) \setminus \overline{D}| \geq \ell - 1. \tag{3}$$

Since $d(\mathbf{y}, \mathbf{c}_1) \leq t$, we have $w(\mathbf{y}) \leq t$. Furthermore, we have $t \geq d(\mathbf{y}, \mathbf{c}_2) \geq |\mathrm{supp}(\mathbf{y})\setminus\overline{D}| + w(\mathbf{c}_2) - |\mathrm{supp}(\mathbf{y})\cap\mathrm{supp}(\mathbf{c}_2)| \geq (\ell-1) + (2e+2) - |\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_2)|$. Hence, $|\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_2)| \geq e+1$. Moreover, $|\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_2)| \leq w(\mathbf{y}) - |\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| \leq e+1$. Therefore,

$$|\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_2)| = e + 1. \tag{4}$$

Thus, $\mathrm{supp}(\mathbf{y}) = (\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_2)) \cup (\mathrm{supp}(\mathbf{y}) \setminus \overline{D})$ and

$$\mathrm{supp}(\mathbf{y}) \cap (\mathrm{supp}(\mathbf{c}_3) \setminus \mathrm{supp}(\mathbf{c}_2)) = \emptyset. \tag{5}$$

Hence, $\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_3) \subseteq \mathrm{supp}(\mathbf{c}_2)$. Moreover, notice that if $\mathrm{supp}(\mathbf{y}) \cap (\mathrm{supp}(\mathbf{c}_2) \setminus \mathrm{supp}(\mathbf{c}_3)) \neq \emptyset$, then $|\mathrm{supp}(\mathbf{y}) \cap (\mathrm{supp}(\mathbf{c}_2) \setminus \mathrm{supp}(\mathbf{c}_3))| \geq 1$. Recall that $|\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| \geq \ell - 1$. Thus, in this case we have $|\mathrm{supp}(\mathbf{c}_3) \cap \mathrm{supp}(\mathbf{y})| \leq w(\mathbf{y}) - |\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| - |\mathrm{supp}(\mathbf{y}) \cap (\mathrm{supp}(\mathbf{c}_2) \setminus \mathrm{supp}(\mathbf{c}_3))| \leq t - (\ell-1) - 1 = e$. Hence, $|\mathrm{supp}(\mathbf{c}_3) \setminus \mathrm{supp}(\mathbf{y})| = w(\mathbf{c}_3) - |\mathrm{supp}(\mathbf{c}_3) \cap \mathrm{supp}(\mathbf{y})| \geq 2e + 1 - e = e + 1$. Furthermore, we have $d(\mathbf{y}, \mathbf{c}_3) \geq |\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| + |\mathrm{supp}(\mathbf{y}) \cap (\mathrm{supp}(\mathbf{c}_2) \setminus \mathrm{supp}(\mathbf{c}_3))| + |\mathrm{supp}(\mathbf{c}_3) \setminus \mathrm{supp}(\mathbf{y})| \geq (\ell-1) + 1 + (e+1) = t + 1 > t$ (a contradiction). Therefore, we have

$$\mathrm{supp}(\mathbf{y}) \cap (\mathrm{supp}(\mathbf{c}_2) \setminus \mathrm{supp}(\mathbf{c}_3)) = \emptyset. \tag{6}$$

Equations (2), (3), (4), (5) and (6) together with $w(\mathbf{y}) \leq t$ give that $\mathrm{supp}(\mathbf{y}) \cap \overline{D} = \mathrm{supp}(\mathbf{c}_2) \cap \mathrm{supp}(\mathbf{c}_3)$.

Furthermore, for each $i \in [4, h]$, we may choose set $\overline{D}$ as set $\overline{D_i}$ in such a way that $\mathrm{supp}(\mathbf{c}_i) \subseteq \overline{D_i}$ since we have defined $|\overline{D}| = b \geq 4e + 4$ and $\mathrm{supp}(\mathbf{c}_i) \cup \mathrm{supp}(\mathbf{c}_1 + \mathbf{c}_2) \cup \mathrm{supp}(\mathbf{c}_1+\mathbf{c}_3) \cup \mathrm{supp}(\mathbf{c}_2+\mathbf{c}_3) = \mathrm{supp}(\mathbf{c}_i) \cup \mathrm{supp}(\mathbf{c}_2) \cup \mathrm{supp}(\mathbf{c}_3)$. Indeed, we have $d(\mathbf{c}_i, \mathbf{c}_j) \in \{2e+1, 2e+2\}$ and $w(\mathbf{c}_j) \in \{2e+1, 2e+2\}$ for each $i, j \in [2, h]$. Hence, $|\mathrm{supp}(\mathbf{c}_i) \setminus (\mathrm{supp}(\mathbf{c}_2) \cup \mathrm{supp}(\mathbf{c}_3))| \leq |\mathrm{supp}(\mathbf{c}_i) \setminus \mathrm{supp}(\mathbf{c}_2)| \leq e+1$. Recall that $|\mathrm{supp}(\mathbf{c}_2) \cup \mathrm{supp}(\mathbf{c}_3)| \leq 3e+3$. Thus, we have $|\mathrm{supp}(\mathbf{c}_2) \cup \mathrm{supp}(\mathbf{c}_3) \cup \mathrm{supp}(\mathbf{c}_i)| \leq (3e+3) + (e+1) = 4e + 4$. Hence, by Lemma 8, for each $i \in [4, h]$, there exists an output word $\mathbf{y}'_i \in Y$ such that $|\mathrm{supp}(\mathbf{c}_1 + \mathbf{y}'_i) \setminus \overline{D_i}| \geq \ell - 1$. Therefore, as above, $\mathrm{supp}(\mathbf{y}'_i) \cap \overline{D_i} = \mathrm{supp}(\mathbf{c}_2) \cap \mathrm{supp}(\mathbf{c}_3)$ and $\mathrm{supp}(\mathbf{y}'_i) \cap \overline{D_i} = \mathrm{supp}(\mathbf{c}_2) \cap \mathrm{supp}(\mathbf{c}_i)$ implying $\mathrm{supp}(\mathbf{c}_2) \cap \mathrm{supp}(\mathbf{c}_i) = \mathrm{supp}(\mathbf{c}_2) \cap \mathrm{supp}(\mathbf{c}_3)$. In particular, we have now shown that $\mathrm{supp}(\mathbf{c}_2) \cap \mathrm{supp}(\mathbf{c}_j)$ (for $j \in [3, h]$) does not depend on our choice for j.

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 $w(\mathbf{c}_j) \in \{e, e+1\}$ (for $j \in [1, h]$) 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 $\mathbf{c}_j$ can have weight e by the minimum distance of C.

Let us then count the number of words in $\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)$. Recall that we have $|Y| = |\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)|$. Clearly, each word $\mathbf{y}$ with $w(\mathbf{y}) \leq \ell - 1$ belongs to the intersection contributing $V(n, \ell-1)$ words to Y. Assume then that $w(\mathbf{y}) = w \geq \ell$. As $d(\mathbf{y}, \mathbf{c}_j) \leq t$ for all $j \in [1, h]$, we have $d(\mathbf{y}, \mathbf{c}_j) = w(\mathbf{y}) + w(\mathbf{c}_j) - 2|\mathrm{supp}(\mathbf{y}) \cap \mathrm{supp}(\mathbf{c}_j)| \leq t$. 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 $d(\mathbf{y}, \mathbf{c}_j) = w + e + 1 - 2i_j \leq t$. Hence, $e + 1 \geq i_j \geq (w + 1 - \ell)/2$. Moreover, $\sum_{j=1}^{h} i_j \leq w$ 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, $\mathbf{y} \in \bigcap_{i=1}^{h} B_t(\mathbf{c}_i)$ if and only if $(i_1, \ldots, i_h) \in W_w$. Recall that the supports of the codewords $\mathbf{c}_i, \mathbf{c}_j$ are disjoint and thus, each i_j is linked to a unique support. Hence, there exist $\sum_{(i_1,\ldots,i_h)\in W_w} \binom{n-h(e+1)}{w-\sum_{j=1}^{h} i_j} \prod_{i_j=1}^{h} \binom{e+1}{i_j}$ words of weight $w \geq \ell$ in $\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)$.

Let us next consider the case where $w(\mathbf{c}_k) = e$ for some k, say $k = 1$, we have $e \geq i_1 \geq (w - \ell)/2$ (recall that $w(\mathbf{c}_i) = e + 1$ for each $i \neq k$). Hence, $\mathbf{y} \in \bigcap_{i=1}^{h} B_t(\mathbf{c}_i)$ if and only if $(i_1, \ldots, i_h) \in W'_w$. Thus, there exist $\sum_{(i_1,\ldots,i_h)\in W'_w} \binom{n+1-h(e+1)}{w-\sum_{j=1}^{h} i_j} \binom{e}{i_1} \prod_{i_j=2}^{h} \binom{e+1}{i_j}$ words of weight $w \geq \ell$ in $\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)$. Together these give the claim. □

Observe by the proof that the bound given in Theorem 11 is tight. Notice also that, compared to [24] which considered only the case $h = \ell + 2$, the output sets giving maximal list size in Theorem 11 are geometrically more complicated than in [24] where a ball of volume $V(n, \ell-1)$ was essential.

If we increase N by one, then $\mathcal{L}$ decreases since we cannot place the output words within the intersection of t-balls centered at codewords in $T(Y)$. Later, in Theorem 13 we see that $\mathcal{L}$ decreases by exactly one (given a suitable e-error correcting code C). 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 $w \leq 2e + \ell$.

In the following theorem, we improve the previous result by showing that the two binomial sums within the max are actually equal.

Theorem 12: Let $n \geq n(e, \ell, b)$, $b \geq \max\{3t, 4e + 4\}$, $\ell \geq 2$, $3 \leq h \leq \ell + 1$. Then

$$\begin{aligned} &N'(n, \ell, e, h) - V(n, \ell - 1) \\ &= \sum_{w\geq\ell} \sum_{(i_1,\ldots,i_h)\in W_w} \binom{n - h(e+1)}{w - \sum_{j=1}^{h} i_j} \prod_{j=1}^{h} \binom{e+1}{i_j} \\ &= \sum_{w\geq\ell} \sum_{(i_1,\ldots,i_h)\in W'_w} \binom{n + 1 - h(e+1)}{w - \sum_{j=1}^{h} i_j} \binom{e}{i_1} \prod_{j=2}^{h} \binom{e+1}{i_j}. \end{aligned}$$

Proof: Observe that the claim follows from Theorem 11 if we can prove that the two binomial sums in the claim are equal. For that, we will be considering subsums

$$S_a = \sum_{w=\ell+2a}^{\ell+2a+1} \sum_{(i_1,\ldots,i_h)\in W_w} \left(\binom{n - h(e+1)}{w - \sum_{j=1}^{h} i_j} \cdot \prod_{j=1}^{h} \binom{e+1}{i_j} \right) \tag{7}$$

and

$$S'_a = \sum_{w=\ell+2a}^{\ell+2a+1} \sum_{(i_1,\ldots,i_h)\in W'_w} \left(\binom{n+1-h(e+1)}{w-\sum_{j=1}^{h} i_j} \cdot \binom{e}{i_1} \prod_{j=2}^{h} \binom{e+1}{i_j} \right). \quad (8)$$

We claim that $S_a = S'_a$ for each non-negative integer a. When $w = \ell + 2a$ we have

$$W_{\ell+2a} = \{(i_1,\ldots,i_h) \mid \text{for each } j : i_j \in \mathbb{N}, e+1 \geq i_j \geq a+1 \text{ and } \ell+2a \geq \sum_{j=1}^{h} i_j\}$$

and

$$W'_{\ell+2a} = \{(i_1,\ldots,i_h) \mid \text{each } i_j \in \mathbb{N}, e \geq i_1 \geq a \text{ and for } j \geq 2 : e+1 \geq i_j \geq a+1 \text{ and } \ell+2a \geq \sum_{j=1}^{h} i_j\}.$$

Moreover, for $w = \ell + 2a + 1$ we have

$$W_{\ell+2a+1} = \{(i_1,\ldots,i_h) \mid \text{for each } j : i_j \in \mathbb{N}, e+1 \geq i_j \geq a+1 \text{ and } \ell+2a+1 \geq \sum_{j=1}^{h} i_j\}$$

and

$$W'_{\ell+2a+1} = \{(i_1,\ldots,i_h) \mid \text{each } i_j \in \mathbb{N}, e \geq i_1 \geq a+1 \text{ and for } j \geq 2 : e+1 \geq i_j \geq a+1 \text{ and } \ell+2a+1 \geq \sum_{j=1}^{h} i_j\}.$$

Assume now that the codewords $\mathbf{c}_i \in T(Y)$ are arranged as in the proof of Theorem 11, that is, each codeword has weight e or $e+1$. Hence, their supports do not intersect. Recall that we have at most one codeword with weight e and if that word exists, then we are using set W'_w in our binomial sum. For further rearranging of the Hamming space, we assume that if each word has weight $e+1$, then $\text{supp}(\mathbf{c}_i) = [(e+1)(i-1)+1,(e+1)i]$ and if there exists a word with weight e, it is denoted by $\mathbf{c}'_1$ (replacing $\mathbf{c}_1$) and we have $\text{supp}(\mathbf{c}'_1) = [2, e+1]$.

