Covering Codes Using Insertions or Deletions

Andreas Lenz, *Student Member, IEEE*, Cyrus Rashtchian, Paul H. Siegel, *Life Fellow, IEEE*, and Eitan Yaakobi, *Senior Member, IEEE*

***Abstract*— A covering code is a set of codewords with the property that the union of balls, suitably defined, around these codewords covers an entire space. Generally, the goal is to find the covering code with the minimum size codebook. While most prior work on covering codes has focused on the Hamming metric, we consider the problem of designing covering codes defined in terms of either insertions or deletions. First, we provide new sphere-covering lower bounds on the minimum possible size of such codes. Then, we provide new existential upper bounds on the size of optimal covering codes for a single insertion or a single deletion that are tight up to a constant factor. Finally, we derive improved upper bounds for covering codes using $R \geq 2$ insertions or deletions. We prove that codes exist with density that is only a factor $O(R \log R)$ larger than the lower bounds for all fixed R. In particular, our upper bounds have an optimal dependence on the word length, and we achieve asymptotic density matching the best known bounds for Hamming distance covering codes.**

***Index Terms*— Covering codes, insertions, deletions.**

I. Introduction

COVERING codes are a core object of study in coding theory and discrete mathematics. They have found applications in diverse areas such as data compression [1], football pools [2], circuit complexity [3], lattice problems [4], and approximate nearest neighbor search [5]. Previous work has mostly studied covering codes with respect to substitutions (i.e., the Hamming distance). Recently, due to the large amount of textual and biological data, there has been a resurgence of interest in the Levenshtein distance and in channels with insertion and deletion errors (e.g., [6]–[14]). Despite this substantial progress, the Levenshtein distance remains poorly understood compared to other metrics on discrete spaces, and many fundamental questions remain open.

In this paper, we study covering codes for either insertions or deletions. Loosely speaking, we aim to cover a space of words by the union of balls around a minimum number of codewords. Let Σ_q^n denote the set of words of length n over a q-ary alphabet. For the case of insertions, a codeword $\boldsymbol{c} \in \Sigma_q^n$ *covers* a word $\boldsymbol{y} \in \Sigma^{n+R}$ at radius R if $\boldsymbol{y}$ can be obtained from $\boldsymbol{c}$ by inserting exactly R symbols. Similarly, for the case of deletions, a codeword $\boldsymbol{c} \in \Sigma_q^n$ *covers* a word $\boldsymbol{y} \in \Sigma^{n-R}$ at radius R if $\boldsymbol{y}$ can be obtained from $\boldsymbol{c}$ by deleting exactly R symbols. This means that the covering problem for insertions deals with finding a small set of codewords of length n such that each word of length $n+R$ is a supersequence of a codeword. Analogously, for the case of deletions, each word of length $n-R$ must be a subsequence of some codeword. In both cases, balls are naturally defined by the set of words obtained by inserting R symbols into or deleting R symbols from some word of length n. Notice, however, that the codewords and the covered words reside in different spaces because they have different lengths. Hence, the covering problem for insertions and deletions is inherently asymmetric.

Although there is a rich literature on covering codes for the Hamming distance [1], as well as recent improvements for insertion/deletion error-correcting codes (e.g., [15]–[20]), much less is known about covering codes using insertions or deletions. Two key challenges are the (ir)regularity of the balls and the asymmetry of the covering problem. Insertion balls are regular, in the sense that for any $\boldsymbol{x} \in \Sigma_q^n$ and $R \geq 1$, there are exactly $\sum_{i=0}^{n+R} \binom{n+R}{i}(q-1)^i$ words of length $n+R$ obtainable by inserting R symbols into $\boldsymbol{x}$ (cf. [21]). In contrast, deletion balls are irregular, and their sizes depend on many properties of their center, such as the number of runs. In fact, a tractable exact formula remains unknown for the size of the deletion balls with radius three or greater. This irregularity and lack of an explicit formula for the ball size mean that, compared to the Hamming distance, it is inherently more challenging to derive bounds on the minimum covering code size, even asymptotically.

In some cases, we caninfer results on covering codes from the theory of error-correcting codes. This is due to the existence of *perfect* error-correcting codes, for which the balls of radius R around all codewords are not only distinct but also cover each word once. For example, the Varshamov-Tenengolts (VT) code is a perfect binary

Manuscript received November 15, 2019; accepted February 24, 2020. Date of publication April 6, 2020; date of current version May 20, 2021. This work was supported in part by the European Research Council through the EU's Horizon 2020 Research and Innovation Programme under Grant 801434, in part by NSF under Grant CCF-BSF-1619053, and in part by the United States-Israel BSF under Grant 2015816. The work of Andreas Lenz was supported by the German–American Fulbright Commission for funding the visit to UCSD. The work of Eitan Yaakobi was supported by the Center for Memory and Recording Research at UCSD. *(Corresponding author: Andreas Lenz.)*

Andreas Lenz is with the Institute for Communications Engineering, Technische Universität München, 80333 Munich, Germany (e-mail: andreas.lenz@mytum.de).

Cyrus Rashtchian is with the Computer Science and Engineering Department, University of California at San Diego, La Jolla, CA 92093 USA, and also with the Qualcomm Institute, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: crashtchian@eng.ucsd.edu).

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

Eitan Yaakobi is with the Computer Science Department, Technion–Israel Institute of Technology, Haifa 3200003, Israel (e-mail: yaakobi@cs.technion.ac.il).

Communicated by R. Gabrys, Guest Editor for the Special Issue: "From Deletion-Correction to Graph Reconstruction: In Memory of Vladimir I. Levenshtein."

Digital Object Identifier 10.1109/TIT.2020.2985691

0018-9448 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

TABLE I

UPPER AND LOWER BOUNDS FOR COVERING CODES $\mathcal{C} \subseteq \Sigma_q^n$ USING SUBSTITUTIONS, INSERTIONS, AND DELETIONS. WE LET c DENOTE A UNIVERSAL CONSTANT. WE DENOTE THE SIZE OF A RADIUS-R HAMMING BALL BY $V_{\mathsf{H}}^q(n,R) = \sum_{i=0}^{R} \binom{n}{i}(q-1)^i$, AND THE SIZE OF A RADIUS-R INSERTION BALL BY $V_{\mathsf{I}}^q(n,R) = \sum_{i=0}^{R} \binom{n+R}{i}(q-1)^i$. ENTRIES MARKED WITH "$(\infty)$" ARE ASYMPTOTIC RESULTS FOR FIXED R AND LARGE n, WHERE A FACTOR OF $1 \pm o(1)$ HAS BEEN OMITTED FOR READABILITY

Covering Code Type	Existence Size		Lower Bound		Reference
1-substitution	$\frac{q^n}{(q-1)n+1}$	(∞)	$\frac{q^n}{(q-1)n+1}$		[25]
R-substitution	$\frac{cR\log R \cdot q^n}{V_{\mathsf{H}}^q(n,R)}$	(∞)	$\frac{q^n}{V_{\mathsf{H}}^q(n,R)}$		[26]
1-insertion	$\frac{7 \cdot q^{n+1}}{(n+1)(q-1)+1}$		$\frac{q^{n+1}}{(n+1)(q-1)+1}$		Theorem 6, Theorem 1
R-insertion	$\frac{cR\log R \cdot q^{n+R}}{V_{\mathsf{I}}^q(n,R)}$	(∞)	$\frac{q^{n+R}}{V_{\mathsf{I}}^q(n,R)}$		Theorem 9, Theorem 1
1-deletion (binary)	$\frac{2^n}{n+1}$		$\frac{2^n}{n+1}$	(∞)	[27], [28]
1-deletion	$\frac{q^n}{(n+1)\lfloor q/2 \rfloor}$		$\frac{q^n(n-2)}{(q-1)n(n+1)}$		Theorem 5, Theorem 3
R-deletion	$\frac{cR\log R \cdot q^n R!}{n^R(q-1)^R}$	(∞)	$\frac{q^n R!}{n^R(q-1)^R}$	(∞)	Theorem 14, Theorem 3

single-deletion-correcting code [22]. It is known that the VT code is the largest single-deletion-correcting code for $n \leq 14$ [23], and this is conjectured to be true for $n > 14$ (see Sloane [24, Conj. 2.6]). This conjecture however remains open. Nevertheless, since the VT code is indeed a perfect single-deletion-correcting code, it is also a single-deletion-covering code.

While it has been shown that an R-deletion-correcting code is equivalent to an R-insertion-and-deletion-correcting code [29], this property does not hold for the case of covering codes. This means that the VT codes are not single-insertion-covering codes and thus also not perfect codes for correcting a single insertion. In fact, it has been shown that the only perfect single-insertion-correcting codes are binary and have length two [22]. Therefore, the best possible size of a single-insertion-covering code is unknown, and constructing optimal covering codes in this case is a highly non-trivial problem, which we address in this paper.

Afrati *et al.* have studied covering codes for insertions and deletions, motivated by designing MapReduce algorithms for similarity joins under the Levenshtein distance [27], [30]. They show the existence of single-insertion-covering codes with size $O(\frac{q^n \log n}{n})$, while they prove a lower bound stating that such codes must have at least $\frac{q^n}{(q-1)n+1}$ codewords. As one of our contributions, we certify that the lower bound is nearly tight by showing the existence of single-insertion-covering codes with $\Theta(\frac{q^n}{n})$ codewords. Afrati *et al.* also provide explicit constructions, albeit using more codewords [27]. They construct single-insertion-covering codes of size $O(q^{n-2})$ and double-insertion-covering codes of size $O(q^{n-3})$. Afrati *et al.* also provide an explicit construction of a single-deletion-covering code with q^n/n codewords. Finally, we note that tensorization arguments may be used to build radius-R covering codes from radius-one codes [1], but this approach introduces a factor of $R^{O(R)}$ with respect to the sphere covering lower bound, leading to larger covering codes than desired.

We mention that covering codes for insertions or deletions are similar in spirit to asymmetric covering codes for substitutions [31]. There, the goal is to find a code $\mathcal{C} \subseteq \{0,1\}^n$ such that the Hamming distance of any word $\boldsymbol{y} \in \{0,1\}^n$ is at most R from some $\boldsymbol{c} \in \mathcal{C}$, while satisfying $\boldsymbol{y} \leq \boldsymbol{c}$ in the binary partial order. Asymmetric covering codes are related to the Erdós-Hanani conjecture on hypergraph coverings, which has been resolved affirmatively by Erdós and Hanani [32] and Rödl [33]. For more information on covering codes, we refer the reader to the book by Cohen *et al.* [1].

A. Our Results

We provide new upper and lower bounds on the minimum size of insertion-covering and deletion-covering codes. We primarily consider the size of such codes for fixed alphabet size q and covering radius R. Table I summarizes our results. The bounds are stated separately for $R = 1$ and general $R \geq 1$ because we obtain tighter bounds in the former case. The first two rows of Table I also recap the best known bounds for substitution-covering codes.[1]

For R-insertion-covering codes, the goal is to cover words of length $n + R$ using words of length n. In the case of $R = 1$, we provide nearly matching upper and lower bounds that differ only by a factor of seven. This improves upon the

[1]We often refer to Hamming distance covering codes as R-substitution-covering codes for consistency.

