Coding for Sequence Reconstruction for Single Edits

Han Mao Kiah*, Tuan Thanh Nguyen†, and Eitan Yaakobi‡

*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371
†Singapore University of Technology and Design, Singapore 487372
‡Department of Computer Science, Technion — Israel Institute of Technology, Haifa, 32000 Israel
Emails: hmkiah@ntu.edu.sg, tuanthanh_nguyen@sutd.edu.sg, yaakobi@cs.technion.ac.il

***Abstract*—The sequence reconstruction problem, introduced by Levenshtein in 2001, considers a communication scenario where the sender transmits a codeword from some codebook and the receiver obtains multiple noisy reads of the codeword. The common setup assumes the codebook to be the entire space and the problem is to determine the minimum number of distinct reads that is required to reconstruct the transmitted codeword.**

Motivated by modern storage devices, we study a variant of the problem where the number of noisy reads N is fixed. Specifically, we design *reconstruction codes* that reconstruct a codeword from N distinct noisy reads. We focus on channels that introduce single edit error (i.e. a single substitution, insertion, or deletion) and their variants, and design reconstruction codes for all values of N. In particular, for the case of a single edit, we show that as the number of noisy reads increases, the number of redundant bits required can be gracefully reduced from $\log n + O(1)$ to $\log\log n + O(1)$, and then to $O(1)$, where n denotes the length of a codeword. We also show that the redundancy of certain reconstruction codes is within one bit of optimality.

I. Introduction

As our data needs surge, new technologies emerge to store these huge datasets. Interestingly, besides promising ultra-high storage density, certain emerging storage media rely on technologies that provide users with *multiple cheap, albeit noisy, reads*. In this paper, we leverage on these multiple reads to increase the information capacity of these next-generation devices, or equivalently, *reduce the number of redundant bits*.

Before we formally state our problem, we list two storage scenarios where multiple cheap reads are available to the user.

(a) **DNA-based data storage**. In these systems [1]–[3], digital information is stored in native or synthetic DNA strands and to read the information, a user typically employs a sequencing platform like the popular Illumina sequencer or more recently, a nanopore sequencer. In most sequencers, a DNA strand undergoes *polymerase chain reaction* (PCR) and multiple copies of the same strand are created. The sequencer then reads all copies and provides multiple (possibly) erroneous reads to the user (see Figure 1). In nanopore sequencers, these reads are often inaccurate and high-complexity read-alignment and consensus algorithms are required to reconstruct the original DNA strand from these noisy reads.

To reduce the read-alignment complexity and improve the read accuracy, one may employ various coding strategies to design DNA information strands. Yazdi *et al.* [4] proposed a simple coding strategy and verified it experimentally. Later, Cheraghchi *et al.* [5] provided a marker-based coding strategy that has provable reconstruction guarantees.

(b) **Racetrack memories**. Based on spintronic technology, a racetrack memory, also known as *domain wall memory*, is composed of cells, also called *domains*, which are positioned on a tape-like strip and are separated by *domain walls* [6], [7]. The magnetization of a domain is programmed to store a single bit value, which can be read by sensing its magnetization direction. The reading mechanism is operated by a read-only port, called a *head*, together with a reference domain. Since the head is fixed, a shift operation is required in order to read all the domains and this is accomplished by applying shift current which moves the domain walls in one direction. Multiple heads can also be used in order to significantly reduce the read access latency of the memory. When these heads read overlapping segments, we have multiple noisy reads. Recently, Chee *et al.* [8] leveraged on these noisy reads to correct shift errors in racetrack memories. They designed an arrangement of heads and devised a corresponding coding strategy to correct such errors with a *constant number* of redundant bits.

Motivated by these applications, we study the following coding problem in a general setting. Consider a data storage scenario where N distinct noisy reads are provided. Our task is to design a *codebook* such that every codeword can be uniquely reconstructed from any N distinct noisy reads. Hence, our fundamental problem is then: how large can this codebook be? Or equivalently, what is the *minimum number of redundancy*?

In this paper, we study in detail the case where the reads are affected by a single *edit* (a substitution, deletion, or insertion) and its variants. In particular, for the case of a single edit, we show that as the number of noisy reads increases, the number of redundant bits required can be gracefully reduced from $\log n + O(1)$ to $\log\log n + O(1)$, and then to $O(1)$, where n denotes the length of a codeword. Due to space limitations, certain technical proofs are omitted. Detailed proofs are in the arXiv version [9].

II. Problem Statement and Contributions

Consider a data storage scenario described by an error-ball function. Formally, given an input space $\mathcal{X}$ and output space $\mathcal{Y}$, an *error-ball* function B maps a *word* $\boldsymbol{x} \in \mathcal{X}$ to a subset of *noisy reads* $B(\boldsymbol{x}) \subset \mathcal{Y}$. Given a code $\mathcal{C} \subseteq \mathcal{X}$, we define the *read coverage* of $\mathcal{C}$, denoted by $\nu(\mathcal{C}; B)$, to be the quantity

$$\nu(\mathcal{C}; B) \triangleq \max\left\{|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| : \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}, \boldsymbol{x} \neq \boldsymbol{y}\right\}. \quad (1)$$

In other words, $\nu(\mathcal{C}; B)$ is the maximum intersection between the error-balls of any two codewords in $\mathcal{C}$. The quantity $\nu(\mathcal{C}; B)$ was introduced by Levenshtein [10], where he showed that the number of reads[1] required to reconstruct a codeword from $\mathcal{C}$ is at least $\nu(\mathcal{C}; B) + 1$. The problem to determine $\nu(\mathcal{C}; B)$ is referred to as the *sequence reconstruction problem*.

The sequence reconstruction problem was studied in a variety of storage and communication scenarios [5], [8], [11]–[15]. In these cases, $\mathcal{C}$ is usually assumed to be the entire space (all binary words of some fixed length) or a classical error-correcting code.

However, in most storage scenarios, the number of noisy reads N is a fixed system parameter and when N is at most $\nu(\mathcal{C}; B)$, we are unable to uniquely reconstruct the codeword. This work

[1] In the original paper, Levenshtein used the term "channels" instead of reads. Here, we used the term "reads" to reflect the data storage scenario.

