On Levenshtein Balls with Radius One

Daniella Bar-Lev
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 3200003, Israel
Email: daniellalev@cs.technion.ac.il

Tuvi Etzion
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 3200003, Israel
Email: etzion@cs.technion.ac.il

Eitan Yaakobi
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 3200003, Israel
Email: yaakobi@cs.technion.ac.il

***Abstract*—The rapid development of DNA storage has brought the deletion and insertion channel, once again, to the front line of research. When the number of deletions is equal to the number of insertions, the *Fixed Length Levenshtein* (*FLL*) metric is the right measure for the distance between two words of the same length. The size of a ball is one of the most fundamental parameters in any metric. The size of the ball with radius one in the FLL metric depends on the number of runs and the length of the alternating segments of the given word. In this work, we find the minimum, maximum, and average size of a ball with radius one, in the FLL metric. The related minimum and maximum sizes of a maximal anticode with diameter one are also calculated.**

I. Introduction

Coding for DNA storage has attracted significant attention in the previous decade due to recent experiments and demonstrations of the viability of storing information in macromolecules [1], [4], [9], [16], [14]. Given the trends in cost decreases of DNA synthesis and sequencing, it is estimated that already within this decade DNA storage may become a highly competitive archiving technology. However, DNA molecules induce error patterns that are fundamentally different from their digital counterparts [6], [10]. This distinction results from the specific error behavior in DNA and it is well known that errors in DNA are typically in the form of substitutions, insertions, and deletions, where most published studies report that deletions are the most prominent ones, depending upon the specific technology for synthesis and sequencing. Hence, coding for insertion and deletion errors has received renewed interest recently due to its high relevance to the error model in DNA storage; see e.g. [2], [5], [13], [15]. This paper takes one more step in advancing this study and its goal is to study the size of the ball in the Fixed Length Levenshtein metric.

If a word $\boldsymbol{x} \in \mathbb{Z}_m^n$ can be transferred to a word $\boldsymbol{y} \in \mathbb{Z}_m^n$ using t deletions and t insertions (and cannot be transferred using a smaller number of deletions and insertions), then their ***Fixed Length Levenshtein (FLL) distance*** is t, which is denoted by $d_\ell(\boldsymbol{x}, \boldsymbol{y}) = t$. Let $G = (V, E)$ be a graph whose set of vertices $V = \mathbb{Z}_m^n$ and two vertices $\boldsymbol{x}, \boldsymbol{y} \in V$ are connected by an edge if $d_\ell(\boldsymbol{x}, \boldsymbol{y}) = 1$. The FLL distance defines a *graphic metric*, i.e., it is a metric and for each $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, $d_\ell(\boldsymbol{x}, \boldsymbol{y}) = t$ if and only if the length of the shortest path between $\boldsymbol{x}$ and $\boldsymbol{y}$ in G is t.

One of the most fundamental parameters in any metric is the size of a ball with a given radius t centered at a word $\boldsymbol{x}$. There are many metrics, e.g. the Hamming metric, the Johnson metric, or the Lee metric, where the size of a ball does not depend on the word $\boldsymbol{x}$. This is not the case in the FLL metric. Moreover, the graph G has a complex structure and it makes it much more difficult to find the exact size of any ball and in particular the size of a ball with minimum or maximum size. In [11], a formula for the size of the ball with radius one in the FLL metric was given. This formula depends on the number of runs in the word and the lengths of its alternating segments. Nevertheless, it is still difficult to compute from this formula what the size of the maximum ball is. In this paper we find explicit expressions for the minimum and maximum sizes of a ball when the ball is of radius one. We also find the average size of a ball when the radius of the ball is one. Finally, we consider the related concept of anticode in the FLL distance and find the maximum size and the minimum size of maximal anticodes with diameter one.

The rest of this paper is organized as follows. Section II introduces some basic concepts and observations required for our exposition. Section III presents basic results and equivalence of codes, for codes correcting deletions and insertions. The minimum size of a ball is discussed in Section IV, while in Section V the minimum size and maximum size of maximal anticodes with diameter one are computed. The maximum size and average size of balls with radius one are computed in Section VI.

II. Preliminaries

In this section we present the definitions and notations as well as several results that will be used throughout the paper.

Let $\mathbb{Z}_m$ denote the set of integers $\{0, 1, \ldots, m-1\}$ and for an integer $n \geq 0$, let $\mathbb{Z}_m^n$ be the set of all sequences (words) of length n over the alphabet $\mathbb{Z}_m$. For an integer t, $0 \leq t \leq n$, a sequence $\boldsymbol{y} \in \mathbb{Z}_m^{n-t}$ is a *t-subsequence* of $\boldsymbol{x} \in \mathbb{Z}_m^n$ if $\boldsymbol{y}$ can be obtained from $\boldsymbol{x}$ by deleting t symbols from $\boldsymbol{x}$. That is, there exist $n - t$ indices $1 \leq i_1 < i_2 < \cdots < i_{n-t} \leq n$ such that $y_j = x_{i_j}$, for all $1 \leq j \leq n - t$. We say that $\boldsymbol{y}$ is a *subsequence* of $\boldsymbol{x}$ if $\boldsymbol{y}$ is a t-subsequence of $\boldsymbol{x}$ for some t. Similarly, a sequence $\boldsymbol{y} \in \mathbb{Z}_m^{n+t}$ is a *t-supersequence* of $\boldsymbol{x} \in \mathbb{Z}_m^n$ if $\boldsymbol{x}$ is a t-subsequence of $\boldsymbol{y}$.