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

upper bound result of Afrati *et al.* [27] by a $\Theta(\log n)$ factor. For $R \geq 2$ insertions, we prove that R-insertion-covering codes exist with size that is off by a factor of $O(R \log R)$ from the lower bound (the dependence on the dimension n and alphabet size q are optimal). We remark that the gap between upper and lower bounds matches the state-of-the-art for R-substitution-covering codes [26], and it seems beyond our current techniques to obtain a tighter bound.

For the case of R-deletion-covering codes, the goal is to cover words of length $n - R$ by deleting R symbols from codewords of length n. First, we provide a new lower bound on the minimum size of R-deletion-covering codes. Then, for words over a q-ary alphabet with $q > 2$, we provide a new explicit construction of single-deletion-covering codes, where the number of codewords is within a factor of two from optimal. Finally, for $R \geq 2$ deletions, we prove that R-deletion-covering codes exist with size that is tight up to a factor of $O(R \log R)$ compared to the lower bound.

We note that our upper bounds for R insertions (resp. R deletions) will depend upon the size of covering codes for a single insertion (resp. single deletion). In particular, establishing a better upper bound for a single insertion/deletion would immediately lead to smaller codes for radii $R > 1$.

Perhaps surprisingly, our existential bounds for insertion and deletion covering codes with radii $R \geq 2$ have the same quantitative overhead as the best known bounds for substitution-covering codes. One reason for this apparent similarity is that our proof techniques are inspired by arguments for substitutions. We observe that the previous arguments are not specific to Hamming distance, but rather, they hold for any space on words that (i) satisfies mild conditions on the distribution of the respective covering ball sizes and (ii) allows a semi-direct sum operation, where covering codes over shorter words may be combined to assemble codes over longer words. As both insertion balls and deletion balls satisfy these two criteria, the same overall proof strategy goes through, even though there are several technical differences. More generally, we believe that this proof technique may be useful for other discrete geometric spaces as well.

While in this work we analyze covering codes using either only insertions or deletions, it would be interesting to study combinations of insertions and deletions as well. We largely leave this as future work, briefly noting certain challenges that arise. Defining the covering problem already leads to flexibility because the length of the covered words may now vary, instead of being fixed. For example, if we wished to use at most R_I insertions and at most R_D deletions, then we could ask: what is the minimum number of length-n codewords that suffice to cover all words with lengths between $n - R_D$ and $n + R_I$? Our methods may extend more easily to formulations where the number of insertions and deletions are fixed individually. For instance, we could consider using exactly R_I insertions and R_D deletions to cover words of length $n + R_I - R_D$. This is in fact an interesting extension of our work, since even when $R_I = R_D = R_S/2$, it is already difficult to understand how this covering problem compares to R_S-substitution-covering codes over words of length exactly n. In general, obtaining nearly-tight upper and lower bounds for these various covering problems would require accurate estimates of the number of words obtained from combinations of insertions and deletions, which is an active area of research that is studied in different contexts as well [34]–[39].

B. Overview of Our Techniques

A lower bound on the necessary size of R-insertion-covering codes can be derived from known bounds on the number of words obtained by R insertions. For example, in the case of a single insertion, each codeword of length n covers $(n + 1)(q - 1) + 1$ distinct words. Thus, one would hope for a code of size $O(q^n/n)$. A natural approach would be to adapt or modify known codes, such as VT codes or Hamming codes. Afrati *et al.* [27] take this approach. After converting q-ary words to binary, they use $O(\log n)$ translates of a Hamming code as the basis for their single-insertion-covering code with $O(q^n \log n/n)$ codewords.[2] To remove the $\log n$ factor, we use a combination of a random construction with a careful inductive argument, inspired by the proof of Cooper, Ellis, and Kahng on asymmetric covering codes [31]. We further refine their technique to obtain explicit values on the resulting code sizes. Our upper bound for $R \geq 2$ insertions uses a generalization of the previous argument. In particular, we will roughly follow the high-level strategy used by Krivelevich, Sudakov, and Vu to obtain the current best bounds on R-substitution-covering codes [26]. The main idea is to first cover a large fraction of words by randomly sampling a subset of possible codewords (where codewords are included with probabilities depending on certain deletion ball sizes). Then, we use a direct-sum-type operation to cover the remaining words (where this operation recursively utilizes covering codes for smaller word lengths or covering radii).

Turning to deletions, we derive a precise lower bound on the minimum size of deletion-covering codes. Obtaining such a bound is somewhat involved because the sizes of deletion balls are different even for words of the same length. Hence, standard sphere-covering arguments do not apply. We use a technique, due to Applegate *et al.* [40], which enables a covering lower bound, even though the sizes of the balls are non-uniform. This technique uses an integer programming approach to analyze a weighted covering. To apply this, we construct a non-uniform weight function, depending on the sizes of deletion balls. This approach is related to bounding the size of deletion-correcting codes, which requires a generalized sphere-packing bound [23], [41]. For our upper bounds, we have already noted that VT codes provide asymptotically optimal single-deletion-covering codes for binary words. Unfortunately, known non-binary single-deletion-correcting codes [42] are not perfect, and thus, it is non-trivial to obtain good covering codes for this case. We provide a new generalization of VT codes to non-binary alphabets, and show that this leads to an explicit construction of nearly optimal covering codes for a single deletion. For $R \geq 2$ deletions, we use an analogous

[2] In fact, it is also possible to achieve the same asymptotic result by analyzing a greedy algorithm that successively adds codewords based on how many new words they cover. This algorithm yields a covering code and the Johnson-Stein-Lovász theorem [1] gives an upper bound of $O(q^n \log(n)/n)$ for the size of single-insertion-covering codes.

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

argument as for insertion-covering codes, combining random sampling with a recursive construction.

The rest of the paper is organized as follows. Section II presents the notations and the definitions of covering codes for insertions and deletions. Section III contains lower bounds on the cardinality of insertion- and deletion-covering codes. Section IV-A is dedicated to the case of single-deletion-covering codes, and Section IV-B presents existence bounds for single-insertion-covering codes. In Section V, we extend the results to multiple insertions and multiple deletions. Lastly, Section VI concludes the paper and discusses open problems.

II. Notations, Definitions, and Preliminaries

For an integer $q \geq 2$, let Σ_q denote the q-ary alphabet $\{0, 1, \ldots, q-1\}$ and $\Sigma_q^* = \bigcup_{\ell \geq 0} \Sigma_q^\ell$. We use $\mathsf{len}(\boldsymbol{x})$ to denote the length of $\boldsymbol{x}$. For $\boldsymbol{x} = (x_1, \ldots, x_n) \in \Sigma_q^n$, we let $\rho(\boldsymbol{x})$ denote the number of runs in $\boldsymbol{x}$, that is,

$$\rho(\boldsymbol{x}) := 1 + |\{1 \leq i < n : x_i \neq x_{i+1}\}|.$$

For $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_q^*$, the notation $\boldsymbol{xy}$ denotes the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$, where $\mathsf{len}(\boldsymbol{xy}) = \mathsf{len}(\boldsymbol{x}) + \mathsf{len}(\boldsymbol{y})$.

For $\boldsymbol{x} \in \Sigma_q^n$, we abbreviate the **radius-t insertion ball** obtained after exactly t insertions by $\mathsf{Ball}_\mathsf{I}^q(\boldsymbol{x}, t)$ and its size is denoted by $V_\mathsf{I}^q(\boldsymbol{x}, t)$. Similarly, the **radius-t deletion ball** obtained after exactly t deletions is denoted by $\mathsf{Ball}_\mathsf{D}^q(\boldsymbol{x}, t)$ and its size is $V_\mathsf{D}^q(\boldsymbol{x}, t)$. It is well known, see e.g. [21], that while the size of a deletion ball depends heavily on its center $\boldsymbol{x}$, insertion balls are regular. Thus, we denote by $V_\mathsf{I}^q(n, t)$ the insertion ball size of length-n words over Σ_q^n.

We will consider two sub-problems, namely covering words with only insertions or only deletions. We begin with insertions, where codewords have length n and they cover words of length $n+R$ after R insertions. We seek codes in Σ_q^n of minimum cardinality such that the union of all radius-R insertion balls around codewords contains the whole hypercube Σ_q^{n+R}. More formally, we have the following.

Definition 1: A code $\mathcal{C} \subseteq \Sigma_q^n$ is an **R-insertion-covering code**, if for every $\boldsymbol{y} \in \Sigma_q^{n+R}$, there exists a codeword $\boldsymbol{c} \in \mathcal{C}$ such that $\boldsymbol{y} \in \mathsf{Ball}_\mathsf{I}^q(\boldsymbol{c}, R)$. That is, $\bigcup_{\boldsymbol{c} \in \mathcal{C}} \mathsf{Ball}_\mathsf{I}^q(\boldsymbol{c}, R) = \Sigma_q^{n+R}$.

An R-deletion-covering code is defined similarly, but codewords cover words of length $n - R$ after R deletions.

Definition 2: A code $\mathcal{C} \subseteq \Sigma_q^n$ is an **R-deletion-covering code**, if for every $\boldsymbol{y} \in \Sigma_q^{n-R}$, there exists a codeword $\boldsymbol{c} \in \mathcal{C}$ such that $\boldsymbol{y} \in \mathsf{Ball}_\mathsf{D}^q(\boldsymbol{c}, R)$. That is, $\bigcup_{\boldsymbol{c} \in \mathcal{C}} \mathsf{Ball}_\mathsf{D}^q(\boldsymbol{c}, R) = \Sigma_q^{n-R}$.

The **insertion (resp. deletion) radius** of a code $\mathcal{C}$ is defined to be the smallest R such that $\mathcal{C}$ is an R-insertion-covering (resp. R-deletion-covering) code. We also denote by $\mathsf{K}_\mathsf{I}^q(n, R)$ (resp. $\mathsf{K}_\mathsf{D}^q(n, R)$) the smallest cardinality of an R-insertion-covering (resp. R-deletion-covering) code, of length n over Σ_q. When discussing the binary case, i.e., $q = 2$, we will typically remove q from the above notations.

III. Lower Bounds

In this section, we establish lower bounds on the size of insertion- and deletion-covering codes based on a sphere covering argument. As in the case of substitution-covering codes, the argument relies on the union bound and an upper bound on the number of words a codeword can cover. For the case of insertions, the insertion-ball size is known and independent of the ball center. The case of deletions is more challenging due to the dependence of the deletion-ball size on the ball center.

We start with the easier case of insertions.