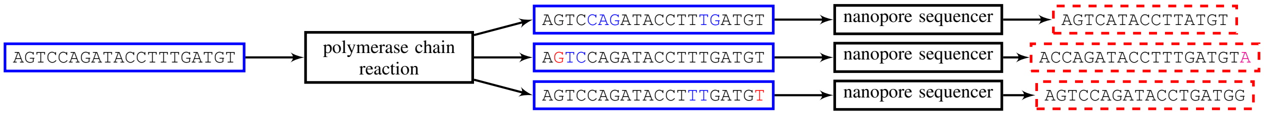

Fig. 1: Noisy reads from a nanopore sequencer. A DNA strand undergoes PCR and multiple copies of the same strand are created. The sequencer then reads all copies and provides multiple errononeous reads. Here, the basepairs coloured in blue were deleted, the basepairs coloured in magenta were inserted, while those coloured in red have substitution errors.

looks at this regime where we design codes whose read coverage is strictly less than N. Specifically, we say that $\mathcal{C}$ is an $(n,N;B)$-*reconstruction code* if $\mathcal{C} \subseteq \{0,1\}^n$ and $\nu(\mathcal{C};B) < N$.

This gives rise to a new quantity of interest that measures the *trade-off between codebook redundancy and read coverage*. Given N and an error-ball function B, we study the quantity

$$\rho(n,N;B) \triangleq \min\left\{n - \log|\mathcal{C}| : \mathcal{C} \subseteq \{0,1\}^n,\, \nu(\mathcal{C};B) < N\right\}. \quad (2)$$

Note that the case $N = 1$ is the classical model which has been studied in the design of error-correcting codes. Thus, the framework studied in this work can be viewed as a natural extension of this classical model.

For a word $\boldsymbol{x} \in \{0,1\}^n$, we consider the following error-ball functions. Let $\mathcal{B}^{\rm I}(\boldsymbol{x})$, $\mathcal{B}^{\rm D}(\boldsymbol{x})$, and $\mathcal{B}^{\rm S}(\boldsymbol{x})$ denote the set of all words obtained from $\boldsymbol{x}$ via one insertion, one deletion, and at most one substitution, respectively. In this work, we study in detail the following error-balls:

$$\mathcal{B}^{\rm SD}(\boldsymbol{x}) \triangleq \mathcal{B}^{\rm S}(\boldsymbol{x}) \cup \mathcal{B}^{\rm D}(\boldsymbol{x}), \qquad \mathcal{B}^{\rm SI}(\boldsymbol{x}) \triangleq \mathcal{B}^{\rm S}(\boldsymbol{x}) \cup \mathcal{B}^{\rm I}(\boldsymbol{x}),$$
$$\mathcal{B}^{\rm ID}(\boldsymbol{x}) \triangleq \mathcal{B}^{\rm I}(\boldsymbol{x}) \cup \mathcal{B}^{\rm D}(\boldsymbol{x}), \qquad \mathcal{B}^{\rm edit}(\boldsymbol{x}) \triangleq \mathcal{B}^{\rm S}(\boldsymbol{x}) \cup \mathcal{B}^{\rm I}(\boldsymbol{x}) \cup \mathcal{B}^{\rm D}(\boldsymbol{x}).$$

Example 1. We consider the single-deletion error-ball $\mathcal{B}^{\rm D}$ and two different codebooks. First, let $\mathcal{C}_{\rm all} = \{0,1\}^n$. Levenshtein in his seminal work [10] showed that $\nu\left(\mathcal{C}_{\rm all};\mathcal{B}^{\rm D}\right) = 2$. In other words, three distinct noisy versions of $\boldsymbol{x}$ allow us to uniquely reconstruct $\boldsymbol{x}$. Hence, $\rho(n,N;\mathcal{B}^{\rm D}) = 0$ for $N \geqslant 3$.

In contrast, to correct a single deletion, we have the classical Varshamov-Tenengolts (VT) code $\mathrm{VT}(n;a)$ whose redundancy is at most $\log(n+1)$ [16] (see also Theorem 4). In this case, $\nu\left(\mathrm{VT}(n;a);\mathcal{B}^{\rm D}\right) = 0$ and one noisy read is sufficient to recover a codeword. Furthermore, it can be shown that $\mathrm{VT}(n;a)$ is asymptotically optimal, or, $\rho\left(n,1;\mathcal{B}^{\rm D}\right) = \log n + \Theta(1)$ (see Theorem 2). A natural question is then: how should we design the codebook when we have only two noisy reads? Or, what is the value of $\rho\left(n,2;\mathcal{B}^{\rm D}\right)$?

Recently, Chee *et al.* constructed a $(n,2;\mathcal{B}^{\rm D})$-reconstruction code with $\log\log n + O(1)$ redundant bits [8]. Hence, $\rho\left(n,2;\mathcal{B}^{\rm D}\right) \leqslant \log\log n + O(1)$. In other words, even though there are only two noisy reads, we can employ a coding strategy that encodes approximately $\log n - \log\log n$ bits of information more than that of the VT code. In this paper, we extend this analysis and design reconstruction codes for other error-balls.

A. Related Work

We first review previous work related to our problem when there is only one noisy read, i.e. $N = 1$. In this case, we recover the usual notion of error-correcting codes. For the error-ball functions studied in this paper, the following results are classical.

Theorem 2. *For $n > 0$,*

(i) $\log(n+1) \leqslant \rho(n,1;\mathcal{B}^{\rm S}) \leqslant \lceil\log(n+1)\rceil$ *[17], [18];*
(ii) $\log(n-1) \leqslant \rho(n,1;\mathcal{B}^{\rm D}) = \rho(n,1;\mathcal{B}^{\rm I}) \leqslant \log(n+1)$ *[16], [19];*
(iii) $\log(n+1) \leqslant \rho(n,1;\mathcal{B}^{\rm edit}) \leqslant 1 + \log n$ *[16].*

Hence, $\rho(n,1;B) = \log n + \Theta(1)$ for $B \in \{\mathcal{B}^{\rm S},\mathcal{B}^{\rm D},\mathcal{B}^{\rm I},\mathcal{B}^{\rm edit}\}$.

For completeness, we present the families of single error-correcting codes provided by Levenshtein [16]. Crucial to these constructions is the concept of *syndrome*.

Definition 3. The *VT syndrome* of a binary sequence $\boldsymbol{x} \in \{0,1\}^n$ is defined to be $\mathrm{Syn}(\boldsymbol{x}) \triangleq \sum_{i=1}^n i x_i$.