For a sequence $\boldsymbol{x} \in \mathbb{Z}_m^n$, let $\boldsymbol{x}_{[i,j]}$ be the subsequence $x_i x_{i+1} \cdots x_j$ and for a set of indices $I \subseteq \{1, \ldots, n\}$, the sequence $\boldsymbol{x}_{|_I}$ is the *projection* of $\boldsymbol{x}$ on the ordered indices of I, which is the subsequence of $\boldsymbol{x}$ received by the symbols

978-1-5386-8209-8/21/$31.00 ©2021 IEEE

in the entries of I. For a symbol $\sigma \in \mathbb{Z}_m$, σ^n denotes the sequence with n consecutive σ's.

The *Hamming weight* of a word $\boldsymbol{x} \in \mathbb{Z}_m^n$ is denoted by $\mathrm{wt}(\boldsymbol{x})$ and is equal to the number of nonzero coordinates in $\boldsymbol{x}$. The *Hamming distance* between two words $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, denoted by $d_H(\boldsymbol{x}, \boldsymbol{y})$, is the number of coordinates in which $\boldsymbol{x}$ and $\boldsymbol{y}$ differ.

Definition 1. *The* **Hamming t-ball** *centered at $\boldsymbol{x} \in \mathbb{Z}_m^n$, $B_t(\boldsymbol{x})$, is defined by*

$$B_t(\boldsymbol{x}) \triangleq \{\boldsymbol{y} \in \mathbb{Z}_m^n \ : \ d_H(\boldsymbol{x}, \boldsymbol{y}) \leq t\}.$$

For $\boldsymbol{x} \in \mathbb{Z}_m^n$, the number of words in each Hamming t-ball is a function of only n, m and t. The number of such words is

$$|B_t(\boldsymbol{x})| = \sum_{i=0}^{t} \binom{n}{i} (m-1)^i. \tag{1}$$

Definition 2. *The* **deletion t-sphere** *centered at $\boldsymbol{x} \in \mathbb{Z}_m^n$, $D_t(\boldsymbol{x}) \subseteq \mathbb{Z}_m^{n-t}$, is the set of all t-subsequences of $\boldsymbol{x}$. The* **insertion t-sphere** *centered at $\boldsymbol{x} \in \mathbb{Z}_m^n$, $I_t(\boldsymbol{x}) \subseteq \mathbb{Z}_m^{n+t}$, is the set of all t-supersequences of $\boldsymbol{x}$.*

The following lemma was proven in [8] and will be used in some of the proofs in the paper.

Lemma 1. *If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_2^n$ are distinct words, then*

$$|D_1(\boldsymbol{x}) \cap D_1(\boldsymbol{y})| \leq 2 \quad \textit{and} \quad |I_1(\boldsymbol{x}) \cap I_1(\boldsymbol{y})| \leq 2.$$

Note that for two sequences $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, the FLL distance between $\boldsymbol{x}$ and $\boldsymbol{y}$, $d_\ell(\boldsymbol{x}, \boldsymbol{y})$, is the smallest t for which there exists a t-subsequence $\boldsymbol{z} \in \mathbb{Z}_m^{n-t}$ of both $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e.,

$$d_\ell(\boldsymbol{x}, \boldsymbol{y}) = \min\{t' : D_{t'}(\boldsymbol{x}) \cap D_{t'}(\boldsymbol{y}) \neq \varnothing\}. \tag{2}$$

In other words, t is the smallest integer such that it is possible to receive $\boldsymbol{y}$ from $\boldsymbol{x}$ by t deletions and t insertions.

A *longest common subsequence* (LCS) of sequences $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{Z}_m^n$ is a subsequence $\boldsymbol{y}$ of $\boldsymbol{x}_i$ for each $1 \leq i \leq p$, where there is no other such subsequence $\boldsymbol{z}$ of a longer length. In other words $\boldsymbol{y} \in \bigcap_{i=1}^{p} D_t(x_i)$ for some t and t is the smallest integer such that $\bigcap_{i=1}^{p} D_t(x_i) \neq \varnothing$. The set of LCSs of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ is denoted by $\mathsf{LCS}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ and the length of any such LCS is denoted by $\ell\ell cs(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$. The following lemma is well known and can be easily verified.

Lemma 2. *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$, $D_t(\boldsymbol{x}) \cap D_t(\boldsymbol{y}) = \varnothing$ if and only if $\ell\ell cs(\boldsymbol{x}, \boldsymbol{y}) < n - t$.*

Combining (2) and Lemma 2 implies the following known result.

Corollary 1. *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_m^n$ we have that*

$$\ell\ell cs(\boldsymbol{x}, \boldsymbol{y}) = n - d_\ell(\boldsymbol{x}, \boldsymbol{y}).$$

For two sequences $\boldsymbol{x} \in \mathbb{Z}_m^{n_1}$ and $\boldsymbol{y} \in \mathbb{Z}_m^{n_2}$, $\ell\ell cs(\boldsymbol{x}, \boldsymbol{y})$ is given by the following recursive formula

$$\ell\ell cs(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 0 & n_1 = 0 \text{ or } n_2 = 0 \\ 1 + \ell\ell cs(\boldsymbol{x}_{[1:n_1-1]}, \boldsymbol{y}_{[1:n_2-1]}) & x_{n_1} = y_{n_2} \\ \max\left\{\ell\ell cs(\boldsymbol{x}_{[1:n_1-1]}, \boldsymbol{y}), \ell\ell cs(\boldsymbol{x}, \boldsymbol{y}_{[1:n_2-1]})\right\} & \text{otherwise} \end{cases} \tag{3}$$

Definition 3. *The (t_1, t_2)-***deletion-insertion sphere** *centered at $\boldsymbol{x} \in \mathbb{Z}_m^n$, $DI_{t_1,t_2}(\boldsymbol{x}) \subseteq \mathbb{Z}_m^{n-t_1+t_2}$, is the set of all the sequences that can be obtained from $\boldsymbol{x}$ by t_1 deletions and t_2 insertions.*

