Coding Over Sets for DNA Storage

Andreas Lenz, *Student Member, IEEE*, Paul H. Siegel, *Life Fellow, IEEE*, Antonia Wachter-Zeh, *Member, IEEE*, and Eitan Yaakobi, *Senior Member, IEEE*

***Abstract*— In this paper we study error-correcting codes for the storage of data in synthetic deoxyribonucleic acid (DNA). We investigate a storage model where a data set is represented by an unordered set of M sequences, each of length L. Errors within that model are a loss of whole sequences and point errors inside the sequences, such as insertions, deletions and substitutions. We derive Gilbert-Varshamov lower bounds and sphere packing upper bounds on achievable cardinalities of error-correcting codes within this storage model. We further propose explicit code constructions than can correct errors in such a storage system that can be encoded and decoded efficiently. Comparing the sizes of these codes to the upper bounds, we show that many of the constructions are close to optimal.**

***Index Terms*— Coding over sets, DNA data storage, Gilbert-Varshamov bound, insertion and deletion errors, sphere packing bound.**

I. Introduction

DNA-BASED storage has attracted significant attention due to recent demonstrations of the viability of storing information in macromolecules. This recent increased interest was paved by substantial progress in synthesis and sequencing technologies. The main advantages of DNA-based storage over classical storage technologies are very high data densities and long-term reliability without electrical supply. Given the trends in cost decreases of DNA synthesis and sequencing, it is now acknowledged that within the next 10–15 years DNA storage may become a highly competitive archiving technology [2].

A DNA storage system consists of three important entities (see Fig. 1): (1) a DNA synthesizer that produces the strands that encode the data to be stored in DNA. In order to produce strands with acceptable error rate the length of the strands is typically limited to no more than 250 nucleotides (cf. [2] and also see Table I for an overview over current experiments); (2) a storage container with compartments that store the DNA strands, although in an unordered manner; (3) a DNA sequencer that reads the strands and transfers them back to digital data. The encoding and decoding stages are external

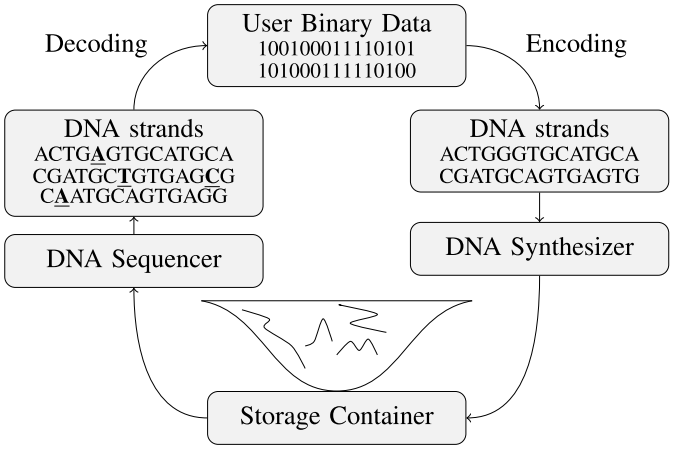


Fig. 1. Illustration of a DNA-based storage system.

TABLE I
SUMMARY OF THE PARAMETERS USED IN RECENT DNA STORAGE EXPERIMENTS. THE STRAND LENGTH IS DEPICTED AS *Code* LENGTH WHICH MATCHES THE LENGTH L IN OUR CHANNEL MODEL

Work	Data Size	Strand Length L	Strands M	$\beta = (\log_4 M)/L$
[10]	0.65MB	115	54,898	0.0685
[11]	0.63MB	117	153,335	0.0736
[12]	0.08MB	117	4,991	0.0525
[14]	0.017MB	1000	32	0.0025
[13]	0.15MB	120	45,652	0.0645
[3]	22MB	190	900,000	0.0521
[4]	2.11MB	152	72,000	0.0531
[15]	0.003MB	1000	17	0.0020
[6]	200.2MB	$150-154$	13,448,372	$0.0769-0.0789$

Manuscript received December 20, 2018; revised October 31, 2019; accepted December 8, 2019. Date of publication December 20, 2019; date of current version March 17, 2020. This work was supported in part by the NSF under Grant CCF-BSF-1619053, in part by the United States–Israel BSF under Grant 2015816, and in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 801434. This work was done in part while A. Lenz and E. Yaakobi were visiting the Center for Memory and Recording Research, University of California San Diego, which also supported the work of E. Yaakobi. This work was presented in part at the 2018 International Symposium on Information Theory, in part at the 2019 Information Theory and Applications Workshop, and in part at the 2019 Non-Volatile Memories Workshop.

A. Lenz and A. Wachter-Zeh are with the Institute for Communications Engineering, Technische Universität München, 80333 Munich, Germany (e-mail: andreas.lenz@mytum.de; antonia.wachter-zeh@tum.de).

P. H. Siegel is with the Electrical and Computer Engineering Department and the Center for Memory and Recording Research, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: psiegel@ucsd.edu).

E. Yaakobi is with the Computer Science Department, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: yaakobi@cs.technion.ac.il).

Communicated by A. Jiang, Associate Editor for Coding Theory.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2961265

processes to the storage system which convert the binary user data into strands of DNA in a way that even in the presence of errors, it is possible to reconstruct the original data.

DNA as a storage system has several attributes which distinguish it from any other storage system. The most prominent one is that the strands are not ordered in the memory and thus it is not possible to know the order in which they were stored. One way to address this problem is using block addresses, also called indices, that are stored as part of the strand. Errors in DNA are typically substitutions, insertions, and deletions, where most published studies report that either substitutions or deletions are the most prominent ones, depending upon the specific technology for synthesis and sequencing [3]–[8]. For example, in column-based DNA oligo synthesis the dominant errors are deletions that result from either failure to remove the dimethoxytrityl (DMT) or combined inefficiencies in the coupling and capping steps [5]. While codes correcting substitution errors were widely studied, much less is known for codes correcting deletions and insertions. The task of error correction becomes even more challenging taking into account the lack of ordering of the strands.

Related work: For a general survey about DNA-based data storage, the reader is referred to [9]. The first large scale experiments that demonstrated the potential of *in vitro* DNA storage were reported by Church *et al.* who recovered 643 KB of data [10] and Goldman *et al.* who accomplished the same task for a 739 KB message [11]. However both of these groups did not recover the entire message successfully due to the lack of using the appropriate coding solutions to correct errors. Church *et al.* had 10 bit errors and Goldman *et al.* lost two strands of 25 nucleotides. Later, in [12], Grass *et al.* reported the first system with usage of error-correcting codes in DNA-based storage and managed to perfectly recover an 81 KB message. Bornholt *et al.* similarly retrieved a 42 KB message [13]. Since then, several groups have built similar systems, storing ever larger amounts of data. Among these, Erlich and Zielinski [4] stored 2.11MB of data with high storage rate, Blawat *et al.* [3] successfully stored 22MB, and more recently Organick *et al.* [6] stored 200MB. Yazdi *et al.* [14], [15] developed a method that offers both random access and rewritable storage. On the other hand, coding theoretic aspects of DNA storage systems have received significant attention recently. The work of [16] discusses error-correcting codes for the DNA sequencing channel, where a possibly erroneous collection of substrings of the original sequence is obtained. In [17], unordered multisets with errors that affect the whole sequence have been discussed. Furthermore, the model proposed in this work has already been adopted in [18], [19]. Namely, codes and bounds for an arbitrary number of substitutions in sets of DNA strands have been derived in [18] and it has been shown that it is possible to construct codes, which have logarithmic redundancy on both, the number of sequences and the length of the sequences. In [19], a distance measure for the DNA storage channel has been discussed and Singleton-like and Plotkin-like code size upper bounds have been derived. In contrast, the goal of this work is to study and to design error-correcting codes which are specifically targeted towards the special structure of DNA storage systems. This goal is accomplished by deriving upper and lower bounds on the achievable size of error-correcting codes and designing constructions over sets that are suitable for data storage in DNA. Errors within this model are a loss of sequences and point errors inside the sequences, such as insertions, deletions, and substitutions. Parts of this work have been published in [1], at the 2019 Information Theory and Applications Workshop, and at the 2019 Non-Volatile Memories Workshop.

The paper is organized as follows. We start by introducing the DNA storage channel model and associated notation. In Sections III and IV we derive generalized Gilbert-Varshamov bounds and sphere packing bounds for the DNA storage channel, which bound the cardinality of optimal error-correcting codes, i.e., codes of largest possible cardinality from below and above. Then, in Section V, we propose code constructions that can correct errors from the DNA storage channel. Lastly, Section VI concludes the paper.

II. Channel Model

A. Notation

We start by introducing the notation that will be used throughout the paper. For any sets $\mathcal{A}, \mathcal{B}$ we write $|\mathcal{A}|$ as the cardinality of $\mathcal{A}$ and $\mathcal{A} \setminus \mathcal{B} = \{x : x \in \mathcal{A} \wedge x \notin \mathcal{B}\}$ as the set difference. We denote by $\mathbb{N}$ and $\mathbb{N}_0$ the sets of natural numbers, where the former consists of the numbers $\{1, 2, 3, \dots\}$ and the latter additionally contains 0. The set $[n] = \{1, 2, \dots, n\}$ contains all natural numbers up to $n \in \mathbb{N}$. Σ_q is a finite alphabet with q elements. In particular, we will write $\Sigma_2 = \{0, 1\}$ for binary sequences and $\Sigma_4 = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ for DNA sequences. A vector of n elements $x_i \in \Sigma_q$ over an alphabet Σ_q is denoted by $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \Sigma_q^n$. Its first, respectively last m elements are denoted by $\mathrm{pref}_m(\mathbf{x})$ and $\mathrm{suff}_m(\mathbf{x})$. The number of runs in $\mathbf{x} \in \Sigma_q^n$, is denoted as $||\mathbf{x}|| \triangleq |\{i : x_i \neq x_{i+1}\}| + 1$. For two vectors $\mathbf{x} \in \Sigma_q^n, \mathbf{y} \in \Sigma_q^m$ we write $(\mathbf{x}, \mathbf{y})$ as the concatenation of $\mathbf{x}$ and $\mathbf{y}$ which has length $n + m$. Throughout the paper, we denote the binary logarithm of a real number $a \in \mathbb{R}^+$ by $\log(a)$ and the natural logarithm by $\ln(a)$. For any integers $n, m \in \mathbb{N}$, $m \leq n$ we write $n! = n \cdot (n-1) \dots 2 \cdot 1$ as the factorial and $n^{\underline{m}} = n(n-1)\dots(n-m+1)$ as the falling factorial. The binomial coefficient is denoted by $\binom{n}{m}$. For the asymptotic behavior of functions, we use the Bachmann-Landau notation, i.e., for $f(n), g(n) : \mathbb{N} \mapsto \mathbb{R}$, we write

- $f(n) = o(g(n))$, if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$,
- $f(n) = \omega(g(n))$, if $\lim_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| = \infty$,
- $f(n) = O(g(n))$, if $\limsup_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| < \infty$,
- $f(n) \sim g(n)$, if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$, and
- $f(n) \gtrsim g(n)$, if $\lim_{n\to\infty} \frac{f(n)}{g(n)} \geq 1$.

B. DNA Channel Model

We consider the DNA storage channel, which is depicted in Fig. 2. In a DNA-based data storage system, data is stored

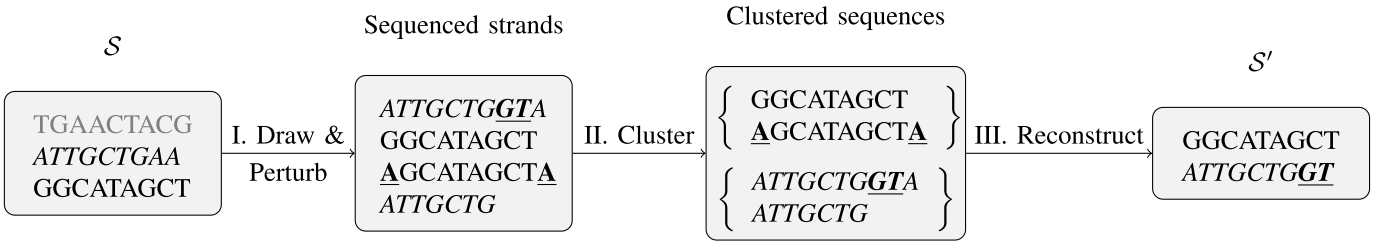


Fig. 2. DNA storage channel model. Sequences with the same text decoration stem from the same original sequence.

in an unordered *set*

$$\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\} \subseteq \Sigma_q^L,$$

with M distinct *sequences* $\mathbf{x}_i \in \Sigma_q^L$, i.e. $\mathbf{x}_i \neq \mathbf{x}_j$ for $i \neq j$. Each sequence $\mathbf{x}_i$ has length L. Here and in the rest of the paper whenever we write the set $\mathcal{S}$ we assume it is a set of M sequences as defined above. Throughout the paper, we will refer to the $\mathbf{x}_i$ by *sequences* or *strands* and to $\mathcal{S}$ by data *sets* or *words*. Representing data words as unordered sets is inherently natural, due to the following two reasons. First, any information about ordering of the data sequences is lost during the storage and second, in the reading process it is not possible to distinguish exactly how many times each sequence was stored, since the sequences are multiplied in the storage medium and not necessarily all of them are read. For more details on the channel model, see [20], [21].

Any such stored data set $\mathcal{S}$ of M sequences is a possible input of the DNA storage channel. Hence, the input space, which comprises all possible data sets is denoted by

$$\mathcal{X}_M^L = \{\mathcal{S} \subseteq \Sigma_q^L : |\mathcal{S}| = M\}.$$

The DNA storage channel can be split into the three following stages, as visualized in Fig. 2.

I. Random sequences are drawn with replacement from the storage medium $\mathcal{S}$ and sequenced, possibly with substitution, insertion or deletion errors.
II. The sequenced strands are clustered according to their Levenshtein distance.[1]
III. The clustered sequences are reconstructed by performing an estimate $\mathbf{x}'$ for each cluster, resulting in the received estimates $\mathcal{S}'$. If two or more reconstructions result in the same estimate $\mathbf{x}'$, we only output a single sequence $\mathbf{x}'$ to avoid possible duplicates of a single stored sequence. Therefore, $\mathcal{S}'$ is a set with distinct elements.

In this work we consider the combination of the above three stages, from the stored sequences $\mathcal{S}$ to the reconstructed sequences $\mathcal{S}'$, as the DNA storage channel. Each sequence $\mathbf{x} \in \mathcal{S}$ is therefore either

- reconstructed correctly, without errors ($\mathbf{x} \in \mathcal{C}$),
- never drawn or its cluster is not identified and thus lost in the storage medium ($\mathbf{x} \in \mathcal{L}$), or
- reconstructed with errors ($\mathbf{x} \in \mathcal{E}$),

where $(\mathcal{C}, \mathcal{L}, \mathcal{E})$ is a partition of $\mathcal{S}$.

[1]This technique was introduced in [6], exploiting the fact that sequences are drawn several times. Other works have either clustered the sequences according to their indices (as in [12]), directly used a code on each sequence to correct insertions and deletions (as in [15]), or simply discarded sequences of incorrect length.

According to the above three cases, we thus associate the following three parameters $(s, t, \epsilon)_{\mathbb{T}}$ that characterize the DNA storage channel. We denote by s the maximum number of sequences that are never drawn (or whose clusters are not identified), by t the maximum number of sequences that have been reconstructed with errors, and by ϵ the maximum number of errors of type $\mathbb{T}$ in each of the latter. Typical error types $\mathbb{T}$ after the reconstruction step are various combinations of insertions, deletions and substitutions, where the latter two are the most prominent ones in DNA storage systems [6]. To be more precise, we define the error balls associated with the channel model. We start with the characterization of point errors inside the sequences.

Definition 1. *The error ball $B_\epsilon^{\mathbb{T}}(\mathbf{x})$ of radius ϵ around a sequence $\mathbf{x} \in \Sigma_q^L$ is defined to be the set of all possible outcomes $\mathbf{x}' \in B_\epsilon^{\mathbb{T}}(\mathbf{x})$, after ϵ (or fewer) errors of type $\mathbb{T}$ in $\mathbf{x}$. Possible types of errors are*

- *Insertions ($\mathbb{I}$),*
- *Deletions ($\mathbb{D}$),*
- *Substitutions ($\mathbb{S}$),*

or combinations of the above, denoted by, e.g., $\mathbb{ID}$ for the case of insertions and deletions. We use the abbreviation $\mathbb{L} \triangleq \mathbb{IDS}$ for insertions, deletions, and substitutions. Similarly, we define the error sphere $S_\epsilon^{\mathbb{T}}(\mathbf{x})$ as the set of possible results $\mathbf{x}' \in S_\epsilon^{\mathbb{T}}(\mathbf{x})$ after exactly ϵ errors of type $\mathbb{T}$. For uniform error balls and spheres, where the size does not depend on the center $\mathbf{x} \in \Sigma_q^L$ we use the abbreviation $B_\epsilon^{\mathbb{T}}(L) \triangleq |B_\epsilon^{\mathbb{T}}(\mathbf{x})|$, respectively $S_\epsilon^{\mathbb{T}}(L) \triangleq |S_\epsilon^{\mathbb{T}}(\mathbf{x})|$. In particular we have

- $S_\epsilon^{\mathbb{I}}(L) = \sum_{i=0}^{\epsilon} \binom{L+\epsilon}{i}(q-1)^i$ *(c.f. [22]),*
- $B_\epsilon^{\mathbb{I}}(L) = \sum_{i=0}^{\epsilon} S_i^{\mathbb{I}}(L)$,
- $S_\epsilon^{\mathbb{S}}(L) = \binom{L}{\epsilon}(q-1)^\epsilon$,
- $B_\epsilon^{\mathbb{S}}(L) = \sum_{i=0}^{\epsilon} \binom{L}{i}(q-1)^i$.

Note that for the case of deletions, such an abbreviation is not possible, since the size of the deletion ball and sphere depends on the center $\mathbf{x}$. The following example illustrates the definitions of error balls for different error types.

Example 1. *Consider the sequence $\mathbf{x} = (\mathrm{AC}) \in \Sigma_4^2$ of length $L = 2$ and a single error, $\epsilon = 1$. The substitution error ball is given by $B_1^{\mathbb{S}}(\mathbf{x}) = \{(\mathrm{AC}), (\mathrm{CC}), (\mathrm{GC}), (\mathrm{TC}), (\mathrm{AA}), (\mathrm{AG}), (\mathrm{AT})\}$. Similarly, the deletion ball around $\mathbf{x}$ is given by $B_1^{\mathbb{D}}(\mathbf{x}) = \{(\mathrm{AC}), (\mathrm{C}), (\mathrm{A})\}$. The insertion sphere around the center $\mathbf{x}$ is $S_1^{\mathbb{I}}(\mathbf{x}) = \{(\mathrm{AAC}), (\mathrm{CAC}), (\mathrm{GAC}), (\mathrm{TAC}), (\mathrm{ACC}), (\mathrm{AGC}), (\mathrm{ATC}), (\mathrm{ACA}), (\mathrm{ACG}), (\mathrm{ACT})\}$.*

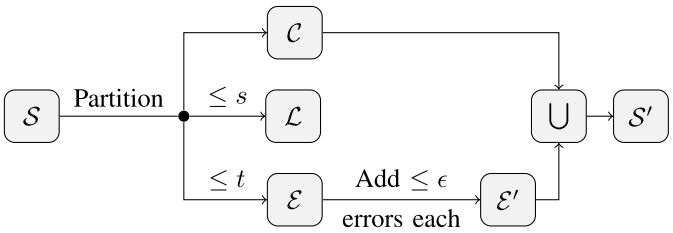


Fig. 3. Illustration of the $(s,t,\epsilon)_{\mathbb{T}}$ channel model.

In a similar fashion it is possible to define the error ball of a data set, as the set of possible received sets after the DNA storage channel.