Theorem 4 (Levenshtein [16]).

(i) For $a \in \mathbb{Z}_{n+1}$, let

$$\mathrm{VT}(n;a) \triangleq \{\boldsymbol{x} \in \{0,1\}^n : \mathrm{Syn}(\boldsymbol{x}) = a \pmod{n+1}\}. \quad (3)$$

Then, the code $\mathrm{VT}(n;a)$ is an $(n,1;B)$-reconstruction code for $B \in \{\mathcal{B}^{\rm D},\mathcal{B}^{\rm I},\mathcal{B}^{\rm ID}\}$.

(ii) For $a \in \mathbb{Z}_{2n}$, let

$$\mathrm{L}(n;a) \triangleq \{\boldsymbol{x} \in \{0,1\}^n : \mathrm{Syn}(\boldsymbol{x}) = a \pmod{2n}\}. \quad (4)$$

Then, the code $\mathrm{L}(n;a)$ is an $(n,1;B)$-reconstruction code for $B \in \{\mathcal{B}^{\rm SI},\mathcal{B}^{\rm SD},\mathcal{B}^{\rm edit}\}$.

When there is more than one noisy read, previous works usually focus on determining the maximum intersection size between two error-balls. When $\mathcal{C} = \{0,1\}^n$ and the error-balls involve insertions only, deletions only and substitutions only, the value of $\nu(\mathcal{C};B)$ was first determined by Levenshtein [10]. Later, Levenshtein's results were extended in [11] for the case where the error-ball involves deletions only and $\mathcal{C}$ is a single-deletion error-correcting code. Recently, the authors of [20] investigated the case where errors are combinations of single substitution and single insertion. Furthermore, they also simplified the reconstruction algorithm when the number of noisy copies exceeds the minimum required, i.e. $N > \nu(\mathcal{C};B) + 1$.

Another recent variant of Levenshtein's sequence reconstruction problem was studied by the authors in [21]. Similar to our model, the authors consider the scenario where the number of reads is not sufficient to reconstruct a unique codeword. As with classical list-decoding, they determined the size of the *list of possible codewords*.

As mentioned above, the sequence reconstruction problem has been studied by Levenshtein and others for several error channels and distances. In many cases, such as for substitutions, the size of the set $B(\boldsymbol{x}) \cap B(\boldsymbol{y})$ does not depend on the specific choice of $\boldsymbol{x}$ and $\boldsymbol{y}$, but only on their distance. In the substitutions case, for any length-n code $\mathcal{C}$ of minimum Hamming distance d and $B = \mathcal{B}^{\rm S}_t$, which is the radius-$t$ substitution ball, it holds that [10]

$$\nu(\mathcal{C};\mathcal{B}^{\rm S}_t) = N^S_n(t,d) \triangleq \sum_{i=0}^{t-\lceil\frac{d}{2}\rceil} \binom{n-d}{i} \sum_{h=d-t+i}^{t-i} \binom{d}{h}. \quad (5)$$

Therefore, the quantity $\rho(n,N;\mathcal{B}^{\rm S}_t)$ can be computed by finding the minimum Hamming distance of the corresponding code. Specifically, if we denote by $A(n,d)$ the size of the largest length-n code of minimum Hamming distance d, the following holds.

$N \setminus B$	$\mathcal{B}^{\rm S}$	$\mathcal{B}^{\rm I}$, $\mathcal{B}^{\rm D}$	$\mathcal{B}^{\rm ID}$	$\mathcal{B}^{\rm SI}$	$\mathcal{B}^{\rm SD}$	$\mathcal{B}^{\rm edit}$
1	$\lceil\log(n+1)\rceil$	$\log(n+1)$	$\log(n+1)$	$1+\log n$	$1+\log n$	$1+\log n$
2	$\lceil\log(n+1)\rceil$	$\log\log n+\Theta(1)$ $\star$	$\log(n+1)$ $\star$	$1+\log n$ $\star$	$1+\log n$ $\star$	$1+\log n$
3	0	0	$\log\log n+\Theta(1)$ $\star$	$\log\log n+\Theta(1)$ $\star$	$\log\log n+\Theta(1)$ $\star$	$\log\log n+\Theta(1)$ $\star$
4	0	0	$\log\log n+\Theta(1)$ $\star$	2 $\star$	1 $\star$	$\log\log n+\Theta(1)$ $\star$
5	0	0	0 $\star$	0 $\star$	0 $\star$	2 $\star$
6	0	0	0 $\star$	0 $\star$	0 $\star$	1 $\star$
$\geqslant 7$	0	0	0 $\star$	0 $\star$	0 $\star$	0 $\star$

TABLE I: Asymptotically *exact* estimates on $\rho(n, N; B)$ for various error-ball functions. Highlighted in blue are entries that are optimal to one bit. Specifically, if the entry is U, then $U-1 \leqslant \rho(n, N; B) \leqslant U$. Marked with $\star$ are entries derived in this work.

Theorem 5. *For all $N \geq 1$, it holds that*

$$\rho(n, N; \mathcal{B}^{\rm S}_t) = n - \log(A(n, d)),$$

where d is smallest integer such that $N^S_n(t, d) < N$.

For odd values of d it holds that $N^S_n(t, d) = N^S_n(t, d-1)$ (see [18, Theorem 10]) and therefore it is enough to consider only odd values of d. Furthermore, for $d = 2t-1$ it holds that $N^S_n(t, d = 2t-1) = \binom{2t}{t}$ [10], which implies that for all $t \geq 1$,

$$\rho\left(n, N = \binom{2t}{t} + 1; \mathcal{B}^{\rm S}_t\right) = n - \log(A(n, 2t-1)),$$

and for all $1 \leq N \leq \binom{2t}{t}$, $\rho\left(n, N; \mathcal{B}^{\rm S}_t\right) = n - \log(A(n, 2t+1))$.