Definition 4. *The* **FLL t-ball** *centered at $\boldsymbol{x} \in \mathbb{Z}_m^n$, $L_t(\boldsymbol{x}) \subseteq \mathbb{Z}_m^n$, is defined by*

$$L_t(\boldsymbol{x}) \triangleq \{\boldsymbol{y} \in \mathbb{Z}_m^n \ : \ d_\ell(\boldsymbol{x}, \boldsymbol{y}) \leq t\}.$$

For a sequence $\boldsymbol{x} \in \mathbb{Z}_m^n$, a *run* of $\boldsymbol{x}$ is a maximal subsequence $\boldsymbol{x}_{[i,j]}$ of identical symbols. The number of runs in $\boldsymbol{x}$ will be denoted by $\rho(\boldsymbol{x})$. We say that a subsequence $\boldsymbol{x}_{[i,j]}$ is an *alternating segment* if $\boldsymbol{x}_{[i,j]}$ is a sequence of alternating distinct symbols $\sigma, \sigma' \in \mathbb{Z}_m$. Note that $\boldsymbol{x}_{[i,j]}$ is a maximal alternating segment if $\boldsymbol{x}_{[i,j]}$ is an alternating segment and $\boldsymbol{x}_{[i-1,j]}, \boldsymbol{x}_{[i,j+1]}$ are not. The number of maximal alternating segments of a sequence $\boldsymbol{x}$ will be denoted by $a(\boldsymbol{x})$. For example, in $\boldsymbol{x} = 00110100$, $\rho(\boldsymbol{x}) = 5$ and $a(\boldsymbol{x}) = 4$, where the four maximal alternating segments are $0, 01, 1010, 0$. Note that for binary sequences, $a(\boldsymbol{x}) + \rho(\boldsymbol{x}) = |\boldsymbol{x}| + 1$. The next lemma states the known result from [11] on the size of the FLL 1-ball.

Lemma 3. *[11]. For all $\boldsymbol{x} \in \mathbb{Z}_m^n$,*

$$|L_1(\boldsymbol{x})| = \rho(\boldsymbol{x}) \cdot (n(m-1) - 1) + 2 - \sum_{i=1}^{a(\boldsymbol{x})} \frac{(s_i - 1)(s_i - 2)}{2},$$

where s_i, $1 \leq i \leq a(\boldsymbol{x})$, denotes the length of the i-th maximal alternating segment of $\boldsymbol{x}$.

An *anticode of diameter t* in $\mathbb{Z}_m^n$ is a subset $\mathcal{A} \subseteq \mathbb{Z}_m^n$ such that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{A}$, $d_\ell(\boldsymbol{x}, \boldsymbol{x}') \leq t$. We say that $\mathcal{A}$ is a *maximal anticode* if there is no other anticode of diameter t in $\mathbb{Z}_m^n$ which contains $\mathcal{A}$.

III. Deletions/Insertions and the ℓ-Distance

A subset $\mathcal{C} \subseteq \Sigma_q^n$ is a *t-deletion-correcting code* (*t-insertion-correcting code*, respectively) if for any two distinct codewords $\boldsymbol{c}, \boldsymbol{c}' \in \mathcal{C}$ we have that $D_t(\boldsymbol{c}) \cap D_t(\boldsymbol{c}') = \varnothing$ ($I_t(\boldsymbol{c}) \cap I_t(\boldsymbol{c}') = \varnothing$, respectively). Similarly, $\mathcal{C}$ is called a *(t_1, t_2)-deletion-insertion-correcting code* if for any two distinct codewords $\boldsymbol{c}, \boldsymbol{c}' \in \mathcal{C}$ we have that $DI_{t_1,t_2}(\boldsymbol{c}) \cap DI_{t_1,t_2}(\boldsymbol{c}') = \varnothing$. Levenshtein [7] proved that $\mathcal{C}$ is a t-deletion-correcting code if and only if $\mathcal{C}$ is a t-insertion-correcting code and if and only if $\mathcal{C}$ is a (t_1, t_2)-deletion-insertion-correcting code for every t_1, t_2 such that $t_1 + t_2 \leq t$. A straightforward generalization is the following result [3].

Lemma 4. *For all $t_1, t_2 \in \mathbb{Z}$, if $\mathcal{C} \subseteq \mathbb{Z}_m^n$ is a (t_1, t_2)-deletion-insertion-correcting code, then $\mathcal{C}$ is also a $(t_1 + t_2)$-deletion-correcting code.*

Corollary 2. *For $\mathcal{C} \subseteq \mathbb{Z}_m^n$, the following statements are equivalent.*

1) *$\mathcal{C}$ is a (t_1, t_2)-deletion-insertion-correcting code.*
2) *$\mathcal{C}$ is a $(t_1 + t_2)$-deletion-correcting code.*
3) *$\mathcal{C}$ is a $(t_1 + t_2)$-insertion-correcting code.*
4) *$\mathcal{C}$ (t_1', t_2')-deletion-insertion-correcting code for any t_1', t_2' such that $t_1' + t_2' = t_1 + t_2$.*

Lemma 5. *A code $\mathcal{C} \in \mathbb{Z}_m^n$ is a $(2t+1)$-deletion-correcting code if and only if*

• $\mathcal{C}$ is a (t, t)-deletion-insertion-correcting code
and also
• if exactly $t+1$ FLL errors (i.e., $t+1$ insertions and $t+1$ deletions) occurred, then $\mathcal{C}$ can detect these $t+1$ FLL errors.

Proof: If $\mathcal{C}$ is a $(2t+1)$-deletion-correcting code, then by definition for any $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}$ we have that

$$D_{2t+1}(\boldsymbol{c}_1) \cap D_{2t+1}(\boldsymbol{c}_2) = \varnothing.$$

Therefore, by Lemma 2 for any two distinct codewords $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}$ we have that $\ell\ell cs(\boldsymbol{c}_1, \boldsymbol{c}_2) \leq n - (2t+1)$. Hence, by Corollary 1, $d_\ell(\boldsymbol{c}_1, \boldsymbol{c}_2) \geq 2(t+1)$. Since the FLL metric is graphic, it follows that $\mathcal{C}$ can correct up to t FLL errors and if exactly $t+1$ FLL errors occurred it can detect them.