Definition 2. *For $\mathcal{S} \in \mathcal{X}_M^L$, the error ball $B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S})$ is defined to be the set of all possible received sets $\mathcal{S}'$ after s (or fewer) sequences have been lost and t (or fewer) sequences of the remaining sequences have been distorted by ϵ (or fewer) errors of type $\mathbb{T} \in \{\mathbb{S}, \mathbb{I}, \mathbb{D}, \mathbb{ID}, \mathbb{IS}, \mathbb{DS}, \mathbb{L}\}$ each.*

More precisely, let $\mathsf{Part}_{s,t}(\mathcal{S})$ be the set of all partitions $(\mathcal{C}, \mathcal{L}, \mathcal{E})$ of $\mathcal{S}$ with $|\mathcal{L}| \leq s$, $|\mathcal{E}| \leq t$ and denote by $\mathcal{E} = \{\mathbf{x}_{e_1}, \mathbf{x}_{e_2}, \dots, \}$ the set of stored sequences, which are received in error. We then define $B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S})$ to be

$$B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}) = \left\{ \mathcal{S}' = \mathcal{C} \cup \mathcal{E}' \;\middle|\; \begin{array}{l} (\mathcal{C}, \mathcal{L}, \mathcal{E}) \in \mathsf{Part}_{s,t}(\mathcal{S}), \\ \mathcal{E}' = \{\mathbf{x}'_1\} \cup \cdots \cup \{\mathbf{x}'_{|\mathcal{E}|}\}, \\ \mathbf{x}'_i \in B^{\mathbb{T}}_{\epsilon}(\mathbf{x}_{e_i}) \backslash \{\mathbf{x}_{e_i}\} \;\forall\; i \in [|\mathcal{E}|] \end{array} \right\}.$$

Hereby $\mathcal{E}'$ satisfies $|\mathcal{E}'| \leq |\mathcal{E}|$ and denotes the set of all distinct erroneous received sequences $\mathbf{x}'_i$, after removing duplicates.

The erroneous sequences $\mathbf{x}'_i$ are not necessarily distinct from each other or from the correct sequences in $\mathcal{C}$ and therefore it is possible that two erroneous sequences or one error-free and one erroneous sequence agree with one another, resulting in a loss of a sequence. The number of distinct received sequences $|\mathcal{S}'|$ therefore satisfies $M - t - s \leq |\mathcal{S}'| \leq M$.

Example 2. *Consider the example in Fig. 2 for the DNA storage channel with $M = 3$ stored sequences, $\mathbf{x}_1 = (\mathrm{TGAACTACG})$, $\mathbf{x}_2 = (\mathrm{ATTGCTGAA})$, and $\mathbf{x}_3 = (\mathrm{GGCATAGCT})$, each of length $L = 9$, i.e., $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \in \mathcal{X}_3^9$. The sequenced strands are clustered and reconstructed, resulting in two estimates $\mathbf{y}_1 = (\mathrm{GGCATAGCT})$ and $\mathbf{y}_2 = (\mathrm{ATTGCTGGT})$. The received set is therefore $\mathcal{S}' = \{\mathbf{y}_1, \mathbf{y}_2\}$. Hereby $\mathbf{x}_3$ was received correctly as $\mathbf{y}_1$, $\mathbf{x}_1$ was lost and $\mathbf{x}_2$ was received in error as $\mathbf{y}_2$. It follows that the set of correct, lost and erroneous sequences is given by*

$$\begin{aligned} \mathcal{C} &= \{\mathbf{x}_3\} = \{(\mathrm{GGCATAGCT})\}, \\ \mathcal{L} &= \{\mathbf{x}_1\} = \{(\mathrm{TGAACTACG})\}, \\ \mathcal{E} &= \{\mathbf{x}_2\} = \{(\mathrm{ATTGCTGAA})\}. \end{aligned}$$

It follows that $s = |\mathcal{L}| = 1$ and $t = |\mathcal{E}| = 1$, where there were $\epsilon = 2$ substitution errors in $\mathbf{x}_2$. Therefore, $\mathcal{S}' \in B^{\mathbb{S}}_{1,1,2}(\mathcal{S})$.

The channel from a stored set $\mathcal{S}$ to a received set $\mathcal{S}'$ is visualized in Fig. 3. Throughout the paper, we will refer to the following definition of an error-correcting code in DNA storage systems.

Definition 3. *A code $\mathscr{C} \subseteq \mathcal{X}_M^L$ is called an $(s,t,\epsilon)_{\mathbb{T}}$-**correcting code**, if it can correct a loss of s (or fewer) sequences and ϵ (or fewer) errors of type $\mathbb{T}$ in each of t (or fewer) sequences, i.e., for any pair $\mathcal{S}_1, \mathcal{S}_2 \in \mathscr{C}$ with $\mathcal{S}_1 \neq \mathcal{S}_2$, it holds that*

$$B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}_1) \cap B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}_2) = \emptyset.$$

We say $\mathscr{C} \subseteq \mathcal{X}_M^L$ is an $(s,t,\bullet)_{\mathbb{T}}$-correcting code if the number of errors ϵ per erroneous sequences can be arbitrarily large.

Note that by this definition, a *code* is a set of *codewords*, where each *codeword* is again a set of M *sequences* of length L. One of the main challenges associated with errors in such codewords is the loss of ordering information about the code sequences. The redundancy of a code is defined as follows.

Definition 4. *The* redundancy *of a code $\mathscr{C} \subseteq \mathcal{X}_M^L$ is*

$$r(\mathscr{C}) = \log |\mathcal{X}_M^L| - \log |\mathscr{C}| = \log \binom{q^L}{M} - \log |\mathscr{C}|.$$

We present the results in this work for binary sequences ($q = 2$), however most or all of them can be extended to the non-binary case (and, in particular, the quaternary case). Our results about the redundancy of the proposed constructions and lower bounds on the redundancy are summarized in Table II.

C. Discussion of the Channel Model

Designing and analyzing codes over sets allows to efficiently combat several important aspects of DNA-based data storage. These include the loss of the ordering information of the sequences and the loss or erroneous reception of some of the stored sequences as described in our channel model. Especially when not all sequences are received with errors (i.e. some sequences are received correctly), it is not obvious at all, whether, e.g., prepending an index to each sequence is optimal and how the stored sequences should be protected from errors. Therefore, discussing the channel model from stored sets to received sets is of relevance when aiming for efficient and error-free data storage in DNA. Such a discussion is not possible when only the channel from a single stored sequence to a single received sequence is analyzed.

The following remarks summarize two further observations about the channel model.

Remark 1. *While in practical DNA-based storage systems, the length of the sequences L is moderate, e.g., in the order of a few hundreds, M is significantly larger. In general, we say that $M = q^{\beta L}$ for some $0 < \beta < 1$. Typical values for the parameters M, L and β can be found in Table I.*

Remark 2. *In view of the underlying DNA storage system, which is visualized in Fig. 2, the parameters s, t, ϵ of the channel model depend on the number of sequences that are drawn from the storage medium and also the reconstruction algorithm. Using an efficient reconstruction algorithm, it can be assumed that s, t, ϵ decrease as the number of draws increases, since the reconstruction can be performed more accurately. In particular, when many more than M sequences are drawn from the storage medium, it can be assumed that there are enough draws per sequence that the sequencing*

errors are corrected by the reconstruction algorithm. Consequently there only remain errors which have been introduced when synthesizing the sequences.

D. Relationship of Insertion- and Deletion-Correcting Codes

In this section, we investigate the relationship between $(s,t,\epsilon)_{\mathbb{I}}$-insertion-correcting and $(s,t,\epsilon)_{\mathbb{D}}$-deletion-correcting codes. It is known [23] that for the case of standard blockcodes, any code can correct ϵ insertions if and only if it can correct any ϵ insertions *and* deletions. Surprisingly such an equivalence does *not* hold for our channel model. Here we show a counterexample that an $(s,t,\epsilon)_{\mathbb{D}}$-correcting code is not necessarily an $(s,t,\epsilon)_{\mathbb{I}}$-correcting code.

Example 3. *Consider the code* $\mathscr{C} = \{\mathcal{S}_1, \mathcal{S}_2\}$*, with* $\mathcal{S}_1 = \{(\mathrm{AACCA}), (\mathrm{AACAA}), (\mathrm{GGTTG})\}$ *and* $\mathcal{S}_2 = \{(\mathrm{ACCAA}), (\mathrm{GGTGG}), (\mathrm{GTTGG})\}$*. We can verify that* $\mathscr{C}$ *is* $(0,3,1)_{\mathbb{D}}$*-correcting. It is however not* $(0,3,1)_{\mathbb{I}}$*-correcting, since* $\{(\mathrm{AACCAA}), (\mathrm{GGTTGG})\} \in B^{\mathbb{I}}_{0,3,1}(\mathcal{S}_1)$ *by editing both* (AACCA) *and* (AACAA) *to become* (AACCAA) *and* (GGTTG) *to become* (GGTTGG)*. Similarly,* $\{(\mathrm{AAC\,CAA}), (\mathrm{GGTTGG})\} \in B^{\mathbb{I}}_{0,3,1}(\mathcal{S}_2)$*, since we can edit* (ACCAA) *to become* (AACCAA) *and both* (GGTGG) *and* (GTTGG) *to become* (GGTTGG)*.*

A counterexample for the other direction, i.e., an $(s,t,\epsilon)_{\mathbb{I}}$-correcting code is not necessarily an $(s,t,\epsilon)_{\mathbb{D}}$-correcting code, can be found analogously.

III. Gilbert-Varshamov Bounds

We start by deriving Gilbert-Varshamov lower bounds on the size (equivalently, upper bounds on the redundancy) of optimal $(s,t,\epsilon)_{\mathbb{S}}$ and $(s,t,\epsilon)_{\mathbb{D}}$-correcting codes. An important entity for the derivation of the Gilbert-Varshamov bounds is the set of words $\tilde{\mathcal{S}} \in \mathcal{X}_M^L$, which have intersecting error balls with some $\mathcal{S} \in \mathcal{X}_M^L$. It is defined as follows.

Definition 5. *For a set* $\mathcal{S} \in \mathcal{X}_M^L$*, we denote by* $V^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S})$ *the set of all sets* $\tilde{\mathcal{S}} \in \mathcal{X}_M^L$*, which have intersecting error balls* $B^{\mathbb{T}}_{s,t,\epsilon}(\cdot)$ *with* $\mathcal{S}$*, that is,*

$$V^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}) = \{\tilde{\mathcal{S}} \in \mathcal{X}_M^L : B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}) \cap B^{\mathbb{T}}_{s,t,\epsilon}(\tilde{\mathcal{S}}) \neq \emptyset\}.$$

Hereby, $|V^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S})|$ *is called the degree of* $\mathcal{S}$*. The average degree of all sets is denoted by*

$$\mathbb{E}[V^{\mathbb{T}}_{s,t,\epsilon}] = \frac{1}{\binom{2^L}{M}} \sum_{\mathcal{S} \in \mathcal{X}_M^L} |V^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S})|.$$

The generalized Gilbert-Varshamov bound (cf. [24], [25]) is based on a graph-theoretic representation of an error-correcting code. We will use this representation to find the generalized Gilbert-Varshamov bound for the DNA storage channel. Consider the simple graph $\mathcal{G}$ with the set of vertices $\mathcal{X}_M^L$. Two vertices $\mathcal{S}_1, \mathcal{S}_2 \in \mathcal{X}_M^L$ are connected, if and only if $B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}_1) \cap B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}_2) = \emptyset$. Note that this definition is slightly different from [24], [25] due to the lack of a distance measure in our case. By construction, a *clique* in $\mathcal{G}$ (collection of vertices, where each pair of vertices is connected) is an $(s,t,\epsilon)_{\mathbb{T}}$-correcting code. Now, it can directly be shown that the total number of edges $\mathcal{G}$ coincides with [25, eq. (2)]. Analogously to [25], it is therefore possible to establish a lower bound on the size of a clique in $\mathcal{G}$ (and therefore an $(s,t,\epsilon)_{\mathbb{T}}$-correcting code).

Theorem 1 (cf. [24], [25])**.** *There exists an* $(s,t,\epsilon)_{\mathbb{T}}$*-correcting code* $\mathscr{C} \subseteq \mathcal{X}_M^L$ *of size at least*

$$|\mathscr{C}| \geq \frac{\binom{2^L}{M}}{\mathbb{E}[V^{\mathbb{T}}_{s,t,\epsilon}]}.$$

Such a code can be constructed by successively selecting words $\mathcal{S}^{(i)}$ with minimum degree from $\mathcal{X}_M^L$ as codewords and removing all words $V^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S}^{(i)})$ as possible candidates for the succeeding codewords. Bounding the denominator in Theorem 1 from above will be the main challenge in this section. We start by stating the bound for the case of an arbitrary number of errors per sequence.

Theorem 2. *There exists an* $(s,t,\bullet)_{\mathbb{L}}$*-correcting code* $\mathscr{C} \subseteq \mathcal{X}_M^L$ *of cardinality at least*

$$|\mathscr{C}| \geq \frac{\binom{2^L}{M}}{\binom{M}{s+2t}\binom{2^L}{s+2t}}.$$

Hence, for fixed $s,t \in \mathbb{N}_0$ *and fixed* $0 < \beta < 1$*, there exists an* $(s,t,\bullet)_{\mathbb{L}}$*-correcting code* $\mathscr{C} \subseteq \mathcal{X}_M^L$ *with redundancy*

$$r(\mathscr{C}) \leq (s+2t)L + (s+2t)\log M - \log((s+2t)!^2) + o(1),$$

when $M \to \infty$ *with* $M = 2^{\beta L}$*.*

Proof. We will find an upper bound on $\mathbb{E}[V^{\mathbb{L}}_{s,t,\bullet}]$ by bounding $|V^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S})|$ from above for all $\mathcal{S} \in \mathcal{X}_M^L$. In the following, let $\tilde{\mathcal{S}} \in V^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S}) \subseteq \mathcal{X}_M^L$ be a set which has an intersecting error ball with $\mathcal{S}$. Start by observing that for any such $\tilde{S}$, there exists $\mathcal{S}' \in B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S}) \cap B^{\mathbb{L}}_{s,t,\bullet}(\tilde{\mathcal{S}})$ with $|\mathcal{S}'| \leq M - s$, since $B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S}) \cap B^{\mathbb{L}}_{s,t,\bullet}(\tilde{\mathcal{S}}) \neq \emptyset$ and for all $\mathcal{S}'' \in B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S}) \cap B^{\mathbb{L}}_{s,t,\bullet}(\tilde{\mathcal{S}})$ with $|\mathcal{S}''| > M - s$ it is possible to construct $\mathcal{S}' \in B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S}) \cap B^{\mathbb{L}}_{s,t,\bullet}(\tilde{\mathcal{S}})$ with $|\mathcal{S}'| = M - s$ by removing any $|\mathcal{S}''| - M + s$ sequences from $\mathcal{S}''$. By Definition 2, $|\mathcal{S} \cap \mathcal{S}'| \geq M - s - t$ and also $|\tilde{\mathcal{S}} \cap \mathcal{S}'| \geq M - s - t$. Further, for any such $\mathcal{S}'$,

$$\begin{aligned} |\mathcal{S} \cap \tilde{\mathcal{S}}| &\geq |\mathcal{S} \cap \tilde{\mathcal{S}} \cap \mathcal{S}'| \overset{(a)}{\geq} |\mathcal{S} \cap \mathcal{S}'| + |\tilde{\mathcal{S}} \cap \mathcal{S}'| - |\mathcal{S}'| \\ &\geq 2(M - s - t) - (M - s) = M - s - 2t, \end{aligned}$$

where we used in (a) that $|\mathcal{S} \cap \tilde{\mathcal{S}} \cap \mathcal{S}'| = |\mathcal{S} \cap \mathcal{S}'| + |\tilde{\mathcal{S}} \cap \mathcal{S}'| - |(\mathcal{S} \cup \tilde{\mathcal{S}}) \cap \mathcal{S}'| \geq |\mathcal{S} \cap \mathcal{S}'| + |\tilde{\mathcal{S}} \cap \mathcal{S}'| - |\mathcal{S}'|$ (for an illustration, refer to Fig. 4). Therefore, any $\tilde{\mathcal{S}}$ has an intersection of size at least $M - s - 2t$ with $\mathcal{S}$. Note that for $2^L \geq M + s + 2t$ this bound is tight, i.e., it is possible to find sets $\mathcal{S}, \tilde{\mathcal{S}} \in \mathcal{X}_M^L$ with $B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S}) \cap B^{\mathbb{L}}_{s,t,\bullet}(\tilde{\mathcal{S}}) \neq \emptyset$ and $\mathcal{S} \cap \tilde{\mathcal{S}} = M - s - 2t$. Each $\tilde{\mathcal{S}}$ can thus be constructed by removing $s + 2t$ sequences from $\mathcal{S}$ and adding $s + 2t$ arbitrary sequences. The total number of such choices is at most $\binom{M}{s+2t}\binom{2^L}{s+2t}$. The bound on the redundancy follows from Definition 4 and the fact that for any fixed $a \in \mathbb{N}_0$, $\log\binom{M}{a} = a\log M - \log a! + o(1)$. □

TABLE II

LOWER AND UPPER BOUNDS ON THE REDUNDANCY OF OPTIMAL $(s,t,\epsilon)_{\mathbb{T}}$-CORRECTING CODES. LOW ORDER TERMS ARE OMITTED

Error correction	Gilbert-Varshamov bound	[Sect. III]	Construction	[Sect. V]	Sphere packing bound	[Sect. IV]
$(s,t,\bullet)_{\mathbb{L}}$	$(s+2t)L+(s+2t)\log M$	[Thm. 2]	$M\log \mathrm{e}+(s+2t)(L-\lceil \log M\rceil)$	[Const. 1]	$(s+t)L+t\log M$	[Cor. 1]
			$(s+2t)L$	[Const. 2]		
			$\frac{(1-c)}{2}M^c\log M$ $+(s+2t)M^{1-c}(L-\log M)$	[Const. 3]		
$(\sigma M,\tau M,\bullet)_{\mathbb{L}}$	$(\sigma+2\tau)(L-\log M)$	[Thm. 2]	$(\sigma+2\tau)M(L-\log M)$	[Const. 2]	$(\sigma+\tau)M(L-\log M)$	[Cor. 1]
$(s,t,\epsilon)_{\mathbb{S}}$	$sL+(s+2t)\log M+2t\epsilon\log L$	[Thm. 3]			$sL+t\log M+t\epsilon\log L$	[Thm. 7]
$(s,t,\epsilon)_{\mathbb{D}}$	$sL+(s+t)\log M+2t\epsilon\log(L/2)$	[Thm. 4]	$(s+t)L$	[Const. 2]	$sL+t\epsilon\log L$	[Thm. 9]
$(0,1,1)_{\mathbb{S}}$	$2\log L$	[Thm. 3]	$2L$	[Const. 2]	$\log(ML)$	[Thm. 7]
$(0,1,1)_{\mathbb{D}}$	$2\log L$	[Thm. 4]	$\log L$	[Const. 5]	$\log L$	[Thm. 9]
$(0,M,\epsilon)_{\mathbb{S}}$	$2M\epsilon\log L$	[Thm. 3]	$M\epsilon\log L$	[Const. 7]	$M\epsilon\log L$	[Thm. 8]
$(0,M,1)_{\mathbb{D}}$	$2M\log L$	[Thm. 4]	$M\log L$	[Const. 6]	$M\log L$	[Thm. 10]

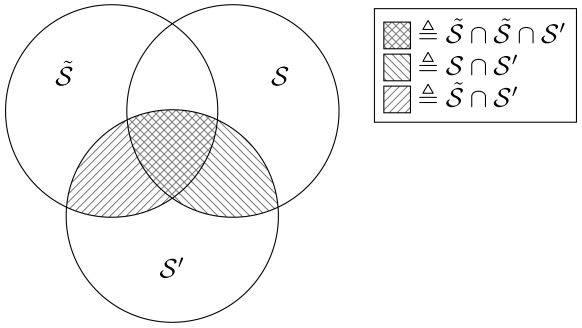


Fig. 4. Illustration for the proof of Theorem 2.