Theorem 1: For all n and R, it holds that

$$\mathsf{K}_\mathsf{I}^q(n, R) \geq \frac{q^{n+R}}{V_\mathsf{I}^q(n, R)} = \frac{q^{n+R}}{\sum_{i=0}^{R} \binom{n+R}{i}(q-1)^i}. \tag{1}$$

Furthermore, for fixed R and large n,

$$\mathsf{K}_\mathsf{I}^q(n, R) \geq \frac{R! q^{n+R}}{n^R (q-1)^R}(1 - o(1)). \tag{2}$$

Proof: Let $\mathcal{C}$ be any R-insertion-covering code. Hence, we have that $\bigcup_{\boldsymbol{c} \in \mathcal{C}} \mathsf{Ball}_\mathsf{I}^q(\boldsymbol{c}, R) = \Sigma_q^{n+R}$. Computing the cardinalities of both sets, we obtain

$$\begin{aligned} q^{n+R} = |\Sigma_q^{n+R}| &= \Big| \bigcup_{\boldsymbol{c} \in \mathcal{C}} \mathsf{Ball}_\mathsf{I}^q(\boldsymbol{c}, R) \Big| \\ &\overset{(a)}{\leq} \sum_{\boldsymbol{c} \in \mathcal{C}} V_\mathsf{I}^q(\boldsymbol{c}, R) = |\mathcal{C}| \sum_{i=0}^{R} \binom{n+R}{i}(q-1)^i, \end{aligned}$$

where we used the union bound in inequality (a) and the fact that for any $\boldsymbol{x} \in \Sigma_q^n$ the size of the insertion balls only depends on the length of $\boldsymbol{x}$ and is given by $V_\mathsf{I}^q(n, R) = \sum_{i=0}^{R} \binom{n+R}{i}(q-1)^i$ [21]. Reading the above inequality from right to left, we obtain the bound (1).

The approximation (2) for large n is obtained by the following standard inequality,

$$\frac{q^{n+R}}{\sum_{i=0}^{R} \binom{n+R}{i}(q-1)^i} \geq \frac{R! q^{n+R}}{n^R (q-1)^R}(1 - o(1)),$$

which is proved in Proposition 18 in the appendix. □

For the case of deletions, deriving a sphere-covering lower bound is more involved due to the fact that the size of the deletion ball $\mathsf{Ball}_\mathsf{D}(\boldsymbol{x}, R)$ can be different for words of the same length. To overcome this difficulty, we use a technique due to Applegate *et al.* [40] that enables the computation of a bound even though the ball sizes are irregular. We restate the lemma from [40] in a form suited to our particular context.

Lemma 2 (cf. [40]): Let $U \subseteq \Sigma_q^*$, $T \subseteq \Sigma_q^*$ be arbitrary sets and $\mathsf{Ball} : U \mapsto \{t : t \subseteq T\}$ be an arbitrary mapping. Further let $w : T \to \mathbb{R}$ be a weight function satisfying

$$\sum_{\boldsymbol{y} \in \mathsf{Ball}(\boldsymbol{x})} w(\boldsymbol{y}) \leq 1 \tag{3}$$

for all $\boldsymbol{x} \in U$. Then any covering code $\mathcal{C} \subseteq U$, which covers T, i.e., $T = \bigcup_{\boldsymbol{c} \in \mathcal{C}} \mathsf{Ball}(\boldsymbol{c})$, satisfies

$$|\mathcal{C}| \geq \sum_{\boldsymbol{y} \in T} w(\boldsymbol{y}).$$

Even though we will derive the sphere-covering lower bound for deletions only, the above statement holds in general for any $\mathsf{Ball}(\boldsymbol{x})$, as has been proven in [40]. We note that for error-correcting codes, in which the space is replaced by Σ_q^*, the

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

analogous bound is called a generalized sphere-packing bound and has been studied in [23], [41].

The maximum possible sum weight $w(\boldsymbol{y})$ that fulfills (3) offers the best lower bound on the size of an R-deletion-covering code. Finding it requires solving a linear programming problem, as specified in Lemma 2. Here, we choose the weight function to approximate the inverse of the deletion-ball size, $w(\boldsymbol{y}) \approx V_{\mathsf{D}}^{q}(\boldsymbol{x}, R)^{-1}$, where we recall that $V_{\mathsf{D}}^{q}(\boldsymbol{x}, R) = |\mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)|$. Under the assumption that for $\boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)$ the deletion balls have approximately the same size, $V_{\mathsf{D}}^{q}(\boldsymbol{x}, R) \approx V_{\mathsf{D}}^{q}(\boldsymbol{y}, R)$, condition (3) is fulfilled with sum-weight close to 1. We will provide a rigorous derivation based upon this intuition in the following theorem.

Theorem 3: For all n and $0 < R < n$, it holds that

$$\mathsf{K}_{\mathsf{D}}^{q}(n, R) \geq q \sum_{r=1}^{n-R} \frac{(q-1)^{r-1}\binom{n-R-1}{r-1}}{\binom{r+3R-1}{R}}.$$

In particular, for $R = 1$ we get that

$$\mathsf{K}_{\mathsf{D}}^{q}(n, 1) \geq \frac{q^n(n-2)}{(q-1)n(n+1)}.$$

Furthermore, for fixed R and large n, we have

$$\mathsf{K}_{\mathsf{D}}^{q}(n, R) \geq \frac{R! q^n}{n^R (q-1)^R} \left(1 - o(1)\right).$$

Proof: We will prove the theorem using Lemma 2. We take $U = \Sigma_q^n$ and $T = \Sigma_q^{n-R}$. Define the weight function $w : T \to \mathbb{R}$ to be

$$w(\boldsymbol{y}) = \frac{1}{\max\limits_{\boldsymbol{x}: \boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} V_{\mathsf{D}}^{q}(\boldsymbol{x}, R)},$$

for all $\boldsymbol{y} \in \Sigma_q^{n-R}$. It follows that for all $\boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)$, $w(\boldsymbol{y}) \leq \frac{1}{V_{\mathsf{D}}^{q}(\boldsymbol{x}, R)}$, which implies that w satisfies (3), since

$$\sum_{\boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} w(\boldsymbol{y}) \leq \sum_{\boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} \frac{1}{V_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} = 1$$

for all $\boldsymbol{x} \in \Sigma_q^n$. By Lemma 2 we obtain a lower bound on the cardinality of any R-deletion-covering code $\mathcal{C}$, as follows.

$$\begin{aligned} |\mathcal{C}| &\geq \sum_{\boldsymbol{y} \in \Sigma_q^{n-R}} w(\boldsymbol{y}) = \sum_{\boldsymbol{y} \in \Sigma_q^{n-R}} \frac{1}{\max\limits_{\boldsymbol{x}: \boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} V_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} \\ &\overset{(a)}{\geq} \sum_{\boldsymbol{y} \in \Sigma_q^{n-R}} \frac{1}{\max\limits_{\boldsymbol{x}: \boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)} \binom{\rho(\boldsymbol{x})+R-1}{R}}, \end{aligned}$$

where in (a) we used the fact that the size of the radius-R deletion ball is at most $V_{\mathsf{D}}^{q}(\boldsymbol{x}, R) \leq \binom{\rho(\boldsymbol{x})+R-1}{R}$ [29]. Moreover, for all $\boldsymbol{y} \in \mathsf{Ball}_{\mathsf{D}}^{q}(\boldsymbol{x}, R)$ we have that $\rho(\boldsymbol{x}) \leq \rho(\boldsymbol{y}) + 2R$, since each deletion can eliminate at most two runs. Thus,

$$|\mathcal{C}| \geq \sum_{\boldsymbol{y} \in \Sigma_q^{n-R}} \frac{1}{\binom{\rho(\boldsymbol{y})+3R-1}{R}} \overset{(b)}{=} q \sum_{r=1}^{n-R} \frac{(q-1)^{r-1}\binom{n-R-1}{r-1}}{\binom{r+3R-1}{R}},$$

where in (b) we used the fact that the number of words $\boldsymbol{x} \in \Sigma_q^n$ with $\rho(\boldsymbol{x}) = r$ runs is given by $q(q-1)^{r-1}\binom{n-1}{r-1}$ [29], and we combined terms corresponding to $\boldsymbol{y} \in \Sigma_q^{n-R}$ with $\rho(\boldsymbol{y}) = r$. This concludes the proof for arbitrary R.

For $R = 1$, we can use the refinement $V_{\mathsf{D}}^{q}(\boldsymbol{x}, 1) = \rho(\boldsymbol{x})$ and obtain, using the same arguments as above,

$$\begin{aligned} |\mathcal{C}| &\geq q \sum_{r=1}^{n-1} \frac{(q-1)^{r-1}\binom{n-2}{r-1}}{r+2} \\ &= q \sum_{r=1}^{n-1} \frac{(q-1)^{r-1}(n-2)!}{(r-1)!(n-r-1)!(r+2)} \\ &= \frac{q}{(n-1)n(n+1)} \sum_{r=1}^{n-1} (q-1)^{r-1} r(r+1) \binom{n+1}{r+2} \\ &\overset{(c)}{=} \frac{q}{(q-1)^2(n-1)n(n+1)} \sum_{r=3}^{n+1} (q-1)^{r-1}(r^2-3r+2)\binom{n+1}{r}, \end{aligned}$$

where (c) follows from a shift of the variable r inside the sum. Using the equalities

$$\sum_{i=0}^{m} \binom{m}{i} x^{i-1} = (x+1)^m/x, \sum_{i=0}^{m} i\binom{m}{i} x^{i-1} = m(x+1)^{m-1},$$

and

$$\sum_{i=0}^{m} i(i-1)\binom{m}{i} x^{i-1} = m(m-1)x(1+x)^{m-1},$$

we obtain by standard transformations after a few steps

$$\mathsf{K}_{\mathsf{D}}^{q}(n, 1) \geq \frac{q^n(n-2)}{(q-1)n(n+1)}.$$

Finally, it remains to derive the asymptotic bound for fixed R and large n. We will use the following inequality,

$$q \sum_{r=1}^{n-R} \frac{(q-1)^{r-1}\binom{n-R-1}{r-1}}{\binom{r+3R-1}{R}} \geq \frac{R! q^n}{n^R (q-1)^R} \left(1 - o(1)\right),$$

whose standard, but rather technical, proof is deferred to Proposition 19 in the appendix. □

IV. Single-Insertion/Deletion-Covering Codes

Having lower bounds on the sizes of R-insertion- and R-deletion-covering codes in hand, we now prove existence of these codes for single insertions and deletions for both binary and non-binary alphabets. Surprisingly, finding covering codes for insertions is harder than for deletions, which is in sharp contrast with error-correcting codes, which have been proven to be equivalent for insertion and deletion errors [29]. For the case of deletions, we prove the existence of codes using explicit constructions. For the case of insertions, due to the lack of small explicit constructions, we resort to proving the existence of codes based on a random construction and a recursive construction.

A. Single-Deletion-Covering Codes

Let us recall the definition of the well-known Varshamov-Tenengolts (VT) [28] codes and their role as single-deletion-covering codes.