B. Main Contributions

In this work, we focus on the case where $2 \leqslant N \leqslant \nu(\{0,1\}^n; B)$ with $B \in \{\mathcal{B}^{\rm S}, \mathcal{B}^{\rm I}, \mathcal{B}^{\rm D}, \mathcal{B}^{\rm ID}, \mathcal{B}^{\rm SI}, \mathcal{B}^{\rm SD}, \mathcal{B}^{\rm edit}\}$. When $N = 2$ and $B = \mathcal{B}^{\rm D}$, we have a recent code construction by Chee *et al.* [8] (see Example 1 and Theorem 16). Specifically, in Section IV, we make suitable modifications to this code construction and show that the resulting codes are $(n, N; B)$-reconstruction code for the error-balls of interest.

To do so, in Section III, we study in detail the intersection of certain error-balls and derive the necessary and sufficient conditions for the size of an intersection. Using these characterizations, we not only design the desired reconstruction codes, we also obtain asymptotically tight lower bounds in Section V and in our companion paper [22]. We summarize our results in Table I and observe that all values of $\rho(n, N; B)$ have been determined asymptotically.

III. Combinatorial Characterization of the Intersection of Error-Balls

In this section, we set $\mathcal{C} = \{0,1\}^n$ and compute the read coverage $\nu(\mathcal{C}; B)$ for the error-ball function $B \in \{\mathcal{B}^{\rm S}, \mathcal{B}^{\rm D}, \mathcal{B}^{\rm I}, \mathcal{B}^{\rm SI}, \mathcal{B}^{\rm SD}, \mathcal{B}^{\rm ID}, \mathcal{B}^{\rm edit}\}$. In addition to determining the read coverage or maximum intersection size, we also characterize when the error-balls of a pair of words have intersection of a certain size. This combinatorial characterization will be crucial in the code construction in Section IV. We provide a detailed proof for the case $B \in \{\mathcal{B}^{\rm D}, \mathcal{B}^{\rm I}\}$. Proofs for the other characterizations are deferred to the arXiv version [9].

First, we recall a result from Levenshtein's seminal work.

Theorem 6 (Levenshtein [10, Theorem 1, Eq. (26)])**.** *Let $B \in \{\mathcal{B}^{\rm S}, \mathcal{B}^{\rm I}, \mathcal{B}^{\rm D}\}$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are distinct binary words of length n, then $|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| \leqslant 2$. Therefore, if we set $\mathcal{C} = \{0,1\}^n$, we have that $\nu(\mathcal{C}; B) = 2$ and $\rho(n, N; B) = 0$ for $N \geqslant 3$.*

Next, we characterize when the error-balls of a pair of words have intersection of size two. The case for single substitution is straightforward consequence of (5).

Lemma 7 (Levenshtein [10])**.** *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct binary words of length n. We have that*

(i) $|\mathcal{B}^{\rm S}(\boldsymbol{x}) \cap \mathcal{B}^{\rm S}(\boldsymbol{y})| = 2$ if and only if the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is at most two.

(ii) $|\mathcal{B}^{\rm S}(\boldsymbol{x}) \cap \mathcal{B}^{\rm S}(\boldsymbol{y})| = 0$ if and only if the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is at least three.

When $B \in \{\mathcal{B}^{\rm D}, \mathcal{B}^{\rm I}\}$, we define the following notion of confusability.

Definition 8. Two words $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n are said to be *Type-A-confusable* if there exists subwords $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ such that the following holds.

(A1) $\boldsymbol{x} = \boldsymbol{acb}$ and $\boldsymbol{y} = \boldsymbol{a\bar{c}b}$, where $\bar{\boldsymbol{c}}$ is the complement of $\boldsymbol{c}$.

(A2) $\boldsymbol{c}$ is one of the following forms: $(01)^m$, $(01)^m0$, $(10)^m$, $(10)^m1$ for some $m \geqslant 1$.

Type-A-confusability characterizes when the intersection size of single-deletion (and single-insertion) balls is two.

Proposition 9. *Let $B \in \{\mathcal{B}^{\rm D}, \mathcal{B}^{\rm I}\}$ and $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct binary words of length n. If the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is at least two, we have that $|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| = 2$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable. On the other hand, if the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is one, we have $|\mathcal{B}^{\rm D}(\boldsymbol{x}) \cap \mathcal{B}^{\rm D}(\boldsymbol{y})| = 1$ while $|\mathcal{B}^{\rm I}(\boldsymbol{x}) \cap \mathcal{B}^{\rm I}(\boldsymbol{y})| = 2$.*

Proof. We first consider the case that the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is at least two and show that $|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| = 2$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable. We present the proof for the case where $B = \mathcal{B}^{\rm D}$ and the case for $B = \mathcal{B}^{\rm I}$ can be similarly proved. Let $\boldsymbol{x} = x_1x_2\cdots x_n$, $\boldsymbol{y} = y_1y_2\cdots y_n$ and $\mathcal{B}^{\rm D}(\boldsymbol{x}) \cap \mathcal{B}^{\rm D}(\boldsymbol{y}) = \{\boldsymbol{z}, \boldsymbol{z}'\}$. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ have Hamming distance at least two, we set i and j be the smallest and largest indices, respectively, where the two words differ.

We first consider $\boldsymbol{z} \in \mathcal{B}^{\rm D}(\boldsymbol{x}) \cap \mathcal{B}^{\rm D}(\boldsymbol{y})$. Let $\boldsymbol{z}$ be obtained from $\boldsymbol{x}$ by deleting index k and from $\boldsymbol{y}$ by deleting index ℓ. We first claim that either $k \leqslant i$ or $k \geqslant j$. Suppose otherwise that $i < k < j$ and we have two cases.

- When $\ell < j$, we consider the $(j-1)$th index of $\boldsymbol{z}$. On one hand, since $k < j$ and we delete x_k from $\boldsymbol{x}$, the $(j-1)$th index of $\boldsymbol{z}$ is x_j. On the other hand, since we delete y_ℓ from $\boldsymbol{y}$, the $(j-1)$th index of $\boldsymbol{z}$ is y_j. Hence, $x_j = y_j$, yielding a contradiction.
- When $\ell \geqslant j$, we consider the ith index of $\boldsymbol{z}$. Proceeding as before, we conclude that $x_i = y_i$, which is not possible.