In a similar fashion, we will now establish the existence of a code for the case of a loss of s sequences and a fixed number of ϵ substitution errors in t sequences.

Theorem 3. *There exists an $(s,t,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C}\subseteq\mathcal{X}_M^L$ with cardinality at least*

$$|\mathscr{C}|\geq\frac{\binom{2^L}{M}}{\binom{M}{s,t}\binom{M+t-1}{t}\binom{2^L}{s}B_\epsilon^{\mathbb{S}}(L)^{2t}}.$$

Hence, for fixed $s,t,\epsilon\in\mathbb{N}_0$ and fixed $0<\beta<1$, there exists an $(s,t,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C}\subseteq\mathcal{X}_M^L$ with redundancy

$$r(\mathscr{C})\leq sL+(s+2t)\log M+2t\epsilon\log L-\log(s!^2t!\epsilon!^{2t})+o(1),$$

when $M\to\infty$ with $M=2^{\beta L}$.

Proof. We will find an upper bound on $|V_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})|$ for all $\mathcal{S}\in\mathcal{X}_M^L$. Let $\mathcal{S}'\in B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})$ with $|\mathcal{S}'|\leq M-s$. The number of such elements $\mathcal{S}'$ is at most $\binom{M}{s,t}B_\epsilon^{\mathbb{S}}(L)^t$, as we can choose s sequences to be lost, t sequences to be erroneous and there are $B_\epsilon^{\mathbb{S}}(L)$ error patterns for each erroneous sequence. Given $\mathcal{S}'\in B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})$, we construct possible $\tilde{\mathcal{S}}$ with $\mathcal{S}'\in B_{s,t,\epsilon}^{\mathbb{S}}(\tilde{\mathcal{S}})$ as follows. For each of the t erroneous sequences it is possible to either add ϵ errors to a sequence $\mathbf{x}\in\mathcal{S}'$ or to create a new sequence inside the error ball $B_\epsilon^{\mathbb{S}}(\mathbf{x})$. There are $\binom{M+t-1}{t}B_\epsilon^{\mathbb{S}}(L)^t$ possible error patterns for this procedure. Finally, the s lost sequences can be arbitrary sequences $\mathbf{x}\in\Sigma_2^L$, and there are at most $\binom{2^L}{s}$ choices for these sequences. Thus,

$$|V_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})|\leq\binom{M}{s,t}B_\epsilon^{\mathbb{S}}(L)^t\binom{M+t-1}{t}B_\epsilon^{\mathbb{S}}(L)^t\binom{2^L}{s}.$$

Applying Theorem 1 and using the definition of the redundancy directly yields the bounds of the theorem. □

For the case of deletion errors, we slightly adapt our arguments since the size of the deletion sphere is non-uniform [26]. As stated in Theorem 1, it is sufficient to find an upper bound on the average degree $\mathbb{E}[V_{s,t,\epsilon}^{\mathbb{D}}]$.

Definition 6. *The average of the t-th power of the deletion sphere size $|S_\epsilon^{\mathbb{D}}(\mathbf{x})|$ over all $\mathbf{x}\in\Sigma_2^L$ is defined to be*

$$\mathbb{E}[S_\epsilon^{\mathbb{D},t}]=\frac{1}{2^L}\sum_{\mathbf{x}\in\Sigma_2^L}|S_\epsilon^{\mathbb{D}}(\mathbf{x})|^t.$$

Based on this definition we formulate the following theorem about the existence of $(s,t,\epsilon)_{\mathbb{D}}$-correcting codes.

Theorem 4. *There exists an $(s,t,\epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C}\subseteq\mathcal{X}_M^L$ with cardinality at least*

$$|\mathscr{C}|\geq\frac{\binom{2^L}{M}}{\binom{M}{s,t}\binom{2^L}{s}B_\epsilon^{\mathbb{S}}(L)^t\mathbb{E}[S_\epsilon^{\mathbb{D},t}]}.$$

Hence, for fixed $s,t,\epsilon \in \mathbb{N}_0$ and fixed $0<\beta<1$, there exists an $(s,t,\epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C}\subseteq\mathcal{X}_M^L$ with redundancy

$$r(\mathscr{C}) \leq sL+(s+t)\log M+2t\epsilon\log L-t\epsilon-\log(s!^2t!^2\epsilon!^{2t})+o(1),$$

when $M\to\infty$ with $M=2^{\beta L}$.

Proof. We will derive an upper bound on $\mathbb{E}[V^{\mathbb{D}}_{s,t,\epsilon}]$. The number of elements in $\mathcal{S}'\in B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})$ after a loss of exactly s sequences and ϵ deletions in t sequences is at most

$$|B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})| \leq \sum_{\mathcal{E}\subseteq\mathcal{S},|\mathcal{E}|=t}\prod_{\mathbf{x}\in\mathcal{E}}|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})|\binom{M-t}{s}.$$

This can be illustrated by the following consideration. First, fix $\mathcal{E}\subseteq\mathcal{S}$ with $|\mathcal{E}|=t$. There are $|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})|$ possible error patterns for each $\mathbf{x}\in\mathcal{E}$ and $\binom{M-t}{s}$ choices of s lost sequences among the remaining $M-t$ error-free sequences. Summing over all possible choices $\mathcal{E}\subseteq\mathcal{S}$ of erroneous sequences yields the bound. Then, for each such set $\mathcal{S}'$, there are at most $\binom{2^L}{s}S^{\mathbb{I}}_{\epsilon}(L-\epsilon)^t$ sets $\tilde{\mathcal{S}}$ with $\mathcal{S}'\in B^{\mathbb{D}}_{s,t,\epsilon}(\tilde{\mathcal{S}})\neq\emptyset$. This is because each erroneous sequence $\mathbf{x}'\in\mathcal{S}'$ has length $L-\epsilon$ and requires ϵ insertions to become a sequence of length L. The s lost sequences can be arbitrary words in $\tilde{\mathcal{S}}$ and therefore

$$|V^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})| \leq |B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})|\binom{2^L}{s}S^{\mathbb{I}}_{\epsilon}(L-\epsilon)^t.$$

Taking the average of $|B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})|$ over all sets $\mathcal{S}\in\mathcal{X}_M^L$ yields

$$\begin{aligned}\sum_{\mathcal{S}\in\mathcal{X}_M^L}\frac{|B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})|}{\binom{2^L}{M}} &\leq \frac{\binom{M-t}{s}}{\binom{2^L}{M}}\sum_{\mathcal{S}\in\mathcal{X}_M^L}\sum_{\mathcal{E}\subseteq\mathcal{S},|\mathcal{E}|=t}\prod_{\mathbf{x}\in\mathcal{E}}|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})| \\ &\overset{(a)}{=} \frac{\binom{M}{s,t}}{\binom{2^L}{t}}\sum_{\mathcal{E}\in\mathcal{X}_t^L}\prod_{\mathbf{x}\in\mathcal{E}}|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})| \overset{(b)}{\leq} \frac{\binom{M}{s,t}}{\binom{2^L}{t}}\sum_{\mathcal{E}\in\mathcal{X}_t^L}\sum_{\mathbf{x}\in\mathcal{E}}\frac{|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})|^t}{t} \\ &\overset{(c)}{=} \binom{M}{s,t}\mathbb{E}[S^{\mathbb{D},t}_{\epsilon}].\end{aligned}$$

Here, for equality (a) we used that each set $\mathcal{E}$ with $|\mathcal{E}|=t$ is contained in exactly $\binom{2^L-t}{M-t}$ sets $\mathcal{S}\in\mathcal{X}_M^L$. It follows from the combination of the arithmetic-geometric mean inequality and Jensen inequality that for any non-negative $a_1,\ldots,a_t\geq 0$ it holds that $a_1\cdot\ldots\cdot a_t\leq\frac{1}{t}(a_1^t+\ldots+a_t^t)$, which has been used in inequality (b). Equality (c) follows from the fact that each $\mathbf{x}\in\Sigma_2^L$ is contained in $\binom{2^L-1}{t-1}$ sets $\mathcal{E}\in\mathcal{X}_t^L$. It is known [26] that $|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})|\leq\binom{||\mathbf{x}||+\epsilon-1}{\epsilon}\leq\frac{(||\mathbf{x}||+\epsilon-1)^{\epsilon}}{\epsilon!}$, which results in

$$\begin{aligned}\mathbb{E}[S^{\mathbb{D},t}_{\epsilon}] &\leq \frac{1}{2^L}\sum_{\mathbf{x}\in\Sigma_2^L}\frac{(||\mathbf{x}||+\epsilon-1)^{t\epsilon}}{\epsilon!^t} \\ &\overset{(a)}{=} \frac{1}{\epsilon!^t}\sum_{i=0}^{L-1}\frac{\binom{L-1}{i}(i+\epsilon)^{t\epsilon}}{2^{L-1}} \overset{(b)}{\lesssim} \frac{1}{\epsilon!^t}\left(\frac{L}{2}\right)^{\epsilon t}.\end{aligned}$$

In equality (a) it has been used that the number of words $\mathbf{x}\in\Sigma_2^L$ with $||\mathbf{x}||=i$ is $2\binom{L-1}{i-1}$. For inequality (b), we identify the sum as the decentralized moment of a binomial distribution with $L-1$ trials and success probability $\frac{1}{2}$ and use [27, eq. (4.10)] for the asymptotic behavior, when $L\to\infty$. □

IV. Sphere-Packing Bounds

A well-known method to find upper bounds on the cardinality of error-correcting codes is the sphere-packing bound. In this section we derive sphere-packing bounds for $(s,t,\epsilon)_{\mathbb{T}}$-correcting codes. These bounds directly imply lower bounds on the redundancy of such codes. One particular observation of the considered DNA storage channel is that it is non-uniform, i.e. the sizes of the error balls $B^{\mathbb{T}}_{s,t,\epsilon}(\mathcal{S})$ depend on the channel input $\mathcal{S}$ for all types of errors $\mathbb{T}$, which hinders the computation of sphere packing bounds. A practical method to find sphere packing bounds for non-uniform error balls is the generalized sphere packing bound [28], [29]. However, due to the complex expressions of the error ball sizes, this method does not yield tractable expressions for the considered channel. Another possibility is to derive the sphere packing bound by finding an upper bound on the error ball size, which we will do in Section IV-A. We will also show that for large M most of the error balls have a similar size, which allows to formulate tighter asymptotic sphere packing bounds in Sections IV-B and IV-C. Note that together with the lower bounds on the achievable size of $(s,t,\epsilon)_{\mathbb{T}}$-correcting codes from the previous section and concrete code constructions in Section V, it can be shown that the sphere packing bounds are asymptotically tight for many channel parameters and provide important insights into the nature of the DNA channel.

A. Non-Asymptotic Bounds

We start by finding an upper bound for $(s,t,\bullet)_{\mathbb{L}}$-correcting codes, which depicts the case of a loss of s sequences and an arbitrary number of edit errors in each of t erroneous sequences.

Theorem 5. *The cardinality of any $(s,t,\bullet)_{\mathbb{L}}$-correcting code $\mathscr{C}\subseteq\mathcal{X}_M^L$ satisfies*

$$|\mathscr{C}| \leq \frac{\binom{2^L}{M-s}}{\binom{M}{t+s}\binom{2^L-M}{t}}.$$

In particular, the redundancy of any $(s,t,\bullet)_{\mathbb{L}}$-correcting code $\mathscr{C}\subseteq\mathcal{X}_M^L$ is therefore at least

$$r(\mathscr{C})\geq(s+t)\log(2^L-M-t)+t\log(M-s-t)-\log(t!(s+t)!).$$

Proof. We prove the theorem by finding a subset of $B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S})$, which gives a lower bound on the sphere size $|B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S})|$ for all $\mathcal{S}\in\mathcal{X}_M^L$. Let $\mathcal{S}'\in B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S})\cap\Sigma_2^L$ denote an element from the error ball of $\mathcal{S}$, which contains only sequences of length L and let $\mathcal{C},\mathcal{E}'$ denote the corresponding error-free, respectively erroneous outcomes of the sequences, i.e. $\mathcal{S}'=\mathcal{C}\cup\mathcal{E}'$, according to Definition 2. We construct such distinct $\mathcal{S}'$ in the following way. Choose $M-s-t$ error-free sequences $\mathcal{C}\subseteq\mathcal{S}$ and choose the t erroneous sequences in $\mathcal{E}'$ to be distinct elements out of the 2^L-M sequences in $\Sigma_2^L\backslash\mathcal{S}$ and let $\mathcal{S}'=\mathcal{C}\cup\mathcal{E}'$. For any such $\mathcal{C}\subseteq\mathcal{S}$ and $\mathcal{E}'\subseteq\Sigma_2^L\backslash\mathcal{S}$ one obtains a unique element from the error ball $B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S})$, since $\mathcal{S}'=\mathcal{C}\cup\mathcal{E}'$ and $\mathcal{C},\mathcal{E}'$ are both subsets of two distinct sets. There are in total $\binom{M}{s+t}$ ways to choose the set $\mathcal{C}$ and $\binom{2^L-M}{t}$ ways to choose

$\mathcal{E}'$ and thus $|B^{\mathbb{L}}_{s,t,\bullet}(\mathcal{S})| \geq \binom{M}{s+t}\binom{2^L-M}{t}$. All such constructed received sets have $|\mathcal{S}'| = |\mathcal{C}| + |\mathcal{E}'| = M - s$ sequences of length L and therefore, we obtain by a sphere packing argument, that any $(s,t,\bullet)_{\mathbb{L}}$-correcting code $\mathscr{C}$ satisfies

$$|\mathscr{C}| \leq \frac{\binom{2^L}{M-s}}{\binom{M}{t+s}\binom{2^L-M}{t}}.$$

Therefore, the redundancy is at least

$$\begin{aligned} r(\mathscr{C}) &= \log\binom{2^L}{M} - \log|\mathscr{C}| \\ &\geq \log\frac{(2^L-M+s)!(M-s)!}{(2^L-M-t)!(M-s-t)!(s+t)!t!} \\ &\geq (s+t)\log(2^L-M-t) + t\log(M-t-s) \\ &\qquad - \log(t!(s+t)!). \end{aligned}$$

□

This non-asymptotic bound directly implies an asymptotic bound, when $M \to \infty$ and $M = 2^{\beta L}$ for fixed $0 < \beta < 1$.

Corollary 1. *For fixed $s,t \in \mathbb{N}_0$ and fixed $0 < \beta < 1$, the redundancy of any $(s,t,\bullet)_{\mathbb{L}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ is asymptotically at least*

$$r(\mathscr{C}) \geq (s+t)L + t\log M - \log(t!(s+t)!) + o(1),$$

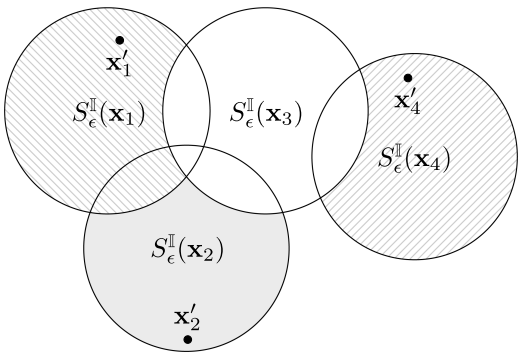
when $M \to \infty$ and $M = 2^{\beta L}$. Further, for any fixed σ, τ with $\sigma > 0$, $\tau > 0$ and $\sigma + \tau < 1$, the redundancy of any $(\sigma M, \tau M, \bullet)_{\mathbb{L}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ satisfies

$$r(\mathscr{C}) \geq (\sigma+\tau)M(L - \log M + \log \mathrm{e}) + MH(\sigma+\tau) + o(M),$$

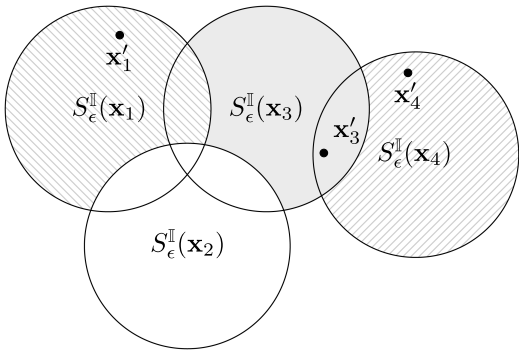
where $H(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

This result is particularly interesting, due to the following consideration. Both lost sequences and erroneous sequences do not carry any useful information, since the erroneous sequences can be distorted by an arbitrary number of errors. However, unlike the lost sequence, the erroneous sequence cannot directly be detected by the decoder and therefore, compared to a loss of sequence, requires additional redundancy of roughly $\log M$ bits to be corrected. This result is analogous to the case of standard binary substitution-correcting block-codes of length n, where erasures require a redundancy of only a single symbol, and errors require roughly $\log n$ symbols of redundancy to be corrected. This analogy becomes particularly visible when sequences are indexed and protected by a standard substitution-correcting code, similarly to Construction 1 (see Section V-B), but also holds for the general case of any $(s,t,\bullet)_{\mathbb{L}}$-correcting code. However, this seems to be not the case, when the number of lost sequences and erroneous sequences scales with M, since the redundancy only depends on $\sigma + \tau$.

In the following, we find code size upper bounds for the case of having a combination of a loss of s sequences and only ϵ insertion errors inside t sequences. We start by defining a quantity that will be useful for the formulation of the bound.



(a) Exemplary case: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4 \in \mathcal{E}$, and $\mathbf{x}_3 \in \mathcal{L}$



(b) Exemplary case: $\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4 \in \mathcal{E}$, and $\mathbf{x}_2 \in \mathcal{L}$

Fig. 5. Illustration for the choice of $\mathcal{E}'$ in the proof of Theorem 6. The erroneous outcomes are chosen out of the corresponding error spheres, which are highlighted in gray.

Definition 7. *The largest intersection of two ϵ-insertion spheres of any two distinct words $\mathbf{x}, \mathbf{y} \in \Sigma_2^L$ is denoted by*

$$N^{\mathbb{I}}_\epsilon(L) = \max_{\mathbf{x},\mathbf{y}\in\Sigma_2^L} |S^{\mathbb{I}}_\epsilon(\mathbf{x}) \cap S^{\mathbb{I}}_\epsilon(\mathbf{y})|.$$

Note that from [30] it is known that $N^{\mathbb{I}}_\epsilon(L) = \sum_{i=0}^{\epsilon-1}\binom{L+\epsilon}{i}(1-(-1)^{\epsilon-i})$. The sphere packing bound is derived in the following theorem.