Recall that w describes the weight of a word $\mathbf{y}_i \in \bigcap_{\mathbf{c}\in T(Y)} B_t(\mathbf{c})$ in the proof of Theorem 11 within the notations W_w and W'_w. When $w(\mathbf{y}_i) = w = 2a + \ell$ and we are considering case S_a, then $|\text{supp}(\mathbf{y}_i) \cap \text{supp}(\mathbf{c}_j)| = i_j \geq a+1$. Let us denote by Y_{2a}, Y_{2a+1}, Y'_{2a} and Y'_{2a+1} the sets of output words contributing to the sums S_a and S'_a, respectively, where Y_w (resp. Y'_w) contains output words of weight $w+\ell$.

We will construct the proof by first showing that $Y_{2a} \subseteq Y'_{2a}$, then that $Y'_{2a+1} \subseteq Y_{2a+1}$ and finally that $|Y'_{2a} \setminus Y_{2a}| = |Y_{2a+1} \setminus Y'_{2a+1}|$. Together, these imply that $S_a = S'_a$.

Assume that $\mathbf{y} \in Y_{2a}$. Thus, $w(\mathbf{y}) = \ell + 2a$ and $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}_i)| \geq a+1$ for each $i \in [1,h]$. Hence, $e \geq |\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}'_1)| \geq a$ since $\text{supp}(\mathbf{c}_1) = \text{supp}(\mathbf{c}'_1) \cup \{1\}$. Hence, $\mathbf{y} \in Y'_{2a}$.

Assume then that $\mathbf{y} \in Y'_{2a+1}$. Thus, $w(\mathbf{y}) = \ell + 2a + 1$ and $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}_i)| \geq a+1$ for each $i \in [2,h]$ and $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}'_1)| \geq a+1$. Hence, $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}_1)| \geq a+1$ and $\mathbf{y} \in Y_{2a+1}$. Therefore, $Y_{2a} \subseteq Y'_{2a}$ and $Y'_{2a+1} \subseteq Y_{2a+1}$.

Let us now consider output word $\mathbf{y}' \in Y'_{2a} \setminus Y_{2a}$. We again have $w(\mathbf{y}') = \ell + 2a$, $|\text{supp}(\mathbf{y}') \cap \text{supp}(\mathbf{c}_j)| = i_j \geq a+1$ for each $j \in [2,h]$. However, $|\text{supp}(\mathbf{y}') \cap \text{supp}(\mathbf{c}_1)| = a = |\text{supp}(\mathbf{y}') \cap \text{supp}(\mathbf{c}'_1)|$. Observe that now especially $\text{supp}(\mathbf{y}') \cap \{1\} = \emptyset$. Thus, $\mathbf{y}' + \mathbf{e}_1 \in Y_{2a+1}$. Consider then output word $\mathbf{y} \in Y_{2a+1} \setminus Y'_{2a+1}$. We again have $w(\mathbf{y}) = \ell + 2a + 1$, $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}_j)| = i_j \geq a+1$ for each $j \in [1,h]$. However, $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}'_1)| = a$; in particular, we have $1 \in \text{supp}(\mathbf{y})$. Indeed, if $1 \notin \text{supp}(\mathbf{y})$, then $\mathbf{y} \notin Y_{2a+1}$ and if $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}'_1)| \geq a+1$, then $\mathbf{y} \in Y'_{2a+1}$. Thus, $\mathbf{y} + \mathbf{e}_1 \in Y'_{2a}$. Therefore, for each output word $\mathbf{y}' \in Y'_{2a} \setminus Y_{2a}$, we have $\mathbf{y}' + \mathbf{e}_1 \in Y_{2a+1} \setminus Y'_{2a+1}$ and for each output word $\mathbf{y} \in Y_{2a+1} \setminus Y'_{2a+1}$, we have $\mathbf{y} + \mathbf{e}_1 \in Y'_{2a} \setminus Y_{2a}$. Thus, $|Y'_{2a} \setminus Y_{2a}| = |Y_{2a+1} \setminus Y'_{2a+1}|$.

Now, we have $S_a = S'_a$ and the claim follows. □

In the following theorem, we show that N_h exists and is unique for each value of $h \in [3, \ell+1]$ when n is large. Hence, we may replace $N'(n,\ell,e,h)$ in Theorems 11 and 12 by $N(n,\ell,e,h)$. As we have seen previously, this does not hold when $h \geq \ell + 2$.

Theorem 13: Let $n \geq n(e,\ell,b)$, $b \geq \max\{3t, 4e+4\}$, $\ell \geq 2$, $3 \leq h \leq \ell$, then $N'(n,\ell,e,h) > N'(n,\ell,e,h+1)$ for each h and thus, each N_h exists and attains unique value.

Proof: Since $N'(n,\ell,e,h+1)$ denotes the maximum value of $|\bigcap_{i=1}^{h+1} B_t(\mathbf{c}_i)|$ over all sets $T(Y) = \{\mathbf{c}_1,\ldots,\mathbf{c}_{h+1}\}$, we clearly have $N'(n,\ell,e,h) \geq N'(n,\ell,e,h+1)$. As we have seen in the previous theorems, the maximum value of $N'(n,\ell,e,h+1)$ is attained when the codewords in $T(Y) = \{\mathbf{c}_1,\ldots,\mathbf{c}_{h+1}\}$ have supports $\text{supp}(\mathbf{c}_i) = [(i-1)(e+1)+1, i(e+1)]$. We show that when we choose h of these codewords, then we can fit more words in the intersection of their t-balls. We clearly have $\bigcap_{i=1}^{h+1} B_t(\mathbf{c}_i) \subseteq \bigcap_{i=1}^{h} B_t(\mathbf{c}_i)$. In the following, we show that there exists a word $\mathbf{y} \in \bigcap_{i=1}^{h} B_t(\mathbf{c}_i) \setminus \bigcap_{i=1}^{h+1} B_t(\mathbf{c}_i)$ and hence, $|\bigcap_{i=1}^{h} B_t(\mathbf{c}_i)| > |\bigcap_{i=1}^{h+1} B_t(\mathbf{c}_i)|$.

Let $w(\mathbf{y}) = \ell + 1$, $|\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}_i)| = 1$ for $i \in [1,h]$, $\text{supp}(\mathbf{y}) \cap \text{supp}(\mathbf{c}_{h+1}) = \emptyset$ and $|\text{supp}(\mathbf{y}) \cap [(h+1)(e+1)+1, n]| = \ell + 1 - h$. We have $d(\mathbf{y},\mathbf{c}_i) = \ell + 1 + e + 1 - 2 = t$ for $i \in [1,h]$ but $d(\mathbf{y},\mathbf{c}_{h+1}) = \ell + 1 + e + 1 = t + 2$. Thus, the claim follows. □

Using Theorem 12, we can improve the bound $\mathcal{L} \leq \ell+1$ of Theorem 6 just by adding a constant number $(e+1)^{\ell+1}$ of channels.

Corollary 14: 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: Assume that $h = \ell + 1$. Now the function $N'(n,\ell,e,h)$ of Theorem 12 may be replaced by $N(n,\ell,e,h)$ due to Theorem 13. In order to calculate $N_{\ell+1}$, we consider the set W_w with $w \geq \ell$. We get that $i_j \geq 1$ for each $j \in [1,h]$ and thus, $w \geq \ell + 1$. On the other hand, $w \leq \ell + 1$ when $W_w \neq \emptyset$. Indeed, if $w \geq \ell + 2$, then $w \geq \sum_{j=1}^{h} i_j \geq \sum_{j=1}^{\ell+1} i_j \geq (\ell+1)(w+1-\ell)/2 = (\ell-1)(w+1-\ell)/2 + w - (\ell-1) > w$ and $W_w = \emptyset$ in this case. Therefore, as $w = \ell + 1$ and $i_j \geq 1$ for each j, we have $W_w = \{(1,1,\ldots,1)\}$. Thus, the sum corresponding

to W_w in Theorem 12 gives $V(n,\ell-1)+(e+1)^{\ell+1}$. Hence, if $N \geq N_{\ell+1}+1 = V(n,\ell-1)+(e+1)^{\ell+1}+1$, then $\mathcal{L} \leq \ell$. □

In the following corollary, we present the asymptotic behaviour of N_h on n. Notice that Corollary 14 considers the case $h=\ell+1$ and Corollary 15 cases $3 \leq h \leq \ell$.

Corollary 15: Let $\ell \geq h \geq 3$ and $e \geq 0$ be fixed, $n \geq n(e,\ell,b)$ and $b \geq \max\{3t, 4e+4\}$. If $N \geq N_h+1$, then $\mathcal{L} \leq h-1$, where

$$N_h \in V(n,\ell-1)+\binom{n-h(e+1)}{\ell+1-h}(e+1)^h+\Theta(n^{\ell-h}).$$

Proof: By Theorem 12, we have $N_h = V(n,\ell-1)+\sum_{w\geq\ell}\sum_{(i_1,\ldots,i_h)\in W_w}\binom{n-h(e+1)}{w-\sum_{j=1}^h i_j}\prod_{j=1}^h\binom{e+1}{i_j}$. When we inspect the set W_w closer, we notice that $\binom{n-h(e+1)}{w-\sum_{j=1}^h i_j}$ attains its maximum value when $w-\sum_{j=1}^h i_j$ is as large as possible since $w \leq 2e+\ell+1$ (recall that e and ℓ are constants with respect to n). For the maximum value of $w-\sum_{j=1}^h i_j$, let $a_j \in \mathbb{R}$ for each $j=1,\ldots,h$ be such that $a_j \geq 0$ and $i_j = (w+1-\ell)/2+a_j$; indeed, such a_j exists as $\lceil (w+1-\ell)/2\rceil \leq i_j \leq e+1$. This implies that

$$\begin{aligned} &w-\sum_{j=1}^h i_j \\ =&w-\left(h\cdot\frac{w+1-\ell}{2}+\sum_{j=1}^h a_j\right) \\ =&w\left(1-\frac{h}{2}\right)+h\cdot\frac{\ell-1}{2}-\sum_{j=1}^h a_j. \end{aligned}$$

Therefore, for the maximum, we require that the values a_j are as small as possible, i.e., $i_j = \lceil (w+1-\ell)/2\rceil$. Furthermore, as $h \geq 3$, the value $w-\sum_{j=1}^h i_j$ is maximized when w is as small as possible. In particular, when $w \in \{\ell,\ell+1\}$, we have $\lceil (w+1-\ell)/2\rceil = 1$ and when $w=\ell+1+a$ for some integer $a \geq 1$, we have $\lceil (w+1-\ell)/2\rceil = 1+\lceil a/2\rceil$. Moreover, when $w=\ell$, we have $w-\sum_{j=1}^h i_j \leq \ell-h$. When $w=\ell+1$, we have $w-\sum_{j=1}^h i_j \leq \ell+1-h$ and when $w=\ell+1+a$, we have $w-\sum_{j=1}^h i_j \leq \ell+1+a-h(1+\lceil a/2\rceil)$. Since $h \geq 3$ and $a \geq 1$, we have $\ell+1+a-h(1+\lceil a/2\rceil) \leq \ell+1-h-\lceil a/2\rceil \leq \ell-h$. Thus, we may concentrate on the case with $w=\ell+1$.

Furthermore, as $w=\ell+1$ and $i_j=1$ for each $j=1,2,\ldots,h$, we have $\prod_{j=1}^h\binom{e+1}{i_j}=(e+1)^h$. Recall that e is constant on n. Hence, for large n, it is enough to consider the binomial coefficient $\binom{n-h(e+1)}{\ell+1-h}$. Indeed, each binomial coefficient has a multiplier bounded by a constant and the number of binomial coefficients is also bounded by a constant since $|W_w|$ is bounded by a constant and $W_w \neq \emptyset$ only when $w \leq 2e+\ell+1$. Moreover, the second largest binomial coefficient is $\binom{n-h(e+1)}{\ell-h}$ and we have $\binom{n-h(e+1)}{k} \in \Theta(n^k)$, when n is large and $k \in \mathbb{N}$. Therefore, we have $N_h \in V(n,\ell-1)+\binom{n-h(e+1)}{\ell+1-h}(e+1)^h+\Theta(n^{\ell-h})$. □

In Theorem 2, a tight bound for the number of channels to certainly attain the list size $\mathcal{L} \leq 2$ is presented when code C has minimum distance d. Observe, that when we choose $h=3$ in Theorem 12, we attain the number of channels required to have $\mathcal{L} \leq 2$. The bound of Theorem 2 by Yaakobi and Bruck looks quite different from the bound in Theorem 12. However, Theorem 2 can be obtained as a corollary from Theorem 12 as shown in Corollary 16. The new presentation in Corollary 16 somewhat simplifies the inequalities for the indices compared to Theorem 2.