Without loss of generality, we assume that $k \leqslant i$. A similar argument shows that $\ell \geqslant j$. Therefore, we have that $y_t = x_{t+1}$ for

$k \leqslant t \leqslant \ell-1$. Recall that $x_t = y_t$ whenever $t \leqslant i-1$ or $t \geqslant j+1$. Hence, we have that $x_k = x_{k+1} = \cdots = x_i$, $y_k = y_{k+1} = \cdots = y_{i-1}$, $x_{j+1} = x_{j+2} = \cdots = x_\ell$, $y_j = y_{j+1} = \cdots = y_\ell$. In summary, if we set $\boldsymbol{a} = x_1x_2\cdots x_{i-1} = y_1y_2\cdots y_{i-1}$ and $\boldsymbol{b} = x_{j+1}x_{j+2}\cdots x_n = y_{j+1}y_{j+2}\cdots y_n$, then

$$\boldsymbol{z} = \boldsymbol{a}x_{i+1}x_{i+2}\cdots x_j\boldsymbol{b} = \boldsymbol{a}y_iy_{i+1}\cdots y_{j-1}\boldsymbol{b}.$$

Now, we consider $\boldsymbol{z}'$, the other word in $\mathcal{B}^{\mathrm{D}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{D}}(\boldsymbol{y})$. Since $\boldsymbol{z}'$ is distinct from $\boldsymbol{z}$ and proceeding as before, we have that

$$\boldsymbol{z}' = \boldsymbol{a}x_ix_{i+1}\cdots x_{j-1}\boldsymbol{b} = \boldsymbol{a}y_{i+1}y_{i+2}\cdots y_j\boldsymbol{b}.$$

Hence, $x_t = y_{t+1} = x_{t+2}$ for $i \leqslant t \leqslant j-2$. Since $\boldsymbol{z} \neq \boldsymbol{z}'$, we have that $x_i \neq x_{i+1}$ and $y_iy_{i+1}\cdots y_j = \overline{x_ix_{i+1}\cdots x_j}$. Therefore, $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable.

Conversely, suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable. Hence, there exist subwords $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ that satisfy conditions (A1) and (A2). We further set $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ to be words obtained by deleting the first and last index from $\boldsymbol{c}$, respectively. Then $\boldsymbol{a}\boldsymbol{c}_1\boldsymbol{b}$ and $\boldsymbol{a}\boldsymbol{c}_2\boldsymbol{b}$ are two distinct subwords that belong to $\mathcal{B}^{\mathrm{D}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{D}}(\boldsymbol{y})$. Since $|\mathcal{B}^{\mathrm{D}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{D}}(\boldsymbol{y})| \leqslant 2$, we have that the intersection size must be exactly two.

When the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is one, i.e. $\boldsymbol{x} = \boldsymbol{a}\alpha\boldsymbol{b}$ and $\boldsymbol{y} = \boldsymbol{a}\beta\boldsymbol{b}$ where $\alpha \neq \beta$, we have $\mathcal{B}^{\mathrm{D}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{D}}(\boldsymbol{y}) = \{\boldsymbol{a}\boldsymbol{b}\}$ and $\mathcal{B}^{\mathrm{I}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{I}}(\boldsymbol{y}) = \{\boldsymbol{a}\alpha\beta\boldsymbol{b}, \boldsymbol{a}\beta\alpha\boldsymbol{b}\}$. $\blacksquare$

Proposition 9 can be extended to characterize the intersection sizes for error-balls involving either a single insertion or deletion.

Proposition 10 (Single ID)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct binary words of length n. Then $|\mathcal{B}^{\mathrm{ID}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{ID}}(\boldsymbol{y})| \in \{0,2,3,4\}$. Furthermore, we have $|\mathcal{B}^{\mathrm{ID}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{ID}}(\boldsymbol{y})| = 4$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable. Therefore, when $\mathcal{C} = \{0,1\}^n$, we have that $\nu(\mathcal{C};\mathcal{B}^{\mathrm{ID}}) = 4$ and $\rho(n,N;\mathcal{B}^{\mathrm{ID}}) = 0$ for $N \geqslant 5$.*

When the error-balls include single substitutions, we require the following notion of confusability.

Definition 11. Two words $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n are said to be *Type-B-confusable* if there exists subwords $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ and $\boldsymbol{c}'$ such that the following hold:

(B1) $\boldsymbol{x} = \boldsymbol{a}\boldsymbol{c}\boldsymbol{b}$ and $\boldsymbol{y} = \boldsymbol{a}\boldsymbol{c}'\boldsymbol{b}$;

(B2) $\{\boldsymbol{c},\boldsymbol{c}'\}$ is one of the following forms: $\{01^m, 1^m0\}$, $\{10^m, 0^m1\}$ for some $m \geqslant 1$.

We use Type-B-confusability to characterize the intersection of error-balls B where $B \in \{\mathcal{B}^{\mathrm{SD}}, \mathcal{B}^{\mathrm{SI}}\}$.

Proposition 12 (Single SD/SI)**.** *Let $B \in \{\mathcal{B}^{\mathrm{SD}}, \mathcal{B}^{\mathrm{SI}}\}$ and $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct binary words of length n. Then $|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| \leqslant 4$. Furthermore, we have the following characterizations.*

- *If the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is two, then*
  - *(i) $|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| = 4$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-B-confusable with condition (B2) satisfied with $m = 1$.*
  - *(ii) $|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| = 3$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-B-confusable with condition (B2) satisfied with $m \geqslant 2$.*
- *If the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is one, then $|\mathcal{B}^{\mathrm{SI}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{SI}}(\boldsymbol{y})| = 4$ while $|\mathcal{B}^{\mathrm{SD}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{SD}}(\boldsymbol{y})| = 3$.*

Therefore, when $\mathcal{C} = \{0,1\}^n$, we have $\nu(\mathcal{C};B) = 4$ and $\rho(n,N;\mathcal{B}) = 0$ for $N \geqslant 5$.

Finally, we characterize the intersection sizes when the error-balls arise from single edits.

Proposition 13 (Single Edit)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct binary words of length n. Then $|\mathcal{B}^{\mathrm{edit}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{edit}}(\boldsymbol{y})| \leqslant 6$. Furthermore, we have the following characterizations.*

(i) $|\mathcal{B}^{\mathrm{edit}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{edit}}(\boldsymbol{y})| = 6$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-B-confusable with condition (B2) satisfied with $m = 1$.

(ii) $|\mathcal{B}^{\mathrm{edit}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{edit}}(\boldsymbol{y})| = 5$ if and only if the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is one.