Definition 3: For all n and $0 \leq a \leq n$, let $\mathcal{C}_{\mathsf{VT}}(n; a) \subseteq \{0,1\}^n$ be the **Varshamov-Tenengolts code**

$$\mathcal{C}_{\mathsf{VT}}(n; a) = \left\{ \boldsymbol{c} \in \{0,1\}^n : \sum_{i=1}^{n} i c_i \equiv a \bmod (n+1) \right\}.$$

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

It is well known [22] that for all n and a the Varshamov-Tenengolts code $\mathcal{C}_{\mathsf{VT}}(n;a)$ is a perfect code under deletions, that is, $\bigcup_{\boldsymbol{c}\in\mathcal{C}_{\mathsf{VT}}(n;a)} \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{c},1) = \{0,1\}^{n-1}$. The next corollary is a direct result of this important property.

Corollary 4: For all $n \geq 1$ it holds that

$$\mathsf{K}_{\mathsf{D}}(n,1) \leq \frac{2^n}{n+1}.$$

It is also known that the largest (resp. smallest) of the VT codes is achieved for $a = 0$ (resp. $a = 1$) [43]. Hence, while for deletion-correcting codes it is common to choose the code $\mathcal{C}_{\mathsf{VT}}(n;0)$, for the purpose of minimizing the size of single-deletion-covering codes, one should choose the code $\mathcal{C}_{\mathsf{VT}}(n;1)$.

Unfortunately, the same property does not hold for insertions, i.e., the VT code is not a perfect code for insertions. In fact, this can be verified by simple counting arguments using the VT code size and the single-insertion ball size: as shown in Theorem 1, a lower bound on the size of any single-insertion-covering code is $2^{n+1}/(n+2)$, which is roughly twice the size of the VT codes. It can further be seen that, while the tasks of correcting a fixed number of insertions, deletions, or a combination of insertions and deletions are all equivalent [44], this sort of equivalence does not extend to covering codes. This makes the problem of finding good single-insertion-covering codes an intriguing question that will be addressed in Section IV-B.

VT codes have a non-binary extension, presented by Tenengolts in [42], which can correct a single deletion in the non-binary case. However, this family of codes is no longer perfect. In fact, their guaranteed size is $\frac{q^n}{qn}$, while the upper bound on a single-deletion-correcting code is approximately $\frac{q^n}{(q-1)n}$. This is also roughly the lower bound on a non-binary single-deletion-covering code we derived in Theorem 3, which confirms that these codes are not perfect and, therefore, are not single-deletion-covering codes. Our main result in this section is another non-binary extension of the binary VT codes, which we will show does satisfy the single-deletion covering property.

For an integer m, we denote by $(m)_2$ the value of $(m \bmod 2)$ and for a vector $\boldsymbol{x} = (x_1,\ldots,x_n)$, let $(\boldsymbol{x})_2 = ((x_1)_2,\ldots,(x_n)_2)$.

Definition 4: For all positive n, $q \geq 2$, $0 \leq a \leq n$, and $0 \leq b < \lfloor q/2 \rfloor$, let $\mathcal{C}^q_{\mathsf{NBVT}}(n;a,b) \subseteq \Sigma_q^n$ be the code

$$\mathcal{C}^q_{\mathsf{NBVT}}(n;a,b) = \Bigg\{ \boldsymbol{c} \in \Sigma_q^n \,:\, \sum_{i=1}^{n} i(c_i)_2 \equiv a \bmod (n+1), \sum_{i=1}^{n} \left\lfloor \frac{c_i}{2} \right\rfloor \equiv b \bmod \left(\left\lfloor \frac{q}{2} \right\rfloor \right) \Bigg\}.$$

The following theorem proves that the code $\mathcal{C}^q_{\mathsf{NBVT}}(n;a,b)$ is indeed a non-binary single-deletion-covering code.

Theorem 5: For all positive n, $q \geq 2$, $0 \leq a \leq n$, and $0 \leq b < \lfloor q/2 \rfloor$, the code $\mathcal{C}^q_{\mathsf{NBVT}}(n;a,b)$ is a single-deletion-covering code. Furthermore,

$$\mathsf{K}^q_{\mathsf{D}}(n,1) \leq \frac{q^n}{(n+1)\lfloor q/2 \rfloor}.$$

Proof: Let $\boldsymbol{x} = (x_1,\ldots,x_{n-1}) \in \Sigma_q^{n-1}$. Since the binary VT code $\mathcal{C}_{\mathsf{VT}}(n;a)$ is a covering code, it follows that there exist $1 \leq i \leq n$ and a binary value d such that

$$(x_1,\ldots,x_{i-1},d,x_i,\ldots,x_{n-1})_2 \in \mathcal{C}_{\mathsf{VT}}(n;a).$$

Let

$$s = \left(b - \sum_{i=1}^{n-1} \left\lfloor \frac{x_i}{2} \right\rfloor \right) \bmod \left(\left\lfloor \frac{q}{2} \right\rfloor \right).$$

Then, it holds that

$$\boldsymbol{c} = (x_1,\ldots,x_{i-1},2s+d,x_i,\ldots,x_{n-1}) \in \mathcal{C}^q_{\mathsf{NBVT}}(n;a,b),$$

and $\boldsymbol{x} \in \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{c},1)$. □

Lastly, we note that this construction improves upon the construction in [27],[3] which provides single-deletion-covering codes of size q^n/n.

B. Single-Insertion-Covering Codes

In this section, we study single-insertion-covering codes. Our main result is stated in the following theorem.

Theorem 6: For all $n \geq 1$ it holds that

$$\mathsf{K}^q_{\mathsf{I}}(n,1) \leq \mu_{\mathsf{I}} \frac{q^{n+1}}{(n+1)(q-1)+1},$$

where $\mu_{\mathsf{I}} \leq 7$.

Note that our result is stated as a fraction of the sphere-covering lower bound in Theorem 1 and implies that the size of optimal single-insertion covering codes is at most a factor of 7 from the theoretical lower limit. Our proof is inspired by and follows the strategy of the existential construction of asymmetric covering codes due to Cooper *et al.* [31]. The argument proceeds in two main steps. First, we use a random subset $S \subseteq \Sigma_q^{n_1}$ of an appropriate size to cover all but a small fraction of words $T \subseteq \Sigma_q^{n_1+1}$ with a single insertion. (This is analogous to the patched covering code in [31].) Then, we "fix up" the set S using a "good" single-insertion-covering code to generate a covering code of larger codeword length. By picking the size of S and T appropriately and using good codes inductively, we show that we will not have to pay too much in efficiency in this process.

We begin by introducing the set operation that will be used in the "fixing up" operation. The main utility of this tensorization is that it allows us to handle the uncovered words in an efficient manner.

Lemma 7: Let $S \subseteq \Sigma_q^{n_1}, T \subseteq \Sigma_q^{n_1+1}$ be such that S covers $\Sigma_q^{n_1+1} \setminus T$ with a single insertion. Let $\mathcal{C}_{n_2} \subseteq \Sigma_q^{n_2}$ be a single-insertion-covering code. Then, the code

$$(S \otimes \Sigma_q^{n_2+1}) \cup (T \otimes \mathcal{C}_{n_2})$$

is a single-insertion-covering code of length n_1+n_2+1 and of size at most $|S| \cdot q^{n_2+1} + |T| \cdot |\mathcal{C}_{n_2}|$, where $A \otimes B = \{\boldsymbol{ab} : \boldsymbol{a} \in A, \boldsymbol{b} \in B\}$ is the tensor product of two sets and $\boldsymbol{ab}$ is the concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$

[3]The result is stated in Corollary 5.5 in [27]. However, note that the authors of this paper refer to deletion-covering codes as insertion-covering codes and the result is stated over length-$(n+1)$ codes.

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

Proof: Consider any $\boldsymbol{xy} \in \Sigma_q^{n_1+n_2+2}$, where $\mathsf{len}(\boldsymbol{x}) = n_1+1$ and $\mathsf{len}(\boldsymbol{y}) = n_2+1$. We consider two cases. If $\boldsymbol{x}$ is covered by $\boldsymbol{s} \in S$, then $\boldsymbol{xy}$ is covered by $\boldsymbol{sy} \in S \otimes \Sigma_q^{n_2+1}$. Otherwise, $\boldsymbol{x} \in T$. In this case, let $\boldsymbol{c} \in \mathcal{C}_{n_2}$ be the word covering $\boldsymbol{y} \in \Sigma_q^{n_2+1}$. Then, $\boldsymbol{xy}$ is covered by $\boldsymbol{xc}$, and $\boldsymbol{xc} \in T \otimes \mathcal{C}_{n_2}$. The size of the code directly follows from the definition of the tensorization and the union bound. □

We next find a suitable (S,T) pair by randomly selecting the subset S. The words in S are non-uniformly sampled from $\Sigma_q^{n_1}$, which will reduce the overall code size by a constant factor compared to uniform sampling. The intuitive motivation for this is that some words in $\Sigma_q^{n_1+1}$ are harder to cover because their single-deletion balls are smaller. Non-uniform sampling ensures that the words in S cover words in $\Sigma_q^{n_1+1}$ in a more equitable fashion.

The following lemma provides a bound on the sizes of S and T. Although we could bound the sizes of S and T directly, the formulation in the lemma scales the size of the uncovered set T by $\mu_{\mathsf{I}}/V_{\mathsf{I}}^q(n_2,1)$ because this is the factor saved by the use of induction later in the construction.

Lemma 8: For all $n \geq 1$ there exist integers n_1, n_2 with $n_1+n_2+1 = n$ and sets $S \subseteq \Sigma_q^{n_1}, T \subseteq \Sigma_q^{n_1+1}$ such that S covers $\Sigma_q^{n_1+1} \setminus T$, that is, $T = \Sigma_q^{n_1+1} \setminus \bigcup_{\boldsymbol{s}\in S} \mathsf{Ball}_{\mathsf{I}}^q(\boldsymbol{s},1)$, while the sizes of S and T satisfy

$$|S| + \frac{\mu_{\mathsf{I}}|T|}{V_{\mathsf{I}}^q(n_2,1)} \leq \frac{\mu_{\mathsf{I}} q^{n_1+1}}{V_{\mathsf{I}}^q(n,1)},$$

where $\mu_{\mathsf{I}} \leq 7$.

Proof: For $n \leq \frac{q\mu_{\mathsf{I}}-q}{q-1}$, the statement is fulfilled by $S = \Sigma_q^{n_1}$ and $T = \emptyset$. Assume that $n > \frac{q\mu_{\mathsf{I}}-q}{q-1}$. We prove the existence of an (S,T) pair with sizes satisfying the lemma by means of a random construction. Include each word $\boldsymbol{x} \in \Sigma_q^{n_1}$ in S with probability $q_{\boldsymbol{x}} \stackrel{\text{def}}{=} cV_{\mathsf{D}}^q(\boldsymbol{x},1)^{-1}$ for a constant $c > 0$ to be set later. Let T be all remaining words that are not covered by S, i.e., $T = \Sigma_q^{n_1+1} \setminus \bigcup_{\boldsymbol{s}\in S} \mathsf{Ball}_{\mathsf{I}}^q(\boldsymbol{s},1)$.