Corollary 16: Let $n \geq n(e,\ell,b)$, $b \geq \max\{3t, 4e+4\}$ and $\ell \geq 2$. If $N \geq N' = \sum_{i_1,i_2,i_3,i_4}\binom{n-3e-3}{i_1}\binom{e+1}{i_2}\binom{e+1}{i_3}\binom{e+1}{i_4}+1$ for

- $0 \leq i_1 \leq \ell-1$,
- $0 \leq i_4 \leq \ell-1-i_1$,
- $0 \leq i_3 \leq \ell-1-i_1$ and
- $i_1+i_3+i_4-(\ell-1) \leq i_2 \leq \ell-1-i_1-|i_4-i_3|$,

then $\mathcal{L} \leq 2$ for any e-error-correcting code C. Moreover, the value of N' is equal to the lower bound obtained in Theorem 2.

Proof: We first show that the formulation of the bound follows from Theorem 12 when $h=3$. This gives the minimum number of channels N required to have $\mathcal{L} \leq 2$. In particular, we have

$$\begin{aligned} N_3 =&V(n,\ell-1)+\sum_{w\geq\ell}\sum_{(i_2,i_3,i_4)\in W_w}\binom{n-3e-3}{w-i_2-i_3-i_4} \\ &\cdot\binom{e+1}{i_2}\binom{e+1}{i_3}\binom{e+1}{i_4}. \end{aligned}$$

We have renamed the indices for convenience (the index i_1 will be saved for later use in the proof). Moreover, we have $W_w = \{(i_2,i_3,i_4) \mid (w+1-\ell)/2 \leq i_j \leq e+1, w \geq i_2+i_3+i_4\}$. Earlier, we have used W_w only with the assumption that $w \geq \ell$. However, we may allow here that $w < \ell$. Hence, we may have some binomial coefficients with $i_j < 0$ for some j. In these cases (and when $i_j > e+1$), we use the common convention that the binomial coefficient attains the value 0. Observe that $N_3 = \sum_{w\geq 0}\sum_{(i_2,i_3,i_4)\in W_w}\binom{n-3e-3}{w-i_2-i_3-i_4}\binom{e+1}{i_2}\binom{e+1}{i_3}\binom{e+1}{i_4}$. Indeed, since $W_w = \{(i_2,i_3,i_4) \mid 0 \leq i_j \leq e+1, w \geq i_2+i_3+i_4\}$ for $w \leq \ell-1$, we have

$$\begin{aligned} &\sum_{w=0}^{\ell-1}\sum_{(i_2,i_3,i_4)\in W_w}\binom{n-3e-3}{w-i_2-i_3-i_4}\binom{e+1}{i_2} \\ &\cdot\binom{e+1}{i_3}\binom{e+1}{i_4} \\ =&\sum_{w=0}^{\ell-1}\sum_{i_2=0}^{e+1}\sum_{i_3=0}^{e+1}\sum_{i_4=0}^{e+1}\binom{n-3e-3}{w-i_2-i_3-i_4}\binom{e+1}{i_2} \\ &\cdot\binom{e+1}{i_3}\binom{e+1}{i_4} \\ =&V(n,\ell-1) \end{aligned}$$

due to binomial identity $\sum_{p=0}^s\binom{m-s}{k-p}\binom{s}{p}=\binom{m}{k}$.

Let us denote by $i_1 = w-i_2-i_3-i_4$. We now get that $2i_2 \geq 2(w+1-\ell)/2 = i_1+i_2+i_3+i_4-\ell+1$ and similar inequalities for i_3 and i_4. Since we do not have to take into account the lower bound $i_1 \geq 0$ (the cases with $i_1 < 0$ increase the binomial sum by 0) or the cases with $i_j > e+1$ for $j \in \{2,3,4\}$, we can consider the following system

of inequalities:

$$i_2 \geq i_1 + i_3 + i_4 - \ell + 1 \tag{9}$$
$$i_3 \geq i_1 + i_2 + i_4 - \ell + 1 \tag{10}$$
$$i_4 \geq i_1 + i_2 + i_3 - \ell + 1. \tag{11}$$

Our goal is to show that this system of inequalities is equivalent to the following system of inequalities:

$$i_4 \leq \ell - 1 - i_1, \tag{12}$$
$$i_3 \leq \ell - 1 - i_1, \tag{13}$$
$$i_1 + i_3 + i_4 - (\ell - 1) \leq i_2 \leq \ell - 1 - i_1 - |i_4 - i_3|. \tag{14}$$

Let us first show that the second system of inequalities follows from the first one.

Inequality (12) follows from

$$i_4 = w - i_1 - i_2 - i_3 \leq w - i_1 - 2(w - \ell + 1)/2 = \ell - 1 - i_1.$$

We obtain Inequality (13) in a similar manner. Moreover, from Inequalities (10) and (11) we obtain $i_2 \leq \ell - 1 - i_1 - i_4 + i_3$ and $i_2 \leq \ell - 1 - i_1 - i_3 + i_4$, respectively. Together, these imply

$$i_2 \leq \ell - 1 - i_1 - |i_4 - i_3|.$$

Finally, the lower bound inequality in (14) follows directly from (9).

Let us then show that the first system of inequalities follows from the second one. First of all, Inequality (9) follows immediately from Inequality (14). Assume first that $i_4 \geq i_3$. Then the upper bound of Inequality (14) is $i_2 \leq \ell - 1 - i_1 - i_4 + i_3$. This implies Inequality (10) and inequality (11) since

$$i_4 \geq i_3 \geq i_1 + i_2 + i_4 - \ell + 1 \geq i_1 + i_2 + i_3 - \ell + 1.$$

The case with $i_3 \geq i_4$ is similar.

Finally, we may add lower bounds $i_j \geq 0$ for all $j \in \{1, 2, 3, 4\}$ due to binomial coefficient context. Similarly we notice that if $i_1 \geq \ell$, then $i_4 < 0$. Thus, we may also add upper bound $i_1 \leq \ell - 1$. Hence, the first part of the claim follows.

Let us then derive the bound of Theorem 2 by Yaakobi and Bruck from this new lower bound. The case with $d = 2e+1$ is included in Appendix and we consider here only the case with $d = 2e + 2$. When we have $d = 2e + 2$, Theorem 2 can be presented in the following way: If $N \geq \sum_{h_1,h_2,h_3,h_4} \binom{n-3e-3}{h_1}\binom{e+1}{h_2}\binom{e+1}{h_3}\binom{e+1}{h_4} + 1$ for

- $0 \leq h_1 \leq \ell - 1$,
- $h_1 - (\ell - 1) \leq h_4 \leq \ell - 1 - h_1$,
- $e + 2 - \ell + h_1 \leq h_3 \leq t - (h_1 + h_4)$ and
- $\max\{h_1 - h_3 - h_4 + 2e + 3 - \ell, h_1 + h_3 + h_4 - \ell + 1\} \leq h_2 \leq \ell - 1 - (h_1 + h_4 - h_3)$,

then $\mathcal{L} \leq 2$ for any e-error-correcting code C (with minimum distance $d = 2e + 2$). Next, we modify the presentation we got for N_3 into the formulation above.

Let us denote by $i'_2 = e + 1 - i_2$ and by $i'_3 = e + 1 - i_3$. Observe that $\binom{e+1}{i'_j} = \binom{e+1}{i_j}$ for $j \in \{2, 3\}$. We can replace the lower bound $i_4 \geq 0$ by $i_4 \geq i_1 - (\ell - 1)$ since $i_4 \geq 0 \geq i_1 - \ell + 1$ and $\binom{e+1}{i_4} = 0$ when $i_4 < 0$. Moreover, we have $0 \leq i_3 \leq \ell - 1 - i_1$. Hence, $e+1 \geq i'_3 \geq e+2-\ell+i_1$. Notice that the upper bound on i'_3 can be replaced by $t-(i_1+i_4)$ since $t - (i_1 + i_4) \geq e + 1$ as $i_1 + i_4 \leq \ell - 1$ and $\binom{e+1}{i'_3} = 0$ when $i'_3 > e + 1$.

For i_2 we have $i_1+i_3+i_4-(\ell-1) \leq i_2 \leq \ell-1-i_1-|i_4-i_3|$ and hence, $\ell-1-(i_1+i_4-i'_3) \geq i'_2 \geq -\ell+e+2+i_1+|i_4-i_3| = e + 2 - \ell + i_1 + |i_4 + i'_3 - e - 1| = \max\{i_1 - i'_3 - i_4 + 2e + 3 - \ell, i_1 + i'_3 + i_4 - \ell + 1\}$. By comparing these inequalities with the bounds used in Theorem 2, we notice that they are identical. The case with $d = 2e + 1$ is similar and is included in Appendix A. Hence, we get the claim. □

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 Theorem 6), for smaller lengths of the codes our best bound is still $\mathcal{L} \leq 2^\ell$ (see Theorem 4) when the number of channels satisfies $N \geq V(n, \ell - 1) + 1$. Although this bound is attained in some cases (see [24]) 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 as we have seen in Theorem 11. 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 [25]. 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 (see [26], [27]). Let $\mathcal{F}$ be a family of subsets of $[1, n]$ (that is, $\mathcal{F}$ is a subset of the power set $2^{[1,n]}$), where n is a fixed positive integer. We say that a subset S of $[1, n]$ is *shattered* by $\mathcal{F}$ if for any subset $E \subseteq S$ there exists a set $F \in \mathcal{F}$ such that $F \cap S = E$. The Sauer-Shelah lemma states that if $|\mathcal{F}| > \sum_{i=0}^{k-1} \binom{n}{i}$, then $\mathcal{F}$ shatters a subset of size (at least) k. Since the subsets of $[1, n]$ can naturally be interpreted as words of $\mathbb{F}^n$, the Sauer-Shelah lemma can be reformulated as follows. Notice that $\sum_{i=0}^{k-1} \binom{n}{i} = V(n, k - 1)$.

Theorem 17 ([26], [27]): 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 $\text{supp}(\mathbf{w}) \subseteq S$ there exists a word $\mathbf{s} \in Y$ satisfying $\text{supp}(\mathbf{w}) = \text{supp}(\mathbf{s}) \cap S$. Here we say that the set S of coordinates is *shattered* by Y.

Observe that if C is an e-error-correcting code then 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 18 [24]: Let $C \subseteq \mathbb{F}^n$ be an e-error-correcting code and N the number of channels. If for any set of output words $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ we have

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

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

The words β_i can (and often do) depend on the set Y of the output words, but their number remains the same. Notice that Lemma 18 also gives a decoding algorithm. Indeed, if the words β_i are known, then there is at most one codeword in

each $B_e(\beta_i)$, the decoding algorithm of C can be used on β_i and the codeword can be added to the list T.

Theorem 19: 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 and Y'' be a set of output words such that $|Y''| \geq V(n, \ell + 2R - 1) + 1$. Due to Theorem 17, we know that there is a subset Y' (of cardinality $2^{\ell+2R}$) of the words in Y'' such that some $\ell + 2R$ coordinates of Y' contain all the $2^{\ell+2R}$ words of length $\ell + 2R$. Let us denote this set of coordinates by S. The words of Y' in these coordinates clearly correspond to the words in $\mathbb{F}^{\ell+2R}$.

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. Consequently, if we remove from $\mathbb{F}^{\ell+2R}$ at most $2^{\ell+2R-\dim(D)} - 1$ words, then at least one of the cosets remains intact.

Therefore, if the set of output words Y satisfies $|Y| \geq V(n, \ell + 2R - 1) + 1 - (2^{\ell+2R-k[\ell+2R,R]} - 1)$, then we can find among the set of output words a subset $Y_1 = \{\mathbf{y}_1, \ldots, \mathbf{y}_{2^{k[\ell+2R,R]}}\} \subseteq Y$ such that the words corresponding to the coordinates of the set S form an R-covering code.

Now let $\mathbf{s} \in \mathbb{F}^n$ be a word such that $\mathrm{supp}(\mathbf{s}) = S$. Denote $\beta_i = \mathbf{s} + \mathbf{y}_i$ for $i = 1, \ldots, 2^{k[\ell+2R,R]}$. 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 $\mathbf{y}_j$, $j \in \{1, \ldots, 2^{k[\ell+2R,R]}\}$, such that the words $\mathbf{y}_j$ and $\mathbf{x} + \mathbf{s}$ differ in at most R places among the coordinates of the set 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. Indeed, notice that $\mathbf{y}_j$ has, among the coordinates of S, at most R bits that are same as in $\mathbf{x}$ and at least $\ell + R$ different bits than in $\mathbf{x}$ and in the word $\mathbf{y}_j + \mathbf{s}$ these bits are changed. Therefore, by Lemma 18, 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 4, 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 $\mathcal{L} \leq 16$.

Notice that we can use in the proof of the previous theorem also the smallest (with respect to the cardinality) known linear R-covering code instead of an optimal code D achieving the exact value of $\dim(D) = k[\ell + 2R, R]$.

V. List Size With Fewer Channels

By the following theorem (of [24]), it is clear that if we have *fewer* 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 20 [24]: Let $V(n, \ell - p - 1) + 1 \leq N \leq V(n, \ell - p)$ where $0 \leq p \leq \ell - 1$. Moreover, let $C \subseteq \mathbb{F}^n$ be such an e-error-correcting code that $\mathcal{L}$ is maximal. Then we have

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

Due to this result, in order to have a smaller list size, let us concentrate on e-error-correcting codes with at most M codewords within any ball of radius $e + a$, for some $a > 0$. An excellent source for these codes is, for example, [28].