(iii) $|\mathcal{B}^{\mathrm{edit}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{edit}}(\boldsymbol{y})| = 4$ if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable with $|\boldsymbol{c}| \geqslant 3$ or Type-B-confusable with condition (B2) satisfied with $m \geqslant 2$.

Therefore, when $\mathcal{C} = \{0,1\}^n$, we have $\nu(\mathcal{C};\mathcal{B}^{\mathrm{edit}}) = 6$ and $\rho(n,N;\mathcal{B}^{\mathrm{edit}}) = 0$ for $N \geqslant 7$.

IV. Reconstruction Codes with $o(\log n)$ Redundancy

An $(n,N;B)$-reconstruction code is also an $(n,N';B)$-reconstruction code for $N' \geqslant N$. Hence, Theorem 2 states that there exists an $(n,N;B)$-reconstruction code with $\log n + O(1)$ redundant bits for all $B \in \{\mathcal{B}^{\mathrm{D}}, \mathcal{B}^{\mathrm{I}}, \mathcal{B}^{\mathrm{ID}}, \mathcal{B}^{\mathrm{SD}}, \mathcal{B}^{\mathrm{SI}}, \mathcal{B}^{\mathrm{edit}}\}$ and $N \geqslant 1$. In this section, we provide reconstruction codes with redundancy $o(\log n)$ when $N > 1$.

We recall a recent construction of an $(n,2;\mathcal{B}^{\mathrm{D}})$-reconstruction code provided in [8] in the context of racetrack memories. Crucial to this construction is the notion of period.

Definition 14. Let ℓ and t be two positive integers where $\ell < t$. Then the word $\boldsymbol{u} = u_1u_2\cdots u_t \in \{0,1\}^t$ is said to have *period* ℓ if $u_i = u_{i+\ell}$ for all $1 \leqslant i \leqslant t-\ell$. We use $\mathcal{R}(n,\ell,t)$ to denote the set of all words $\boldsymbol{c}$ of length n such that the length of any subword of $\boldsymbol{c}$ with period $\leqslant \ell$ is at most t.

In [8], a lower bound on the size of $\mathcal{R}(n,\ell,t)$ is given.

Proposition 15 ([8])**.** *For $\ell \in \{1,2\}$, if $t \geqslant \lceil \log n \rceil + \ell$, we have that the size of $\mathcal{R}(n,\ell,t)$ is at least 2^{n-1}.*

We are now ready to present the following construction of an $(n,2;\mathcal{B}^{\mathrm{D}})$-reconstruction code from [8]. Here, we demonstrate its correctness for completeness and also because the key ideas are crucial to the constructions in Theorems 18 and 20. Furthermore, we improve the construction from [8] and reduce the redundancy by approximately one bit.

Theorem 16 (Single Deletion, $N = 2$ [8])**.** *For $n, P > 0$ with P even, let $c \in \mathbb{Z}_{1+P/2}$ and $d \in \mathbb{Z}_2$. Define $\mathcal{C}_{\mathrm{D}}(n;c,d)$ to be the set of all words $\boldsymbol{x} = x_1x_2\cdots x_n$ such that the following holds.*

(i) $\mathrm{Syn}(\boldsymbol{x}) = c \pmod{1+P/2}$.

(ii) $\sum_{i=1}^{n} x_i = d \pmod 2$.

(iii) $\boldsymbol{x}$ belongs to $\mathcal{R}(n,2,P)$.

Then $\mathcal{C}_{\mathrm{D}}(n;c,d)$ is an $(n,2;\mathcal{B}^{\mathrm{D}})$-reconstruction code. Furthermore, if we set $P = \lceil \log n \rceil + 2$, the code $\mathcal{C}_{\mathrm{D}}(n;c,d)$ has redundancy $1 + \log(\lceil \log n \rceil + 4) = \log\log n + O(1)$ for some choice of c and d.

Proof. We prove by contradiction. Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are two distinct words in $\mathcal{C}_{\mathrm{D}}(n;c,d)$ with $|\mathcal{B}^{\mathrm{D}}(\boldsymbol{x}) \cap \mathcal{B}^{\mathrm{D}}(\boldsymbol{y})| = 2$. Then Proposition 9 states that $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-A-confusable. In other words, there exist substrings $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ such that $\boldsymbol{x} = \boldsymbol{a}\boldsymbol{c}\boldsymbol{b}$, $\boldsymbol{y} = \boldsymbol{a}\bar{\boldsymbol{c}}\boldsymbol{b}$ and $\boldsymbol{c}$ has period at most two.

Note that since the weights of $\boldsymbol{x}$ and $\boldsymbol{y}$ have the same parity, we have $\boldsymbol{c} \in \{(01)^m, (10)^m\}$ for some $m \geqslant 1$. First, suppose that $\boldsymbol{c} = (01)^m$. Then by construction,

$$\mathrm{Syn}(\boldsymbol{x}) - \mathrm{Syn}(\boldsymbol{y}) = 0 \pmod{1+P/2}. \tag{6}$$

On the other hand, since $\boldsymbol{x} = \boldsymbol{acb}$ and $\boldsymbol{y} = \boldsymbol{a\bar{c}b}$, the left-hand side of (6) evaluates to m. However, since $\boldsymbol{c}$ is a subword of $\boldsymbol{x}$ with period at most two, we have that $2m \leqslant P$, and so, $m \neq 0 \pmod{1+P/2}$, arriving at a contradiction. When $\boldsymbol{c} = (10)^m$, the left-hand side of (6) evaluates to $-m \neq 0 \pmod{1+P/2}$. ■

We can similarly show that the code $\mathcal{C}_{\rm D}(n;c,d)$ is capable of reconstructing codewords from noisy reads affected by single insertions or deletions. We omit the proof due to space constraints.

Corollary 17 (Single Insertion/Deletion, $N \in \{3,4\}$). *Let $\mathcal{C}_{\rm D}(n;c,d)$ be as defined in Theorem 16. Then $\mathcal{C}_{\rm D}(n;c,d)$ is an $(n,N;\mathcal{B}^{\rm ID})$-reconstruction code for $N \in \{3,4\}$.*

When $B \in \{\mathcal{B}^{\rm SD}, \mathcal{B}^{\rm SI}\}$, we make suitable modifications to the code $\mathcal{C}_{\rm D}(n;c,d)$ to correct (possibly) a single substitution.