Theorem 6. *The cardinality of any $(s,t,\epsilon)_{\mathbb{I}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ satisfies*

$$|\mathscr{C}| \leq \frac{\binom{2^L}{M-s-t}\binom{2^{L+\epsilon}}{t}}{\binom{M}{s,t}\prod_{i=0}^{t-1}(S^{\mathbb{I}}_\epsilon(L) - (s+i)N^{\mathbb{I}}_\epsilon(L)}.$$

Proof. We prove the theorem by bounding the error ball size $|B^{\mathbb{I}}_{s,t,\epsilon}(\mathcal{S})|$ from below for all $\mathcal{S}$, which yields an upper bound on the cardinality of $(s,t,\epsilon)_{\mathbb{I}}$-correcting codes by a sphere packing argument. Distinct elements $S' \in B^{\mathbb{I}}_{s,t,\epsilon}(\mathcal{S})$ of the error ball can be found in the following way. First, choose two distinct sets $\mathcal{L}, \mathcal{E} = \{\mathbf{x}_{e_1}, \ldots, \mathbf{x}_{e_t}\} \subseteq \mathcal{S}$ with $|\mathcal{L}| = s$ and $|\mathcal{E}| = t$. Further choose the set of erroneous sequences $\mathcal{E}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_t\}$ such that

$$\mathbf{x}'_i \in S^{\mathbb{I}}_\epsilon(\mathbf{x}_{e_i}) \Bigg\backslash \left(\bigcup_{\mathbf{y}\in\mathcal{P}_i} S^{\mathbb{I}}_\epsilon(\mathbf{y})\right)$$

as illustrated in Fig. 5, where $\mathcal{P}_i = \mathcal{L} \cup \{\mathbf{x}_{e_1}, \ldots, \mathbf{x}_{e_{i-1}}\}$. The received set $\mathcal{S}'$ is then constructed by $\mathcal{S}' = \mathcal{C} \cup \mathcal{E}'$, where $\mathcal{C} = \mathcal{S} \setminus (\mathcal{L} \cup \mathcal{E})$ are the error-free sequences, as in Definition 2. We will show that each choice $\mathcal{L}, \mathcal{E}, \mathcal{E}'$ leads to a unique element in $B^{\mathbb{I}}_{s,t,\epsilon}(\mathcal{S})$. Denote by $\mathcal{L}, \mathcal{E}, \mathcal{E}'$ and $\tilde{\mathcal{L}}, \tilde{\mathcal{E}}, \tilde{\mathcal{E}}'$ two different choices and let $\mathcal{S}'$ and $\tilde{\mathcal{S}}'$ be the corresponding received sets. If $\mathcal{L} \cup \mathcal{E} \neq \tilde{\mathcal{L}} \cup \tilde{\mathcal{F}}$, it directly follows that $\mathcal{S}' \neq \tilde{\mathcal{S}}'$, since the error-free sequences are different. However, if $\mathcal{L} \cup \mathcal{E} = \tilde{\mathcal{L}} \cup \tilde{\mathcal{F}}$, it follows that $\mathcal{E}' \neq \tilde{\mathcal{E}}'$ due to the choice of the sequences in the set $\mathcal{E}'$. Therefore, two different choices of the sets $\mathcal{L}, \mathcal{E}, \mathcal{E}'$ yield different elements in $B^{\mathbb{I}}_{s,t,\epsilon}(\mathcal{S})$. The number of possible sets $\mathcal{L}, \mathcal{E}$ is $\binom{M}{s,t}$. For each $\mathbf{x}_{e_i} \in \mathcal{E}$, we have at least $S^{\mathbb{I}}_{\epsilon}(L) - (s+i)N^{\mathbb{I}}_{\epsilon}(L)$ possibilities to choose the erroneous outcome $\mathbf{x}'_i$, since there are $S^{\mathbb{I}}_{\epsilon}(L)$ sequences in $S^{\mathbb{I}}_{\epsilon}(\mathbf{x}_{e_i})$ and at most $(s+i)N^{\mathbb{I}}_{\epsilon}(L)$ of them are in common with elements of the insertion spheres of $\mathcal{P}_i$. Hence, in total, there are $\binom{M}{s,t}\prod_{i=0}^{t-1}(S^{\mathbb{I}}_{\epsilon}(L) - (s+i)N^{\mathbb{I}}_{\epsilon}(L))$ ways to choose $\mathcal{L}, \mathcal{E}, \mathcal{E}'$ and therefore $|B^{\mathbb{I}}_{s,t,\epsilon}(\mathcal{S})| \geq \binom{M}{s,t}\prod_{i=0}^{t-1}(S^{\mathbb{I}}_{\epsilon}(L) - (s+i)N^{\mathbb{I}}_{\epsilon}(L))$ for all $\mathcal{S} \in \mathcal{X}^L_M$. Each such created received set $\mathcal{S}'$ consists of $M - s - t$ sequences of length L and t sequences of length $L+\epsilon$. There are in total $\binom{2^L}{M-s-t}\binom{2^{L+\epsilon}}{t}$ such sets, which yields the theorem by a sphere packing argument. □

Note that, Theorem 6 provides a valid upper bound for any parameter M, L, s, t, ϵ. For the case of deletion errors or combinations of insertions and deletions, formulating a sphere packing bound based on the minimum error ball size yields a weak bound, since the minimum deletion ball size is $|B^{\mathbb{D}}_{\epsilon}(\mathbf{0})| = \epsilon + 1$. Therefore, a conservative analysis similar to Theorem 6 would yield unsatisfactory results. However, an asymptotic analysis, which yields asymptotically tighter bounds is possible, as we will see in Theorem 9.

B. Asymptotic Bounds for Substitution Errors

We now derive asymptotic sphere packing bounds for large numbers of sequences M on the code size for $(s,t,\epsilon)_{\mathbb{S}}$-correcting codes, which depicts the case of only substitution errors inside the sequences. As discussed before, the error ball sizes depend on the center $\mathcal{S}$. However, as it turns out, asymptotically the error balls have similar sizes. We will start by finding a lower bound on the error ball size for a set $\mathcal{S}$.

Lemma 1. *Let $\mathcal{Y} \subseteq \mathcal{S} \in \mathcal{X}^L_M$ be an ϵ-substitution-correcting code, i.e. $B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_1) \cap B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_2) = \emptyset$ for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ and $\mathbf{y}_1 \neq \mathbf{y}_2$. Further, let $s + t \leq |\mathcal{Y}|$. Then,*

$$|B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})| \geq \binom{|\mathcal{Y}|}{s,t}\left(B^{\mathbb{S}}_{\epsilon}(L) - 1\right)^t.$$

Proof. A lower bound for $|B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})|$ will be proven by identifying and counting specific patterns of a loss of sequences and errors in sequences that lead to distinct channel outputs $\mathcal{S}' \in B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})$. Throughout this proof, we impose a lexicographic ordering onto the sequences in Σ_2^L, which means that, writing $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_{|\mathcal{A}|}\}$ for any set $\mathcal{A} \subseteq \Sigma_2^L$ uniquely determines each element $\mathbf{a}_i$. The sets of stored sequences in the error balls around the elements in $\mathcal{Y}$ are denoted by $\mathcal{Y}_i = \mathcal{S} \cap B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_i)$. Similarly, the sets of received sequences in these error balls are $\mathcal{Y}'_i = \mathcal{S}' \cap B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_i)$. Note that the sets $B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_i)$ and thus also the sets $\mathcal{Y}_i$ are distinct, since $\mathcal{Y}$ is an ϵ-substitution-correcting code. We further define the selector function for sequences $\mathbf{a}, \mathbf{b}, \mathbf{x} \in \Sigma_2^L$ as

$$\mathbb{I}^{\mathcal{S}}_{\mathbf{x}}(\mathbf{a}, \mathbf{b}) = \begin{cases} \mathbf{a}, & \text{if } \mathbf{x} \notin \mathcal{S} \\ \mathbf{b}, & \text{otherwise.} \end{cases}$$

The distinct channel outputs $\mathcal{S}' \in B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})$ are obtained in the following manner. First, choose two distinct sets $\mathcal{L} \subseteq \mathcal{Y}$ with $|\mathcal{L}| = s$ and $\mathcal{E} = \{\mathbf{y}_{e_1}, \ldots, \mathbf{y}_{e_t}\} \subseteq \mathcal{Y}$ with $|\mathcal{E}| = t$ and a collection of error vectors $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_t)$, where $\mathbf{e}_j \in \Sigma_2^L$ are non-zero error vectors of weight at most ϵ. We will show that for each choice of $\mathcal{L}, \mathcal{E}$, and $\mathbf{E}$ we obtain a unique point $\mathcal{S}' \in B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})$ in the following manner. First, all sequences in $\mathcal{L}$ are lost. Let $\mathbf{y}'_i \triangleq \mathbf{y}_{e_i} + \mathbf{e}_i$. The set $\mathcal{E}$ of erroneous sequences is chosen as

$$\mathcal{E} = \bigcup_{i=1}^{t}\left\{\mathbb{I}^{\mathcal{S}}_{\mathbf{y}'_i}(\mathbf{y}_{e_i}, \mathbf{y}'_i)\right\}.$$

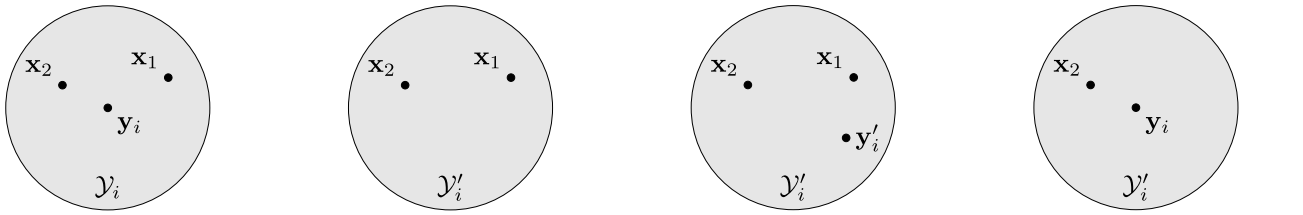
In other words, if $\mathbf{y}'_i \notin \mathcal{S}$ we choose the sequence, which will be distorted by errors to be $\mathbf{y}_{e_i}$ and otherwise we choose it to be exactly $\mathbf{y}'_i$. The erroneous outcomes of the sequences in $\mathcal{E}$ are now constructed by

$$\mathcal{E}' = \bigcup_{i=1}^{t}\left\{\mathbb{I}^{\mathcal{S}}_{\mathbf{y}'_i}(\mathbf{y}'_i, \mathbf{y}_{e_i})\right\}.$$

That is if $\mathbf{y}'_i \notin \mathcal{S}$, we have $\mathbf{y}_{e_i} \in \mathcal{E}$ and we add $\mathbf{e}_i$ to that sequence to obtain $\mathbf{y}'_i \in \mathcal{E}'$. If $\mathbf{y}'_i \in \mathcal{S}$, $\mathbf{y}'_i \in \mathcal{E}$ is the sequence which is distorted and we add $-\mathbf{e}_i$, resulting in $\mathbf{y}_{e_i} \in \mathcal{E}'$. It is very important to note that by this choice of error patterns, the erroneous sequence $\mathbf{y}'_i \in B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_{e_i})$ and therefore will never be present in another error ball $B^{\mathbb{S}}_{\epsilon}(\mathbf{y}), \mathbf{y} \in \mathcal{Y}\setminus\{\mathbf{y}_{e_i}\}$, since $\mathcal{Y}$ is an ϵ-substitution-correcting code. The received set is now $\mathcal{S}' = \mathcal{C} \cup \mathcal{E}'$, where $\mathcal{C} = \mathcal{S}\setminus(\mathcal{L}\cup\mathcal{E})$ are the error-free sequences, as in Definition 2. We will show now that two choices $\mathcal{L}, \mathcal{E}, \mathbf{E}$ and $\tilde{\mathcal{L}}, \tilde{\mathcal{E}}, \tilde{\mathbf{E}}$ yield different received sets $\mathcal{S}'$ and $\tilde{\mathcal{S}}'$, if (and only if) they differ in at least one of the components, i.e., $\mathcal{L} \neq \tilde{\mathcal{L}}$, $\mathcal{E} \neq \tilde{\mathcal{E}}$, or $\mathbf{E} \neq \tilde{\mathbf{E}}$. We distinguish between the following three different cases (visualized in Fig. 6) and the resulting received parts $\mathcal{Y}'_i$

- $\mathbf{y}_i \in \mathcal{Y} \setminus (\mathcal{L} \cup \mathcal{E}) : \mathcal{Y}'_i = \mathcal{Y}_i$,
- $\mathbf{y}_i \in \mathcal{L} : \mathcal{Y}'_i = \mathcal{Y}_i \setminus \{\mathbf{y}_i\}$,
- $\mathbf{y}_i \in \mathcal{E} : \quad \mathcal{Y}'_i = (\mathcal{Y}_i \setminus \{\mathbf{y}_i\}) \cup \{\mathbf{y}'_i\}$ or
  $\mathcal{Y}'_i = \mathcal{Y}_i \setminus \{\mathbf{x}\}$,

where $\mathbf{y}'_i \in B^{\mathbb{S}}_{\epsilon}(\mathbf{y}_i) \setminus \mathcal{S}$ and $\mathbf{x} \in \mathcal{Y}_i \setminus \{\mathbf{y}_i\}$. By comparing the outputs $\mathcal{Y}'_i$ for these three cases, it is verified that for any two different cases, $\mathcal{Y}'_i$ can never be the same. Now, if $\mathcal{L} \neq \tilde{\mathcal{L}}$ there is at least one i such that $\mathbf{y}_i \in \mathcal{L}$ and $\mathbf{y}_i \notin \tilde{\mathcal{L}}$ and if $\mathcal{E} \neq \tilde{\mathcal{E}}$ there is at least one i such that $\mathbf{y}_i \in \mathcal{E}$ and $\mathbf{y}_i \notin \tilde{\mathcal{E}}$ and therefore it follows that $\mathcal{Y}'_i \neq \tilde{\mathcal{Y}}'_i$ and $\mathcal{S}' \neq \tilde{\mathcal{S}}'$. Further, if $\mathcal{L} = \tilde{\mathcal{L}}$ and $\mathcal{E} = \tilde{\mathcal{E}}$, but $\mathbf{E} \neq \tilde{\mathbf{E}}$, there is at least one i with $\mathbf{e}_i \neq \tilde{\mathbf{e}}_i$ and thus $\mathcal{Y}'_i \neq \tilde{\mathcal{Y}}'_i$. This proves that each $\mathcal{L}, \mathcal{E}, \mathbf{E}$ yields a unique point in $B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})$. Finally, there are $\binom{|\mathcal{Y}|}{s,t}$ possible solutions to choose the sets $\mathcal{L}$ and $\mathcal{E}$ and $(B^{\mathbb{S}}_{\epsilon}(L) - 1)^t$ error patterns $\mathbf{E}$. □



(a) No errors (b) $\mathbf{y}_i$ is lost (c) $\mathbf{y}_i$ is erroneous, $\mathbf{y}_i' \neq \mathbf{x}_j$ (d) $\mathbf{x} \neq \mathbf{y}_i$ is erroneous, $\mathbf{x}' = \mathbf{y}_i$

Fig. 6. Cases for error patterns in Lemma 1.

This means, that if a set $\mathcal{S} \in \mathcal{X}_M^L$ contains an ϵ-substitution-correcting code $\mathcal{Y}$ with cardinality $|\mathcal{Y}|$, the error ball has size at least $|B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})| \geq \binom{|\mathcal{Y}|}{s,t} B_\epsilon^{\mathbb{S}}(L)^t$. Interestingly, for an appropriate choice of parameters, most of the sets $\mathcal{S} \in \mathcal{X}_M^L$ have the property of containing a large ϵ-error-correcting code. To establish the fact, we need the following lemma.

Lemma 2. *Let $\mathcal{Y} \subseteq \mathcal{S}$ be the largest ϵ-error-correcting code (error type $\mathbb{T}$) with $B_\epsilon^{\mathbb{T}}(\mathbf{y}_1) \cap B_\epsilon^{\mathbb{T}}(\mathbf{y}_2) = \emptyset$ for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ and $\mathbf{y}_1 \neq \mathbf{y}_2$. The number of sets $\mathcal{S} \subseteq \mathcal{X}_M^L$ with $|\mathcal{Y}| \leq K$, denoted as $D(K)$, is at most*

$$D(K) \leq \binom{2^L}{K} \binom{KV_\epsilon^{\mathbb{T}}}{M-K},$$

where

$$V_\epsilon^{\mathbb{T}} = \max_{\mathbf{x} \in \Sigma_2^L} |\{\mathbf{y} \in \Sigma_2^L : B_\epsilon^{\mathbb{T}}(\mathbf{x}) \cap B_\epsilon^{\mathbb{T}}(\mathbf{y}) \neq \emptyset\}|$$

is the maximum number of sequences $\mathbf{y} \in \Sigma_2^L$ that have intersecting error balls $B_\epsilon^{\mathbb{T}}(\cdot)$ with any $\mathbf{x} \in \Sigma_2^L$.

Proof. Consider the following procedure on a set $\mathcal{S} \in \mathcal{X}_M^L$ whose largest ϵ-error-correcting subset $\mathcal{Y} \subseteq \mathcal{S}$ has size at most K. Write $\mathcal{S}^{(1)} \triangleq \mathcal{S}$. Take an arbitrary word $\mathbf{x}^{(1)} \in \mathcal{S}^{(1)}$ and remove all words $\mathbf{y} \in \Sigma_2^L$ with intersecting error balls, i.e. $B_\epsilon^{\mathbb{T}}(\mathbf{x}) \cap B_\epsilon^{\mathbb{T}}(\mathbf{y}) \neq \emptyset$ from $\mathcal{S}^{(1)}$. Then select an arbitrary sequence from the resulting set $\mathcal{S}^{(2)}$, and, again, remove all elements with intersecting error balls. Continue this procedure until $S^{(j+1)} = \emptyset$. This procedure will stop after at most $j \leq K$ steps, since otherwise $\mathbf{x}_1, \ldots, \mathbf{x}_{K+1}$ would form an ϵ-error-correcting code. Hence, each such set $\mathcal{S}$ can be constructed by first selecting K arbitrary, distinct words $\mathbf{x}_1, \ldots, \mathbf{x}_K$ and then choosing the remaining $M - K$ words to have intersecting error balls with at least one of the $\mathbf{x}_1, \ldots, \mathbf{x}_K$. □

While the bound from Lemma 2 may not seem particularly strong, it can be used to show that the number of sets that do not contain an ϵ-substitution-correcting code of large size is negligible with respect to the sets that do contain an ϵ-substitution-correcting code. We will elaborate this result and use it in the following to prove an upper bound on the size of $(s,t,\epsilon)_{\mathbb{S}}$-correcting codes.