Theorem 21: 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,

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

Proof: Assume that we received the set Y of output words from the channels where $|Y| \geq V(n, \ell - a - 1) + 1$. Due to Theorem 17, we know that there exists a subset $Y' \subseteq Y$ of size $|Y'| = 2^{\ell-a}$ such that these words have in some $\ell - a$ coordinates all possible words of length $\ell - a$. Denote this set of coordinates by S. Suppose $\mathbf{x}$ is the input word and denote $Y' = \{\mathbf{y}_1, \ldots, \mathbf{y}_{2^{\ell-a}}\}$. Let $\beta_i = \mathbf{y}_i + \mathbf{s}$, $i = 1, \ldots, 2^{\ell-a}$ where $\mathrm{supp}(\mathbf{s}) = S$. It is easy to check that $\mathbf{x}$ has distance at most $e + a$ from one of the β_i's. Since there are at most $2^{\ell-a}$ β_i's and there are at most M codewords within distance $e + a$ from each of them, we obtain the bound $\mathcal{L} \leq 2^{\ell-a} M$. □

The previous result is useful when our e-error-correcting code is a code for traditional list-decoding, see [28]. Here we allow e and ℓ to depend on n. For the number of channels being smaller than $V(n, \ell - 1) + 1$, it also gives, for every e-error-correcting code with suitable a, a small exponent for n compared to Theorem 20 (see Corollary 23(ii) below), or even constant bounds (see Corollary 23(i)) on $\mathcal{L}$.

Let us denote

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

and

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

Notice that $1/2 \leq r(n, e) - e \leq e + 1$. In the following theorem, we reformulate the result of [28, Theorem 3.2] using our notations.

Theorem 22 [28, Theorem 3.2]: Let n, M and e be positive integers where $2e + 1 \leq n/2$. Moreover, let C be an e-error-correcting code and r be an integer with $r \geq 1$. If $r \leq r(n, e, M)$, then $|B_r(\mathbf{x}) \cap C| \leq M$ for every $\mathbf{x} \in \mathbb{F}^n$. If $r < r(n, e)$, then $|B_r(\mathbf{x}) \cap C| \leq n$ for every $\mathbf{x} \in \mathbb{F}^n$.

Corollary 23: Let C be an e-error-correcting code, $t = e + \ell$, an integer $M \geq 1$ and $2e + 1 < n/2$ ($n \geq 8$). We have

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

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

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

$$\mathcal{L} \leq 2^{t-r(n,e)+1} n.$$

Proof: The claim follows straightforwardly from applying the results in Theorem 22 to Theorem 21. Indeed, in Theorem 21 for the case (i) choose $a = r(n, e, M) - e$ ($0 \leq a \leq \ell - 1$) and for the case (ii) $a = r(n, e) - e - 1$. Notice also in (ii) that the condition $r(n, e) \geq 2$ follows from $e > 1$. □

We remark that the result $\mathcal{L} \leq 2^\ell$ in Theorem 4 requires at least $V(n, \ell-1)+1$ channels. In what follows, we continue our study of the list size for codes with $|B_{e+a}(\mathbf{u}) \cap C| \leq M$ for every $\mathbf{u} \in \mathbb{F}^n$. Here we consider e and ℓ to be again constants with respect to n. First, we introduce two technical lemmas. Lemma 24 can be seen as an easy extension of [24, Lemma 13] for smaller number of channels and Lemma 25 as an extension of Lemma 9.

Lemma 24: Let $N \geq V(n, \ell-a-1)+1$, $\ell-1 \geq a \geq 1$ and $n \geq (\ell-a-1)^2 2^b + \ell - a - 2$. Then for any codeword $\mathbf{c} \in T(Y)$ and any set $\overline{D} \subseteq [1, n]$ with cardinality $|\overline{D}| = b$ there exists an output word $\mathbf{y} \in Y$ such that $|\mathrm{supp}(\mathbf{y}+\mathbf{c}) \setminus \overline{D}| \geq \ell - a - 1$.

Proof: Let us assume without loss of generality that $\mathbf{c} = \mathbf{0}$. Thus, $w(\mathbf{y}) \leq t$ for each $\mathbf{y} \in Y$. Let us count the number of binary words of weight at most t which have $|\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| < \ell - a - 1$. There are at most

$$\begin{aligned}
&\sum_{j=0}^{\ell-a-2} \sum_{i=0}^{\min\{b, t-j\}} \binom{b}{i} \binom{n-b}{j} \\
\leq &(\ell-a-1) 2^b \binom{n}{\ell-a-2} \\
= &(\ell-a-1) 2^b \frac{\ell-a-1}{n-\ell+a+2} \binom{n}{\ell-a-1} \\
\leq &\binom{n}{\ell-a-1}
\end{aligned}$$

such words, which is strictly less than N and hence, the claim holds. □

Lemma 25: Let $N \geq V(n, \ell-a-1)+1$, $\ell-1 \geq a \geq 1$, $b = 2t$, $n \geq (\ell-a-1)^2 2^b + \ell - a - 2$ and $\mathcal{L} \geq 2$. Then for any two codewords $\mathbf{c}_1$ and $\mathbf{c}_2$ in $T(Y)$ it holds that

$$2e+1 \leq d(\mathbf{c}_1, \mathbf{c}_2) \leq 2e+2a+2.$$

Proof: Let us assume without loss of generality that $\mathbf{c}_1 = \mathbf{0}$. Notice that $w(\mathbf{c}_2) \leq 2t$. Choose $\overline{D} = \mathrm{supp}(\mathbf{c}_2)$. Now, by Lemma 24, there exists an output word $\mathbf{y} \in Y$ such that $|\mathrm{supp}(\mathbf{y}) \setminus \overline{D}| \geq \ell - a - 1$. Since $d(\mathbf{c}_2, \mathbf{y}) \leq t$ and $w(\mathbf{y}) \leq t$, we have $t \geq \ell - a - 1 + (w(\mathbf{c}_2) - (t - (\ell - a - 1))) = \ell - 2a - 2 - e + w(\mathbf{c}_2)$ and hence, $w(\mathbf{c}_2) \leq 2e + 2a + 2$. Therefore, $d(\mathbf{c}_1, \mathbf{c}_2) \leq 2e + 2a + 2$. Since we consider an e-error-correcting code, the lower bound follows. □

Together with the two previous lemmas, we can now prove an upper bound for $\mathcal{L}$, depending on the maximum number M of codewords in an $(e+a)$-radius ball. The result requires a large n and it would be interesting to know whether a similar result holds for smaller values of n as well. Recall that ε denotes the Napier's constant.

Theorem 26: Let $N \geq V(n, \ell-a-1)+1$, $\ell-1 \geq a \geq 1$, $b = \left\lceil (2e+2a+2)^{\varepsilon \cdot (e+a+1)!} \right\rceil$, $n \geq (\ell-a-1)^2 2^b + \ell - a - 2$. Moreover, let C be such an e-error-correcting code that there are at most M codewords in any $(e+a)$-radius ball. Then

$$\mathcal{L} \leq \max\{(t+1)M, b/(2e+2a+2)\}.$$

Proof: Let us denote $T(Y) = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_\mathcal{L}\}$. If $\mathcal{L} \leq b/(2e+2a+2)$, then the claim follows. Assume now that $\mathcal{L} > b/(2e+2a+2)$. Assume then, without loss of generality, that $\mathbf{c}_1 = \mathbf{0}$. By Lemma 25, we have $w(\mathbf{c}_i) \leq 2e+2a+2$ for each $i \in [1, \mathcal{L}]$.

Let Z be a subset of $\mathbb{F}^n$. We say that $\mathbf{w} \in \mathbb{F}^n$ is a *central word* with respect to Z if $d(\mathbf{w}, \mathbf{z}) \leq e+a+1$ for all $\mathbf{z} \in Z$. Moreover, the set of all central words with respect to Z is denoted by W_Z. In what follows, we first show a useful observation stating that if a subset $C_S \subseteq T(Y)$ is such that $\mathbf{0} \in C_S$ and $|C_S| \leq b/(2e+2a+2)$, then there exists a word $\mathbf{w} \in \mathbb{F}^n$ such that $w(\mathbf{w}) \leq e+a+1$ and $d(\mathbf{w}, \mathbf{c}) \leq e+a+1$ for any $\mathbf{c} \in C_S$, i.e., $\mathbf{w} \in W_{C_S} \neq \emptyset$. Since $w(\mathbf{c}) \leq 2e+2a+2$ for any $\mathbf{c} \in C_S$, we have

$$\left| \bigcup_{\mathbf{c} \in C_S} \mathrm{supp}(\mathbf{c}) \right| \leq (2e+2a+2) \cdot \frac{b}{2e+2a+2} = b.$$

Therefore, by Lemma 24, there exists an output word $\mathbf{y} \in Y$ such that

$$|\mathrm{supp}(\mathbf{y}) \setminus \bigcup_{\mathbf{c} \in C_S} \mathrm{supp}(\mathbf{c})| \geq \ell - 1 - a.$$

Let then $\mathbf{w} \in \mathbb{F}^n$ be such that

$$\mathrm{supp}(\mathbf{w}) = \mathrm{supp}(\mathbf{y}) \cap \bigcup_{\mathbf{c} \in C_S} \mathrm{supp}(\mathbf{c}).$$

Now, as $d(\mathbf{y}, \mathbf{c}) \leq t = e + \ell$ for any $\mathbf{c} \in C_S$, we have $d(\mathbf{w}, \mathbf{c}) \leq e+a+1$. Moreover, $w(\mathbf{w}) \leq e+a+1$ since $\mathbf{0} \in C_S$.

In what follows, we show using an iterative approach that there exists a central word $\mathbf{w} \in \mathbb{F}^n$ with respect to $T(Y)$. We begin the iterative process by considering a subset $C_0 = \{\mathbf{c}_1, \mathbf{c}_2\} \subseteq T(Y)$ such that $\mathbf{c}_1 = \mathbf{0}$ and $w(\mathbf{c}_2) = e+a+1+p_1$ with $1 \leq p_1 \leq e+a+1$. Indeed, we may assume that such a codeword $\mathbf{c}_2$ exists since otherwise we are immediately done due to $\mathbf{w} = \mathbf{0}$ being the searched central word. Observe that the weight of a central word $\mathbf{w} \in W_{C_0}$ satisfies $p_1 \leq w(\mathbf{w}) \leq e+a+1$. In particular, there are exactly $\binom{e+a+1+p_1}{p_1}$ central words $\mathbf{w} \in W_{C_0}$ of weight p_1. For each such central word $\mathbf{w}$, we may assume that there exists a codeword $\mathbf{c} \in T(Y)$ such that $d(\mathbf{w}, \mathbf{c}) > e+a+1$ as otherwise $\mathbf{w} \in W_{T(Y)}$ and we are done. Now we form a new code C_1 by adding such a codeword $\mathbf{c}$ for each $\mathbf{w} \in W_{C_0}$ with $w(\mathbf{w}) = p_1$. The number of added codewords is at most

$$\binom{e+a+1+p_1}{p_1} \leq (2e+2a+2)^{p_1} - 1.$$

Therefore, we have $|C_1| \leq (2e+2a+2)^{p_1} + 1$. Notice that there are no central words in W_{C_1} of weight at most p_1. Furthermore, by the previous observation for W_{C_S}, W_{C_1} is nonempty as $|C_1| \leq (2e+2a+2)^{p_1} + 1 \leq b/(2e+2a+2)$.

Assume that $p_2 > p_1$ is now the smallest weight of a central word in W_{C_1}. Let $\mathbf{w}$ be a central word with respect to C_1 of weight p_2. The support of $\mathbf{w}$ is a subset of $\bigcup_{\mathbf{c} \in C_1} \mathrm{supp}(\mathbf{c})$ since otherwise there exists a central word $\mathbf{w}' \in W_{C_1}$ with $w(\mathbf{w}') < w(\mathbf{w})$ (a contradiction). Therefore, as ($\mathbf{c}_1 = \mathbf{0} \in C_1$ implies) $\left|\bigcup_{\mathbf{c} \in C_1} \mathrm{supp}(\mathbf{c})\right| \leq (2e+2a+2)(2e+2a+2)^{p_1}$,

the number of central words of weight p_2 in W_{C_1} is at most

$$\begin{aligned}&\binom{(2e+2a+2)(2e+2a+2)^{p_1}}{p_2}\\ =&\binom{(2e+2a+2)^{p_1+1}}{p_2}\\ \leq&\frac{(2e+2a+2)^{p_1p_2+p_2}}{2}.\end{aligned}$$

Again, for each central word $\mathbf{w} \in W_{C_1}$ of weight p_2, there exists a codeword $\mathbf{c} \in T(Y)$ such that $d(\mathbf{w}, \mathbf{c}) > e + a + 1$. Now we form a new code C_2 by adding such a codeword $\mathbf{c}$ for each $\mathbf{w} \in W_{C_1}$ with $w(\mathbf{w}) = p_2$. Thus, we have $|C_2| \leq (2e + 2a + 2)^{p_1} + 1 + (2e + 2a + 2)^{p_1p_2+p_2}/2 \leq (2e + 2a + 2)^{p_1p_2+p_2}/2 + (2e+2a+2)^{p_1p_2+p_2}/2 = (2e+2a+2)^{p_1p_2+p_2}$.