Theorem 18 (Single Substitution/Deletion, $N = 3$). *For $n, P > 0$, let $c \in \mathbb{Z}_{1+P}$ and $d \in \mathbb{Z}_2$. Define $\mathcal{C}_{\rm SD}(n;c,d)$ to be the set of all words $\boldsymbol{x} = x_1x_2\cdots x_n$ such that the following holds.*

(i) $\mathrm{Syn}(\boldsymbol{x}) = c \pmod{1+P}$.
(ii) $\sum_{i=1}^{n} x_i = d \pmod 2$.
(iii) *$\boldsymbol{x}$ belongs to $\mathcal{R}(n,1,P)$.*

Then $\mathcal{C}_{\rm SD}(n;c,d)$ is an $(n,3;B)$-reconstruction code for $B \in \{\mathcal{B}^{\rm SD}, \mathcal{B}^{\rm SI}\}$. Furthermore, if we set $P = \lceil \log n \rceil + 1$, the code $\mathcal{C}_{\rm D}(n;c,d)$ has redundancy $2 + \log(\lceil \log n \rceil + 1) = \log\log n + O(1)$ for some choice of c and d.

Proof. We prove for the error-ball function $\mathcal{B}^{\rm SD}$ and prove by contradiction. Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are two distinct words in $\mathcal{C}_{\rm SD}(n;c,d)$ with $|\mathcal{B}^{\rm SD}(\boldsymbol{x}) \cap \mathcal{B}^{\rm SD}(\boldsymbol{y})| \geqslant 3$. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ have the same parity, the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$ is at least two. Then Proposition 12 states that $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-B-confusable. Without loss of generality, let $\boldsymbol{x} = \boldsymbol{a}01^m\boldsymbol{b}$, $\boldsymbol{y} = \boldsymbol{a}1^m0\boldsymbol{b}$. As before, we have

$$\mathrm{Syn}(\boldsymbol{x}) - \mathrm{Syn}(\boldsymbol{y}) = 0 \pmod{1+P}. \tag{7}$$

Now, the left-hand side of (7) evaluates to m. However, since $\boldsymbol{x}$ belongs to $\mathcal{R}(n,1,P)$, we have that $m \leqslant P$, a contradiction. ■

To correct a single edit with three or four reads, we make a small modification to $\mathcal{C}_{\rm SD}(n;c,d)$. The proof is omitted due to space constraints.

Corollary 19 (Single Edit, $N \in \{3,4\}$). *For $n, P > 0$, let $c \in \mathbb{Z}_{1+P}$ and $d \in \mathbb{Z}_2$. Define $\mathcal{C}_{\rm edit}(n;c,d)$ to be the set of all words $\boldsymbol{x} = x_1x_2\cdots x_n$ such that the following holds.*

(i) $\mathrm{Syn}(\boldsymbol{x}) = c \pmod{1+P}$.
(ii) $\sum_{i=1}^{n} x_i = d \pmod 2$.
(iii) *$\boldsymbol{x}$ belongs to $\mathcal{R}(n,2,P)$.*

Then $\mathcal{C}_{\rm edit}(n;c,d)$ is an $(n,N;\mathcal{B}^{\rm edit})$-reconstruction code for $N \in \{3,4\}$. Furthermore, if we set $P = \lceil \log n \rceil + 2$, the code $\mathcal{C}_{\rm edit}(n;c,d)$ has redundancy $2 + \log(\lceil \log n \rceil + 2) = \log\log n + O(1)$ for some choice of c and d.

Our final code constructions introduce one and two bits of redundancy, respectively. Instead of taking the parity bit of all coordinates, we take the parity of all *even* coordinates.

Theorem 20 (Single Substitution/Deletion, $N = 4$). *Let*

$$\mathcal{C}_1 = \left\{ x_1x_2\cdots x_n \in \{0,1\}^n : \sum_{i=1}^{\lfloor n/2 \rfloor} x_{2i} = 0 \pmod 2 \right\}.$$

Then $\mathcal{C}_1$ is an $(n,4;\mathcal{B}^{\rm SD})$-reconstruction code with one redundant bit.

Proof. We prove by contradiction. Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are two distinct words in $\mathcal{C}_{\rm SD}(n;c,d)$ with $|\mathcal{B}^{\rm SD}(\boldsymbol{x}) \cap \mathcal{B}^{\rm SD}(\boldsymbol{y})| = 4$. Then Proposition 12 states that $\boldsymbol{x}$ and $\boldsymbol{y}$ are Type-B-confusable with $m = 1$. Without loss of generality, let $\boldsymbol{x} = \boldsymbol{a}01\boldsymbol{b}$, $\boldsymbol{y} = \boldsymbol{a}10\boldsymbol{b}$. Then $\sum_{i=1}^{\lfloor n/2 \rfloor} x_{2i} - y_{2i} = 1 \neq 0 \pmod 2$, a contradiction. ■

The next construction takes another bit of redundancy, that is, the parity bit of all coordinates.

Theorem 21 (Single Substitution/Insertion, $N = 4$). *Let*

$$\mathcal{C}_2 = \left\{ x_1x_2\cdots x_n \in \{0,1\}^n : \sum_{i=1}^{n} x_i = \sum_{i=1}^{\lfloor n/2 \rfloor} x_{2i} = 0 \pmod 2 \right\}.$$

Then $\mathcal{C}_2$ is an $(n,4;\mathcal{B}^{\rm SI})$-reconstruction code with two redundant bits.

Finally, it can be shown that $\mathcal{C}_1$ can correct a single edit whenever we have at least five noisy reads.

Corollary 22 (Single Edit, $N \in \{5,6\}$). *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be as defined in Theorems 20 and 21, respectively. Then $\mathcal{C}_2$ is an $(n,5;\mathcal{B}^{\rm edit})$-reconstruction code and $\mathcal{C}_1$ is an $(n,6;\mathcal{B}^{\rm edit})$-reconstruction code.*

V. Lower Bounds for the Redunancy of Reconstruction Codes

In this section, we state without proof some straightforward lower bounds for the redundancy of reconstruction codes when $N = 2$. Detailed proofs can be found in [9] and [22].

The first theorem demonstrates that the codes from Theorem 2 are essentially optimal.