For the other direction, assume that $\mathcal{C}$ is a (t, t)-deletion-insertion-correcting code and if exactly $t+1$ FLL errors occurred, then $\mathcal{C}$ can detect them. By Lemma 4, $\mathcal{C}$ is a $2t$-deletion-correcting code which implies that $D_{2t}(\boldsymbol{c}_1) \cap D_{2t}(\boldsymbol{c}_2) = \varnothing$ for all $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}$, and hence by (2) we have that

$$\forall \boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}: \quad d_\ell(\boldsymbol{c}_1, \boldsymbol{c}_2) > 2t.$$

Let us assume to the contrary that there are two codewords $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}$ such that $d_\ell(\boldsymbol{c}_1, \boldsymbol{c}_2) = 2t+1$. Since the FLL metric is graphic, it follows that there exists a word $\boldsymbol{y} \in \mathbb{Z}_m^n$ such that $d_\ell(\boldsymbol{c}_1, \boldsymbol{y}) = t$ and $d_\ell(\boldsymbol{y}, \boldsymbol{c}_2) = t+1$ which contradicts the fact that up to t FLL errors can be corrected and exactly $t+1$ FLL errors can be detected. Hence,

$$\forall \boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathcal{C}: \quad d_\ell(\boldsymbol{c}_1, \boldsymbol{c}_2) > 2t+1,$$

and by definition, $\mathcal{C}$ can correct $2t+1$ deletions. ∎

IV. Minimum Size of a Ball

The minimum size of the FLL t-ball is derived from the size of the Hamming ball. Changing the symbol in the i-th position from σ to σ' in any sequence can be done by first deleting σ and then inserting σ'. Hence we have the following lemma.

Lemma 6. *If $n \geq t \geq 0$ are integers and $\boldsymbol{x} \in \mathbb{Z}_m^n$, then $B_t(\boldsymbol{x}) \subseteq L_t(\boldsymbol{x})$ and hence $|B_t(\boldsymbol{x})| \leq |L_t(\boldsymbol{x})|$.*

Lemma 7. *If $n > t \geq 0$ are integers, then $B_t(\boldsymbol{x}) = L_t(\boldsymbol{x})$ if and only if $\boldsymbol{x} = \sigma^n$ for $\sigma \in \mathbb{Z}_m$.*

Proof: Assume w.l.o.g that $\boldsymbol{x} = 0^n$ and let $\boldsymbol{y} \in L_t(\boldsymbol{x})$ be a sequence obtained from $\boldsymbol{x}$ by at most t insertions and t deletions. Hence, $\mathrm{wt}(\boldsymbol{y}) \leq t$ and $\boldsymbol{y} \in B_t(\boldsymbol{x})$. Therefore, Lemma 6 implies that $B_t(\boldsymbol{x}) = L_t(\boldsymbol{x})$.

For the other direction, let $\boldsymbol{x} \in \mathbb{Z}_m^n$ were $\boldsymbol{x} \neq \sigma^n$ for all $\sigma \in \mathbb{Z}_m$. Since by Lemma 6, $B_t(\boldsymbol{x}) \subseteq L_t(\boldsymbol{x})$, to complete the proof, it is sufficient to show that there exists a sequence $\boldsymbol{y} \in L_t(\boldsymbol{x}) \backslash B_t(\boldsymbol{x})$. Denote $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and let i be the smallest index for which $x_i \neq x_{i+1}$. Let $\boldsymbol{y}$ be the sequence defined by

$$\boldsymbol{y} \triangleq (y_1, y_2, \ldots, y_{i-1}, x_{i+1}, x_i, y_{i+2}, \ldots, y_n),$$

where $y_j \neq x_j$ for the first $t-1$ indices (for which $j \notin \{i, i+1\}$) and $y_j = x_j$ otherwise. Clearly, $\boldsymbol{y}$ differs from $\boldsymbol{x}$ in $t+1$ indices and therefore $\boldsymbol{y} \notin B_t(\boldsymbol{x})$. On the other hand, $\boldsymbol{y}$ can be obtained from $\boldsymbol{x}$ by deleting x_i and inserting it to the right of x_{i+1} and then applying $t-1$ deletions and $t-1$ insertions whenever $y_j \neq x_j$ (for $j \notin \{i, i+1\}$). Thus, $\boldsymbol{y} \in L_t(\boldsymbol{x}) \backslash B_t(\boldsymbol{x})$ which completes the proof. ∎

Corollary 3. *If $n > t \geq 0$ and $m > 1$ are integers, then the size of the minimum FLL t-ball is*

$$\min_{\boldsymbol{x} \in \mathbb{Z}_m^n} |L_t(\boldsymbol{x})| = \sum_{i=0}^{t} \binom{n}{i} (m-1)^i,$$

and the minimum is obtained only by the balls centered at $\boldsymbol{x} = \sigma^n$ for any $\sigma \in \mathbb{Z}_m$.

V. Binary Anticodes with Diameter one

In this section we present tight lower and upper bounds on the size of maximal binary anticodes of diameter one in the FLL metric. To prove these bounds we need some useful properties of anticodes with diameter one in the FLL metric.