The process can be iteratively continued by forming a new code C_i based on the previous code C_{i-1}, until we have reached $p_i = e+a+1$ or we have already found a central word with respect to $T(Y)$. In what follows, the iterative process is explained in more detail:

- Assume that $p_i > p_{i-1}$ is the smallest weight of a central word in $W_{C_{i-1}}$.
- Assume that $|C_{i-1}| \leq (2e + 2a + 2)^{\sum_{h=1}^{i-1} \prod_{k=h}^{i-1} p_k}$. By the observation above for W_{C_S}, this implies that $W_{C_{i-1}}$ is nonempty as $|C_{i-1}| \leq b/(2e + 2a + 2)$ (see also Equation (16)). Furthermore, we have $|\bigcup_{\mathbf{c}\in C_{i-1}} \mathrm{supp}(\mathbf{c})| \leq (2e + 2a + 2)|C_{i-1}| = (2e + 2a + 2)^{(\sum_{h=1}^{i-1} \prod_{k=h}^{i-1} p_k)+1}$.
- Let $\mathbf{w} \in W_{C_{i-1}}$ be of weight p_i. As previously, the support of $\mathbf{w}$ is a subset of $\bigcup_{\mathbf{c}\in C_{i-1}} \mathrm{supp}(\mathbf{c})$ since otherwise there exists a central word $\mathbf{w}' \in W_{C_{i-1}}$ with $w(\mathbf{w}') < w(\mathbf{w})$ (a contradiction). Therefore, as $|\bigcup_{\mathbf{c}\in C_{i-1}} \mathrm{supp}(\mathbf{c})| \leq (2e + 2a + 2)|C_{i-1}| = (2e + 2a + 2)^{(\sum_{h=1}^{i-1} \prod_{k=h}^{i-1} p_k)+1}$, the number of central words in $W_{C_{i-1}}$ of weight p_i is at most $\binom{(2e+2a+2)^{(\sum_{h=1}^{i-1} \prod_{k=h}^{i-1} p_k)+1}}{p_i}$.
- Again, for each central word $\mathbf{w} \in W_{C_{i-1}}$ of weight p_i, there exists a codeword $\mathbf{c} \in T(Y)$ such that $d(\mathbf{w}, \mathbf{c}) > e+a+1$. Now we form a new code C_i by adding such a codeword $\mathbf{c}$ for each $\mathbf{w} \in W_{C_{i-1}}$ with $w(\mathbf{w}) = p_i$. Thus, we have

$$\begin{aligned}|C_i| \leq&(2e + 2a + 2)^{\sum_{h=1}^{i-1} \prod_{k=h}^{i-1} p_k}\\ &+\binom{(2e + 2a + 2)^{1+(\sum_{h=1}^{i-1} \prod_{k=h}^{i-1} p_k)}}{p_i}\\ \leq&(2e + 2a + 2)^{\sum_{h=1}^{i} \prod_{k=h}^{i} p_k}/2\\ &+(2e + 2a + 2)^{\sum_{h=1}^{i} \prod_{k=h}^{i} p_k}/2 \qquad (15)\\ =&(2e + 2a + 2)^{\sum_{h=1}^{i} \prod_{k=h}^{i} p_k}.\end{aligned}$$

Notice that since $1 \leq p_1 < p_2 < \cdots < p_i \leq e + a + 1$, we reach $p_j = e + a + 1$ at some point (or the central word $\mathbf{w}$ with respect to $T(Y)$ has already been found in an earlier step). By (15), we have

$$\begin{aligned}|C_j| \leq&(2e + 2a + 2)^{\sum_{h=1}^{j} \prod_{k=h}^{j} p_k}\\ \leq&(2e + 2a + 2)^{\sum_{h=1}^{e+a+1} \prod_{k=h}^{e+a+1} k}\\ =&(2e + 2a + 2)^{\sum_{h=0}^{e+a} (e+a+1)!/h!}\\ \leq&(2e + 2a + 2)^{\varepsilon\cdot(e+a+1)!-1}\\ \leq&\frac{b}{2e + 2a + 2} \qquad (16)\end{aligned}$$

since $\sum_{h=0}^{e+a+1} 1/h! < \sum_{h=0}^{\infty} 1/h! = \varepsilon$. Therefore, by the previous observation, W_{C_j} is nonempty. Thus, in conclusion, there exists a central word $\mathbf{w} \in \mathbb{F}^n$ with respect to $T(Y)$.

Let us now translate the Hamming space so that $\mathbf{w} = \mathbf{0}$ and thus, $T(Y) \subseteq B_{e+a+1}(\mathbf{0})$. In other words, we have $w(\mathbf{c}_i) \leq e+a+1$ for each i. Recall that we have $N \geq V(n, \ell-a-1)+1$. Thus, there exists a word $\mathbf{y} \in Y$ such that $w(\mathbf{y}) \geq \ell - a$. Moreover, since $d(\mathbf{c}_i, \mathbf{y}) \leq t$ for each i, we have $w(\mathbf{y}) \leq t + e + a + 1$. The proof now divides into the following three cases depending on the weight of $w(\mathbf{y})$.

(i) Assume first that $\ell - a \leq w(\mathbf{y}) \leq t$. Now the support of each $\mathbf{c}_i \in T(Y)$ of weight $e + a + 1$ intersects with $\mathrm{supp}(\mathbf{y})$ since otherwise $d(\mathbf{y}, \mathbf{c}_i) \geq (e + a + 1) + (\ell - a) = t + 1$ (a contradiction). Hence,

$$T(Y) \subseteq B_{e+a}(\mathbf{0}) \cup \bigcup_{i\in\mathrm{supp}(\mathbf{y})} B_{e+a}(\mathbf{e}_i).$$

(ii) Assume then that $t \leq w(\mathbf{y}) \leq t + e + a$. Let $\mathbf{y}_s$ be a word such that $\mathrm{supp}(\mathbf{y}_s) \subseteq \mathrm{supp}(\mathbf{y})$ and $w(\mathbf{y}_s) = t$. Now the support of each $\mathbf{c}_i \in T(Y)$ of weight $e+a+1$ intersects with $\mathrm{supp}(\mathbf{y}_s)$ since otherwise $d(\mathbf{y}, \mathbf{c}_i) \geq t + 1$ (a contradiction). Hence,

$$T(Y) \subseteq B_{e+a}(\mathbf{0}) \cup \bigcup_{i\in\mathrm{supp}(\mathbf{y}_s)} B_{e+a}(\mathbf{e}_i).$$

(iii) Assume finally that $w(\mathbf{y}) = t+e+a+1$. Then we have $w(\mathbf{c}_i) = e + a + 1$ for any $\mathbf{c}_i \in T(Y)$ as $d(\mathbf{y}, \mathbf{c}_i) \leq t$. Let $\mathbf{y}_s$ be a word such that $\mathrm{supp}(\mathbf{y}_s) \subseteq \mathrm{supp}(\mathbf{y})$ and $w(\mathbf{y}_s) = t + 1$. Again the support of each $\mathbf{c}_i \in T(Y)$ of weight $e + a + 1$ intersects with $\mathrm{supp}(\mathbf{y}_s)$. Observing that $w(\mathbf{c}_i) = e+a+1$ for any $\mathbf{c}_i \in T(Y)$ as $d(\mathbf{y}, \mathbf{c}_i) \leq t$, we have

$$T(Y) \subseteq \bigcup_{i\in\mathrm{supp}(\mathbf{y}_s)} B_{e+a}(\mathbf{e}_i).$$

Based on the cases (i)–(iii), the set of codewords $T(Y)$ is always contained in a union of $t + 1$ balls of radius $e + a$. Thus, as each ball of radius $e + a$ has at most M codewords, we obtain that $\mathcal{L} \leq (t + 1)M$. □

VI. Decoding With Majority Algorithm

In this section, we focus on decoding the transmitted word $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in C$ based on the set Y of the output words using a majority algorithm. For the rest of the section, we assume that n, t and e are constants (compared to N) and 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. Probabilistic set-up has been studied for different error types, for example, in [22] and [23]. Our approach differs from these articles in that we have given an upper limit for the possible number of errors in any single channel. This allows us to have a verifiability property in Theorem 30 unlike, for example, in [22] and [23]. With the verifiability property we mean that

although we cannot be certain whether we can deduce the transmitted word correctly before looking at the output words, we can sometimes deduce the transmitted word with complete certainty after seeing the output words. In other words, some output word sets have properties which allow us to know the transmitted word with certainty.

First we describe the (well-known) majority algorithm using similar terminology and notation as in [15]. The coordinates of the output words $\mathbf{y}_j \in Y$ are denoted by $\mathbf{y}_j = (y_{j,1}, y_{j,2}, \ldots, y_{j,n})$. 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 $\mathbf{z}_Y = \mathbf{z} = (z_1, z_2, \ldots, z_n) \in \mathbb{F}^n$, where

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

In other words, for each coordinate of $\mathbf{z}$, we choose ?, 0 or 1 based on whether the numbers of 0's and 1's are equal or which one occurs more frequently. Observe that the coordinate z_i outputted by the majority algorithm is equal to x_i if and only if at most $\lceil N/2 \rceil - 1$ errors occur in the ith coordinates of Y. Observe that the complexity of the majority algorithm is $\Theta(Nn)$ and since reading all the output words takes $\Theta(Nn)$ time, majority algorithm has optimal time complexity.

In [15, Example 1], it is shown that the majority algorithm does not always output the correct transmitted word $\mathbf{x}$ when the number of channels N is equal to the value in Theorem 1 even though we take the e-error-correction capability of C into account. In [7], 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 there exists among the output words of the modified algorithm one belonging to $B_e(\mathbf{x})$ and this word can be uniquely decoded to $\mathbf{x}$. In what follows, we demonstrate that *with high probability* the word $\mathbf{z}$ is within distance e from $\mathbf{x}$ with significantly smaller number of channels (than in [7]). For example, the Monte Carlo simulations for the values $t = 5$ and $n = 28$ are illustrated in Table I.

For this purpose, we first consider a variant of the so called *multiple birthday problem*; the multiple birthday problem has been studied, for example, in [29] and [30]. Here we assume that s, q, n and t are integers satisfying $2 \leq s \leq q$ and $0 \leq t \leq n$. A *throw* consists of placing t balls randomly into n buckets in such a way that each ball lands in a different bucket and each subset of the buckets of size t for a throw has an equal probability. Denote then by $C_t(n, q, s)$ the event that after q throws, at least one bucket contains at least s balls. Observe that if $t = 1$, then we are actually considering the multiple birthday problem; furthermore, if ($t = 1$ and) $s = 2$, then the case is the (regular) birthday problem. Furthermore, by $Pr[C_t(n, q, s)]$ we denote the probability that there exist at least s balls in a bucket. Notice that an output word of Y with *exactly* t errors can be interpreted in terms of the described variant of the multiple birthday problem as follows: an output $\mathbf{y}_i \in Y$ with exactly t errors can be considered as a throw of t balls into n buckets. In the following theorem, we present (based on this interpretations and the probability $Pr[C_t(n, q, s)]$) a lower bound on the probability that the output $\mathbf{z}$ of the majority algorithm is equal to $\mathbf{x}$.

Theorem 27: Let $\mathbf{x}$ be the transmitted codeword of $C \subseteq \mathbb{F}^n$ and N the number of channels. The probability that the output $\mathbf{z}$ of the majority algorithm is equal to $\mathbf{x}$ is at least

$$1 - Pr[C_t(n, N, \lceil N/2 \rceil)].$$

Proof: Let P_1 denote the probability that some coordinate of the outputs Y contains at least $\lceil N/2 \rceil$ errors when *at most* t errors (chosen uniformly and randomly from $B_t(\mathbf{0})$) occur in each channel and P_2 the probability that a coordinate of the outputs Y contains at least $\lceil N/2 \rceil$ errors when *exactly* t errors (chosen uniformly and randomly from $S_t(\mathbf{0})$) occur in each channel. As there occur in a channel exactly t errors in the case of P_2 and at most t errors in the case of P_1, it is immediate that $P_1 \leq P_2$. Based on the previous interpretation, each output word of Y with *exactly* t errors can be represented as a throw of t balls into n buckets, and there are obviously N throws in total. Therefore, $P_2 = Pr[C_t(n, N, \lceil N/2 \rceil)]$ and the claim follows. □

In order to obtain a lower bound on the probability $1 - Pr[C_t(n, N, \lceil N/2 \rceil)]$, we require an upper bound on $Pr[C_t(n, N, \lceil N/2 \rceil)]$. In the following lemma, we present such an upper bound loosely based on a recursive idea introduced in [29] for computing the exact probability of the multiple birthday problem.