For a fixed word $\boldsymbol{y} \in \Sigma_q^{n_1+1}$, we have that $\boldsymbol{y}$ is covered by S unless all of the words covering $\boldsymbol{y}$ fail to be included in S. The number of words that can cover $\boldsymbol{y}$ is exactly $V_{\mathsf{D}}^q(\boldsymbol{y},1)$, the size of the single-deletion ball. Note that $V_{\mathsf{D}}^q(\boldsymbol{y},1) = \rho(\boldsymbol{y})$ [29], and observe that for any $\boldsymbol{x} \in \mathsf{Ball}_{\mathsf{D}}^q(\boldsymbol{y},1)$ the number of runs cannot increase as a result of the deletion, i.e., $\rho(\boldsymbol{x}) \leq \rho(\boldsymbol{y})$. Hence, $q_{\boldsymbol{x}} = cV_{\mathsf{D}}^q(\boldsymbol{x},1)^{-1} = c\rho(\boldsymbol{x})^{-1} \geq c\rho(\boldsymbol{y})^{-1} \stackrel{\text{def}}{=} q_{\boldsymbol{y}}$. We bound the probability that S misses $\boldsymbol{y}$ as follows:

$$\begin{aligned}\mathrm{P}[\boldsymbol{y} \text{ is uncovered}] &= \prod_{\boldsymbol{x}\in\mathsf{Ball}_{\mathsf{D}}^q(\boldsymbol{y},1)} (1-q_{\boldsymbol{x}}) \\ &\stackrel{(a)}{\leq} \prod_{\boldsymbol{x}\in\mathsf{Ball}_{\mathsf{D}}^q(\boldsymbol{y},1)} (1-q_{\boldsymbol{y}}) = (1-q_{\boldsymbol{y}})^{V_{\mathsf{D}}^q(\boldsymbol{y},1)},\end{aligned}$$

where (a) uses that $q_{\boldsymbol{x}} \geq q_{\boldsymbol{y}}$, as discussed above.

We now compute the expected weighted size of S and T under the above random selection.

$$\begin{aligned}W &\stackrel{\text{def}}{=} \mathrm{E}\left[|S| + \frac{\mu_{\mathsf{I}}|T|}{V_{\mathsf{I}}^q(n_2,1)}\right] = \mathrm{E}[|S|] + \frac{\mu_{\mathsf{I}}\mathrm{E}[|T|]}{V_{\mathsf{I}}^q(n_2,1)} \\ &= \sum_{\boldsymbol{x}\in\Sigma_q^{n_2}} q_{\boldsymbol{x}} + \frac{\mu_{\mathsf{I}}}{V_{\mathsf{I}}^q(n_2,1)} \sum_{\boldsymbol{y}\in\Sigma_q^{n_2+1}} \mathrm{P}[\boldsymbol{y} \text{ is uncovered}].\end{aligned}$$

Plugging in the bound for $\mathrm{P}[\boldsymbol{y}$ is uncovered$]$ and recalling that $q_{\boldsymbol{x}} = c\rho(\boldsymbol{x})^{-1}$, we obtain

$$W \leq \sum_{\boldsymbol{x}\in\Sigma_q^{n_1}} \frac{c}{\rho(\boldsymbol{x})} + \frac{\mu_{\mathsf{I}}}{V_{\mathsf{I}}^q(n_2,1)} \sum_{\boldsymbol{y}\in\Sigma_q^{n_1+1}} (1-q_{\boldsymbol{y}})^{V_{\mathsf{D}}^q(\boldsymbol{y},1)}.$$

It is well-known [29] that the number of words $\boldsymbol{x} \in \Sigma_q^{n_1}$ with $\rho(\boldsymbol{x}) = r$ is given by $q\binom{n_1-1}{r-1}(q-1)^{r-1}$, which allows us to group terms in the first sum by $\rho(\boldsymbol{x}) = r$. Using that $1-z \leq \mathrm{e}^{-z}$ for all $z \in \mathbb{R}$, we have that $(1-q_{\boldsymbol{y}})^{V_{\mathsf{D}}^q(\boldsymbol{y},1)} \leq \mathrm{e}^{-c}$ and we bound W by

$$\begin{aligned}W &\leq qc\sum_{r=1}^{n_1} \frac{\binom{n_1-1}{r-1}(q-1)^{r-1}}{r} + \frac{\mu_{\mathsf{I}}}{V_{\mathsf{I}}^q(n_2,1)} \sum_{\boldsymbol{y}\in\Sigma_q^{n_1+1}} \mathrm{e}^{-c} \\ &\stackrel{(b)}{=} qc\sum_{r=1}^{n_1} \frac{(n_1-1)!(q-1)^{r-1}}{r!(n_1-r)!} + \frac{q^{n_1+1}\mu_{\mathsf{I}}\mathrm{e}^{-c}}{V_{\mathsf{I}}^q(n_2,1)} \\ &\stackrel{(c)}{=} \frac{qc}{n_1(q-1)}\sum_{r=1}^{n_1}\binom{n_1}{r}(q-1)^r + \frac{q^{n_1+1}\mu_{\mathsf{I}}\mathrm{e}^{-c}}{V_{\mathsf{I}}^q(n_2,1)},\end{aligned}$$

where in equality (b) and (c) we used the definition of the binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Finally, we use the binomial identity $\sum_{k=0}^n \binom{n}{k}x^k = (1+x)^n$ and obtain

$$\begin{aligned}W &\leq \frac{cq^{n_1+1}}{n_1(q-1)} + \frac{q^{n_1+1}\mu_{\mathsf{I}}\mathrm{e}^{-c}}{V_{\mathsf{I}}^q(n_2,1)} \\ &= \frac{\mu_{\mathsf{I}}q^{n_1+1}}{V_{\mathsf{I}}^q(n,1)}\left(\frac{cV_{\mathsf{I}}^q(n,1)}{\mu_{\mathsf{I}}n_1(q-1)} + \frac{V_{\mathsf{I}}^q(n,1)e^{-c}}{V_{\mathsf{I}}^q(n_2,1)}\right).\end{aligned}$$

Abbreviating the term in round brackets by γ and setting $n_1 = \lfloor \beta n \rfloor$ for some $0 \leq \beta \leq 1$, we derive the upper bound

$$\begin{aligned}\gamma &\stackrel{\text{def}}{=} \frac{cV_{\mathsf{I}}^q(n,1)}{\mu_{\mathsf{I}}n_1(q-1)} + \frac{V_{\mathsf{I}}^q(n,1)e^{-c}}{V_{\mathsf{I}}^q(n_2,1)} \\ &\leq \frac{V_{\mathsf{I}}^q(n,1)}{n(q-1)}\left(\frac{cn}{\mu_{\mathsf{I}}n_1} + \frac{e^{-c}n}{n-n_1}\right) \\ &\stackrel{(d)}{\leq} \frac{\mu_{\mathsf{I}}}{\mu_{\mathsf{I}}-1}\left(\frac{cn}{\mu_{\mathsf{I}}\lfloor\beta n\rfloor} + \frac{e^{-c}n}{n-\lfloor\beta n\rfloor}\right),\end{aligned}$$

where we used in equality (d) that $V_{\mathsf{I}}^q(n,1)/(n(q-1))$ is monotonically decreasing in n and thus $V_{\mathsf{I}}^q(n,1)/(n(q-1)) \leq \mu_{\mathsf{I}}/(\mu_{\mathsf{I}}-1)$ for all $n > (q\mu_{\mathsf{I}}-q)/(q-1)$. Note that this bound is convenient to handle as it is independent of q. To conclude, we find the smallest μ_{I} such that there exists some $c > 0$ and $0 \leq \beta \leq 1$ for which $\gamma \leq 1$ for all $n > (q\mu_{\mathsf{I}}-q)/(q-1)$. A quick computer search yields that $\mu_{\mathsf{I}} = 7$, $c = 3$ and $\beta = \frac{3}{4}$ fulfills this requirement. By definition of the random sets S and T, any realization of them will have the desired property that S covers $\Sigma_q^{n_1+1} \setminus T$. As the expected weighted size W is at most $\frac{\mu_{\mathsf{I}}q^{n_1+1}}{V_{\mathsf{I}}^q(n,1)}$, it follows that there exists an (S,T) pair satisfying the desired bound. □

Putting everything together, we prove Theorem 6 for single-insertion-covering codes.

Proof of Theorem 6: We proceed by induction on n. As the base case, for all $n \leq \frac{q\mu_{\mathsf{I}}-q}{q-1}$, it suffices to take $\mathcal{C}_n = \Sigma_q^n$. Assume now that the statement is correct for all lengths up to $n-1$, so that there exist codes $\mathcal{C}_{n_2}$ with size at most $\mu_{\mathsf{I}}q^{n_2+1}/V_{\mathsf{I}}^q(n_2,1)$ for all $1 \leq n_2 \leq n-1$. Let $n_1+n_2+1 = n$

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

and $S \subseteq \Sigma_q^{n_1}$ and $T \subseteq \Sigma_q^{n_1+1}$ denote sets guaranteed by Lemma 8. Note that clearly $n_2 < n$ in Lemma 8, which will be useful later. As these sets S and T fulfill the requirement of Lemma 7, we define

$$\mathcal{C}_n = (S \otimes \Sigma_q^{n_2+1}) \cup (T \otimes \mathcal{C}_{n_2}),$$

and we have that there exists a single-insertion-covering code $\mathcal{C}_n \subseteq \Sigma_q^n$ of size

$$|C_n| \leq q^{n_2+1}|S| + |T| \cdot |\mathcal{C}_{n_2}| \overset{(e)}{\leq} q^{n_2+1}\left(|S| + \frac{\mu_{\mathsf{I}}}{V_{\mathsf{I}}^q(n_2,1)}|T|\right),$$

where in (e) we used the existence of a covering code of length $n_2 < n$ and size $\mu_{\mathsf{I}} q^{n_2+1}/V_{\mathsf{I}}^q(n_2,1)$ by the induction hypothesis. Using the existence of good sets S and T from Lemma 8, we obtain the desired bound on the code size

$$|\mathcal{C}_n| \leq q^{n_2+1}\frac{\mu_{\mathsf{I}} q^{n_1+1}}{V_{\mathsf{I}}^q(n,1)} = \frac{\mu_{\mathsf{I}} q^{n+1}}{V_{\mathsf{I}}^q(n,1)}.$$

□

Together with our existence result from Theorem 6, we can infer that the size of the smallest single-insertion-covering code lies between $q^{n+1}/V_{\mathsf{I}}^q(n,1)$ and $7q^{n+1}/V_{\mathsf{I}}^q(n,1)$ and thus is known up to a constant factor of 7.

V. Multiple-Insertion/Deletion-Covering Codes

We now turn to the discussion of multiple-insertion/deletion covering codes. We begin by defining the optimal density of insertion- and deletion-covering codes, by analogy with the notion of density often used in the context of classical covering codes.