Theorem 7. *For fixed $s, t, \epsilon \in \mathbb{N}_0$ and $0 < \beta < 1$, any $(s,t,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ satisfies*

$$|\mathscr{C}| \lesssim \frac{\binom{2^L}{M-s}}{\binom{M}{s,t}\binom{L}{\epsilon}^t},$$

when $M \to \infty$ with $M = 2^{\beta L}$. The redundancy is at least

$$r(\mathscr{C}) \geq sL + t \log M + t\epsilon \log L - \log\left(s!t!\epsilon!^t\right) + o(1),$$

Proof. Denote by $\mathcal{D} \subseteq \mathcal{X}_M^L$, the set of all $\mathcal{S} \in \mathcal{X}_M^L$, which contain an ϵ-substitution-correcting code $\mathcal{Y} \subseteq \mathcal{S}$ of size larger than $|\mathcal{Y}| > M - y(M)$, where we define $y(M) = M/\log M$. The remaining sets are comprised in $\mathcal{D}^{\mathrm{C}} = \mathcal{X}_M^L \setminus \mathcal{D}$. With the partition $\mathcal{D} \cup \mathcal{D}^{\mathrm{C}} = \mathcal{X}_M^L$, it follows that the cardinality of any $(s,t,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ is at most

$$|\mathscr{C}| = |\mathscr{C} \cap \mathcal{D}| + |\mathscr{C} \cap \mathcal{D}^{\mathrm{C}}| \leq \frac{\left|\bigcup_{\mathcal{S} \in \mathcal{D}} B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})\right|}{\min_{\mathcal{S} \in \mathcal{D}} |B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})|} + |\mathcal{D}^{\mathrm{C}}|.$$

The first term follows from a sphere packing bound on all sets $\mathcal{S} \in \mathcal{D}$. The numerator counts the total number of possible channel outputs and the denominator is a lower bound on the error ball size for all sets $\mathcal{S} \in \mathcal{D}$. Since each channel output is a set of sequences of size $M - s - t$ up to $M - s$, we have

$$\left|\bigcup_{\mathcal{S} \in \mathcal{D}} B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})\right| \leq \sum_{i=0}^{t} \binom{2^L}{M-s-i}.$$

From Lemma 1 it is known that

$$\min_{\mathcal{S} \in \mathcal{D}} |B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})| \geq \binom{M - y(M)}{s,t} (B_\epsilon^{\mathbb{S}}(L) - 1)^t,$$

and applying Lemma 2, we find that $|\mathcal{D}^{\mathrm{C}}| \leq D(M - y(M))$. It follows that

$$\begin{aligned} |\mathscr{C}| &\leq \frac{\sum_{i=0}^{t} \binom{2^L}{M-s-i}}{\binom{M-y(M)}{s,t}(B_\epsilon^{\mathbb{S}}(L) - 1)^t} + D(M - y(M)) \\ &= \frac{\sum_{i=0}^{t} \binom{2^L}{M-s-i}}{\binom{M-y(M)}{s,t}(B_\epsilon^{\mathbb{S}}(L) - 1)^t}(1 + \Delta), \end{aligned}$$

where Δ accounts for $D(M - y(M))$ and is defined implicitly as in the following equation. We will show that for our choice of $y(M)$, the first summand dominates the bound, i.e. $\Delta \to 0$ for $M \to \infty$. We obtain

$$\begin{aligned} \log \Delta &= \log \frac{D(M - y(M))\binom{M-y(M)}{s,t}(B_\epsilon^{\mathbb{S}}(L) - 1)^t}{\sum_{i=0}^{t} \binom{2^L}{M-s-i}} \\ &\overset{(a)}{\leq} \log \frac{\binom{2^L}{M-y(M)}\binom{(M-y(M))B_{2\epsilon}^{\mathbb{S}}(L)}{y(M)}}{\binom{2^L}{M-s}} + O(L) \\ &\overset{(b)}{\leq} -\frac{1-\beta}{\beta} M + o(M), \end{aligned}$$

where for inequality (a) we used $|V_\epsilon^{\mathbb{S}}(\mathbf{x})| = B_{2\epsilon}^{\mathbb{S}}(L)$ for all $\mathbf{x} \in \Sigma_2^L$, $\log\binom{M-y(M)}{s,t} = O(L)$ and $t\log(B_\epsilon^{\mathbb{S}}(L)-1) = O(\log L)$. Inequality (b) follows from an application of Lemma 14 with $z(L) = 2^L/((M-y(M))B_{2\epsilon}^{\mathbb{S}}(L))$. Therefore, $\Delta \to 0$, as $M \to \infty$ and $D(M-y(M))$ is asymptotically negligible. We obtain for any $(s,t,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$

$$|\mathscr{C}| \lesssim \frac{\sum_{i=0}^{t}\binom{2^L}{M-s-i}}{\binom{M-y(M)}{s,t}(B_\epsilon^{\mathbb{S}}(L)-1)^t} \sim \frac{\binom{2^L}{M-s}}{\binom{M}{s,t}\binom{L}{\epsilon}^t}.$$

The redundancy is asymptotically at least

$$\begin{aligned} r(\mathscr{C}) &= \log\frac{\binom{2^L}{M}}{|\mathscr{C}|} \geq \log\frac{\binom{2^L}{M}\binom{M}{s}\binom{M-s}{t}\binom{L}{\epsilon}^t}{\binom{2^L}{M-s}} + o(1) \\ &\geq s\log(2^L-M) + t\log(ML^\epsilon) - \log(s!t!\epsilon!^t) + o(1) \\ &= sL + t\log M + t\epsilon\log L - \log\left(s!t!\epsilon!^t\right) + o(1), \end{aligned}$$

where we used that $\log\binom{M}{s} = s\log M - \log s! + o(1)$, $\log\binom{M-s}{t} = t\log M - \log t! + o(1)$ and $t\log\binom{L}{\epsilon} = t(\epsilon\log L - \log\epsilon!) + o(1)$. □

In particular, for $s=0$ and $\epsilon=1$, the redundancy of any $(0,t,1)_{\mathbb{S}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ is at least $t\log(ML) - \log t!$ bits. Note that this coincides with the results from [18] for $t=1$. Comparing the bound on the redundancy stated in Theorem 7 with the well known sphere packing bound for conventional ϵ-substitution-correcting block codes, $\log B_\epsilon^{\mathbb{S}}(L)$, yields an interesting interpretation of the $(0,t,1)_{\mathbb{S}}$ channel. While it seems intuitive that the redundancy required is at least $t\log(ML) - \log t!$ bits, since there are t errors inside a total of ML symbols, it is interesting that from a sphere packing point of view, the fact the sequences are not ordered does appear to require as much redundancy as not knowing the distribution of the errors in an ordered array. While Theorem 7 is formulated for a fixed number of errors s,t, we will find a bound for the case, when number of erroneous sequences t is scaling with M in the following.

Theorem 8. *For fixed $s,\epsilon \in \mathbb{N}_0$ and fixed $0<\beta<1$, any $(s,M-s,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ satisfies*

$$r(\mathscr{C}) \geq M\epsilon\log L + O(M),$$

when $M \to \infty$ with $M = 2^{\beta L}$.

Proof. We follow a similar outline as in the proof for Theorem 7. Denote by $\mathcal{D} \subseteq \mathcal{X}_M^L$, the set of all $\mathcal{S} \in \mathcal{X}_M^L$, which contain an ϵ-substitution-correcting code $\mathcal{Y} \subseteq \mathcal{S}$ of size $|\mathcal{Y}| > M - y(M)$, where we define $y(M) = M/\log\log M$ and $\mathcal{D}^{\mathrm{C}} = \mathcal{X}_M^L \backslash \mathcal{D}$. Allowing only $t = M-s-y(M)$ erroneous sequences, we can apply Lemma 1 and obtain

$$|B_{s,t,\epsilon}^{\mathbb{S}}(\mathcal{S})| \geq \binom{M-y(M)}{s}(B_\epsilon^{\mathbb{S}}(L)-1)^{M-y(M)-s},$$

for all $\mathcal{S} \in \mathcal{D}$. It follows that

$$|\mathscr{C}| \leq \frac{\sum_{i=0}^{M-y(M)-s}\binom{2^L}{M-s-i}}{\binom{M-y(M)}{s}(B_\epsilon^{\mathbb{S}}(L)-1)^{M-y(M)-s}}(1+\Delta).$$

We will show that $\Delta \to 0$ for $M \to \infty$. We obtain

$$\begin{aligned} \log\Delta &= \log\frac{\binom{M-y(M)}{s}(B_\epsilon^{\mathbb{S}}(L)-1)^{M-y(M)-s}D(M-y(M))}{\sum_{i=0}^{M-y(M)-s}\binom{2^L}{M-s-i}} \\ &\overset{(a)}{\leq} \log\frac{\binom{2^L}{M-y(M)}\binom{(M-y(M))B_{2\epsilon}^{\mathbb{S}}(L)}{y(M)}}{\binom{2^L}{M-s}} + M\epsilon\log L + O(L) \\ &\overset{(b)}{\leq} -\frac{ML}{\log\log M} + M\epsilon\log L + o\left(\frac{M}{\log\log M}\right) \\ &= -\frac{ML}{\log(\beta L)} + O(M\log L) \end{aligned}$$

where in inequality (a) we used $\log\binom{M-y(M)}{s} = O(L)$. For inequality (b) we applied Lemma 14 with $z(L) = 2^L/((M-y(M))B_{2\epsilon}^{\mathbb{S}}(L))$. Therefore, $\Delta \to 0$, as $M \to \infty$. We obtain for any $(s,M-s,\epsilon)_{\mathbb{S}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$

$$|\mathscr{C}| \lesssim \frac{\sum_{i=0}^{M-y(M)-s}\binom{2^L}{M-s-i}}{\binom{M-y(M)}{s}(B_\epsilon^{\mathbb{S}}(L)-1)^{M-y(M)-s}} \lesssim \frac{\binom{2^L}{M-s}}{\binom{M}{s}\binom{L}{\epsilon}^{M-y(M)-s}}.$$

Therefore, the redundancy satisfies

$$\begin{aligned} r(\mathscr{C}) &= \log\frac{\binom{2^L}{M}}{|\mathscr{C}|} \geq \log\frac{\binom{2^L}{M}\binom{M}{s}\binom{L}{\epsilon}^{M-y(M)-s}}{\binom{2^L}{M-s}} + o(1) \\ &\geq sL + (M-y(M)-s)\epsilon\log(L/\epsilon) - \log s! + o(1). \end{aligned}$$

□

C. Asymptotic Bounds for Deletion Errors

We will now turn to derive an asymptotic bound on the cardinality of $(s,t,\epsilon)_{\mathbb{D}}$-correcting codes. Note that it is possible to use the technique that we present here also for insertion errors, however this is deferred to future work. Since the deletion ball is non-uniform, it is not directly possible to use an analogue of Lemma 1 as in Theorem 7. We will therefore slightly adapt our arguments and use the fact that, although the deletion ball size is non-uniform, most of the deletion balls have a similar size. It has been shown in [26] that

$$|S_\epsilon^{\mathbb{D}}(\mathbf{x})| \geq \binom{||\mathbf{x}|| - \epsilon + 1}{\epsilon}$$

and most words $\mathbf{x} \in \Sigma_2^n$ have roughly $L/2$ runs. We will elaborate this result in the following.

Lemma 3. *Let $\rho \in \mathbb{N}$. The number of words with less than $L/2-\rho$ runs satisfies*

$$\left|\left\{\mathbf{x} \in \Sigma_2^L : ||\mathbf{x}|| < \frac{L}{2} - \rho\right\}\right| \leq \frac{2^L}{\mathrm{e}^{\frac{2\rho^2}{L}}}.$$

Proof. The number of words $\mathbf{x} \in \Sigma_2^L$ with exactly i runs, i.e., $||\mathbf{x}|| = i$ is given by $2\binom{L-1}{i-1}$. Therefore, the number of words with less than $L/2-\rho$ runs is given by

$$\begin{aligned} |\{\mathbf{x} \in \Sigma_2^L : ||\mathbf{x}|| < L/2-\rho\}| &= 2\sum_{i=1}^{L/2-\rho-1}\binom{L-1}{i-1} \\ &\overset{(a)}{\leq} \sum_{i=1}^{L/2-\rho}\binom{L}{i} \overset{(b)}{\leq} \frac{2^L}{\mathrm{e}^{\frac{2\rho^2}{L}}}, \end{aligned}$$

where we used $\binom{L-1}{i-1} \leq \frac{1}{2}\binom{L}{i}$ for $i \leq \frac{L}{2}$ in inequality (a) and Hoeffding's inequality [31] on the binomial sum in (b). □

Next, we find a lower bound on the ball size $B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})$, for sets, which contain a deletion-correcting code.

Lemma 4. *Let $\mathcal{Y} \subseteq \mathcal{S} \in \mathcal{X}^L_M$ be an ϵ-deletion-correcting code, i.e. $B^{\mathbb{D}}_{\epsilon}(\mathbf{y}_1) \cap B^{\mathbb{D}}_{\epsilon}(\mathbf{y}_2) = \emptyset$ for all $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ and $\mathbf{y}_1 \neq \mathbf{y}_2$. Further, let $s + t \leq |\mathcal{Y}|$. Then,*

$$|B^{\mathbb{S}}_{s,t,\epsilon}(\mathcal{S})| \geq \sum_{\substack{\mathcal{E},\mathcal{L}\subseteq\mathcal{Y},\mathcal{E}\cap\mathcal{L}=\emptyset,\\ |\mathcal{L}|=s,|\mathcal{E}|=t}} \prod_{\mathbf{y}\in\mathcal{E}} |S^{\mathbb{D}}_{\epsilon}(\mathbf{y})|,$$

Proof. We will find a lower bound on the number of words inside the error ball $|B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})|$ by counting distinct elements $S' \in B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})$ in the following way. Choose two arbitrary distinct sets $\mathcal{L}, \mathcal{E} = \{\mathbf{y}_{e_1}, \ldots, \mathbf{y}_{e_t}\} \subseteq \mathcal{Y}$ with $|\mathcal{L}| = s$ and $|\mathcal{E}| = t$ and choose a set of erroneous outcomes $\mathcal{E}' = \{\mathbf{y}'_1, \ldots, \mathbf{y}'_t\}$, where $\mathbf{y}'_i \in S^{\mathbb{D}}_{\epsilon}(\mathbf{y}_{e_i})$. Note that we delete exactly ϵ symbols from each $\mathbf{y}_{e_i}$ and thus $\mathbf{y}'_i \in \Sigma_2^{L-\epsilon}$. Denote by $\mathcal{L}, \mathcal{E}, \mathcal{E}'$ and $\tilde{\mathcal{L}}, \tilde{\mathcal{E}}, \tilde{\mathcal{E}}'$ two different choices of error realizations and let $\mathcal{S}'$ and $\tilde{\mathcal{S}}'$ be the corresponding received sets. If $\mathcal{L} \cup \mathcal{E} \neq \tilde{\mathcal{L}} \cup \tilde{\mathcal{E}}$, then $\mathcal{S}' \neq \tilde{\mathcal{S}}'$, as the resulting error-free sequences in $\mathcal{S}'$ and $\tilde{\mathcal{S}}$ of length L are different. In the case $\mathcal{L} \cup \mathcal{E} = \tilde{\mathcal{L}} \cup \tilde{\mathcal{E}}$ and $\mathcal{E} \neq \tilde{\mathcal{E}}$, it follows that $\mathcal{E}' \neq \tilde{\mathcal{E}}'$, as the erroneous outcomes are chosen out of the ϵ deletion spheres from an ϵ-deletion-correcting code. Finally, if $\mathcal{L} \cup \mathcal{E} = \tilde{\mathcal{L}} \cup \tilde{\mathcal{E}}$ and $\mathcal{E} = \tilde{\mathcal{E}}$ it follows that $\mathcal{L} = \tilde{\mathcal{L}}$ and thus $\mathcal{E}' \neq \tilde{\mathcal{E}}'$ as we chose $\mathcal{L}, \mathcal{E}, \mathcal{E}'$ and $\tilde{\mathcal{L}}, \tilde{\mathcal{E}}, \tilde{\mathcal{E}}'$ to be different. Hence, for each choice of error realizations $\mathcal{L}, \mathcal{E}, \mathcal{E}'$, we obtain a unique element in $B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})$. Counting the number of choices yields the lemma. □

This allows to formulate the following theorem.

Theorem 9. *For fixed $s, t, \epsilon \in \mathbb{N}_0$ and $0 < \beta < 1$, any $(s, t, \epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}^L_M$ satisfies*

$$|\mathscr{C}| \lesssim \frac{\binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t}}{\binom{M}{s,t}\binom{L/2}{\epsilon}^t}$$

when $M \to \infty$ with $M = 2^{\beta L}$. The redundancy is at least

$$r(\mathscr{C}) \geq sL + t\epsilon \log L - \log(s!\epsilon!^t) + o(1).$$

Proof. Denote by $\mathcal{D}_{\mathrm{r}} \subseteq \mathcal{X}^L_M$, the set of all $\mathcal{S} \in \mathcal{X}^L_M$, which contain more than $M - y(M)$ sequences with $||\mathbf{x}|| \geq L/2 - \rho(L)$, where we choose $y(M) = M/\log M$ and $\rho(L) = \sqrt{L \ln L}$. Further, let $\mathcal{D}_{\mathrm{e}} \subseteq \mathcal{X}^L_M$ be all sets $\mathcal{S} \in \mathcal{X}^L_M$ that contain an ϵ-deletion-correcting code $\mathcal{Y} \subseteq \mathcal{S}$ of size $|\mathcal{Y}| > M - y(M)$ and let $\mathcal{D} = \mathcal{D}_{\mathrm{r}} \cap \mathcal{D}_{\mathrm{e}}$. The remaining sets are comprised in $\mathcal{D}^{\mathrm{C}} = \mathcal{X}^L_M \setminus \mathcal{D}$. Since $\mathcal{D}$ and $\mathcal{D}^{\mathrm{C}}$ are a partition of $\mathcal{X}^L_M$, every $(s, t, \epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}^L_M$ satisfies

$$|\mathscr{C}| = |\mathscr{C} \cap \mathcal{D}| + |\mathscr{C} \cap \mathcal{D}^{\mathrm{C}}| \leq \frac{\left|\bigcup_{\mathcal{S}\in\mathcal{D}} B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})\right|}{\min_{\mathcal{S}\in\mathcal{D}} |B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})|} + |\mathcal{D}^{\mathrm{C}}|.$$

The number of received sets after a loss of exactly s sequences and t sequences with exactly ϵ deletions each is at most

$$\left|\bigcup_{\mathcal{S}\in\mathcal{D}} B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})\right| \leq \binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t},$$

as each received set consists of $M - s - t$ error-free sequences and t sequences of length $L - \epsilon$. Each $\mathcal{S} \in \mathcal{D}$ contains less than $y(M)$ sequences, which do not belong to the ϵ-deletion-correcting code $\mathcal{Y}$ and less than $y(M)$ (possibly different) sequences with $||\mathbf{x}|| < L/2 - \rho(L)$. Thus, at least $M - 2y(M)$ sequences form an ϵ-deletion-correcting code and satisfy $||\mathbf{x}|| \geq L/2 - \rho(L)$ and by Lemma 4, we have

$$|B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})| \geq \binom{M-2y(M)}{s,t}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^t$$

for each $\mathcal{S} \in \mathcal{D}$. The number of remaining sets $\mathcal{S} \notin \mathcal{D}$ satisfies $|\mathcal{D}^{\mathrm{C}}| = |\mathcal{X}^L_M \setminus \mathcal{D}| \leq |\mathcal{X}^L_M \setminus \mathcal{D}_{\mathrm{r}}| + |\mathcal{X}^L_M \setminus \mathcal{D}_{\mathrm{e}}|$. Each such set contains at least $y(M)$ sequences with $||\mathbf{x}|| < L/2 - \rho(L)$ or does not contain an ϵ-deletion-correcting code of size more than $M - y(M)$. By Lemma 3, we have that

$$|\mathcal{X}^L_M \setminus \mathcal{D}_{\mathrm{r}}| \leq \binom{2^L}{M-y(M)}\binom{2^L/L^2}{y(M)},$$

for large enough L, as each $\mathcal{S} \in \mathcal{X}^L_M \setminus \mathcal{D}_{\mathrm{r}}$ can be constructed by choosing $y(M)$ sequences to have less than $L/2 - \rho(L)$ runs and the remaining sequences are chosen arbitrarily. Further, using Lemma 2, it follows that

$$|\mathcal{X}^L_M \setminus \mathcal{D}_{\mathrm{e}}| \leq \binom{2^L}{M-y(M)}\binom{KV^{\mathbb{D}}_{\epsilon}}{y(M)},$$

where $V^{\mathbb{D}}_{\epsilon} = \max_{\mathbf{x}\in\Sigma_2^L} |\{\mathbf{y} \in \Sigma_2^L : B^{\mathbb{D}}_{\epsilon}(\mathbf{x}) \cap B^{\mathbb{D}}_{\epsilon}(\mathbf{y}) \neq \emptyset\}|$. This number can be bounded from above by the following consideration. Given $\mathbf{x} \in \Sigma_2^L$, each $\mathbf{y} \in \Sigma_2^L$ can be constructed by first deleting ϵ symbols from $\mathbf{x}$ and then inserting ϵ arbitrary symbols to the result. Using $|S^{\mathbb{D}}_{\epsilon}(\mathbf{x})| \leq \binom{L}{\epsilon}$ for all $\mathbf{x} \in \Sigma_2^L$ and $|S^{\mathbb{I}}_{\epsilon}(\mathbf{x}')| = \sum_{i=0}^{\epsilon}\binom{L}{i} = B^{\mathbb{S}}_{\epsilon}$ for all $\mathbf{x}' \in \Sigma_2^{L-\epsilon}$ yields $V^{\mathbb{D}}_{\epsilon} \leq \binom{L}{\epsilon}B^{\mathbb{S}}_{\epsilon}$. It follows that the size of any $(s, t, \epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}^L_M$ is at most

$$\begin{aligned}|\mathscr{C}| \leq& \frac{\binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t}}{\binom{M-2y(M)}{s,t}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^t} \\ &+ \binom{2^L}{M-y(M)}\left(\binom{2^L/L^2}{y(M)} + \binom{K\binom{L}{\epsilon}B^{\mathbb{S}}_{\epsilon}}{y(M)}\right) \\ =& \frac{\binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t}}{\binom{M-2y(M)}{s,t}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^t}(1 + \Delta_{\mathrm{r}} + \Delta_{\mathrm{e}}).\end{aligned}$$

We will show now that $\Delta_{\mathrm{r}} \to 0$ and $\Delta_{\mathrm{e}} \to 0$ for $M \to \infty$.

$$\begin{aligned}\log \Delta_{\mathrm{r}} &= \log \frac{\binom{M}{s,t}\binom{\frac{L}{\epsilon}}{\epsilon}^t\binom{2^L/L^2}{y(M)}\binom{2^L}{M-y(M)}}{\binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t}} + o(1) \\ &= \log \frac{\binom{2^L/L^2}{y(M)}\binom{2^L}{M-y(M)}}{\binom{2^L}{M-s-t}} + O(L) \\ &\overset{(a)}{\leq} -\frac{M}{\log M}\log\log M + O\left(\frac{M}{\log M}\right),\end{aligned}$$

where we applied Lemma 14 in inequality (a). Hence, $\Delta_{\mathrm{r}} \to 0$ for $M \to \infty$. Analogous to the proof of Theorem 7, it can be shown that $\Delta_{\mathrm{e}} \to 0$ for $M \to \infty$. We obtain for the maximum size of a $(s,t,\epsilon)_{\mathbb{D}}$-correcting code

$$|\mathscr{C}| \lesssim \frac{\binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t}}{\binom{M}{s,t}\binom{L/2}{\epsilon}^t}.$$

The redundancy is consequently at least

$$\begin{aligned} r(\mathscr{C}) &= \log\binom{2^L}{M} - \log|\mathscr{C}| \geq \log\frac{\binom{2^L}{M}\binom{M}{s,t}\binom{L/2}{\epsilon}^t}{\binom{2^L}{M-s-t}\binom{2^{L-\epsilon}}{t}} + o(1) \\ &= sL + t\epsilon\log L - \log(s!\epsilon!^t) + o(1). \end{aligned}$$

□

The result of Theorem 9 is particularly interesting, when comparing with Theorem 7, which depicts the case of substitution errors inside the sequences. It can be seen that correcting substitutions requires $t\log M - \log t!$ more bits of redundancy as compared to insertion or deletion errors only. While this seems surprising, there is a practical reason for this phenomena. For the case of insertion or deletion errors, it is directly possible to identify erroneous sequences, by checking their length to be different from L. This is not possible for substitution errors, and erroneous sequences can be confused with correct sequences, which means that additional redundancy is required for detecting the erroneous sequences. In fact, we will show in Construction 5, how to constructively exploit the identification of erroneous sequences for the case of $(0,1,1)_{\mathbb{D}}$ deletion errors and obtain a code that asymptotically achieves the bound from Theorem 9. In the following we derive a sphere packing bound for the case when the number of erroneous sequences scales with M.