Lemma 28: Let s, q, n and t be integers satisfying $2 \leq s \leq q$ and $0 \leq t \leq n$. (i) Now the probability $Pr[C_t(n, q, s)]$ is at most

$$\frac{t^s}{n^{s-1}} \sum_{i=s}^{q} \left(\binom{i-1}{s-1} \left(\frac{n-t}{n} \right)^{i-s} (1 - Pr[C_{t-1}(n-1, i-1, s)]) \right),$$

where $Pr[C_0(n, q, s)] = 0$ and $Pr[C_t(n, q, s)] = 0$ if $q < s$. (ii) Furthermore, we obtain that

$$Pr[C_t(n, q, s)] \leq \frac{t^s}{n^{s-1}} \binom{q}{s}.$$

Proof: (i) Observe first that we clearly have $Pr[C_0(n, q, s)] = 0$. Let then i be an integer such that $s \leq i \leq q$. Denote by $C_t(n, q, s, i)$ the event that the ith throw is the first throw of t balls such that there exists at least one bucket with s balls; notice that after the ith throw it is possible that s balls appear in multiple buckets. Using this notation, we have

$$Pr[C_t(n, q, s)] = \sum_{i=s}^{q} Pr[C_t(n, q, s, i)]. \tag{17}$$

The probability $Pr[C_t(n, q, s, i)]$ can be calculated based on the following facts:

(i) Let B be one of the buckets that first attains s balls. Clearly, the bucket B can be chosen in n ways.

(ii) The $s-1$ throws placing balls into B before the ith throw can be chosen from the previous $i-1$ throws in $\binom{i-1}{s-1}$ ways.

(iii) As the probability that a ball of a single throw lands into B is $\binom{n-1}{t-1}/\binom{n}{t}$, the probability of the event that the s selected throws put balls into B is equal to

$$\left(\frac{\binom{n-1}{t-1}}{\binom{n}{t}}\right)^s = \left(\frac{t}{n}\right)^s.$$

(iv) As the probability that no ball of a single throw lands into B is $\binom{n-1}{t}/\binom{n}{t}$, the probability of the event that no other throw (than the s selected ones) puts a ball into B is equal to

$$\left(\frac{\binom{n-1}{t}}{\binom{n}{t}}\right)^{i-s} = \left(\frac{n-t}{n}\right)^{i-s}.$$

(v) Finally, let B' denote the set of $n-1$ buckets other than B and P_i denote the probability that no bucket in B' contains at least s balls after the first $i-1$ throws with the conditional assumption that the events of (iii) and (iv) occur. Observe that if a ball of a throw lands into B, then the throw puts $t-1$ balls into the buckets of B', and otherwise t balls land into B'.

Thus, in conclusion, we have

$$Pr[C_t(n,q,s,i)] \leq n \cdot \binom{i-1}{s-1} \cdot \left(\frac{t}{n}\right)^s \cdot \left(\frac{n-t}{n}\right)^{i-s} \cdot P_i.$$

Therefore, by (17), we obtain that $Pr[C_t(n,q,s)]$ is at most

$$\frac{t^s}{n^{s-1}} \sum_{i=s}^{q} \left(P_i \cdot \binom{i-1}{s-1} \left(\frac{n-t}{n}\right)^{i-s} \right).$$

Finally, notice that P_i is equal to the probability that no bucket in B' contains at least s balls after $s-1$ throws with $t-1$ balls and $i-s$ throws with t balls have been performed in the buckets of B' (corresponding to the events of (iii) and (iv), respectively). Therefore, we obtain that $P_i \leq 1 - Pr[C_{t-1}(n-1, i-1, s)]$ since at least $t-1$ balls are thrown $i-1$ times into the $n-1$ buckets of B' (by the observation in (v)). Hence, the claim immediately follows.

(ii) For the second upper bound, we first notice that by the so called *hockey stick identity* for the binomial coefficients, we have

$$\sum_{i=s}^{q} \binom{i-1}{s-1} = \binom{q}{s}.$$

Therefore, as $(n-t)/n \leq 1$ and the probability $P_i \leq 1$, the second claim immediately follows. □

In the following theorem, the upper bound (ii) of the previous lemma is applied to estimate the probability in Theorem 27. Observe that according to the second claim of the theorem, the probability that the majority algorithm outputs the transmitted word is as close to one as required when the rather weak condition $n > 4t$ is satisfied and N is large enough.

Theorem 29: Let C be an e-error-correcting code and $\mathbf{x}$ the transmitted word of C. The probability that the output $\mathbf{z}$ of the majority algorithm is equal to the transmitted word $\mathbf{x}$ is at least

$$1 - \frac{t^{\lceil N/2 \rceil}}{n^{\lceil N/2 \rceil - 1}} \binom{N}{\lceil N/2 \rceil}.$$

TABLE I

THE LOWER BOUND OF THEOREM 27 TOGETHER WITH LEMMA 28 AND THEOREM 29 AS WELL AS THE MONTE CARLO APPROXIMATIONS WITH 100000 SAMPLES OF THE PROBABILITY THAT $\mathbf{z} = \mathbf{x}$ WHEN $n = 28$, $t = 5$ AND $N = 11, 21, 31, 41, 101$

N	Theorem 27 and Lemma 28	Theorem 29	Simulation
11	0.8199	0.5806	0.8615
21	0.9901	0.9419	0.9929
31	0.9994	0.9910	0.9997
41	0.9997	0.9985	1.000
101	1.000	1.000	1.000

Furthermore, if $n > 4t$, then

$$\lim_{N \to \infty} (1 - Pr[C_t(n, N, \lceil N/2 \rceil)]) = 1.$$

Proof: The first claim follows immediately by applying Theorem 27 and Lemma 28(ii). Since $\binom{N}{\lceil N/2 \rceil} < 2^N$, we obtain that

$$\begin{aligned} Pr[C_t(n, N, \lceil N/2 \rceil)] &\leq \frac{t^{\lceil N/2 \rceil}}{n^{\lceil N/2 \rceil - 1}} \binom{N}{\lceil N/2 \rceil} \\ &< \frac{t^{\lceil N/2 \rceil}}{n^{\lceil N/2 \rceil - 1}} \cdot 2^N \\ &\leq 4t \left(\frac{4t}{n}\right)^{\lceil N/2 \rceil - 1} \to 0 \end{aligned}$$

as $N \to \infty$ since $n > 4t$. Thus, the second claim follows. □

In Table I, we illustrate the various approaches to approximate the probability that the output $\mathbf{z}$ of the majority algorithm is equal to the transmitted word $\mathbf{x}$ when $t = 5$ and $n = 28$ (satisfying the condition of Theorem 29): the lower bound on the probabilities of Theorem 27 together with the recursive formula of Lemma 28(i) and Theorem 29 as well as the Monte Carlo simulations with 100000 samples. For the simulations, notice that the probabilities do not depend on the choice of $\mathbf{x}$ due to the fact that any word of $B_t(\mathbf{x})$ is outputted from a channel with equal probability. Hence, for simplicity, the simulations are performed using the word $\mathbf{x} = \mathbf{0}$. By Table I, we may observe that the majority algorithm outputs $\mathbf{z} = \mathbf{x}$ with high probability for significantly smaller number N of channels compared to Theorem 1, for which the required number of channels is 41709 when $e = 0$, $t = 5$ and $n = 28$. For explaining the choice of $e = 0$, recall that in Theorem 1 if a certain number of channels is given, then based on them a unique word in $B_e(\mathbf{x})$ can be determined and the transmitted word $\mathbf{x}$ can be obtained due to the e-error-correction capability of the underlying code C. Hence, we chose $e = 0$ in applying the theorem in order to ensure that $\mathbf{x}$ $(= \mathbf{z})$ is outputted from the algorithm (rather than just a word of $B_e(\mathbf{x})$).

Above, we saw that it is highly probable that the majority algorithm works correctly with significantly smaller number of channels compared to the one given in Theorem 1, which was required in the algorithm presented in [7]. In what follows, we take another approach on the majority algorithm and show that if a certain criteria (see Theorem 30) is met for the outputs Y, then we can verify that the output $\mathbf{z}$ of the majority algorithm belongs to $B_e(\mathbf{x})$. For this purpose, notice first that the total number of errors occurring in the ith coordinates of the outputs Y is at least $m_i = \min\{m_{i,0}, m_{i,1}\}$. On the other

TABLE II
THE MONTE CARLO APPROXIMATIONS WITH 100000 SAMPLES OF THE PROBABILITY FOR e SATISFYING THE CONDITION OF THEOREM 30 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

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

$$\sum_{i=1}^{n} m_i \leq tN. \tag{18}$$

Furthermore, if $\mathbf{x} = \mathbf{z}$, then the number of errors is exactly $\sum_{i=1}^{n} m_i$. In addition, if $\mathbf{x} \neq \mathbf{z}$, 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 (19)) has to satisfy Inequality (18).

Theorem 30: Let $\mathbf{x}$ be the transmitted word, m_i' be the integers m_i ordered in such a way that $m_1' \geq m_2' \geq \cdots \geq m_n'$ and $\mathbf{z}$ be the output word of the majority algorithm. We have $d(\mathbf{x}, \mathbf{z}) \leq k$ if k is a nonnegative integer such that

$$\sum_{i=1}^{k+1}(N - m_i') + \sum_{i=k+2}^{n} m_i' > tN. \tag{19}$$

Proof: Let k be a positive integer satisfying (19). Suppose to the contrary that $d(\mathbf{x}, \mathbf{z}) \geq k + 1$. This implies that for at least $k + 1$ coordinates i, the number of errors is $\max\{m_{i,0}, m_{i,1}\} = N - m_i$. Therefore, by the ordering of m_i', the number of errors is at least

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

Thus, due to (19), we have a contradiction with the maximum number of errors being tN and the claim follows. □

Observe that (19) allows us to estimate the accuracy of $\mathbf{z}$. In particular, if there exists a nonnegative integer k (for given $\mathbf{x}$ and $\mathbf{z}$) such that $k \leq e$, then $d(\mathbf{x}, \mathbf{z}) \leq k \leq e$ and the word $\mathbf{z}$ can be decoded to $\mathbf{x}$ as C is an e-error-correcting code. Furthermore, if $k > e$, then $\mathbf{x} \in C \cap B_k(\mathbf{z})$ 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 [28]). In conclusion, the theorem gives us a condition guaranteeing that the transmitted word can be decoded uniquely or with certain accuracy. Moreover, it should be noted that the theorem is rather efficient to use. Indeed, we only need to order the integers m_i, to perform approximately $n+k$ additions/subtractions and decode $\mathbf{z}$ to the closest codeword (or to a list of codewords if $k > e$).

The probability that a given k satisfies the property (19) of Theorem 30 can be analysed analytically as shown below, but first we approximate it using Monte Carlo simulations. In Table II, the probability is approximated using 100000 samples for $n = 24$, $t = 7$, $e = 2, 3, 4$ and varying numbers of channels N; here we choose $k = e$ and we strive for the exact transmitted word $\mathbf{x}$. From the table, we can notice that as the number of channels increases it becomes very likely that Inequality (19) is satisfied and that the majority algorithm together with the e-error-correction capability of C verifiably outputs the transmitted word $\mathbf{x}$.

In what follows, we further study analytically the probability that for a given integer k there exists a set Y of output words satisfying the conditions of Theorem 30. For this purpose, let $C_t'(n, q, s)$ denote the event that after q random throws of t balls, each of the n buckets contains at most s balls. Observe that $C_t'(n, q, s)$ is the complement of the event $C_t(n, q, s+1)$, and hence, for the probabilities, we have $Pr[C_t'(n, q, s)] = 1 - Pr[C_t(n, q, s+1)]$. Furthermore, regarding the number of errors occurring in the channels (in total), let $Er_t(r, q)$ denote the event that in total at least r balls are placed into buckets after q throws when each throw consists of at most t balls. Moreover, let $p(N)$ denote the parity of N, i.e., $p(N) = 1$ if N is odd, and otherwise $p(N) = 0$. Now we are ready to formulate the following theorem.

Theorem 31: Let α be a positive integer smaller than $\lceil N/2 \rceil$. The probability that a positive integer k satisfies (19) is at least

$$Pr[C_t'(n, N, \lceil N/2 \rceil - \alpha) \cap Er_t(tN - (k+1)(2\alpha - p(N)) + 1, N)].$$

Proof: Assume that (i) at most $\lceil N/2 \rceil - \alpha$ balls are placed into each of the n buckets and that (ii) in total at least $tN - (k+1)(2\alpha - p(N)) + 1$ balls are placed into buckets after N throws (when each throw consists of at most t balls). By the interpretation stated above Theorem 27, the assumptions (i) and (ii) can be reformulated as follows: (i) at most $\lceil N/2 \rceil - \alpha$ errors occur in each coordinate of the outputs Y and (ii) in total at least $tN - (k+1)(2\alpha - p(N)) + 1$ errors occur in channels (when at most t errors occur in each channel). Now the difference $(N - m_i) - m_i = N - 2m_i$ gives the number of additional errors occurring in a coordinate in the case that $\max\{m_{i,0}, m_{i,1}\} = N - m_i$ errors happen instead of $m_i = \min\{m_{i,0}, m_{i,1}\}$ errors. The difference $(N - m_i) - m_i$ can be estimated based on the parity of N as follows:

- If N is even, then $N - 2m_i \geq N - 2(\lceil N/2 \rceil - \alpha) = 2\alpha = 2\alpha - p(N)$.
- If N is odd, then $N - 2m_i \geq N - 2(\lceil N/2 \rceil - \alpha) = 2\alpha - 1 = 2\alpha - p(N)$.