Lemma 8. *If an anticode $\mathcal{A}$ of diameter one contains three distinct words with the suffix 00 then there is at most one word in $\mathcal{A}$ with the suffix 01.*

Proof: Let $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{a}'' \in \mathcal{A}$ be three words with the suffix 00. Assume to the contrary that there exist two distinct words $\boldsymbol{b}, \boldsymbol{b}' \in \mathcal{A}$ with the suffix 01. Let $\boldsymbol{y} \in \mathsf{LCS}(\boldsymbol{a}, \boldsymbol{b})$. By Corollary 1 the length of $\boldsymbol{y}$ is $n-1$ and since $\boldsymbol{a}$ ends with 00, $\boldsymbol{y}$ must end with 0 which implies that $\boldsymbol{y} = \boldsymbol{b}_{[1,n-1]}$. By the same arguments $\boldsymbol{y} \in \mathsf{LCS}(\boldsymbol{b}, \boldsymbol{a}')$ and $\boldsymbol{y} \in \mathsf{LCS}(\boldsymbol{b}, \boldsymbol{a}'')$. Similarly,

$$\boldsymbol{y}' = \boldsymbol{b}'_{[1,n-1]} \in \mathsf{LCS}(\boldsymbol{b}', \boldsymbol{a}, \boldsymbol{a}', \boldsymbol{a}'').$$

Hence, $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{a}'' \in I_1(\boldsymbol{y}) \cap I_1(\boldsymbol{y}')$ which is a contradiction to Lemma 1 since $\boldsymbol{y} \neq \boldsymbol{y}'$. Thus, $\mathcal{A}$ contains at most one word with the suffix 01. ∎

Lemma 9. *If an anticode $\mathcal{A}$ of diameter one contains three distinct words with the suffix 01, then there is at most one word in $\mathcal{A}$ with the suffix 00.*

Proof: Let $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{a}'' \in \mathcal{A}$ be three words with the suffix 01. Assume to the contrary that there exist two distinct words $\boldsymbol{b}, \boldsymbol{b}' \in \mathcal{A}$ with the suffix 00. For $\boldsymbol{y} \in \mathsf{LCS}(\boldsymbol{a}, \boldsymbol{b})$, by Corollary 1 the length of $\boldsymbol{y}$ is $n-1$ and since $\boldsymbol{b}$ ends with 00, $\boldsymbol{y}$ must end with 0 which implies that $\boldsymbol{y} = \boldsymbol{a}_{[1,n-1]}$. By the same arguments $\boldsymbol{y} \in \mathsf{LCS}(\boldsymbol{a}, \boldsymbol{b}')$. Similarly,

$$\boldsymbol{y}' = \boldsymbol{a}'_{[1,n-1]} \in \mathsf{LCS}(\boldsymbol{a}', \boldsymbol{b}, \boldsymbol{b}')$$
$$\boldsymbol{y}'' = \boldsymbol{a}''_{[1,n-1]} \in \mathsf{LCS}(\boldsymbol{a}'', \boldsymbol{b}, \boldsymbol{b}').$$

Hence, $\boldsymbol{y}, \boldsymbol{y}', \boldsymbol{y}'' \in D_1(\boldsymbol{b}) \cap D_1(\boldsymbol{b}')$ which is a contradiction to Lemma 1. Thus, $\mathcal{A}$ contains at most one word with the suffix 00. ■

Lemma 10. *Let $\mathcal{A}$ be an anticode of diameter one. If $\boldsymbol{a}, \boldsymbol{a}' \in \mathcal{A}$ are two distinct words that end with 00 and $\boldsymbol{b}, \boldsymbol{b}' \in \mathcal{A}$ are two distinct words that end with 01, then $\boldsymbol{a}_{[1,n-1]} \neq \boldsymbol{b}_{[1,n-1]}$ or $\boldsymbol{a}'_{[1,n-1]} \neq \boldsymbol{b}'_{[1,n-1]}$.*

Proof: Assume to the contrary that there exist $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{b}, \boldsymbol{b}' \in \mathcal{A}$ such that $\boldsymbol{a}_{[1,n-1]} = \boldsymbol{b}_{[1,n-1]} = \boldsymbol{y}0$ and $\boldsymbol{a}'_{[1,n-1]} = \boldsymbol{b}'_{[1,n-1]} = \boldsymbol{y}'0$, $\boldsymbol{a}, \boldsymbol{a}'$ end with 00, and $\boldsymbol{b}, \boldsymbol{b}'$ end with 01. Let,

$$\begin{aligned}
\boldsymbol{a} &= a_1\ a_2 \cdots a_{n-2}\ 0\ 0 = \boldsymbol{y}\ \ 0\ 0 \\
\boldsymbol{a}' &= a'_1\ a'_2 \cdots a'_{n-2}\ 0\ 0 = \boldsymbol{y}'\ 0\ 0 \\
\boldsymbol{b} &= a_1\ a_2 \cdots a_{n-2}\ 0\ 1 = \boldsymbol{y}\ \ 0\ 1 \\
\boldsymbol{b}' &= a'_1\ a'_2 \cdots a'_{n-2}\ 0\ 1 = \boldsymbol{y}'\ 0\ 1.
\end{aligned}$$

Notice that since the FLL distance between any two words in $\mathcal{A}$ is one, the Hamming weight of any two words can differ by at most one, which implies that $\mathrm{wt}(\boldsymbol{y}) = \mathrm{wt}(\boldsymbol{y}')$ (by considering the pairs $\boldsymbol{a}, \boldsymbol{b}'$ and $\boldsymbol{a}', \boldsymbol{b}$). Clearly, $\boldsymbol{y}0 \in \mathsf{LCS}(\boldsymbol{a}', \boldsymbol{b})$ which implies that $\boldsymbol{a}'$ can be obtained from $\boldsymbol{b}$ by deleting the last 1 of $\boldsymbol{b}$ and then inserting 0 into the LCS. Hence, there exists an index $0 \leq j \leq n-2$ such that

$$a_1 a_2 \cdots a_j 0 a_{j+1} \cdots a_{n-2} 0 = a'_1 a'_2 \cdots a'_j a'_{j+1} \cdots a'_{n-2} 00 \quad (4)$$