Proposition 23. *Let $n > 0$.*

(i) $\rho(n,2;\mathcal{B}^{\rm S}) = \rho(n,1;\mathcal{B}^{\rm S})$.
(ii) $\rho(n,2;\mathcal{B}^{\rm ID}) = \rho(n,1;\mathcal{B}^{\rm D})$.
(iii) $\rho(n,2;\mathcal{B}^{\rm SD}) \geqslant \rho(n,1;\mathcal{B}^{\rm S})$.
(iv) $\rho(n,2;\mathcal{B}^{\rm SI}) \geqslant \rho(n,1;\mathcal{B}^{\rm S})$.
(v) $\rho(n,2;\mathcal{B}^{\rm edit}) \geqslant \rho(n,1;\mathcal{B}^{\rm S})$.

Therefore, we have that $\rho(n,2;B) = \log n + \Theta(1)$ for $B \in \{\mathcal{B}^{\rm S}, \mathcal{B}^{\rm ID}, \mathcal{B}^{\rm SD}, \mathcal{B}^{\rm SI}, \mathcal{B}^{\rm edit}\}$.

The next proposition shows that we need at least one bit of redundancy for certain instances.

Proposition 24. *$\rho(n,4;\mathcal{B}^{\rm SI}) \geqslant 1$ and $\rho(n,5;\mathcal{B}^{\rm edit}) \geqslant 1$.*

Finally, with the following proposition, we completely determined the asymptotic values of $\rho(n,N;B)$ for all our error-balls of interest.

Proposition 25. *For $(N,B) \in \Big\{(2,\mathcal{B}^{\rm D}), (2,\mathcal{B}^{\rm I}), (3,\mathcal{B}^{\rm ID}), (4,\mathcal{B}^{\rm ID}), (3,\mathcal{B}^{\rm SI}), (3,\mathcal{B}^{\rm SD}), (3,\mathcal{B}^{\rm edit}), (4,\mathcal{B}^{\rm edit})\Big\}$, we have that $\rho(n,N;B) \geqslant \log\log n - O(1)$.*

VI. Conclusion

We studied the sequence reconstruction problem in the context when the number of noisy reads N is fixed. Specifically, for a variety of error-balls B, we designed $(n,N;B)$-reconstruction codes for $2 \leqslant N \leqslant \nu(\{0,1\}^n;B)$ and derived their corresponding lower bounds. Of significance, our code constructions use $o(\log n)$ bits of redundancy and in certain cases are within one bit of optimality. Our results for $\rho(n,N;B)$ are summarized in Table I.

REFERENCES

[1] G. M. Church, Y. Gao, and S. Kosuri. “Next-generation digital information storage in DNA,” *Science*, 337(6102):1628–1628, 2012.

[2] 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*, 494:77–80, 2013.

[3] S. Yazdi, H. M. Kiah, E. R. Garcia, J. Ma, H. Zhao, and O. Milenkovic. DNA-based storage: Trends and methods. *IEEE Trans. Molecular, Biological, Multi-Scale Commun.*, 1(3):230–248, 2015.

[4] S. H. T. Yazdi, R. Gabrys, and O. Milenkovic. Portable and error-free DNA-based data storage. *Scientific reports*, 7(1):5011, 2017.

[5] M. Cheraghchi, R. Gabrys, O. Milenkovic and J. Ribeiro, “Coded trace reconstruction,” *arXiv preprint arxiv:1903.09992*, 2019

[6] S. S. Parkin, M. Hayashi, and L. Thomas, “Magnetic domain-wall racetrack memory,” *Science*, vol. 320, pp. 190–194, 2008.

[7] Y. Zhang, C. Zhang, J. Nan, Z. Zhang, X. Zhang, J.-O. Klein, D. Ravelosona, G. Sun, and W. Zhao. “Perspectives of racetrack memory for large-capacity on-chip memory: From device to system,” *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 63, no. 5, pp. 629–638, 2016.

[8] Y. M. Chee, H. M. Kiah, A. Vardy, E. Yaakobi, and V. K. Vu. “Coding for racetrack memories,” *IEEE Trans. on Information Theory*, 2018.

[9] H. M. Kiah, T. T. Nguyen and E. Yaakobi, “Coding for Sequence Reconstruction for Single Edits,” *arXiv preprint arxiv:2001.01376*, 2020

[10] V. I. Levenshtein, “Efficient reconstruction of sequences,” *IEEE Trans. on Information Theory*, 47(1), pp. 2–22, 2001.

[11] R. Gabrys, and E. Yaakobi. “Sequence reconstruction over the deletion channel,” *IEEE Trans. on Information Theory*, 64(4), pp.2924-2931, 2018.

[12] E. Konstantinova, ”On reconstruction of signed permutations distorted by reversal errors,” *Discrete Math.*, vol. 308, pp. 974–984, Mar. 2008.

[13] V. I. Levenshtein and J. Siemons, ”Error graphs and the reconstruction of elements in groups,” *J. Combinat. Theory, A*, vol. 116, no. 4, pp. 795–815, 2009.

[14] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, ”Exact reconstruction from insertions in synchronization codes,” *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2428–2445, Apr. 2017.

[15] Y. Yehezkeally and M. Schwartz. ”Reconstruction codes for DNA sequences with uniform tandem-duplication errors,” In *Information Theory (ISIT), 2018 IEEE International Symposium on*, pages 2535–2539. IEEE, 2018.

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

[17] R. W. Hamming, “Error detecting and error correcting codes,” *Bell System Technical Journal*, vol. 29 no. 2, pp. 147–160, 1950.

[18] F. J. MacWilliams, and N. J. A. Sloane, The theory of error-correcting codes (Vol. 16). Elsevier, 1977

[19] A. A. Kulkarni and N. Kiyavash, “Nonasymptotic upper bounds for deletion correcting codes,” *IEEE Trans. on Inform. Theory*, vol. 59, no. 8, pp. 5115–5130, 2013.

[20] M. Abu Sini, and E. Yaakobi, “Reconstruction of Sequences in DNA Storage”. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 2019.

[21] V. Junnnila, T. Laihonen, and T. Lehtila, “The Levenshtein’s channel and the list size in information retrieval” In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 2019.

[22] J. Chrisnata, H. M. Kiah, and E. Yaakobi, “Coding for Sequence Reconstruction for Single Edits,” *arXiv preprint arxiv:2004.06032*, 2020