Theorem 10. *For fixed $s,\epsilon \in \mathbb{N}_0$ and fixed $0<\beta<1$, any $(s, M-s, \epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ satisfies*

$$r(\mathscr{C}) \geq M\epsilon\log L + O(M),$$

when $M \to \infty$ with $M = 2^{\beta L}$.

Proof. The proof is similar to that of Theorem 9 and we use the same notation for $\mathcal{D} = \mathcal{D}_{\mathrm{r}} \cap \mathcal{D}_{\mathrm{e}}$ for sets that contain an ϵ-deletion-correcting code of size $|\mathcal{Y}| > M - y(M)$ and more than $M - y(M)$ sequences with at least $||\mathbf{x}|| \geq L/2 - \rho(L)$ runs, where $y(M) = M/\log\log M$ and $\rho(L) = \sqrt{L/2\ln L\log^2 L}$. With Lemma 4, it follows

$$|B^{\mathbb{D}}_{s,t,\epsilon}(\mathcal{S})| \geq \binom{M-2y(M)}{s}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^{M-2y(M)-s}$$

for all $\mathcal{S} \in \mathcal{D}$. It follows that the size of any $(s,t,\epsilon)_{\mathbb{D}}$-correcting code $\mathscr{C} \subseteq \mathcal{X}_M^L$ is at most

$$|\mathscr{C}| \leq \frac{\binom{2^L}{2y(M)}\binom{2^{L-\epsilon}}{M-2y(M)-s}}{\binom{M-2y(M)}{s}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^{M-y(M)-s}}(1+\Delta_{\mathrm{r}}+\Delta_{\mathrm{e}}).$$

We will show now that $\Delta_{\mathrm{r}} \to 0$ and $\Delta_{\mathrm{e}} \to 0$ for $M \to \infty$.

$$\begin{aligned} \log\Delta_{\mathrm{r}} &\leq \frac{\binom{L/2-\rho(L)-\epsilon}{\epsilon}^M\binom{2^L/L^{\log^2 L}}{y(M)}\binom{2^L}{M-y(M)}}{\binom{2^L}{2y(M)}\binom{2^{L-\epsilon}}{M-2y(M)-s}} + O(L) \\ &\leq \log\frac{\binom{2^L/L^{\log^2 L}}{y(M)}\binom{2^L}{M-y(M)}}{\binom{2^L}{2y(M)}\binom{2^{L-\epsilon}}{M-2y(M)-s}} + M\epsilon\log L + O(L) \\ &\overset{(a)}{\leq} \log\frac{\binom{2^L/L^{\log^2 L}}{y(M)}\binom{2^L}{M-y(M)}}{\binom{2^L}{M-s}} + M\epsilon\log L + O(M) \\ &\overset{(b)}{\leq} -\frac{M\log^3 L}{\log(\beta L)} + O(M\log L) \end{aligned}$$

where for inequality (a) we used that

$$\log\frac{\binom{2^L}{M-s}}{\binom{2^L}{2y(M)}\binom{2^{L-\epsilon}}{M-2y(M)-s}} \leq O(M)$$

and applied Lemma 14 in inequality (b). Hence, $\Delta_{\mathrm{r}} \to 0$ for $M \to \infty$. Analogous to the proof of Theorem 8, it can be shown that $\Delta_{\mathrm{e}} \to 0$ for $M \to \infty$. We obtain for the maximum size of a $(s, M-s, \epsilon)_{\mathbb{D}}$-correcting code

$$|\mathscr{C}| \lesssim \frac{\binom{2^L}{2y(M)}\binom{2^{L-\epsilon}}{M-2y(M)-s}}{\binom{M}{s}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^{M-y(M)-s}}.$$

The redundancy is consequently at least

$$\begin{aligned} r(\mathscr{C}) &\geq \log\frac{\binom{2^L}{M}\binom{M}{s}\binom{L/2-\rho(L)-\epsilon}{\epsilon}^{M-y(M)-s}}{\binom{2^L}{2y(M)}\binom{2^{L-\epsilon}}{M-2y(M)-s}} + o(1) \\ &\geq M\epsilon\log L + O(M). \end{aligned}$$

□

V. Code Constructions

Having available suitable bounds on the redundancy of $(s,t,\epsilon)_{\mathbb{T}}$-correcting codes, we now present several code constructions for DNA storage systems that are suitable for different types of errors $\mathbb{T}$ and choices of parameters s, t and ϵ. Note that constructing codes for the considered channel model is surprisingly challenging. This can be explained as follows. In order to find efficient codes, one has to deal with both, the fact that sequences are received in an unordered fashion and also that sequences are distorted by random errors. Especially for the case when only few sequences contain errors, concatenated schemes, where each strand is protected by an inner code and an outer code is used over all strands, are suboptimal. This is because the inner code is "wasted" for the correct sequences, as they do not contain any errors and no inner code would be needed to obtain the correct strand. Therefore, it is important to design codes over the whole set of strands that use redundancy over different sequences and allow to correct errors in a subset of strands. We start with constructions that are suitable for an arbitrary number of errors per sequence and will elaborate more specialized constructions towards the end of this section.

A. Indexing Sequences

A common efficient way to combat the loss of ordering of sequences is to prepend an index to each sequence, which contains the position i of the sequence. This approach has been discussed in different settings, e.g. [17], [20]. The set of all possible sets of sequences with indexing is given by

$$\mathscr{C}_{\mathrm{I}}(M,L)=\{\mathcal{S}\in\mathcal{X}_M^L:\mathbf{x}_i=(\mathbf{I}(i),\mathbf{u}_i),i\ \in\{1,2,\ldots,M\},$$

where $\mathbf{I}(i)\in\Sigma_2^{\lceil\log M\rceil}$ denotes the binary representation of $i-1$ and $\mathbf{u}_i\in\Sigma_2^{L-\lceil\log M\rceil}$ are arbitrary information vectors. Note that by this definition the prefix of an indexed sequence contains the index of a sequence $\mathrm{pref}_{\lceil\log M\rceil}(\mathbf{x}_i)=\mathbf{I}(i)$.

This requires an index $\mathbf{I}(i)$ of $\lceil\log M\rceil$ bits in each sequence so the maximum number of information bits that can be stored this way is $M(L-\lceil\log M\rceil)$ without any error correction. While this solution is attractive for its simplicity, it introduces already a redundancy, which increases linearly in M, which is stated in the following theorem.

Theorem 11. *For fixed $0<\beta<1$, the redundancy required for indexing sequences is given by*

$$r(\mathscr{C}_{\mathrm{I}}(M,L))=M(\lceil\log M\rceil-\log M+\log\mathrm{e})+o(M),$$

when $M\to\infty$ with $M=2^{\beta L}$.

Proof. From $M=2^{\beta L}$ with $0<\beta<1$, we have that $M=o(2^L)$ and $M=\omega(1)$, when $M\to\infty$. Therefore,

$$\begin{aligned}r(\mathscr{C}_{\mathrm{I}}(M,L))&=\log\binom{2^L}{M}-M(L-\lceil\log M\rceil)\\&=M(\lceil\log M\rceil-\log M+\log\mathrm{e})+o(M),\end{aligned}$$

where we used Lemma 13, which is derived in Appendix A, to characterize the binomial coefficient. □

This means that every construction which uses indexing already incurs a redundancy of at best roughly $M\log\mathrm{e}$ bits. Note that this amount can be significant, as the number of sequences M is significantly larger than their length L, as explained in Remark 1. However, in terms of code rate, it has been shown in [20] that for the case of no errors inside the sequences, the indexing approach is capacity achieving.

The following function, which collects all indices of a set of sequences will be useful for our constructions that are based on indexing.

Definition 8. *For any set $\mathcal{A}\subseteq\mathcal{X}_M^L$ we define*

$$\mathcal{I}(\mathcal{A})=\bigcup_{\mathbf{x}\in\mathcal{A}}\{\mathrm{pref}_{\lceil\log M\rceil}(\mathbf{x})\}$$

to be the set of indices of the sequences in $\mathcal{A}$.

Note that it is possible that $|\mathcal{I}(\mathcal{A})|<|\mathcal{A}|$, if one (or more) of the indices appear multiple times because of errors.

B. An Index-Based Construction Using MDS Codes

The following construction is based on adding an index in front of all sequences $\mathbf{x}_i$ and using an MDS code over the M sequences for error correction. For all n and k, where $k\leq n$ we denote by $\mathsf{MDS}[n,k]$ an MDS code over any field of size at least $n-1$.

In Construction 1, the sequences $\mathbf{x}_i=(\mathbf{I}(i),\mathbf{u}_i)$ of each codeword set are constructed by writing a binary representation of the index, $\mathbf{I}(i)$, of length $\lceil\log M\rceil$ in the first part of each sequence. Then, the remaining part $\mathbf{u}_i$ is viewed as a symbol over the extension field $\mathbb{F}_{2^{L-\lceil\log M\rceil}}$, and $(\mathbf{u}_1,\ldots,\mathbf{u}_M)$ will form a codeword in some MDS code.[2] A similar construction has been used in [20], where index-based constructions are analyzed for the correction of only a loss of sequences.

Construction 1. *For all M,L, and a positive integer δ, let $\mathscr{C}_1(M,L,\delta)$ be the code defined by*

$$\begin{aligned}\mathscr{C}_1(M,L,\delta)=\{\mathcal{S}\in\mathcal{X}_M^L:&\mathbf{x}_i=(\mathbf{I}(i),\mathbf{u}_i),\\&(\mathbf{u}_1,\ldots,\mathbf{u}_M)\in\mathsf{MDS}[M,M-\delta]\}.\end{aligned}$$

This code provides a direct construction to correct a loss of sequences and erroneous sequences with an arbitrary amount of errors each. The error correction capability for several types of errors is summarized in the following lemma.

Lemma 5. *For all M,L,δ, the code $\mathscr{C}_1(M,L,\delta)$ is*

- *$(s,t,\bullet)_{\mathbb{L}}$-correcting for all $s+2t\leq\delta$,*
- *$(s,t,\bullet)_{\mathbb{I}}$-correcting for all $s+t\leq\delta$,*
- *$(s,t,\bullet)_{\mathbb{D}}$-correcting for all $s+t\leq\delta$.*

Proof. Denote by $\mathcal{S}'$ the received set after a loss of sequences and errors. We start with proving the lemma for the case of arbitrary edit errors. According to Definition 2, we write $\mathcal{C},\mathcal{L},\mathcal{E}$ as the sets of error-free, lost, and erroneous sequences, and $\mathcal{E}'$ are the erroneous outcomes of the sequences in $\mathcal{E}$. First, we observe that if we can recover the MDS codeword $\mathbf{U}=(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_M)$, we can also recover $\mathcal{S}$ by prepending the index $\mathbf{I}(i)$ in front of each $\mathbf{u}_i$. Given $\mathcal{S}'$, we then create the received estimate word $\mathbf{U}'$ by declaring all positions i with

$$|\{\mathbf{x}'\in\mathcal{S}':\mathrm{pref}_{\lceil\log M\rceil}(\mathbf{x}')=\mathbf{I}(i)\}|\neq 1,$$

i.e., for which there is not exactly one index in $\mathcal{S}'$, as erasures. The remaining positions in $\mathbf{U}'$ are filled with the corresponding symbols $\mathbf{u}_i'$. We will show that the number of erasures s' and the number of errors t' in $\mathbf{U}'$ satisfy $s'+2t'\leq\delta$ by the following consideration. Consider a genie, which first only adds the error-free sequences $\mathcal{C}$ to $\mathbf{U}'$. Since $|\mathcal{C}|\geq M-s-t$ at least $M-s-t$ positions in $\mathbf{U}'$ have been filled with correct symbols, and thus there remain at most $s'\leq s+t$ erasures and $t'=0$ errors inside $\mathbf{U}'$ up to this point. Now, the genie successively adds the t erroneous sequences $\mathcal{E}'$ to $\mathbf{U}'$. Each of the t erroneous sequences $\mathbf{x}'\in\mathcal{E}'$ can have an arbitrary index $\mathrm{pref}_{\lceil\log M\rceil}(\mathbf{x}')$ from 0 to $M-1$. If the erroneous sequence $\mathbf{x}'$ has an index of a position which is not occupied yet, t' increases by one and s' decreases by one, as this position is not declared as an erasure anymore and contains now an erroneous symbol. If the erroneous sequence $\mathbf{x}'$ has an index, which is already present, this position is declared as an erasure as explained above. Consequently, s' increases

[2]Note that we assume $M\leq\sqrt{2^L}$ in this section to guarantee the existence of the MDS code [32, ch. 11]. However, the case $M>\sqrt{2^L}$ can always be used by employing non-MDS codes.

by one for this case. Hence, the number of erasures in $\mathbf{U}'$ is bounded from above by $s' \leq s+t-t'+|\mathcal{I}(\mathcal{E}')\cap\mathcal{I}(\mathcal{C})|$ where $|\mathcal{I}(\mathcal{E}')\cap\mathcal{I}(\mathcal{C})|$ accounts for the situation when an erroneous sequence has the same index as an error-free one. The number of errors is at most $t' \leq |\mathcal{I}(\mathcal{E}')\cap(\mathcal{I}(\mathcal{E})\cup\mathcal{I}(\mathcal{L}))|$. Hence, $s'+2t' \leq s+t+t'+|\mathcal{I}(\mathcal{E}')\cap\mathcal{I}(\mathcal{C})| \leq s+2t \leq \delta$, which proves the error correcting capability.

For the case of only insertion ($\mathbb{I}$) and only deletion ($\mathbb{D}$) errors, it is possible to identify the erroneous sequences by checking their length to be larger (respectively smaller) than L. If these sequences are discarded, there are in total $s+t$ erasures inside the MDS codeword, which can be corrected, if $s+t \leq \delta$. □

Note that for the practically important case of a loss of sequences and combinations of substitution and deletion errors, $\mathscr{C}_1(M,L,\delta)$ can correct all errors, if $s+2t_{\mathbb{S}}+t_{\mathbb{D}} \leq \delta$, where $t_{\mathbb{S}}$ is the number of sequences suffering from substitution errors only and $t_{\mathbb{D}}$ is the number of sequences with deletion errors. The same also holds for combinations of substitution and insertion errors. However, this is not true for combinations of substitutions, insertions *and* deletions as a sequence that contains insertions and deletions might have length exactly L and therefore cannot be erased. In this case, as elaborated in the proof, $s+2t \leq \delta$ has to hold. More generally, erroneous sequences which have length exactly L require 2 redundancy symbols inside the MDS codeword to be correctable, while sequences which have a different length only require a single symbol, as they can be detected as erroneous.

The redundancy of Construction 1 is stated in the following theorem.

Theorem 12. *For all M, L, δ, the redundancy of the code $\mathscr{C}_1(M,L,\delta)$ is*

$$r(\mathscr{C}_1(M,L,\delta)) = r(\mathscr{C}_1(M,L)) + \delta(L - \lceil \log M \rceil).$$

Proof. First, indexing the sequences requires a redundancy of $r(\mathscr{C}_1(M,L))$, which is derived in Theorem 11. Second, the MDS code has δ redundant symbols and thus there are $\delta(L-\lceil\log M\rceil)$ additional redundancy bits. □

While the redundancy of Construction 1 can be very large, especially for the case $M \gg L$, it provides some very useful features. First, it is possible to efficiently encode and decode this code using standard encoders and decoders for MDS codes. Second, it is not necessary to design the code for a specific number of errors s and t, but rather their sum $s+2t$, which allows for a flexible decoding procedure.

C. A Construction Based on Constant Weight Codes

Imposing an ordering (e.g., lexicographic) onto the sequences in Σ_2^L, every data set $S \in \mathcal{X}_M^L$ can be represented by a binary vector $\mathbf{v}(\mathcal{S})$ of length 2^L, where each non-zero entry in $\mathbf{v}(\mathcal{S})$ indicates that a specific sequence is contained in the set $\mathcal{S}$.

The set of possible data sets can therefore be represented[3] by constant-weight binary vectors of length 2^L

$$\mathcal{V}_M^L = \{\mathbf{v} \in \{0,1\}^{2^L} : \mathrm{wt}(\mathbf{v}) = M\},$$

where $\mathrm{wt}(\mathbf{v})$ denotes the *Hamming weight* of $\mathbf{v}$, i.e., the number of non-zero entries inside the vector $\mathbf{v}$. That is, the mapping $\mathbf{v}$ defines an isomorphism between $\mathcal{X}_M^L$ and $\mathcal{V}_M^L$ and thus $\mathbf{v}^{-1}$ is well-defined. Using this representation, a loss of a sequence $\mathbf{x} \in \mathcal{S}$ corresponds to an asymmetric $1 \to 0$ error inside $\mathbf{v}(\mathcal{S})$ at the position corresponding to $\mathbf{x}$. Substitution errors inside a sequence $\mathbf{x} \in \mathcal{S}$ translate to single errors in the Johnson graph in $\mathbf{v}(\mathcal{S})$, i.e. a single $1 \to 0$ at the corresponding position of the original sequence $\mathbf{x}$, and a single $0 \to 1$ error at the position of its erroneous outcome $\mathbf{x}'$. In case, the erroneous outcome $\mathbf{x}'$ is already present in $\mathcal{S}'$, the $0 \to 1$ error is omitted and there is only a single asymmetric $1 \to 0$ error at the position of the original sequence $\mathbf{x}$, similar to a loss of a sequence. For codes in the Johnson graph, the reader is referred to, e.g., [33]

Example 4. *Consider the following $M=3$ stored sequences $\mathcal{S} = \{(001),(010),(110)\}$, each of length $L=3$. We choose $\mathbf{v}(\mathcal{S})$ to map each sequence $\mathbf{x} \in \mathcal{S}$ to its decimal equivalent by standard base conversion and let $\mathbf{v}(\mathcal{S})$ be non-zero at exactly these indices. Hence, e.g., the sequence (110) is mapped to $1\cdot 2^2 + 1\cdot 2^1 + 0\cdot 2^0 + 1 = 7$ and thus $\mathbf{v}(\mathcal{S})$ will be non-zero at index 7. Note that we additionally add 1, since we index vectors starting by 1. Therefore, $\mathbf{v}(\mathcal{S}) = (01100010)$. Assume now, the set $\mathcal{S}$ is transmitted over a $(1,1,2)_{\mathbb{S}}$ channel, resulting in $\mathcal{S}' = \{(001),(111)\}$, where the sequence (110) was lost and the sequence (010) has been perturbed by two substitution errors. The corresponding binary representation is $\mathbf{v}(\mathcal{S}') = (01000001)$, where there was a single $1 \to 0$ at position 7 due to the loss of the sequence (110) and $1 \to 0$ and $0 \to 1$ errors at positions 3, respectively 8, since the sequence (010) was distorted to the sequence (111).*

With this principle in mind, we define a code that can correct asymmetric errors and errors in the Johnson graph.