In conclusion, we have $N - 2m_i \geq 2\alpha - p(N)$. Therefore, using the notation of Theorem 30, we have

$$\sum_{i=1}^{k+1}(N - m_i') + \sum_{i=k+2}^{n} m_i' \geq \sum_{i=1}^{n} m_i' + (k+1)(2\alpha - p(N)) > tN$$

since $\sum_{i=1}^{n} m_i' \geq tN - (k+1)(2\alpha - 2p(N)) + 1$ by the assumption (ii). Thus, the integer k meets the condition of Theorem 30 when the assumptions (i) and (ii) are satisfied.

In order to estimate the probability of the events of the assumptions (i) and (ii) occurring simultaneously, we denote

by A the event that at most $\lceil N/2 \rceil - \alpha$ errors occur in each coordinate of the outputs Y when at most t errors happen in each channel. Clearly, since the event $C'_t(n, N, \lceil N/2 \rceil - \alpha)$ can be interpreted as the outputs Y of the channels as explained earlier, we have $C'_t(n, N, \lceil N/2 \rceil - \alpha) \subseteq A$. Hence, the probability of the events of the assumptions (i) and (ii) occurring simultaneously is at least $Pr[A \cap Er_t(tN - (k+1)(2\alpha - p(N)) + 1, N)] \geq Pr[C'_t(n, N, \lceil N/2 \rceil - \alpha) \cap Er_t(tN - (k+1)(2\alpha - p(N)) + 1, N)]$ and the claim follows. □

Observe that if A and B are events (independent or dependent), then the probability of the event $A \cap B$ can be estimated as follows by the inclusion-exclusion principle:

$$Pr[A \cap B] = Pr[A] + Pr[B] - Pr[A \cup B] \geq Pr[A] + Pr[B] - 1.$$

By applying this lower bound to Theorem 31, the following corollary is immediately obtained.

Corollary 32: Let α be a positive integer. The probability that a positive integer k satisfies (19) is at least $Pr[C'_t(n, N, \lceil N/2 \rceil - \alpha)] + Pr[Er_t(tN - (k+1)(2\alpha - p(N)) + 1, N)] - 1$.

In the following, we discuss how to choose α in the corollary so that the lower bound is as close to 1 as desired. Consider first the estimation of the probability $Pr[Er_t(tN - (k+1)(2\alpha - p(N)) + 1, N)]$. Let X be a random variable representing the number of errors occurring in a channel. Hence, for $r = 0, 1, \ldots, t$, the probability

$$Pr[X = r] = p_r = \frac{\binom{n}{r}}{\sum_{i=0}^{t} \binom{n}{i}}.$$

The expected value of the distribution corresponding to X is

$$\mu = E(X) = \sum_{r=0}^{t} r p_r$$

and the variance is

$$\sigma^2 = Var(X) = E(X^2) - \mu^2 = \sum_{r=0}^{t} r^2 p_r - \mu^2.$$

Let

$$S_N = X_1 + X_2 + \cdots + X_N$$

be the sum of independent random variables X_i each equally distributed to X. Clearly, we have $Pr[Er_t(tN - (k+1)(2\alpha - p(N)) + 1, N)] = Pr[S_N \geq tN - (k+1)(2\alpha - p(N)) + 1]$. The expected value and the standard deviation of S_N are respectively equal to μN and $\sigma\sqrt{N}$. A usual approach to estimate distributions such as S_N is to use the concept of *confidence intervals*. There are numerous results designed to estimate the probabilities using confidence intervals. One of the most fundamental of such results is the well-known Chebyshev's inequality, according to which $Pr[|S_N - \mu N| \geq h \cdot \sigma\sqrt{N}] \leq 1/h^2$. Therefore, we have

$$\begin{aligned} &Pr[\mu N - h \cdot \sigma\sqrt{N} \leq S_N \leq \mu N + h \cdot \sigma\sqrt{N}] \\ =&1 - Pr[|S_N - \mu N| \geq h \cdot \sigma\sqrt{N}] \\ \geq&1 - \frac{1}{h^2}. \end{aligned}$$

Thus, the lower bound approaches 1 as h increases (as well as the probability on the left hand side); for example, the lower bound is equal to 0.9975 when $h = 20$. By straightforward calculations, we obtain that $tN - (k+1)(2\alpha - p(N)) + 1 \leq \mu N - h \cdot \sigma\sqrt{N}$ if and only if

$$\alpha \geq \frac{(t - \mu)N + h \cdot \sigma\sqrt{N} + 1 + p(N)(k+1)}{2(k+1)} (> 0).$$

For any α satisfying the lower bound, we have $Pr[S_N \geq tN - (k+1)(2\alpha - p(N)) + 1] \geq Pr[\mu N - h \cdot \sigma\sqrt{N} \leq S_N \leq \mu N + h \cdot \sigma\sqrt{N}] \geq 1 - 1/h^2$. The final lower bound $1 - 1/h^2$ based on Chebyshev's inequality is rather rough but enough for our purposes. Instead of the previous lower bound, a rather good *approximation* could be obtained using the Central Limit Theorem (for example, see [31, Section 5.4]), by which the sum S_N could be approximated by the normal distribution $\mathcal{N}(\mu N, \sigma\sqrt{N})$ (for large enough values of N). Denoting by Z the random variable distributed according to $\mathcal{N}(\mu N, \sigma\sqrt{N})$, we have $Pr[\mu N - h \cdot \sigma\sqrt{N} \leq S_N \leq \mu N + h \cdot \sigma\sqrt{N}] \approx Pr[\mu N - h \cdot \sigma\sqrt{N} \leq Z \leq \mu N + h \cdot \sigma\sqrt{N}] = 0.9999376\ldots$ by the properties of the normal distribution when $h = 4$. For $h = 4$, compare the *approximation* $0.9999376\ldots$ to the corresponding *lower bound* 0.9375 based on Chebyshev's inequality.

In order to estimate the lower bound of Corollary 32 for some fixed n, e and t, we first choose a suitable positive integer h such that the previous lower bound of $Pr[S_N \geq tN - (k+1)(2\alpha - p(N)) + 1]$ is as large as desired. Then, according to the following lemma, we have a lower bound on the probability $Pr[C'_t(n, tN, \lceil N/2 \rceil - \alpha)]$, which approaches 1 as the number N of channels tends to infinity (assuming a minor restriction on n). Hence, for carefully chosen h, α and N, the probability discussed in Corollary 32 gets as close to 1 as desired.

Lemma 33: Let h and k be (fixed) positive integers and $\alpha = \lceil ((t - \mu)N + h \cdot \sigma\sqrt{N} + 1 + p(N)(k+1))/(2(k+1)) \rceil$, where μ and σ are respectively the expected value and the standard deviation of the random variable X. Assume further that $n \geq 2t^2 + t - 1$ and $N \geq \max\{(k+2)^2, (h \cdot \sigma + 1)^2, 8\}$. Then we have the following results:

(i) Now the expected value $\mu > t - 1/2$. This further implies that $\alpha \leq \lceil N/2 \rceil$.

(ii) The probability $Pr[C'_t(n, N, \lceil N/2 \rceil - \alpha)]$ is at least

$$1 - \frac{t^{\lceil N/2 \rceil - \alpha + 1}}{n^{\lceil N/2 \rceil - \alpha}} \binom{N}{\lceil N/2 \rceil - \alpha + 1}.$$

(iii) Finally, if $n > 8t \cdot \varepsilon$, then

$$\lim_{N \to \infty} Pr[C'_t(n, N, \lceil N/2 \rceil - \alpha)] = 1.$$

Proof: (i) For the lower bound on the expected value μ, we first observe that for any $a \in [1, t]$ we have

$$\binom{n}{t-a} \leq \frac{t^a}{\prod_{i=1}^{a} n - t + i} \binom{n}{t} \leq \frac{t^a}{(n - t + 1)^a} \binom{n}{t}.$$

Thus, we obtain

$$\begin{aligned}\mu =&\sum_{r=0}^{t}\left(r\cdot\frac{\binom{n}{r}}{\sum_{i=0}^{t}\binom{n}{i}}\right)\\ >&\frac{t\binom{n}{t}}{\sum_{i=0}^{t}\binom{n}{i}}\\ \geq&\frac{t}{\sum_{i=0}^{t}\frac{t^i}{(n-t+1)^i}}\\ \geq&\frac{t}{\sum_{i=0}^{t}\frac{1}{(2t)^i}}\\ >&\frac{t}{\sum_{i=0}^{\infty}\frac{1}{(2t)^i}}=t-\frac{1}{2},\end{aligned}$$

where the assumption $n \geq 2t^2+t-1$ is used in the third inequality. In order to provide an upper bound on α, we first notice that due to the assumption $N \geq \max\{(k+2)^2,(h\cdot\sigma+1)^2,8\}$ and the previous lower bound on μ we have

$$\begin{aligned}&(t-\mu)N+h\cdot\sigma\sqrt{N}+1+p(N)(k+1)\\ <&\frac{1}{2}N+h\cdot\sigma\sqrt{N}+\sqrt{N}\\ =&\frac{1}{2}N+(h\cdot\sigma+1)\sqrt{N}\\ \leq&\frac{3}{2}N.\end{aligned}$$

Therefore, by straightforward calculations, we obtain

$$\alpha \leq \frac{3N/2}{2(k+1)}+1 \leq 3N/8+1 \leq \frac{N}{2} \leq \lceil N/2\rceil, \qquad (20)$$

where the third inequality follows from the fact that $N \geq 8$.

(ii) The lower bound on $Pr[C'_t(n,N,\lceil N/2\rceil-\alpha)] = 1-Pr[C_t(n,N,\lceil N/2\rceil-\alpha+1)]$ immediately follows by Lemma 28(ii) and hence, the second claim follows.

(iii) First assume that $n > 8t\cdot\varepsilon$. As $\alpha \leq 3N/8+1$ by Inequality (20), we obtain

$$1-\frac{2(\alpha-1)}{N} \geq 1-\frac{2((3N/8+1)-1)}{N} = 1-\frac{3}{4}=\frac{1}{4}.$$

In order to estimate the binomial coefficient appearing in the lower bound (ii), we recall the following well-known upper bound on the binomial coefficient: if m and r are integers such that $1 \leq r \leq m$, then

$$\binom{m}{r} < \left(\frac{m\cdot\varepsilon}{r}\right)^r.$$

These two inequalities together imply that

$$\begin{aligned}&Pr[C'_t(n,N,\lceil N/2\rceil-\alpha)]\\ >&1-\frac{t^{\lceil N/2\rceil-\alpha+1}}{n^{\lceil N/2\rceil-\alpha}}\left(\frac{N\cdot\varepsilon}{N/2-\alpha+1}\right)^{\lceil N/2\rceil-\alpha+1}\\ =&1-2t\varepsilon\left(\frac{2t\varepsilon}{n}\right)^{\lceil N/2\rceil-\alpha}\left(\frac{1}{1-\frac{2(\alpha-1)}{N}}\right)^{\lceil N/2\rceil-\alpha+1}\\ \geq&1-2t\varepsilon\left(\frac{2t\varepsilon}{n}\right)^{\lceil N/2\rceil-\alpha}4^{\lceil N/2\rceil-\alpha+1}\\ =&1-8t\varepsilon\left(\frac{8t\varepsilon}{n}\right)^{\lceil N/2\rceil-\alpha}.\end{aligned}$$

By Inequality (20), the probability

$$\begin{aligned}&Pr[C'_t(n,N,\lceil N/2\rceil-\alpha)]\\ >&1-8t\varepsilon\left(\frac{8t\varepsilon}{n}\right)^{\lceil N/2\rceil-\alpha}\\ >&1-8t\varepsilon\left(\frac{8t\varepsilon}{n}\right)^{N/2-(3N/8+1)}\\ >&1-8t\varepsilon\left(\frac{8t\varepsilon}{n}\right)^{N/8-1}\end{aligned}$$

and the lower bound approaches one as N tends to infinity since $n > 8t\cdot\varepsilon$ and $N/8-1 \to \infty$. Thus, the third claim follows. □

Observe that Lemma 33 is usually applied for $k=e$ as then the output $\mathbf{z}$ of the majority algorithm can be uniquely decoded to the transmitted word $\mathbf{x}$ according to Theorem 30. Notice also that the conditions $n \geq 2t^2+t-1$ and $N \geq \max\{(k+2)^2,(h\cdot\sigma+1)^2,8\}$ of the previous lemma are rather undemanding. In addition, the estimations of the proof are quite crude and smaller choices for n and N would be possible; in particular, this is true for n concerning the limit of the case (iii).

As explained before the previous lemma, we can get the lower bound of Corollary 32 as close to 1 as desired for carefully chosen h, α and N by combining Lemma 33 with the discussions on the probability $Pr[Er_t(tN-(k+1)(2\alpha-p(N))+1,N)] = Pr[S_N \geq tN-(k+1)(2\alpha-p(N))+1]$. For example, if $n=153$, $t=7$, $k=e=2$, $h=20$ and $N=31$, then we have $\mu \approx 6.951$, $\sigma \approx 0.226$ and $\alpha=6$ implying the conditions of Lemma 33 are satisfied. By Lemma 33(ii), we obtain the lower bound $Pr[C'_t(n,N,\lceil N/2\rceil-\alpha)] \geq 0.99997$. Furthermore, by the lower bound based on Chebyshev's inequality, we have $Pr[Er_t(tN-(k+1)(2\alpha-p(N))+1,N)] \geq 1-1/20^2 = 0.9975$. Combining these two lower bounds and Corollary 32, the probability for $k=2$ satisfying (19) can be bounded from below by $Pr[C'_t(n,N,\lceil N/2\rceil-\alpha)]+Pr[Er_t(tN-(k+1)(2\alpha-p(N))+1,N)]-1 \geq 0.99747$.