Definition 5: For R-insertion-covering codes of length n, the **optimal density** $\mu_{\mathsf{I}}^q(n,R)$ is defined as

$$\mu_{\mathsf{I}}^q(n,R) = \frac{\mathsf{K}_{\mathsf{I}}^q(n,R)V_{\mathsf{I}}^q(n,R)}{q^{n+R}}.$$

For R-deletion-covering codes of length n, we define the **optimal density** $\mu_{\mathsf{D}}^q(n,R)$ as

$$\mu_{\mathsf{D}}^q(n,R) = \frac{\mathsf{K}_{\mathsf{D}}^q(n,R)n^R(q-1)^R}{q^n R!}.$$

Finally, for fixed R, we define the corresponding **asymptotic optimal densities** $\mu_{\mathsf{I}}^{q,*}(R)$ and $\mu_{\mathsf{D}}^{q,*}(R)$ as

$$\mu_{\mathsf{I}}^{q,*}(R) = \limsup_{n\to\infty} \mu_{\mathsf{I}}^q(n,R)$$

and

$$\mu_{\mathsf{D}}^{q,*}(R) = \limsup_{n\to\infty} \mu_{\mathsf{D}}^q(n,R).$$

Note that we define the optimal density of deletion-covering codes slightly differently than that of insertion-covering codes. This is due to the fact that for deletions, the deletion balls are non-uniform and the density is thus defined with respect to the lower bound obtained in Theorem 3 for large n. A powerful tool in building covering codes of larger radius is to take the tensor product of two short covering codes of small radius. For example, taking the tensor product of two covering codes of length n and radius 1 gives a covering code of length $2n$ and radius 2. However, a straightforward application of this technique only gives covering codes whose density is at least exponential in R. We therefore refine this technique to obtain codes that have a density that is almost linear in R. Note that in the following sections we prove our results for binary words for simplicity. The proofs for $q > 2$ are obtained by only a slight modification, which will be explained in more detail in Remark 1 at the end of Section V-B.

A. Multiple-Insertion-Covering Codes

Our main result about R-insertion-covering codes is stated in the following theorem.

Theorem 9: For any fixed $R \geq 2$ and $q \geq 2$,

$$\mu_{\mathsf{I}}^{q,*}(R) \leq \mathrm{e}(R\log R + \sqrt{2R\log R} + 1)\mu_{\mathsf{I}}^{q,*}(1).$$

Recall that according to Theorem 6, we have that $\mu_{\mathsf{I}}^{q,*}(1) \leq 7$. Before proving the theorem, we give a short outline of the proof, along with the intuition behind it. As in the proof of the upper bound for single-insertion-covering codes, we start by proving in Lemma 10 the existence of a small *almost-covering* code S, i.e., a code that covers all words in $\{0,1\}^{n+R}$ except for a small subset T. Then, in Lemmas 11 and 12, we combine this code with small covering codes to recursively build larger codes. By computing the size of the resulting codes, we can then prove Theorem 9.

The proof of the existence of small almost-covering codes is again based on a random coding argument. Since we are building covering codes for insertions, we must take into account the fact that each word $\boldsymbol{y} \in \{0,1\}^{n+R}$ is covered by a different number of potential codewords $\boldsymbol{x} \in \{0,1\}^n$. This is because the number of words that can cover $\boldsymbol{y}$ is given by $V_{\mathsf{D}}(\boldsymbol{y},R)$, which is known to depend on $\boldsymbol{y}$. In our random selection of codewords, we therefore need to favor codewords that cover words with small $V_{\mathsf{D}}(\boldsymbol{y},R)$ to ensure that each word is covered with high enough probability. Our proof follows the general idea of a recursive covering code construction presented in [26], here modified to work for insertions. In particular, we need to adapt the arguments for the random construction and the recursive combination of almost-covering codes with existing covering codes.

The following lemma gives an upper bound on the sizes of the almost-covering code S and the complement T of its coverage.

Lemma 10: For every $n \geq R$ and every positive constant $c > 0$ there exists a set $S \subseteq \{0,1\}^{n-R}$ of size at most

$$|S| \leq \frac{c2^n}{V_{\mathsf{I}}(n-R,R)}f_{n,R}$$

such that S covers $\{0,1\}^n \setminus T$ with R insertions for some set $T \subseteq \{0,1\}^n$ of size at most

$$|T| \leq \mathrm{e}^{-c}2^n,$$

for some function $f_{n,R}$ with $\lim_{n\to\infty} f_{n,R} = 1$.

Proof: We prove the lemma by choosing a random set S and computing the expected number of words that are not covered by such a random choice. Let $\mathcal{X}_i = \{\boldsymbol{x} \in \{0,1\}^{n-R} : V_{\mathsf{D}}(\boldsymbol{x},R) = i\}$ be the set of all strings of length n, which have a deletion ball size of exactly i. We construct S by

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

choosing $S = S_1 \cup S_2 \cup \cdots \cup S_m$, where $S_i \subseteq \mathcal{X}_i$ and $m \leq \binom{n}{R}$ is the maximum size of the deletion ball of any $x \in \{0,1\}^{n-R}$. Denoting $m_i = |\mathcal{X}_i|$, each S_i is a uniformly chosen random subset of $\mathcal{X}_i$ of cardinality $|S_i| = \lceil cm_i/i \rceil$, if $cm_i/i \leq m_i$, and m_i otherwise. Hereby each such subset has the same probability. By this choice of the sets $S_1, \ldots, S_m$, the probability that any $\boldsymbol{y} \in \{0,1\}^n$ is not covered by S can be bounded from above as follows. First, note that

$$\mathrm{P}[\boldsymbol{y} \text{ is uncovered}] = \prod_{i=1}^{m} \mathrm{P}[S_i \cap \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{y}, R) = \emptyset],$$

since the random sets S_i are independent. Denote by $\gamma_i = |\{\mathcal{X}_i \cap \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{y}, R)\}|$ the number of words in $\mathcal{X}_i$ which can cover $\boldsymbol{y}$. With this notation, the individual probabilities in the product can be expressed as

$$\mathrm{P}[S_i \cap \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{y}, R) = \emptyset] = \frac{\binom{m_i - \gamma_i}{|S_i|}}{\binom{m_i}{|S_i|}} = \frac{\prod_{j=0}^{|S_i|}(m_i - \gamma_i - j)}{\prod_{j=0}^{|S_i|}(m_i - j)}.$$

From the fact that $\frac{m_i - \gamma_i - j}{m_i - j} \leq \frac{m_i - \gamma_i}{m_i}$ for any $0 \leq j < m_i - \gamma_i$, we obtain

$$\begin{aligned}\mathrm{P}[\boldsymbol{y} \text{ is uncovered}] &\leq \prod_{i=1}^{m} \left(\frac{m_i - \gamma_i}{m_i}\right)^{|S_i|} = \prod_{i=1}^{m} \left(1 - \frac{\gamma_i}{m_i}\right)^{|S_i|} \\ &\overset{(a)}{\leq} \prod_{i=1}^{m} \mathrm{e}^{-\frac{\gamma_i}{m_i}|S_i|} \leq \mathrm{e}^{-\sum_{i=1}^{m} c\frac{\gamma_i}{i}},\end{aligned}$$

where we used in (a) that $1 - x \leq \mathrm{e}^{-x}$ for any $x \in \mathbb{R}$. Let $\mu(\boldsymbol{y}) = \max_{\boldsymbol{x} \in \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{y}, R)} V_{\mathsf{D}}(\boldsymbol{x}, R)$ be the maximum deletion ball size of any $\boldsymbol{x} \in \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{y}, R)$, which is obtained from $\boldsymbol{y}$ by R deletions. Since $\gamma_i = 0$ for all $i > \mu(\boldsymbol{y})$, we can bound the exponent from below by

$$\sum_{i=1}^{m} \frac{\gamma_i}{i} = \sum_{i=1}^{\mu(\boldsymbol{y})} \frac{\gamma_i}{i} \geq \sum_{i=1}^{\mu(\boldsymbol{y})} \frac{\gamma_i}{\mu(\boldsymbol{y})} = \frac{\sum_{i=1}^{\mu(\boldsymbol{y})} \gamma_i}{\mu(\boldsymbol{y})} = \frac{V_{\mathsf{D}}(\boldsymbol{y}, R)}{\mu(\boldsymbol{y})} \overset{(b)}{\geq} 1,$$

where (b) follows from the fact that $V_{\mathsf{D}}(\boldsymbol{x}, R) \leq V_{\mathsf{D}}(\boldsymbol{y}, R)$ for any $\boldsymbol{x} \in \mathsf{Ball}_{\mathsf{D}}(\boldsymbol{y}, R)$. Hence, $\mathrm{P}[\boldsymbol{y} \text{ is uncovered}] \leq \mathrm{e}^{-c}$ and the expected size of T is consequently at most $\mathrm{E}[|T|] \leq \mathrm{e}^{-c}2^n$. Thus, there must exist a set S for which $|T| \leq \mathrm{e}^{-c}2^n$.

It remains to compute the size of S. By construction

$$|S| = \sum_{i=1}^{m} |S_i| \leq cm + c\sum_{i=1}^{m} \frac{m_i}{i} = cm + \sum_{\boldsymbol{x} \in \{0,1\}^{n-R}} \frac{c}{V_{\mathsf{D}}(\boldsymbol{x}, R)}.$$

Noting that $V_{\mathsf{D}}(\boldsymbol{x}, R) \geq \binom{\rho(\boldsymbol{x}) - R}{R}$ and recalling that the number of words of length n with r runs is $2\binom{n-1}{r-1}$, we obtain

$$\begin{aligned}|S| &\leq cm + c\sum_{r=1}^{n-R} \frac{2\binom{n-R-1}{r-1}}{\max\left\{1, \binom{r-R}{R}\right\}} \\ &\leq c\binom{n}{R} + 2c\sum_{r=1}^{r^*} \binom{n-R-1}{r} + 2c\sum_{r=r^*}^{n-R} \frac{\binom{n-R-1}{r-1}}{\binom{r^*-R}{R}} \\ &\leq c\binom{n}{R} + 2c\sum_{r=1}^{r^*} \binom{n-R-1}{r} + c\frac{2^{n-R}}{\binom{r^*-R}{R}},\end{aligned}$$

where $r^* = \max\{2R, \frac{n}{2} - \sqrt{Rn\log n}\}$. For $r^* = 2R$, we can directly bound the second term by

$$2c\sum_{r=1}^{r^*} \binom{n-R-1}{r} \leq 2c\binom{n-R+r^*}{r^*} = 2c\binom{n+R}{2R}.$$

On the other hand, for $r^* = \frac{n}{2} - \sqrt{Rn\log n}$, applying Chernoff's inequality to the binomial tail, we obtain

$$\sum_{r=1}^{r^*} \binom{n-R-1}{r} \leq \sum_{r=1}^{r^*} \binom{n}{r} \leq 2^n \mathrm{e}^{-2\frac{(n/2 - r^*)^2}{n}} = \frac{2^n}{n^{2R}}.$$

Hence, the overall size of S is bounded from above by

$$|S| \leq c\frac{2^n}{V_{\mathsf{I}}(n-R, R)} f_{n,R},$$

where $f_{n,R}$ is at most

$$\begin{aligned}f_{n,R} \leq\ & \frac{\binom{n}{R}\binom{n+R}{R}}{2^n} + 2\binom{n+R}{R} \max\left\{\frac{\binom{n+R}{2R}}{2^n}, \frac{1}{n^{2R}}\right\} \\ &+ \frac{2^{-R}\binom{n+R}{R}}{\binom{r^*-R}{R}},\end{aligned}$$

where we used that $V_{\mathsf{I}}(n-R, R) \leq \binom{n+R}{R}$.