Definition 9. *For all M, L and positive integers s, t, we define $\mathscr{C}_M^L(s,t) \subseteq \mathcal{V}_M^L$ to be a code of length 2^L that consists of codewords with constant Hamming weight M, which corrects s asymmetric $1 \to 0$ errors and t errors in the Johnson graph.*

With such a code $\mathscr{C}_M^L(s,t) \subseteq \mathcal{V}_M^L$ in hand that can correct asymmetric errors and errors in the Johnson graph, we can construct a code for the DNA storage channel.

Construction 2. *For all M, L, we define the following code*

$$\mathscr{C}_2(M,L,s,t) = \{\mathcal{S} \in \mathcal{X}_M^L : \mathbf{v}(\mathcal{S}) \in \mathscr{C}_M^L(s,t)\}.$$

By this construction, given a constant-weight code $\mathscr{C}_M^L(s,t)$, we construct the DNA storage code $\mathscr{C}_2(M,L,s,t)$ by mapping each $\mathbf{c} \in \mathscr{C}_M^L(s,t)$ to its corresponding set $\mathcal{S} = \mathbf{v}^{-1}(\mathbf{c})$. Note that this mapping can be efficiently implemented, by, e.g., a decimal to binary mapping of the non-zero positions in $\mathbf{c}$,

[3]This representation has been used as a proof technique in [20].

as illustrated in Example 4. The correctness of the construction is established in the following lemma.

Lemma 6. *For all M, L and positive integers s, t, the code $\mathscr{C}_2(M, L, s, t)$ is an $(s, t, \bullet)_{\mathbb{L}}$-correcting code.*

Proof. Denote by $\mathcal{S}'$ the received set after a loss of at most s sequences and errors in at most t sequences. Let s' be the number of asymmetric errors and t' be the number of errors in $\mathbf{v}(\mathcal{S})$ with $s' + t' \leq s + t$ and $t' \leq t$. Note that $s' = M - \mathrm{wt}(\mathbf{v}(\mathcal{S}'))$ is detectable by the decoder. If $s' \leq s$, then the decoder can directly decode the loss of $s' \leq s$ sequences and $t' \leq t$ errors in the Johnson graph. If $s' > s$, the decoder adds $s' - s$ (arbitrarily placed) ones to $\mathbf{v}(\mathcal{S}')$, resulting in a loss of exactly s sequences and at most $t' + s' - s \leq t$ errors in the Johnson graph. □

To obtain a code based on Construction 2, we use the fact that an asymmetric error can be represented by a single substitution error and an error in the Johnson graph can be represented by two substitution errors. With an appropriate minimum distance, it is therefore possible to employ standard codes, which will be done in the following theorem.

Theorem 13. *There exists a construction of the code $\mathscr{C}_2(M, L, s, t)$ with redundancy at most*

$$r(\mathscr{C}_2(M, L, s, t)) \leq (s + 2\ t)L.$$

Proof. By Lemma 6, it is sufficient to find a sufficiently large M-constant-weight code which can correct $s + 2t$ substitution errors. This is since each loss in $\mathcal{S}$ causes an $1 \to 0$ asymmetric error in $\mathbf{v}(\mathcal{S})$ and can be represented as a single substitution error and every error in a sequence in $\mathcal{S}$ will cause at most one $1 \to 0$ and one $0 \to 1$ error in $\mathbf{v}(\mathcal{S})$ and thus can be represented by two substitution errors. Next, it is known, that there exists a τ-substitution-correcting binary alternant code of length 2^L and dimension $2^L - \tau L$, cf. [32]. Due to the pigeonhole principle and since the alternant code has at most $2^{\tau L}$ cosets, there is one coset of the alternant code that contains at least $\binom{2^L}{M}/2^{\tau L}$ words with constant weight M, and therefore there exists a code $\mathscr{C}_2(M, L, s, t)$ of cardinality at least $\binom{2^L}{M}/2^{\tau L}$. With this alternant code, the redundancy of Construction 2 is therefore at most

$$r(\mathscr{C}_2(M, L, s, t)) \leq \log \binom{2^L}{M} - \log \frac{\binom{2^L}{M}}{2^{\tau L}} = \tau L.$$

Using $\tau = s + 2t$ yields the theorem. □

The redundancy of Construction 2 is lower than that of Construction 1, especially for the considered case $M = 2^{\beta L}$. However, for Construction 1 there exist efficient encoders and decoders while this is unclear for Construction 2, also since the code length of the constant-weight code is exponential in L.

D. An Improved Indexed-Based Construction

Construction 1, which uses indexing, is beneficial for its simplicity in the encoding and decoding procedures, however its redundancy is significantly larger than the one achieved by Construction 2. On the other hand, Construction 2 does not provide an efficient encoder and decoder due to the lack of ordering in the set $\mathcal{S}$. In this section, we present a construction which introduces ideas from both of these methods.

The main idea of this construction is to reduce the number of bits allocated for indexing each sequence. This allows a trade-off in redundancy with respect to L and M. To simplify notation, we assume here that $M = 2^z$ for some $z \in \mathbb{N}$.

Construction 3. *Denote by $\mathbf{I}_c(i) \in \Sigma_2^{c \log M}$ the $c \log M$ most significant bits of the binary representation $\mathbf{I}(i)$ of i, where $0 \leq c < 1$ and $c \log M \in \mathbb{N}_0$. Further, for $1 \leq i \leq M^c$, let $\mathbf{U}_i = \{\mathbf{u}_{(i-1)M^{1-c}+1}, \ldots, \mathbf{u}_{iM^{1-c}}\}$ denote a set of distinct sequences with the same index $\mathbf{I}_c(i)$, which are ordered lexicographically and form a symbol over a field of size $\binom{2^L M^{-c}}{M^{1-c}}$, where $\mathbf{u}_j \in \Sigma_2^{L - c \log M}$.*

For $\delta \geq 0$, let $\mathscr{C}_3(M, L, c, \delta)$ be the code defined by

$$\mathscr{C}_3(M, L, c, \delta) = \{\mathcal{S} \in \mathcal{X}_M^L : \mathbf{x}_i = (\mathbf{I}_c(i), \mathbf{u}_i), \\ (\mathbf{U}_1, \ldots, \mathbf{U}_{M^c}) \in \mathsf{MDS}[M^c, M^c - \delta]\}.$$

To guarantee existence of the MDS code, we require $M^c \leq \binom{2^L M^{-c}}{M^{1-c}}$ [32]. For $M = 2^{\beta L}$, $c \leq 1 + \frac{\log \frac{1-\beta}{\beta}}{\beta L}$ is sufficient.

Note that there are M^c groups of sequences which use the same index and each group contains M^{1-c} sequences.

Lemma 7. *For all M, L, δ, the code $\mathscr{C}_3(M, L, c, \delta)$ is*

- *$(s, t, \bullet)_{\mathbb{L}}$-correcting for all $s + 2t \leq \delta$,*
- *$(s, t, \bullet)_{\mathbb{I}}$-correcting for all $s + t \leq \delta$,*
- *$(s, t, \bullet)_{\mathbb{D}}$-correcting for all $s + t \leq \delta$,*

Proof. The proof follows the same idea as that for Lemma 5. We will show that the MDS codeword $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{M^c})$ can be recovered from $\mathbf{U}' = (\mathbf{U}'_1, \mathbf{U}'_2, \ldots, \mathbf{U}'_{M^c})$, where $\mathbf{U}'_i$ collects all sequences in $\mathcal{S}'$ which have the same index i, i.e. $\mathbf{U}'_i = \{\mathrm{suff}_{L - c \log M}(\mathbf{x}') : \mathbf{x}' \in \mathcal{S}', \mathrm{pref}_{c \log M}(\mathbf{x}) = \mathbf{I}_c(i)\}$. Given $\mathcal{S}'$, we create the received estimate word $\mathbf{U}'$ by declaring all positions i with

$$|\mathbf{U}'_i| \neq M^{1-c},$$

as erasures. The remaining positions in $\mathbf{U}'$ are filled with the corresponding symbols $\mathbf{U}'_i$. We will show that the number of erasures s' and the number of errors t' in $\mathbf{U}'$ satisfy $s' + 2t' \leq \delta$ by the following consideration. First, insert all error-free sequences $\mathbf{x} \in \mathcal{C}$ into $\mathbf{U}'$. Up to this point $s' \leq s + t$ and $t' = 0$, since there are $s + t$ sequences missing and all inserted sequences are error-free. Therefore, the $s + t$ affected groups, which contain less than M^{1-c} sequences can be detected and declared erasures. Now, each of the t erroneous sequences $\mathbf{x}' \in \mathcal{E}'$ is inserted to $\mathbf{U}'$ and can have an arbitrary index i due to errors. If the erroneous sequence $\mathbf{x}'$ has an index i of an index group with $|\mathbf{U}'_i| = M^{1-c} - 1$ elements, this group cannot be detected as erroneous anymore, as it contains now exactly M^{1-c} sequences. Consequently t' increases by one and s' decreases by one, as the group is erroneous but is not declared as an erasure in $\mathbf{U}'$ anymore. If the erroneous sequence $\mathbf{x}'$ has an index of an index group with $|\mathbf{U}'_i| = M^{1-c}$, this group will contain $M^{1-c} + 1$ sequences afterwards and can be detected as erroneous and thus declared as erasure. In this case the number of erasures s' increases by one. In all

other cases neither s' nor t' change. Since t sequences of $\mathcal{S}$ are erroneous the sum $s'+2t'$ can increase at most by t with respect to the starting point $s+t$ and thus $s'+2t' \leq s+2t \leq \delta$, which proves the error correcting capability.

For the case of only insertion ($\mathbb{I}$) and only deletion ($\mathbb{D}$) errors, it is possible to identify the erroneous groups by checking the length of the respective sequences to be larger (respectively smaller) than L. If these sequences are discarded and the corresponding groups declared as erasures, there are in total at most $s+t$ erasures inside the MDS codeword, which can be corrected, if $s+t \leq \delta$. □

The redundancy of Construction 3 is stated in the following theorem.

Theorem 14. *The redundancy of Construction 3 is given by*

$$r(\mathscr{C}_3(M,L,c,\delta)) = \log \binom{2^L}{M} - (M^c-\delta)\log \binom{2^L M^{-c}}{M^{1-c}}.$$

For fixed $0<c<1$, $\delta \in \mathbb{N}_0$ and $0<\beta<1$, the redundancy of $\mathscr{C}_3(M,L,c,\delta)$ is asymptotically

$$r(\mathscr{C}_3(M,L,c,\delta)) = \frac{(1-c)}{2}M^c \log M + \frac{\log 2\pi}{2}M^c + \delta M^{1-c}(L-\log M+\log \mathrm{e}) + o(M^c+M^{1-c}),$$

when $M \to \infty$ with $M = 2^{\beta L}$.

The proof is given in Appendix B. Note that the last summand in the asymptotic expression for $\mathscr{C}_3(M,L,c,\delta)$ in Theorem 14 quantifies the redundancy from the MDS construction, since it is multiplied by δ, the redundancy of the MDS code. The two remaining terms therefore quantify the redundancy required for indexing. This shows that, asymptotically, for $c > 0.5$ the redundancy needed for indexing dominates, as the terms for indexing scale as M^c and the term for the MDS construction scales as M^{1-c} and for $c < 0.5$ the redundancy from the MDS construction dominates the redundancy of the overall construction.

E. Concatenated Constructions

Since the input of the DNA storage channel, $\mathcal{S} \in \mathcal{X}_M^L$ is a set of M sequences, each of which has length L, it is possible to use a concatenated coding scheme to correct both a loss of sequences and errors inside the sequences. The concatenation can be constructed by choosing a set $\mathcal{S}_\mathrm{o}$ as a codeword from an outer code $\mathscr{C}_\mathrm{o} \subseteq \mathcal{X}_M^{L_\mathrm{o}}$, where $L_\mathrm{o} < L$. Then, each sequence $\mathbf{x}_\mathrm{o} \in \mathcal{S}_\mathrm{o}$ is encoded with some inner block-code $\mathscr{C}_\mathrm{i} \subseteq \Sigma_2^L$ of dimension L_o and length L. This procedure is formalized in the following construction.

Construction 4. *For all M, L, $L_\mathrm{o} < L$ and positive integers s,t, let $\mathscr{C}_\mathrm{o} \subseteq \mathcal{X}_M^{L_\mathrm{o}}$ be an outer code and $\mathscr{C}_\mathrm{i} \subseteq \Sigma_2^L$ be a standard block-code of dimension L_o and length L. Further, $\mathrm{en}(\cdot) : \Sigma_2^{L_\mathrm{o}} \mapsto \Sigma_2^L$ is an encoder of the code $\mathscr{C}_\mathrm{i}$. We define the concatenated construction as*

$$\mathscr{C}_4(M,L,\mathscr{C}_\mathrm{i},\mathscr{C}_\mathrm{o}) = \left\{\mathcal{S} \in \mathcal{X}_M^L : \mathcal{S} = \bigcup_{\mathbf{x}_\mathrm{o} \in \mathcal{S}_\mathrm{o}} \mathrm{en}(\mathbf{x}_\mathrm{o}), \mathcal{S}_\mathrm{o} \in \mathscr{C}_\mathrm{o}\right\}.$$

As outer code $\mathscr{C}_\mathrm{o}$ it is in principle possible to use any code over $\mathcal{X}_M^L$. However, using the proposed Constructions 1, 2, or 3 it is possible to enhance the inner code to additionally correct a loss of sequences. This is done as follows.

Lemma 8. *Let $\mathscr{C}_\mathrm{o} \subseteq \mathcal{X}_M^{L_\mathrm{o}}$ be an $(s,0,0)_\mathbb{T}$-correcting code and $\mathscr{C}_i \subseteq \Sigma_2^L$ be a block-code that can correct ϵ errors of type $\mathbb{T}$. Then, $\mathscr{C}_4(M,L,\mathscr{C}_\mathrm{i},\mathscr{C}_\mathrm{o})$ is $(s,M-s,\epsilon)_\mathbb{T}$-correcting.*

Proof. The proof is immediate, since the inner code can correct all errors of type $\mathbb{T}$ inside the sequences. After correcting these errors, it is possible to correct the lost sequences using the outer code. □

Note that such concatenated constructions are highly relevant in practice, as in the case that there are some sequences, which experienced more than ϵ errors can be corrected by the outer code, since Constructions 1, 2, or 3 can correct both a loss of sequences and errors in sequences, as long as $s+2t \leq \delta$. Such a construction has been used in [12], where a Reed-Solomon code has been used as inner code and an indexed Reed-Solomon code has been used as outer code.

F. Special Constructions

In this section, we suggest constructions that can correct errors for some special cases of errors in the DNA storage channel. These constructions are interesting, since they provide insights about the channel and can likely be generalized to more general error types.

The following $(0,1,1)_\mathbb{D}$-correcting construction is based on Varshamov-Tenengolts (VT) codes [26], [34] that can correct a single insertion/deletion in one of the M sequences. The VT code is defined to be all sequences which have the same checksum, that is defined as follows.

Definition 10. *The Varshamov-Tenegolts checksum $s_L(\mathbf{x})$ of $\mathbf{x} \in \Sigma_2^L$ is defined by*

$$s_L(\mathbf{x}) = \sum_{i=1}^{L} i x_i \bmod (L+1).$$

Our construction now employs the idea of using a single-erasure-correcting code over the checksums of all sequences. The insertion/deletion can then be corrected by first recovering the checksum of the distorted sequence and then using this checksum to correct the insertion/deletion. Note that this idea is similar to the concept of tensor product codes [35].

Construction 5. *For an integer a, with $0 \leq a \leq L$, the code construction $\mathscr{C}_5(M,L,a)$ is given by*

$$\mathscr{C}_5(M,L,a) = \left\{\mathcal{S} \in \mathcal{X}_M^L : \sum_{i=1}^{M} s_L(\mathbf{x}_i) \equiv a \bmod (L+1)\right\}.$$

Note that the code can be extended to an arbitrary alphabet size q by applying non-binary VT codes [36].

Lemma 9. *For all M, L, a, the code $\mathscr{C}_5(M,L,a)$ is an $(0,1,1)_\mathbb{ID}$-correcting code.*

Proof. Assume there has been a single insertion or deletion in the k-th sequence, for $1 \leq k \leq M$. After the reading process, the $M-1$ error-free sequences can be identified as they have length exactly L. The checksum deficiency is given by

$$a - \sum_{i \in \mathcal{C}} s_L(\mathbf{x}_i) \bmod (L+1) = s_L(\mathbf{x}_k).$$

The error in $\mathbf{x}_k$ is corrected by decoding in the VT code with checksum $s_L(\mathbf{x}_k)$. □

The redundancy of Construction 5 is established in the following theorem.

Theorem 15. *There exists $0 \leq a \leq L$ such that the redundancy of Construction 5 is at most*

$$r(\mathscr{C}_5(M, L, a)) \leq \log(L+1).$$

Proof. The codes $r(\mathscr{C}_5(M, L, a))$ form a partition over $\mathcal{X}_M^L$ for all $0 \leq a \leq L$. Since, there are $L+1$ distinct values for a, based on the pigeonhole principle there exists $0 \leq a \leq L$ such that the cardinality of the code $\mathscr{C}_5(M, L, a)$ satisfies $|\mathscr{C}_5(M, L, a)| \geq \binom{2^L}{M}/(L+1)$ and thus its redundancy is at most $\log(L+1)$. □

As we show in Theorem 9, the redundancy of any $(0, 1, 1)_{\mathbb{D}}$-correcting code is at least $\log(L) + o(1)$, and thus Construction 5 is asymptotically optimal.

Using VT codes, we propose another construction of $(0, M, 1)_{\mathbb{ID}}$-correcting codes. That is, the code can correct a single deletion or insertion in every sequence.

Construction 6. *Let $a \in \mathbb{N}_0$, with $0 \leq a \leq L$. Then,*

$$\mathscr{C}_6(M, L, a) = \{\mathcal{S} \in \mathcal{X}_M^L : s_L(\mathbf{x}_i) \equiv a \bmod (L+1), \forall\, 1 \leq i \leq M\}.$$

Lemma 10. *The code $\mathscr{C}_6(M, L, a)$ is an $(0, M, 1)_{\mathbb{ID}}$-correcting code.*

Proof. All erroneous sequences can be detected by checking their length. If a sequence is erroneous, it can be corrected by decoding in the VT code with checksum a. Note that two distinct sequences cannot have the same erroneous outcome since they are different and belong to a single-deletion-correcting code. □

By Construction 6, all sequences $\mathbf{x}_i$ have the same checksum a, which allows to correct a single insertion or a single deletion in each sequence. The redundancy of Construction 6 is computed in the following lemma.