Similarly, $\boldsymbol{a}$ can be obtained from $\boldsymbol{b}'$, i.e., there exists an index $0 \leq i \leq n-2$ such that

$$a'_1 a'_2 \ldots a'_i 0 a'_{i+1} \cdots a'_{n-2} 0 = a_1 a_2 \ldots a_i a_{i+1} \cdots a_{n-2} 00. \quad (5)$$

Assume w.l.o.g. that $i \leq j$. (4) implies that $a_r = a_{r'}$ for $1 \leq r \leq j$. In addition, $a_{n-2} = 0$ by (4) and $a'_{n-2} = 0$ by (5). By assigning $a_{n-2} = a'_{n-2} = 0$ into (4) and (5) we obtain that $a_{n-3} = a'_{n-3} = 0$. Repeating this process implies that $a_r = a_{r'} = 0$ for $j+1 \leq r \leq n-2$. Thus, it must be that $\boldsymbol{y} = \boldsymbol{y}'$ which is a contradiction. ■

Definition 5. *For an anticode $\mathcal{A} \subseteq \mathbb{Z}_2^n$, the* **puncturing of $\mathcal{A}$ in the n-th coordinate***, $\mathcal{A}'$, is defined by*

$$\mathcal{A}' \triangleq \left\{ \boldsymbol{a}_{[1:n-1]} \ : \ \boldsymbol{a} \in \mathcal{A} \right\}.$$

The following lemma is a conclusion of (3) and Corollary 1.

Lemma 11. *Let $\mathcal{A} \subseteq \mathbb{Z}_2^n$ be an anticode of diameter one. If the last symbol in all the words in $\mathcal{A}$ is the same symbol $\sigma \in \mathbb{Z}_2^n$, then $\mathcal{A}'$ is an anticode of diameter one and $|\mathcal{A}'| = |\mathcal{A}|$.*

Lemma 12. *Let $\mathcal{A}$ be an anticode of diameter one. If the suffix of each word in $\mathcal{A}$ is either 01 or 10, then $\mathcal{A}'$ is an anticode of diameter one and $|\mathcal{A}'| = |\mathcal{A}|$.*

Proof: Let $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}$ be two different words and let $\boldsymbol{y} \in \mathsf{LCS}(\boldsymbol{a}_{[1:n-1]}, \boldsymbol{b}_{[1:n-1]})$. By (3), $\ell\ell cs(\boldsymbol{a}, \boldsymbol{b}) \leq |\boldsymbol{y}| + 1$ and since $d_\ell(\boldsymbol{a}, \boldsymbol{b}) = 1$ it follows that $|\boldsymbol{y}| \geq n-2$ and that $d_\ell(\boldsymbol{a}_{[1:n-1]}, \boldsymbol{b}_{[1:n-1]}) \leq 1$. Thus, $\mathcal{A}'$ is an anticode of diameter one. It is readily verified that $|\mathcal{A}'| = |\mathcal{A}|$. ■

Theorem 1. *Let $n > 1$ be an integer and let $\mathcal{A} \subseteq \mathbb{Z}_2^n$ be a maximal anticode of diameter one. Then, $|\mathcal{A}| \leq n+1$, and there exists a maximal anticode with exactly $n+1$ codewords.*

Proof: Since two words $\boldsymbol{x}, \boldsymbol{y}$ such that $\boldsymbol{x}$ ends with 00 and $\boldsymbol{y}$ ends with 11 are at FLL distance at least 2, w.l.o.g. assume that $\mathcal{A}$ does not contain codewords that end with 11. It is easy to verify that the theorem holds for $n \in \{2, 3, 4\}$. Assume that the theorem does not hold and let $n^* > 4$ be the smallest integer such that there exists an anticode $\mathcal{A} \subseteq \mathbb{Z}_2^{n^*}$ such that $|\mathcal{A}| = n^* + 2$. Since there are only three possible options for the last two symbols of codewords in $\mathcal{A}$, there exist three different codewords in $\mathcal{A}$ with the same suffix of two symbols.

Case 1 - Assume $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{A}$ are three different words with the suffix 00. By Lemma 8, there exists at most one codeword in $\mathcal{A}$ with the suffix 01 and since $\mathcal{A}$ does not contain codewords with the suffix 11, there exists at most one codeword in $\mathcal{A}$ that ends with the symbol 1. That is, there are n^*+1 codewords with 0 as the last symbol. Denote this set of n^*+1 codewords by $\mathcal{A}_1$. As a subset of the anticode $\mathcal{A}$, $\mathcal{A}_1$ is also an anticode and hence by Lemma 11, $\mathcal{A}'_1$ is an anticode of length n^*-1 and size n^*+1 which is a contradiction to the minimality of n^*.

Case 2 - Assume $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{A}$ are three different words with the suffix 01. By Lemma 9, there exists at most one codeword in $\mathcal{A}$ with the suffix 00 and since $\mathcal{A}$ does not contain codewords with the suffix 11 there exist n^*+1 codewords that end with either 01 or 10. Denote this set of n^*+1 codewords as $\mathcal{A}_1$. As a subset of the anticode $\mathcal{A}$, $\mathcal{A}_1$ is also an anticode and hence by Lemma 12, $\mathcal{A}'_1$ is an anticode of length n^*-1 and size n^*+1 which is a contradiction to the minimality of n^*.