Lastly, we bound the third summand in the bound on $|S|$. For $r^* = 2R$ it is trivially bounded by $2^{-R}\binom{n+R}{R}$. When $r^* = \frac{n}{2} - \sqrt{Rn\log n}$, a quick calculation yields

$$\begin{aligned}\frac{2^{-R}\binom{n+R}{R}}{\binom{r^*-R}{R}} &\leq \frac{2^{-R}\binom{n+R}{R}R!}{(n/2 - \sqrt{Rn\log n} - 2R)^R} \\ &\overset{(a)}{\leq} \frac{2^{-R}n^R\mathrm{e}^{R^2/n}}{(\frac{n}{2})^R(1 - 2R\sqrt{R\log n}/\sqrt{n} - 4R^2/n)} \\ &= \frac{n\mathrm{e}^{R^2/n}}{n - 2R\sqrt{Rn\log n} - 4R^2},\end{aligned}$$

where in inequality (a) we used that $(1+x)^R \geq 1 + Rx$ for any $x \geq -1$. Finally we obtain for $\frac{n}{2} - \sqrt{Rn\log n} \leq 2R$,

$$f_{n,R} \leq \frac{(n+R)^{2R}}{2^n} + \frac{2(n+R)^{3R}}{2^n} + \frac{\binom{n+R}{R}}{2^R},$$

and for $\frac{n}{2} - \sqrt{Rn\log n} > 2R$,

$$f_{n,R} \leq \frac{(n+R)^{2R}}{2^n} + \frac{2}{n^R} + \frac{n\mathrm{e}^{R^2/n}}{n - 2R\sqrt{Rn\log n} - 4R^2}.$$

Here we additionally used $\binom{n}{R} \leq \binom{n+R}{R} \leq (n+R)^R$. Note that for large enough n and any fixed R, $\frac{n}{2} - \sqrt{Rn\log n} > 2R$ and it is directly verified that $\lim_{n\to\infty} f_{n,R} = 1$. □

Note that while the expression of $f_{n,R}$ looks quite involved, we are interested in its asymptotic behavior and it will only be important in the following that it approaches 1 for large n.

Lemma 11: Let $S \subseteq \{0,1\}^{n_1 - R_1}, T \subseteq \{0,1\}^{n_1}$ be such that S covers $\{0,1\}^{n_1} \setminus T$ with R_1 insertions. Denote by $\mathcal{C}_1 \subseteq \{0,1\}^{n_2 + R_1}$ an R_2-insertion-covering code of length $n_2 + R_1$ and by $\mathcal{C}_2 \subseteq \{0,1\}^{n_2}$ an R-insertion-covering code of length n_2. We have that

$$(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)$$

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

is an $R = R_1 + R_2$-insertion-covering code of length $n = n_1 + n_2$ with size at most $|S| \cdot |\mathcal{C}_1| + |T| \cdot |\mathcal{C}_2|$.

Proof: Consider any $\boldsymbol{xy} \in \{0,1\}^{n+R}$, where $\mathsf{len}(\boldsymbol{x}) = n_1$ and $\mathsf{len}(\boldsymbol{y}) = n_2 + R$. We distinguish between two cases. First consider the case where $\boldsymbol{x}$ is covered by $\boldsymbol{s} \in S$ with R insertions. Denote by $\boldsymbol{c}_1 \in \mathcal{C}_1$ the word that covers $\boldsymbol{y} \in \{0,1\}^{n_2+R}$ with R_2 insertions. Note that such a word always exists, as $\mathcal{C}_1$ is an R_2-insertion-covering code. Then $\boldsymbol{xy}$ is covered by $\boldsymbol{sc}_1 \in S \otimes \mathcal{C}_1$ with a total of $R = R_1 + R_2$ insertions. Otherwise, $\boldsymbol{x} \in T$. In this case, let $\boldsymbol{c}_2 \in \mathcal{C}_2$ be the string covering $\boldsymbol{y} \in \{0,1\}^{n_2+R}$ with R insertions. Then, $\boldsymbol{xy}$ is covered by $\boldsymbol{xc}_2$, and $\boldsymbol{xc}_2 \in T \otimes \mathcal{C}_2$. The size of the code directly follows from the union bound. □

Lemma 12: For any $n \geq R$ and $c > 0$,

$$\mu_{\mathsf{I}}(n,R) \leq c\mathrm{e}\mu_{\mathsf{I}}(n/R + R - 1, 1)\frac{(1+2R/n)^R}{1-R^2/n} f_{\frac{R-1}{R}n,R-1} + R^R \mathrm{e}^{-c}\mu_{\mathsf{I}}(n/R, R)(1+2R/n)^R.$$

Proof: Let $S \subseteq \{0,1\}^{n_1-R_1}, T \subseteq \{0,1\}^{n_1}$ with $n_1 \geq R_1$ be such that S covers $\{0,1\}^{n_1} \setminus T$ with R_1 insertions. Denote by $\mathcal{C}_1 \subseteq \{0,1\}^{n_2+R_1}$ an R_2-insertion-covering code of length $n_2 + R_1$ and by $\mathcal{C}_2 \subseteq \{0,1\}^{n_2}$ an R-insertion-covering code of length n_2, where $n_1 + n_2 = n$, $n_1 = \frac{y-1}{y}n$, $n_2 = \frac{n}{y}$, and $R_1 + R_2 = R$. We compute the size of the tensorization $(S \otimes \mathcal{C}_1) \cup (T \cup \mathcal{C}_2)$, which, by Lemma 11, is an R-insertion covering code of length n. To begin with, $|(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)| \leq |S| \cdot |\mathcal{C}_1| + |T| \cdot |\mathcal{C}_2|$. Using Lemma 10 and optimal codes $\mathcal{C}_1$ and $\mathcal{C}_2$, we can bound the size of S and T to obtain

$$|(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)| \leq \frac{c2^{n+R} f_{n_1,R_1}\mu_{\mathsf{I}}(n_2+R_1,R_2)}{V_{\mathsf{I}}(n_1-R_1,R_1)V_{\mathsf{I}}(n_2+R_1,R_2)} + \frac{\mathrm{e}^{-c}2^{n+R}\mu_{\mathsf{I}}(n_2,R)}{V_{\mathsf{I}}(n_2,R)}.$$

Since $(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)$ is a covering code of length n and covering radius R, we obtain

$$\begin{aligned}\mu_{\mathsf{I}}(n,R) &\leq \frac{cV_{\mathsf{I}}(n,R)f_{n_1,R_1}\mu_{\mathsf{I}}(n_2+R_1,R_2)}{V_{\mathsf{I}}(n_1-R_1,R_1)V_{\mathsf{I}}(n_2+R_1,R_2)} + \frac{V_{\mathsf{I}}(n,R)\mu_{\mathsf{I}}(n_2,R)}{\mathrm{e}^c V_{\mathsf{I}}(n_2,R)} \\ &\overset{(a)}{\leq} \frac{cR_1!R_2!(n+2R)^R\mu_{\mathsf{I}}(n_2+R_1,R_2)}{R!(n_1-R_1)^{R_1}(n_2+R_1)^{R_2}} f_{n_1,R_1} + \mathrm{e}^{-c}\left(\frac{n+2R}{n_2}\right)^R \mu_{\mathsf{I}}(n_2,R) \\ &\leq \frac{cn^R\mu_{\mathsf{I}}(n_2+R_1,R_2)}{n_1^{R_1}n_2^{R_2}\binom{R}{R_1}} \frac{(1+2R/n)^R}{1-R_1^2/n_1} f_{n_1,R_1} + \mathrm{e}^{-c}\left(\frac{n}{n_2}\right)^R (1+2R/n)^R \mu_{\mathsf{I}}(n_2,R),\end{aligned}$$

where in (a) we used the well-known inequalities $\binom{n+R}{R} \leq V_{\mathsf{I}}(n,R) \leq \binom{n+2R}{R}$ and $(n-R)^R/R! \leq \binom{n}{R} \leq n^R/R!$. Inserting $R_1 = R - 1$, $R_2 = 1$, and $y = R$ yields the lemma. □

With this recursive expression, we are ready to prove the theorem with the help of the following lemma.

Lemma 13 (cf. [26]): Let $(\mu_n), (\mu'_n), (a_n)$ and (b_n), $n \in \mathbb{N}$ be sequences of positive numbers with

$$\limsup_{n\to\infty} \mu'_n \leq \mu', \quad \limsup_{n\to\infty} a_n \leq a \quad \limsup_{n\to\infty} b_n \leq b$$

and

$$\mu_n \leq a_n\mu'_{n/R} + b_n\mu_{n/R},$$

where $R > 1$. Then

$$\limsup_{n\to\infty} \mu_n \leq \frac{a\mu'}{1-b}.$$

One convenient property of this lemma is that it incorporates the recursive assembly of the covering codes, without having to perform a thorough analysis of the induction start. We are now in a position to prove Theorem 9.

Proof of Theorem 9: Using Lemmas 12 and 13, we obtain $\mu_{\mathsf{I}}^*(R) \leq \frac{c\mathrm{e}}{1-R^R\mathrm{e}^{-c}}\mu_{\mathsf{I}}^*(1)$. Minimizing $\frac{c\mathrm{e}}{1-R^R\mathrm{e}^{-c}}$ over c, we can directly verify that $\min_c \frac{c\mathrm{e}}{1-R^R\mathrm{e}^{-c}} = \mathrm{e}(c_0+1)$, where c_0 is the solution to $c + 1 = \mathrm{e}^c R^{-R}$. Using standard bounds on c_0, we obtain the theorem. □

B. Multiple-Deletion-Covering Codes

Proving the existence of small covering codes for deletions follows basically the same steps as the proof for the case of insertions. However, there are some subtle differences, such as the different definition of density for deletion-covering codes. Our main result is as follows.

Theorem 14: For any fixed $R \geq 2$ and $q \geq 2$,

$$\mu_{\mathsf{D}}^{q,*}(R) \leq \mathrm{e}(R\log R + \sqrt{2R\log R} + 1)\mu_{\mathsf{D}}^{q,*}(1).$$

In particular, for $q = 2$,

$$\mu_{\mathsf{D}}^{*}(R) \leq \mathrm{e}(R\log R + \sqrt{2R\log R} + 1).$$

Note that compared to the case of insertions, we could use in Theorem 14 that for the binary case $\mu_{\mathsf{D}}^*(1) = 1$, which results in a tighter bound also for the case of $R > 1$. Since the outline of the proof for the case of deletions is similar to that of insertions, we merely state the ingredients of the proof to establish the analogy to the case of insertions. Again, we prove the theorem for the binary case, however the extension to non-binary alphabets is straightforward.