APPENDIX

Proof: [Proof of Corollary 16] Let us now consider the proof of Corollary 16 in the case where $d=2e+1$. The proof is quite similar to the case with $d=2e+2$. Although the first half of the proof in Corollary 16 works for $d=2e+1$, we start from the beginning since using set W'_w instead of W_w is better suited for our goal.

First we study the second binomial sum in the claim of Theorem 12 when $h=3$. Then we show that it gives equivalent result with Corollary 16 when $d=2e+1$. Since the binomial sum in Equation (8) is equivalent with the binomial sum in Equation (7), the claim follows.

Let us now consider the value we get for N_3 in the second binomial sum of Theorem 12 when $h=3$. We have

$$\begin{aligned}N_3 =&V(n,\ell-1)+\sum_{w\geq\ell}\sum_{(i_2,i_3,i_4)\in W'_w}\binom{n-3e-2}{w-i_2-i_3-i_4}\\ &\cdot\binom{e+1}{i_2}\binom{e+1}{i_3}\binom{e}{i_4}.\end{aligned}$$

In what follows, we have renamed the indices for convenience in such a way that i_4 corresponds to i_1 of Theorem 12 (the index i_1 will be saved for later use in the proof). Moreover, we have $W'_w = \{(i_2, i_3, i_4) \mid \text{for } j \in [2,3] : (w+1-\ell)/2 \le i_j \le e+1 \text{ and } (w-\ell)/2 \le i_4 \le e, w \ge i_2+i_3+i_4\}$. Again, if we have some binomial coefficients with $i_j < 0$ for some j, then we use the common convention that the binomial coefficient attains the value 0. Then, as in the case $d = 2e+2$, we obtain that $N_3 = \sum_{w\ge 0} \sum_{(i_2,i_3,i_4)\in W'_w} \binom{n-3e-2}{w-i_2-i_3-i_4}\binom{e+1}{i_2}\binom{e+1}{i_3}\binom{e}{i_4}$.

Let us denote by $i_1 = w - i_2 - i_3 - i_4$. We again get that $2i_2 \ge i_1+i_2+i_3+i_4-\ell+1$ and similar inequality for i_3. Moreover, for i_4, we have $2i_4 \ge i_1+i_2+i_3+i_4-\ell$. Since we do not have to take into account the lower bound $i_1 \ge 0$ (the cases with $i_1 < 0$ increase the binomial sum by 0) or the cases with $i_j > e+1$ for $j \in \{2,3,4\}$, we can consider following system of inequalities:

$$i_2 \ge i_1+i_3+i_4-\ell+1 \tag{21}$$
$$i_3 \ge i_1+i_2+i_4-\ell+1 \tag{22}$$
$$i_4 \ge i_1+i_2+i_3-\ell. \tag{23}$$

Our goal is to show that this system of inequalities is equivalent with the following system of inequalities:

$$i_4 \le \ell-1-i_1, \tag{24}$$
$$i_3 \le \ell-1-i_1, \tag{25}$$
$$i_1+i_3+i_4-(\ell-1) \le i_2 \le \ell-1/2-i_1-|i_4-i_3+1/2|. \tag{26}$$

Let us first show that the second system of inequalities follows from the first one.

Inequality (24) follows from

$$i_4 = w-i_1-i_2-i_3 \le w-i_1-2(w-\ell+1)/2 = \ell-1-i_1.$$

We obtain Inequality (25) in similar manner. Indeed,

$$\begin{aligned} i_3 &= w-i_1-i_2-i_4 \\ &\le w-i_1-(w-\ell+1)/2-(w-\ell)/2 \\ &= \ell-1/2-i_1 \end{aligned}$$

and the upper bound follows from the fact that i_3 is an integer. Moreover, from Inequalities (22) and (23) we obtain $i_2 \le \ell-1/2-i_1-i_4-1/2+i_3$ and $i_2 \le \ell-1/2-i_1-i_3+i_4+1/2$, respectively. Together, these imply

$$i_2 \le \ell-1/2-i_1-|i_4+1/2-i_3|.$$

Finally, the lower bound inequality in (26) follows directly from (21).

Let us then show that the first system of inequalities follows from the second one. First of all, Inequality (21) follows immediately from Inequality (26). Assume first that $i_4+1/2 > i_3$. Then the upper bound of Inequality (26) is $i_2 \le \ell-1-i_1-i_4+i_3$. This implies Inequality (22) and inequality (23) since

$$i_4 \ge i_3 \ge i_1+i_2+i_4-\ell+1 \ge i_1+i_2+i_3-\ell+1.$$

Notice that we cannot attain the lower bound in Inequality (23) in this case.

When $i_3 > i_4+1/2$, then the upper bound of Inequality (26) is $i_2 \le \ell-i_1-i_3+i_4$. In this case $i_4 \ge i_1+i_2+i_3-\ell$ and since $i_3 > i_4$, both lower bounds, (22) and (23), follow.

Finally, we may add lower bounds $i_j \ge 0$ for all $j \in \{1,2,3,4\}$ due to binomial coefficient context. Similarly we notice that if $i_1 \ge \ell$, then $i_3 < 0$. Thus, we may also add upper bound $i_1 \le \ell-1$. Now, we are ready to compare this bound with the bound of Theorem 2 by Yaakobi and Bruck.

When we have $d = 2e+1$, Theorem 2 can be presented in the following way: Let $N \ge \sum_{h_1,h_2,h_3,h_4} \binom{n-3e-2}{h_1}\binom{e+1}{h_2}\binom{e+1}{h_3}\binom{e}{h_4} + 1$ for

- $0 \le h_1 \le \ell-1$,
- $h_1-\ell \le h_4 \le \ell-1-h_1$,
- $e+2-\ell+h_1 \le h_3 \le t-(h_1+h_4)$ and
- $\max\{h_1-h_3-h_4+2e+2-\ell, h_1+h_3+h_4-\ell+1\} \le h_2 \le \ell-1-(h_1+h_4-h_3)$,

then $\mathcal{L} \le 2$ for any e-error-correcting code C. Next, we modify the presentation we got for N_3 into the formulation above.

Let us denote by $i'_2 = e+1-i_2$ and by $i'_3 = e+1-i_3$. Observe that $\binom{e+1}{i'_j} = \binom{e+1}{i_j}$ for $j \in \{2,3\}$. Notice that we have $i_4 \ge 0 \ge i_1-\ell$ and $\binom{e+1}{i_4} = 0$ when $i_4 < 0$. Hence, we may just replace this lower bound by $i_1-\ell$. Moreover, we have $0 \le i_3 \le \ell-1-i_1$. Hence, $e+1 \ge i'_3 \ge i_1+e+2-\ell$. Notice that $t-(i_1+i_4) \ge e+1$ since $i_1+i_4 \le \ell-1$ by Inequality (24), and $\binom{e+1}{i'_3} = 0$ when $i'_3 > e+1$.

For i_2 we have $i_1+i_3+i_4-(\ell-1) \le i_2 \le \ell-1/2-i_1-|i_4+1/2-i_3|$ and hence, $\ell-1-(i_1+i_4-i'_3) \ge i'_2 \ge -\ell+e+3/2+i_1+|i_4+1/2-i_3| = e+3/2-\ell+i_1+|i_4+i'_3-e-1/2| = \max\{i_1-i'_3-i_4+2e+2-\ell, i_1+i'_3+i_4-\ell+1\}$. Hence, we get the claim. □

REFERENCES

[1] V. Junnila, T. Laihonen, and T. Lehtilä, "On the list size in the Levenshtein's sequence reconstruction problem," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 510–515.

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

[3] V. Levenshtein, E. Konstantinova, E. Konstantinov, and S. Molodtsov, "Reconstruction of a graph from 2-vicinities of its vertices," *Discrete Appl. Math.*, vol. 156, no. 9, pp. 1399–1406, May 2008.

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

[5] M. Horovitz and E. Yaakobi, "Reconstruction of sequences over non-identical channels," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1267–1286, Feb. 2019.

[6] M. Abu-Sini and E. Yaakobi, "On list decoding of insertions and deletions under the reconstruction model," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 1706–1711.

[7] M. Abu-Sini and E. Yaakobi, "On Levenshtein's reconstruction problem under insertions, deletions, and substitutions," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 7132–7158, Nov. 2021.

[8] V. I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences," *J. Combinat. Theory A*, vol. 93, no. 2, pp. 310–332, Feb. 2001.

[9] J. Chrisnata, H. M. Kiah, and E. Yaakobi, "Correcting deletions with multiple reads," *IEEE Trans. Inf. Theory*, vol. 68, no. 11, pp. 7141–7158, Nov. 2022.

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

[11] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," *ACM SIGPLAN Notices*, vol. 51, no. 4, pp. 637–649, Jun. 2016.

[12] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, Sep. 2012.
[13] 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. Ed.*, vol. 54, no. 8, pp. 2552–2555, Feb. 2015.
[14] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Trans. Mol., Biol. Multi-Scale Commun.*, vol. 1, no. 3, pp. 230–248, Sep. 2015.
[15] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2155–2165, Apr. 2019.
[16] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 106–110.
[17] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6084–6103, Oct. 2020.
[18] V. Junnila and T. Laihonen, "Information retrieval with varying number of input clues," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 625–638, Feb. 2016.
[19] V. Junnila and T. Laihonen, "Codes for information retrieval with small uncertainty," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 976–985, Feb. 2014.
[20] T. Laihonen and T. Lehtilä, "Improved codes for list decoding in the Levenshtein's channel and information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2643–2647.
[21] T. Laihonen, "On t-revealing codes in binary Hamming spaces," *Inf. Comput.*, vol. 268, Oct. 2019, Art. no. 104455.
[22] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proc. 15th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2004, pp. 910–918.
[23] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *Proc. 19th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2008, pp. 399–408.
[24] V. Junnila, T. Laihonen, and T. Lehtilä, "On Levenshtein's channel and list size in information retrieval," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3322–3341, Jun. 2021.
[25] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes* (North-Holland Mathematical Library). Amsterdam, The Netherlands: North Holland, 1997, vol. 54.
[26] N. Sauer, "On the density of families of sets," *J. Combinat. Theory A*, vol. 13, no. 1, pp. 145–147, Jul. 1972.
[27] S. Shelah, "A combinatorial problem; stability and order for models and theories in infinitary languages," *Pacific J. Math.*, vol. 41, no. 1, pp. 247–261, Apr. 1972.
[28] V. Guruswami, "List decoding of error-correcting codes," Ph.D. thesis, Massachusetts Inst. Technol., Ann Arbor, MI, USA, 2001.
[29] K. Suzuki, D. Tonien, K. Kurosawa, and K. Toyota, "Birthday paradox for multi-collisions," in *Information Security and Cryptology—ICISC 2006*, M. S. Rhee and B. Lee, Eds. Berlin, Germany: Springer, 2006, pp. 29–40.
[30] M. Kounavis, S. Deutsch, D. Durham, and S. Komijani, "Non-recursive computation of the probability of more than two people having the same birthday," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 1263–1270.
[31] H. Tijms, *Understanding Probability*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

Ville Junnila was born in Turku, Finland, in 1981. He received the M.Sc. and Ph.D. degrees in mathematics from the University of Turku, Finland, in 2007 and 2011, respectively.

From 2007 to 2011, he was a Doctoral Student with the Department of Mathematics and Statistics, University of Turku. From 2011 to 2014, he was a Post-Doctoral Researcher on a grant. In 2014, he joined the Faculty of the Department of Mathematics and Statistics, University of Turku, where he was a Post-Doctoral Researcher and has been a University Lecturer since 2020. His research interests include combinatorial coding and graph theory as well as related areas of discrete mathematics. He has around 30 journal articles in these topics.

Tero Laihonen received the M.Sc. and Ph.D. degrees in mathematics from the University of Turku, Turku, Finland, in 1995 and 1998, respectively. From 1999 to 2002, he was a Post-Doctoral Researcher with the Academy of Finland. From 2003 to 2008, he was an Academy Research Fellow with the Academy of Finland. He joined the Faculty of the Department of Mathematics and Statistics, University of Turku, in 2008, where he is currently a Professor of discrete mathematics and theoretical computer science. His research interests include coding theory and graph theory with applications to DNA data storage, information retrieval, and sensor networks.

Tuomo Lehtilä received the M.Sc. and Ph.D. degrees in mathematics from the University of Turku, Turku, Finland, in 2016 and 2020, respectively. He was a Post-Doctoral Researcher with LIRIS, Université Claude Bernard Lyon 1, Lyon, France, from 2021 to 2022, and a Post-Doctoral Researcher with the University of Turku from 2022 to 2023. He is currently a Post-Doctoral Researcher with the Department of Computer Science, University of Helsinki, Helsinki, Finland. His current research interests include coding theory, graph theory, related areas of discrete mathematics, and security aspects of mobile networks and related areas.