Case 3 - Assume $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{A}$ are three different words with the suffix 10. By the previous two cases, there exist at most two codewords in $\mathcal{A}$ with the suffix 00 and at most two codewords with the suffix 01. Since there are no codewords with the suffix 11, it follows that the number of words that end with 1 is at most two. If there exist at most one codeword in $\mathcal{A}$ that ends with 1, then there are n^*+1 codewords in $\mathcal{A}$ that end with 0 and as in the first case, this leads to a contradiction. Otherwise there are exactly two codewords in $\mathcal{A}$ with the suffix 01. If there are less than two codewords with the suffix 00, then, the number of codewords with suffixes 01 and 10 is at least n^*+1 and similarly to the second case, this is a contradiction to the minimality of n^*. Hence, there exist exactly two codewords in $\mathcal{A}$ with the suffix 00. There are exactly n^*-2 codewords in $\mathcal{A}$ with the suffix 10 and two more codewords with the suffix 01. By Lemma 12 the words in $\mathcal{A}'$ that were obtained from these n^* codewords are all different and have FLL distance one from each other. In addition, by Lemma 10, the prefix of length n^*-1 of at least one of the codewords that end with 00 is different from the prefixes of length n^*-1 of the codewords that end with 01. This prefix also differs from the prefixes of the codewords that end with 10. Therefore, $\mathcal{A}'$ is an anticode with n^*+1 different codewords which is a contradiction to

the minimality of n^*.

Note that the set $\mathcal{A} = \{\boldsymbol{a} \in \Sigma_2^n : \mathrm{wt}(\boldsymbol{a}) \leq 1\}$ is an anticode of diameter one with exactly $n+1$ codewords. Thus, the maximum size of an anticode of diameter one is $n+1$. ∎

Theorem 2. *Let $n > 2$ be a positive integer and let $\mathcal{A} \subseteq \mathbb{Z}_2^n$ be a maximal anticode of diameter one. Then, $|\mathcal{A}| \geq 4$ and there exists a maximal anticode with exactly 4 codewords.*

VI. Balls with Radius one

A. Balls with Maximum Size

The following theorem was proved in [12].

Theorem 3. *If n and $m > 2$ are positive integers, then the maximal FLL 1-balls are the balls centered at $\boldsymbol{x} \in \mathbb{Z}_m^n$, such that the number of runs in $\boldsymbol{x}$ is n and $x_i \neq x_{i+2}$ for all $1 \leq i \leq n-2$. In addition, the size of such a maximum 1-ball is,*

$$\max_{\boldsymbol{x} \in \mathbb{Z}_m^n} |L_1(\boldsymbol{x})| = n^2(m-1) - n + 2.$$

For the binary case, there exists a ball whose size is larger than the ball considered in [12]. The analysis to find such a ball is slightly more difficult, since by definition of a run, there is no sequence $\boldsymbol{x}$ with n runs such that $x_i \neq x_{i+2}$ for some i. The following lemmas lead to the main theorem of this subsection.

Definition 6. *For $t \in \mathbb{N}$, $\boldsymbol{x} \in \mathbb{Z}_2^n$ is an α-balanced sequence if $a(\boldsymbol{x}) = \alpha$ and $s_i \in \{\lceil \frac{n}{\alpha} \rceil, \lceil \frac{n}{\alpha} \rceil - 1\}$ for all $i \in \{1, \ldots, \alpha\}$, where s_i was defined in Lemma 3.*

Lemma 13. *If n and $1 \leq \alpha \leq n$ are positive integers then*

$$\arg\max_{\substack{\boldsymbol{x} \in \mathbb{Z}_2^n \\ a(\boldsymbol{x}) = \alpha}} |L_1(\boldsymbol{x})| = \{\boldsymbol{x} \in \mathbb{Z}_2^n : \boldsymbol{x} \text{ is an } \alpha\text{-balanced sequence}\}.$$

Lemma 14. *If $\boldsymbol{x}^{(\alpha)}$ is an α-balanced sequence of length n then*

$$|L_1(\boldsymbol{x}^{(\alpha)})| = (n+1-\alpha)(n-1) + 2 - \frac{k}{2}\left(\left\lceil \frac{n}{\alpha} \right\rceil - 1\right)\left(\left\lceil \frac{n}{\alpha} \right\rceil - 2\right) - \frac{\alpha - k}{2}\left(\left\lceil \frac{n}{\alpha} \right\rceil - 2\right)\left(\left\lceil \frac{n}{\alpha} \right\rceil - 3\right),$$

where $k \equiv n \pmod{\alpha}$ and $1 \leq k \leq \alpha$.

By Lemma 13

$$\max_{x \in \mathbb{Z}_2^n} |L_1(\boldsymbol{x})| = \max_{1 \leq \alpha \leq n} \left\{ \max_{\substack{\boldsymbol{x} \in \mathbb{Z}_2^n \\ \alpha(\boldsymbol{x}) = \alpha}} |L_1(\boldsymbol{x})| \right\} = \max_{1 \leq \alpha \leq n} \left| L_1(\boldsymbol{x}^{(\alpha)}) \right|,$$

and the size $|L_1(\boldsymbol{x}^{(\alpha)})|$ for $1 \leq \alpha \leq n$ is given by Lemma 14. Hence, our goal is to find the set

$$\mathsf{T}(n) \triangleq \arg\max_{1 \leq \alpha \leq n} \left| L_1(\boldsymbol{x}^{(\alpha)}) \right|.$$

Lemma 15. *If $\boldsymbol{x}^{(\alpha)}$ is an α-balanced sequence of length n, then*

$$|L_1(\boldsymbol{x}^{(\alpha)})| > |L_1(\boldsymbol{x}^{(\alpha-1)})|$$

if and only if $n > 2(t-1)t$.