Theorem 16. *For fixed $0 < \beta < 1$, the redundancy of the code $\mathscr{C}_6(M, L, 0)$ satisfies asymptotically*

$$r(\mathscr{C}_6(M, L, 0)) \leq M \log(L+1) + o(M),$$

when $M \to \infty$ with $M = 2^{\beta L}$.

Proof. It is known [26] that the number of words satisfying $s_L(\mathbf{x}) = 0 \bmod (L+1)$ is at least $2^L/(L+1)$. Each codeword of $\mathscr{C}_6(M, L, a)$ is a subset of a VT code with cardinality M. Therefore the redundancy of Construction 6 is at most

$$\begin{aligned} r(\mathscr{C}_6(M, L, 0)) &\leq \log \binom{2^L}{M} - \log \binom{\frac{2^L}{L+1}}{M} \\ &\leq M \log(L+1) + \frac{M^2 \log \mathrm{e}}{2^L/(L+1) - M}. \end{aligned}$$

For $M = 2^{\beta L}$, $0 < \beta < 1$ the second term is $o(M)$, which concludes the proof. □

Interestingly, as has been shown in Theorem 10, the redundancy of this construction is asymptotically optimal in terms of scaling with the parameters M and L. Note that there is a non-asymptotic expression for the redundancy in the proof.

The next construction can be used to correct ϵ substitution errors in each sequence.

Construction 7. *Let $\mathscr{C}[L, \epsilon] \subseteq \Sigma_2^L$ denote a binary ϵ-substitution-correcting code of length L. For all $M \leq |\mathscr{C}[L, \epsilon]|, L$, and ϵ we define the code*

$$\mathscr{C}_7(M, L, \epsilon) = \{\mathcal{S} \in \mathcal{X}_M^L : \mathcal{S} \subseteq \mathscr{C}[L, \epsilon]\}.$$

Lemma 11. *The code $\mathscr{C}_7(M, L, \epsilon)$ is an $(0, M, \epsilon)_{\mathbb{S}}$-correcting code.*

The proof is immediate, since every sequence is a codeword of a code that can correct ϵ substitutions. Using binary alternant codes, it is possible to find a lower bound on the redundancy of Construction 7.

Theorem 17. *There exists a construction for which the code $\mathscr{C}_7(M, L, \epsilon)$ with fixed $\epsilon \in \mathbb{N}_0$ and $0 < \beta < 1$ has an asymptotic redundancy of at most*

$$r(\mathscr{C}_7(M, L, \epsilon)) \leq M \epsilon \lceil \log L \rceil + o(M),$$

when $M \to \infty$ with $M = 2^{\beta L}$.

Proof. For $\mathscr{C}[L, \epsilon]$ in Construction 7 we use a binary ϵ-substitution-correcting alternant code of length L, which has redundancy at most $\epsilon \lceil \log L \rceil$, cf. [32, Ch. 5.5] and thus obtain a code $\mathscr{C}_7(M, L, \epsilon)$ with redundancy at most

$$\begin{aligned} r(\mathscr{C}_7(M, L, \epsilon)) &\leq \log \binom{2^L}{M} - \log \binom{2^{L - \epsilon \lceil \log L \rceil}}{M} \\ &\leq M \epsilon \lceil \log(L+1) \rceil + \frac{M^2 \log \mathrm{e}}{2^{L - \epsilon \lceil \log L \rceil} - M}. \end{aligned}$$

For $M = 2^{\beta L}$, $0 < \beta < 1$ the second term is $o(M)$, which concludes the proof. □

Note that Theorem 8 implies that for fixed ϵ this construction is close to optimality.

VI. Conclusion

In this paper, we set the foundations for codes over sets for DNA storage applications. After presenting the channel model and a new family of error-correcting codes over sets, we derived several bounds and constructions. Our bounds consist of extensions of the Gilbert-Varshamov and sphere packing bounds for the studied codes in the paper. We also proposed

several constructions which can be either with or without indices or a reduced version of the indices. Lastly, we derived several more special constructions for a specific set of parameters. It has been illustrated that many of the proposed constructions are close to optimal, such as for the case of substitution, respectively single insertion or deletion errors inside all of the strands. We further have proposed several constructions that can cope with combinations of a loss of sequences and errors inside the sequences. By analyzing the sphere packing bounds and comparing them to our constructions, we have found important insights about the nature of the DNA storage channel. These include the surprising fact that correcting insertions or deletions requires less redundancy than correcting substitution errors inside the sequences.

APPENDIX A
AUXILIARY LEMMAS

Lemma 12. *Let $f(n), g(n) : \mathbb{N} \mapsto \mathbb{R}$ be two arbitrary functions with $f(n) = o(1)$ for $n \to \infty$. Then,*

$$g(n)\ln\left(1+f(n)\right) = g(n)f(n) + O\left(g(n)f^2(n)\right).$$

Proof. We use the standard bound on the natural logarithm

$$\frac{x}{x+1} \leq \ln(1+x) \leq x,$$

for all $x > -1$. Since $f(n) = o(1)$, there exists $n_0 \in \mathbb{N}$, such that $|f(n)| < 1$ for all $n \geq n_0$ and therefore

$$g(n)\frac{f(n)}{f(n)+1} \leq g(n)\ln\left(1+f(n)\right) \leq g(n)f(n),$$

for all $n \geq n_0$. This allows to find an upper bound to the following limit of the first order approximation

$$\lim_{n\to\infty}\left|\frac{g(n)\ln\left(1+f(n)\right) - g(n)f(n)}{g(n)f^2(n)}\right| \leq 1,$$

by plugging in the lower and upper bound on $g(n)\ln(1+f(n))$, which proves the statement. □

Lemma 13. *Let $f(n), g(n) : \mathbb{N} \mapsto \mathbb{N}$ be two arbitrary functions with $g(n) = o(f(n))$ and $g(n) = \omega(1)$, when $n \to \infty$. The binomial coefficient satisfies*

$$\log\binom{f(n)}{g(n)} = g(n)\log\frac{\mathrm{e}f(n)}{g(n)} + o(g(n)),$$

when $n \to \infty$.

Proof. Note that $g(n) = o(f(n))$ and $g(n) = \omega(1)$ automatically implies $f(n) = \omega(1)$. The binomial coefficient satisfies

$$\begin{aligned}\log\binom{f(n)}{g(n)} &= \log\frac{f(n)!}{(f(n)-g(n))!g(n)!}\\ &= g(n)\log\frac{f(n)}{g(n)} - \frac{1}{2}\log g(n)\\ &\quad - \left(f(n)-g(n)+\frac{1}{2}\right)\log\left(1-\frac{g(n)}{f(n)}\right) + \gamma,\end{aligned}$$

where $\gamma = -\log\sqrt{2\pi} + O(\frac{1}{g(n)})$. Here we used a refinement [37] of Stirling's approximation, which states that

$$\sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n \mathrm{e}^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n \mathrm{e}^{\frac{1}{12n}},$$

for any $n \in \mathbb{N}$. Using Lemma 12, we obtain

$$\begin{aligned}&-\left(f(n)-g(n)+\frac{1}{2}\right)\log\left(1-\frac{g(n)}{f(n)}\right)\\ &= \log\mathrm{e}\left(g(n) - \frac{g^2(n)}{f(n)} + \frac{g(n)}{2f(n)}\right) + O\left(\frac{g^2(n)}{f(n)}\right)\\ &= g(n)\log\mathrm{e} + o(g(n)),\end{aligned}$$

where we used that $\frac{g(n)}{f(n)} = o(1)$. Plugging this result into the expression of the binomial coefficient and using further $\log g(n) = o(g(n))$ and $\gamma = o(g(n))$ proves the lemma. □

Lemma 14. *For any fixed integer $\delta \in \mathbb{N}_0$ and any integer functions $y(M) \leq M$ and $z(L)$ with $z(L) \leq 2^L/y(M)$ for large enough M, the following asymptotic property holds*

$$\begin{aligned}\log\frac{\binom{2^L}{M-y(M)}\binom{2^L/z(L)}{y(M)}}{\binom{2^L}{M-\delta}} &\leq -y(M)\log\frac{z(L)y(M)}{\mathrm{e}M}\\ &\quad + O\left(\frac{My(M)}{2^L}\right) + O(L),\end{aligned}$$

when $M \to \infty$ and $M = 2^{\beta L}$ with $0 < \beta < 1$.

Proof. The lemma can be shown directly by calculating the expression for the binomial coefficient

$$\begin{aligned}&\log\frac{\binom{2^L}{M-y(M)}\binom{2^L/z(L)}{y(M)}}{\binom{2^L}{M-\delta}}\\ &= \log\frac{(2^L/z(L))^{\underline{y(M)}}(2^L-M+\delta)^{\underline{\delta}}}{(2^L-M+y(M))^{\underline{y(M)}}M^{\underline{\delta}}} + \log\binom{M}{y(M)}\\ &\leq y(M)\log\frac{2^L/z(L)}{2^L-M} + \log\binom{M}{y(M)} + O(L)\\ &\overset{(a)}{\leq} y(M)\log\frac{\mathrm{e}M}{z(L)y(M)} + O\left(\frac{My(M)}{2^L}\right) + O(L),\end{aligned}$$

where, $n^{\underline{m}} = n\cdot(n-1)\ldots(n-m+1)$ for $n, m \in \mathbb{N}_0$ denotes the falling factorial. In inequality (a), we used Lemma 12 for the approximation of the logarithm and $\binom{n}{k} \leq \left(\frac{\mathrm{e}n}{k}\right)^k$ as an upper bound for the binomial coefficient. □

APPENDIX B
PROOF OF THEOREM 14

The cardinality of Construction 3 can be computed as follows. Each group $\mathbf{U}_i$ consists of M^{1-c} unordered, distinct sequences, which share the same index $\mathbf{I}_c(i)$. In total, there are $M^c - \delta$ information groups, since δ groups are redundancy symbols of the MDS codeword. Therefore, the redundancy is

$$r(\mathcal{C}_3(M,L,c,\delta)) = \log\binom{2^L}{M} - \log\binom{2^LM^{-c}}{M^{1-c}}^{M^c-\delta}.$$

Applying Stirling's approximation [37] onto the binomial coefficients yields

$$\begin{aligned} r(\mathcal{C}_3(M,L,c,\delta)) &= \log\binom{2^L}{M} - (M^c-\delta)\log\binom{2^L M^{-c}}{M^{1-c}} \\ &= \frac{1-c}{2}M^c\log M + \frac{M^c-1}{2}\log\left(1-\frac{M}{2^L}\right) - \gamma_2 M^c + \gamma_1 \\ &\quad + \delta\Bigg(M^{1-c}L - M^{1-c}\log M - \frac{1-c}{2}\log M \\ &\qquad - \left(2^L M^{-c} - M^{1-c} + \frac{1}{2}\right)\log\left(1-\frac{M}{2^L}\right) + \gamma_2\Bigg), \end{aligned}$$

where $\gamma_1 = -\log\sqrt{2\pi} + o(1)$ and $\gamma_2 = -\log\sqrt{2\pi} + o(1)$, when $c < 1$. Note that it can be verified that for $c = 1$, γ_2 has a different asymptotic behavior, i.e., $\gamma_2 = -\log \mathrm{e} + o(1)$. Therefore, for $c = 1$, the expression for $r(\mathcal{C}_3(M,L,c,\delta))$ yields the same redundancy as in Theorem 12. Employing Lemma 12 onto the two logarithmic terms yields

$$\begin{aligned} r(\mathcal{C}_3(M,L,c,\delta)) &= \frac{1-c}{2}M^c\log M + \frac{\log 2\pi}{2}M^c \\ &\quad + \delta M^{1-c}\left(L - \log M + \log \mathrm{e}\right) + o(M^c + M^{1-c}). \end{aligned}$$

□

REFERENCES

[1] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobit, "Coding over sets for DNA storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 2411–2415.

[2] D. Carmean, L. Ceze, G. Seelig, K. Stewart, K. Strauss, and M. Willsey, "DNA data storage and hybrid molecular-electronic computing," *Proc. IEEE*, vol. 107, no. 1, pp. 63–72, Jan. 2019.

[3] M. Blawat *et al.*, "Forward error correction for DNA data storage," in *Proc. Int. Conf. Comput. Sci.*, San Diego, CA, USA, Jun. 2016, pp. 1011–1022.

[4] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, Mar. 2017.

[5] S. Kosuri and G. Church, "Large-scale de novo DNA synthesis: Technologies and applications," *Nature Methods*, vol. 11, no. 5, pp. 499–507, May 2014.

[6] L. Organick *et al.*, "Random access in large-scale DNA data storage," *Nature Biotechnol.*, vol. 36, pp. 242–248, Feb. 2018.

[7] M. G. Ross *et al.*, "Characterizing and measuring bias in sequence data," *Genome Biol.*, vol. 14, no. 5, p. R51, May 2013.

[8] A. K. Yim *et al.*, "The essential component in DNA-based information storage system: Robust error-tolerating module," *Front. Bioeng. Biotechnol.*, vol. 2, no. 49, pp. 1–5, Nov. 2014.

[9] 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.

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

[11] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Jan. 2013.

[12] 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.

[13] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," in *Proc. 21st Int. Conf. Architectural Support Program. Lang. Operating Syst.*, Atlanta, Georgia, Apr. 2016, pp. 637–649.

[14] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Nature Sci. Rep.*, vol. 5, Aug. 2015, Art. no. 14138.

[15] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Nature Sci. Rep.*, vol. 7, no. 1, p. 5011, Jul. 2017.

[16] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[17] M. Kovačević and V. Y. F. Tan, "Codes in the space of multisets—Coding for permutation channels with impairments," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5156–5169, Jul. 2018.

[18] J. Sima, N. Raviv, and J. Bruck, "On coding over sliced information," 2018, *arXiv:1809.02716*. [Online]. Available: https://arxiv.org/abs/1809.02716

[19] W. Song and K. Cai, "Sequence-subset distance and coding for error control in DNA-based data storage," 2018, *arXiv:1809.05821*. [Online]. Available: https://arxiv.org/abs/1809.05821

[20] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proc. IEEE Int. Symp. Inf. Theory*, Aachen, Germany, Jun. 2017, pp. 3130–3134.

[21] R. Heckel, G. Mikutis, and R. Grass, "A characterization of the DNA data storage channel," 2018, *arXiv:1803.03322*. [Online]. Available: https://arxiv.org/abs/1803.03322

[22] V. I. Levenshtein, "Elements of coding theory," in *Diskretnaya Matematika i Matematicheskie Voprosy Kibernetiki*. Moscow, Russia: Nauka, 1974, pp. 207–305.

[23] V. I. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones," *Prob. Inf. Trans.*, vol. 1, no. 1, pp. 8–17, Jan. 1965.

[24] J. Gu and T. Fuja, "A generalized Gilbert-Varshamov bound derived via analysis of a code-search algorithm," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 1089–1093, May 1993.

[25] L. Tolhuizen, "The generalized Gilbert-Varshamov bound is implied by Turan's theorem [code construction]," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1605–1606, Sep. 1997.

[26] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and revers'als," *Sov. Phys. Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[27] A. Knoblauch, "Closed-form expressions for the moments of the binomial distribution," *SIAM J. Appl. Math.*, vol. 69, no. 1, pp. 197–204, Jan. 2008.

[28] A. Fazeli, A. Vardy, and E. Yaakobi, "Generalized sphere packing bound," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2313–2334, May 2015.

[29] A. A. Kulkarni and N. Kiyavash, "Nonasymptotic upper bounds for deletion correcting codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 5115–5130, Aug. 2013.

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

[31] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

[32] R. M. Roth, *Introduction to Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[33] A. E. Brouwer, J. B. Shearer, N. J. A. Sloane, and W. D. Smith, "A new table of constant weight codes," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1334–1380, Nov. 1990.

[34] R. R. Varshamov and G. M. Tenengolts, "Codes which correct single asymmetric errors," *Autom. Remote Control*, vol. 26, no. 2, pp. 286–290, 1965.

[35] J. K. Wolf, "An introduction to tensor product codes and applications to digital storage systems," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Oct. 2006, pp. 6–10.

[36] G. Tenengolts, "Nonbinary codes, correcting single deletion or insertion," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 5, pp. 766–769, Sep. 1984.

[37] H. Robbins, "A remark on Stirling's formula," *Amer. Math. Monthly*, vol. 62, no. 1, pp. 26–29, 1955.

Andreas Lenz (S'17) received the B.Sc. and M.Sc. degrees (Hons.) in electrical engineering and information technology from the Technische Universität München (TUM), Germany, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the Coding for Communications and Data Storage (COD) Group, with the focus on research about coding theory for insertion and deletion errors and modern data storage systems. His research interests include parameter estimation, communications, and circuit theory. In 2014, he received the Leo-Brandt Award *Master of Navigation* from the German society of positioning and navigation.

Paul H. Siegel (M'82–SM'90–F'97–LF'19) received the S.B. and Ph.D. degrees in mathematics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1975 and 1979, respectively. He held a Chaim Weizmann Post-Doctoral Fellowship with the Courant Institute, New York University, New York, NY, USA. He was with the IBM Research Division, San Jose, CA, USA, from 1980 to 1995. He joined the faculty with the University of California, San Diego, La Jolla, CA, USA, in 1995, where he is currently a Distinguished Professor of electrical and computer engineering with the Jacobs School of Engineering. He is also affiliated with the Center for Memory and Recording Research, where he holds the Endowed Chair and served as the Director from 2000 to 2011. His research interests include information theory and communications, particularly coding and modulation techniques, with applications to digital data storage and transmission. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996 and from 2009 to 2014. He is a member of the National Academy of Engineering. He was the 2015 Padovani Lecturer of the IEEE Information Theory Society. He was a recipient of the 2007 Best Paper Award in Signal Processing and Coding for Data Storage from the Data Storage Technical Committee of the IEEE Communications Society. He was a co-recipient of the 1992 IEEE Information Theory Society Paper Award and the 1993 IEEE Communications Society Leonard G. Abraham Prize Paper Award. He served as a Co-Guest Editor for the 1991 Special Issue on Coding for Storage Devices of the IEEE TRANSACTIONS ON INFORMATION THEORY. He served as an Associate Editor of Coding Techniques for the IEEE TRANSACTIONS ON INFORMATION THEORY from 1992 to 1995 and the Editor-in-Chief from 2001 to 2004. He was also a Co-Guest Editor for the 2001 two-part issue on The Turbo Principle: From Theory to Practice and the 2016 issue on Recent Advances in Capacity Approaching Codes of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.

Antonia Wachter-Zeh (S'10–M'14) received the B.S. degree in electrical engineering from the University of Applied Sciences Ravensburg, Germany, in 2007, the M.S. degree in communications technology from Ulm University, Germany, in 2009, and the Ph.D. degree from the Institute of Communications Engineering, University of Ulm, Germany, and the Institut de recherche math ématique de Rennes (IRMAR), Universite de Rennes 1, Rennes, France, in 2013. From 2013 to 2016, she was a Post-Doctoral Researcher with the Technion–Israel Institute of Technology, Haifa, Israel. She is currently an Assistant Professor with the Technische Universität München (TUM), Munich, Germany. Her research interests are coding theory, information theory and cryptography and their application to storage, communications, and security.

Eitan Yaakobi (S'07–M'12–SM'17) received the B.A. degree in computer science and mathematics and the M.Sc. degree in computer science from the Technion–Israel Institute of Technology, Haifa, Israel, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California, San Diego, La Jolla, CA, USA, in 2011. From 2011 to 2013, he was a Post-Doctoral Researcher with the Department of Electrical Engineering, California Institute of Technology, and with the Center for Memory and Recording Research, University of California, San Diego. He is currently an Associate Professor with the Computer Science Department, Technion–Israel Institute of Technology. His research interests include information and coding theory with applications to nonvolatile memories, associative memories, DNA storage, data storage and retrieval, and private information retrieval. He received the Marconi Society Young Scholar in 2009 and the Intel Ph.D. Fellowship in 2010 and 2011.