Theorem 4. *Let n be an integer. It holds that*

$$\mathsf{T}(n) = \arg\min_{\alpha \in \mathbb{N}} \left| \alpha - \frac{1}{2}\sqrt{1+2n} \right|.$$

Consequently, the maximum FLL 1-balls are the balls centered at the α-balanced sequences of length n, for any $\alpha \in \mathsf{T}(n)$ and

$$\max_{\boldsymbol{x} \in \mathbb{Z}_2^n} |L_1(\boldsymbol{x})| = n^2 - n(\alpha+1) + \alpha + 2 - \frac{k}{2}\left(\left\lceil \frac{n}{\alpha} \right\rceil - 1\right)\left(\left\lceil \frac{n}{\alpha} \right\rceil - 2\right) - \frac{\alpha - k}{2}\left(\left\lceil \frac{n}{\alpha} \right\rceil - 2\right)\left(\left\lceil \frac{n}{\alpha} \right\rceil - 3\right),$$

where $k \equiv n \pmod{\alpha}$ and $1 \leq k \leq \alpha$.

Note: $\mathsf{T}(n)$ in Theorem 4 can be either of size one or of size two.

Corollary 4. *If n is a sufficiently large integer, then*

$$\max_{x \in \Sigma_2^n} \{|L_1(x)|\} = n^2 - \sqrt{2}n^{\frac{3}{2}} + O(n).$$

B. The Average Size of a Ball

By Lemma 3, for any $\boldsymbol{x} \in \mathbb{Z}_m^n$

$$|L_1(\boldsymbol{x})| = \rho(\boldsymbol{x})(nm - n - 1) + 2 - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{3}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i - a(\boldsymbol{x}).$$

Thus, the average size of the FLL 1-ball is

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathbb{Z}_m^n}\left[\rho(\boldsymbol{x})(n(m-1)-1) + 2 - \frac{1}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i^2 + \frac{3}{2}\sum_{i=1}^{a(\boldsymbol{x})} s_i - a(\boldsymbol{x})\right].$$

The following lemma leads to our main theorem of this subsection.

Lemma 16. *For any integers $n, m > 1$, we have that*

1) $\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathbb{Z}_m^n}\left[\sum_{i=1}^{a(\boldsymbol{x})} s_i\right] = n + (n-2) \cdot \frac{(m-1)(m-2)}{m^2}$,

2) $\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathbb{Z}_m^n}[a(\boldsymbol{x})] = 1 + \frac{(n-2)(m-1)(m-2)}{m^2} + \frac{n-1}{m}$,

3) $\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathbb{Z}_m^n}[\rho(\boldsymbol{x})] = n - \frac{n-1}{m}$,

4) $\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathbb{Z}_m^n}\left[\sum_{i=1}^{a(\boldsymbol{x})} s_i^2\right] = \frac{n(4m^2 - 3m + 2)}{m^2} + \frac{6m-4}{m^2} - 4 - \frac{2}{m-1}\left(1 - \frac{1}{m^n}\right)$.

Theorem 5. *For any integers $n, m > 1$, we have that*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathbb{Z}_m^n}[|L_1(\boldsymbol{x})|] = n^2\left(m + \frac{1}{m} - 2\right) - \frac{n}{m} - \frac{(m-1)(m-2)}{m^2} + 3 - \frac{3}{m} + \frac{2}{m^2} + \frac{m^n - 1}{m^n(m-1)}.$$

Acknowledgment

D. Bar-Lev and T. Etzion were supported by ISF grant no. 222/19. E. Yaakobi was supported by the United States-Israel BSF grant no. 2018048.

REFERENCES

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

[2] R. Con and A. Shpilka, "Explicit and efficient constructions of coding schemes for the binary deletion channel and the poisson repeat channel," *IEEE International Symposium on Information Theory (ISIT)*, pp. 84-89, 2020.

[3] D. Cullina and N. Kiyavash, "An improvement to Levenshtein's upper bound on the cardinality of deletion correcting codes," *2013 IEEE International Symposium on Information Theory*, 2013, pp. 699–703,

[4] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[5] V. Guruswami and C. Wang, "Deletion codes in the high-noise and high-rate regimes," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 1961–1970, 2017.

[6] R. Heckel, G. Mikutis, and R.N. Grass, "A characterization of the DNA data storage channel," *Scientific Reports*, vol. 9, sp. 9663, 2019.

[7] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[8] V.I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences," *Journal of Combinatorial Theory, Series A*, 93(2): 310–332, 2001.

[9] L. Organick, S.D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M.Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, and K. Strauss, "Scaling up DNA data storage and random access retrieval," *bioRxiv*, Mar. 2017.

[10] O. Sabary, Y. Orlev, R. Shafir L. Anavy, E. Yaakobi, and Z.Yakhini, "SOLQC: Synthetic oligo library quality control Tool," *Bioinformatics*, btaa740, https://doi.org/10.1093/bioinformatics/btaa740, Mar. 2021

[11] F. Sala and L. Dolecek, "Counting sequences obtained from the synchronization channel," *Proc. Int. Symp. Inf. Theory*, Istanbul, Turkey, pp. 2925–2929, Jul. 2013.

[12] F. Sala, R. Gabrys, and L. Dolecek, "Gilbert-Varshamov-like lower bounds for deletion-correcting codes," *IEEE Information Theory Workshop*, Hobart, TAS, pp. 147–151, 2014.

[13] J. Sima and J. Bruck, "Optimal k-deletion correcting codes," *Proc. IEEE Int. Symp. Inf. Theory*, Paris, France, pp. 847–851, Jul. 2019.

[14] S.K. Tabatabaei, B. Wang, N.B.M. Athreya, B. Enghiad, A.G. Hernandez, J.-P. Leburton, D. Soloveichik, H. Zhao, and O. Milenkovic, "DNA punch cards: Encoding data on native DNA sequences via topological modifications," *BioRxiv*, June 2019.

[15] I. Tal, H. D. Pfister, A. Fazeli, and A. Vardy, "Polar codes for the deletion channel: Weak and strong polarization," *Proc. IEEE Int. Symp. Inf. Theory*, Paris, France, pp. 1362–1366, Jul. 2019.

[16] S. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific Reports*, vol. 7, sp. 5011, Jul. 2017.