Lemma 15: For every n and R and every positive constant $c > 0$ there exist sets $S \subseteq \{0,1\}^{n+R}$ and $T \subseteq \{0,1\}^n$ with

$$|S| \leq \frac{c2^{n+R}}{\binom{n}{R}}$$

such that S covers $\{0,1\}^n \setminus T$ with R deletions for some set T of size at most

$$|T| \leq \mathrm{e}^{-c}\mathrm{e}^{\frac{V_{\mathsf{I}}(n,R)}{2^{n+R}}}2^n.$$

Proof: We prove the lemma using a random set S and then compute the expected number of words that are not covered by such a random choice. We choose S to be a uniformly random set of cardinality $|S| = \left\lfloor c2^{n+R}/V_{\mathsf{I}}(n,R)\right\rfloor$, where each subset has the same probability. By this choice of S, the probability for any $\boldsymbol{y} \in \{0,1\}^n$ to be not covered by S is given by

$$\begin{aligned}\mathrm{P}[\boldsymbol{y} \text{ is uncovered}] &= \frac{\binom{2^{n+R}-V_{\mathsf{I}}(n,R)}{|S|}}{\binom{2^{n+R}}{|S|}} \leq \left(\frac{2^{n+R}-V_{\mathsf{I}}(n,R)}{2^{n+R}}\right)^{|S|} \\ &= \left(1 - \frac{V_{\mathsf{I}}(n,R)}{2^{n+R}}\right)^{|S|} \leq \mathrm{e}^{-c}\mathrm{e}^{\frac{V_{\mathsf{I}}(n,R)}{2^{n+R}}}.\end{aligned}$$

The bound on the size of S follows from $V_{\mathsf{I}}(n,R) \geq \binom{n}{R}$. □

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

Lemma 16: Let $S \subseteq \{0,1\}^{n_1+R_1}, T \subseteq \{0,1\}^{n_1}$ be such that S covers $\{0,1\}^{n_1} \setminus T$ with R_1 deletions. Denote by $\mathcal{C}_1 \subseteq \{0,1\}^{n_2-R_1}$ an R_2-deletion-covering code of length n_2 and by $\mathcal{C}_2 \subseteq \{0,1\}^{n_2}$ an R-insertion-covering code of length n_2. We have that

$$(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)$$

is an $R = R_1 + R_2$-deletion-covering code of length $n = n_1 + n_2$ with size at most $|S| \cdot |\mathcal{C}_1| + |T| \cdot |\mathcal{C}_2|$.

We omit the proof here as it is proven in the same manner as Lemma 11. Using this construction of codes, we can again prove the following asymptotic relationship.

Lemma 17: For any $n \geq R$ and $c > 0$,

$$\begin{aligned}\mu_{\mathsf{D}}(n,R) \leq\ & ce\mu_{\mathsf{D}}(n/R + R - 1, 1)\gamma_{\mathsf{D}}(n,R) \\ & + R^R \mathrm{e}^{-c}\mu_{\mathsf{D}}(n/R, R)\gamma'_{\mathsf{D}}(n,R),\end{aligned}$$

for some functions $\gamma_{\mathsf{D}}(n,R)$ and $\gamma'_{\mathsf{D}}(n,R)$ with

$$\lim_{n\to\infty} \gamma_{\mathsf{D}}(n,R) = \lim_{n\to\infty} \gamma'_{\mathsf{D}}(n,R) = 1.$$

Proof: Let $S \subseteq \{0,1\}^{n_1+R_1}, T \subseteq \{0,1\}^{n_1}$ be such that S covers $\{0,1\}^{n_1} \setminus T$ with R_1 deletions. Denote by $\mathcal{C}_1 \subseteq \{0,1\}^{n_2-R_1}$ an R_2-deletion-covering code of length $n_2 - R_1$ and by $\mathcal{C}_2 \subseteq \{0,1\}^{n_2}$ an R-deletion-covering code of length n_2, where $n_1 + n_2 = n$, $n_1 = \frac{y-1}{y}n$, $n_2 = \frac{n}{y}$, and $R_1 + R_2 = R$. We compute the size of the tensorization $(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)$, which, by Lemma 16, is an R-deletion-covering code of length n. To begin with, $|(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)| \leq |S| \cdot |\mathcal{C}_1| + |T| \cdot |\mathcal{C}_2|$. Using Lemma 15, we can bound the size of S and T to obtain

$$\begin{aligned}\mathsf{K}_{\mathsf{D}}(n,R) \leq\ & \frac{c2^n \mu_{\mathsf{D}}(n_2 - R_1, R_2) R_2!}{\binom{n_1}{R_1}(n_2 - R_1)^{R_2}} \\ & + \frac{\mathrm{e}^{-c}\mathrm{e}^{\frac{V_{\mathsf{I}}(n,R)}{2^{n+R}}} 2^n \mu_{\mathsf{D}}(n_2, R) R!}{n_2^R}.\end{aligned}$$

Since $(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)$ is a covering code of length n and covering radius R, we obtain

$$\begin{aligned}\mu_{\mathsf{D}}(n,R) \leq & \frac{c\mu_{\mathsf{D}}(n_2 - R_1, R_2) R_2! n^R}{\binom{n_1}{R_1}(n_2 - R_1)^{R_2} R!} + \frac{\mathrm{e}^{-c}\mathrm{e}^{\frac{V_{\mathsf{I}}(n,R)}{2^{n+R}}} \mu_{\mathsf{D}}(n_2, R) n^R}{n_2^R} \\ \leq & \frac{cn^R \mu_{\mathsf{D}}(n_2 - R_1, R_2)}{n_1^{R_1} n_2^{R_2} \binom{R}{R_1}} \frac{1}{(1 - R_1^2/n_1)(1 - R_1 R_2/n_2)} \\ & + \mathrm{e}^{-c} \left(\frac{n}{n_2}\right)^R \mu_{\mathsf{D}}(n_2, R) \mathrm{e}^{\frac{V_{\mathsf{I}}(n,R)}{2^{n+R}}}.\end{aligned}$$

Inserting $R_1 = R - 1$, $R_2 = 1$, and $y = R$ yields the lemma. □

Remark 1: Proving Theorem 9 for non-binary words follows basically the same steps as for the binary case with only slight differences. To start with, the q-ary version of Lemma 10 is obtained by replacing the binary expressions with their q-ary analogues, and letting $f_{n,R}$ depend also on q, but still converge to 1 for $n \to \infty$. Lemma 11 directly extends to the non-binary case and Lemma 12 can be extended by allowing additional factors in the recursive expression that depend on q and converge to 1 and using standard bounds on $V_{\mathsf{I}}^q(n,R)$. The remaining steps are equivalent to the binary case. The proof of Theorem 14 for q-ary words is obtained analogously.

Remark 2: We note that Theorems 9 and 14 imply the asymptotic bounds in Table I. More precisely, by the properties of the $\limsup$, for any $\varepsilon > 0$, there exists a value n_0, such that for all $n > n_0$, we have that $\mu_{\mathsf{I}}^q(n,R) \leq \mu_{\mathsf{I}}^{q,*}(R)(1+\varepsilon)$ and $\mu_{\mathsf{D}}^q(n,R) \leq \mu_{\mathsf{D}}^{q,*}(R)(1+\varepsilon)$.

VI. Conclusion

This paper studied covering codes for insertions or deletions. We proved general sphere-covering lower bounds on the size of insertion- and deletion-covering codes. We gave constructions for single-deletion- and single-insertion-covering codes that implied improved upper bounds on the code size. Finally, we presented upper bounds on the optimal density of multiple-insertion- and multiple-deletion-covering codes.

There are many avenues for future work. There are gaps between our lower and upper bounds for covering codes. For large covering radius R, we expect that much smaller codes should be possible, e.g., for $R = \varepsilon n$ with $\varepsilon \in (0, 1/2)$. Many of our proofs are existential in nature; it would be nice to have explicit constructions. Establishing the exact size of deletion balls is a long-standing open question with many implications. On the practical side, there may be interesting applications of covering codes based on insertions and deletions. Finally, it would be worthwhile to extend these results to edit distance, in which insertions, deletions, and substitutions are considered.

Appendix

Proposition 18: For fixed R and large n, it holds that

$$\frac{q^{n+R}}{V_{\mathsf{I}}^q(n,R)} \geq \frac{R! q^{n+R}}{n^R (q-1)^R}(1 - o(1)).$$

Proof:

$$\begin{aligned}&\frac{q^{n+R}}{\sum_{i=0}^{R} \binom{n+R}{i}(q-1)^i} \\ &= \frac{q^{n+R}}{\binom{n+R}{R}(q-1)^R \left(1 + \sum_{i=0}^{R-1} \frac{\binom{n+R}{i}(q-1)^{i-R}}{\binom{n+R}{R}}\right)} \\ &\overset{(a)}{\geq} \frac{q^{n+R}}{\binom{n+R}{R}(q-1)^R} \left(1 - \sum_{i=0}^{R-1} \frac{\binom{n+R}{i}(q-1)^{i-R}}{\binom{n+R}{R}}\right) \\ &\overset{(b)}{\geq} \frac{q^{n+R}}{\binom{n+R}{R}(q-1)^R} \left(1 - R^R \sum_{i=0}^{R-1} \frac{(n+R)^i (q-1)^{i-R}}{(n+R)^R}\right) \\ &\overset{(c)}{\geq} \frac{q^{n+R}}{\binom{n+R}{R}(q-1)^R} \left(1 - \frac{R^{R+1}(q-1)^{-1}}{n+R}\right) \\ &\overset{(d)}{\geq} \frac{R! q^{n+R}}{(n+R)^R (q-1)^R} \left(1 - \frac{R^{R+1}(q-1)^{-1}}{n+R}\right) \\ &= \frac{R! q^{n+R}}{n^R (1 + \frac{R}{n})^R (q-1)^R} \left(1 - \frac{R^{R+1}(q-1)^{-1}}{n+R}\right) \\ &\overset{(e)}{\geq} \frac{R! q^{n+R}}{n^R (q-1)^R} \left(1 - \frac{R^2}{n}\right) \left(1 - \frac{R^{R+1}(q-1)^{-1}}{n+R}\right) \\ &= \frac{R! q^{n+R}}{n^R (q-1)^R}(1 - o(1)),\end{aligned}$$

where we used in $(a), (e)$ the inequality $(1+x)^r \geq 1 + rx$ for any $x > -1$ and any $r \leq 0$ or $r \geq 1$. In inequalities $(b), (d)$

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

we used that $n^R/R^R \leq \binom{n}{R} \leq n^R/R!$. Further, in (c) we used that the largest of the terms in the sum is $i = R-1$ to bound the sum. Finally, the statement holds for fixed R and large n. □

Proposition 19: For fixed R and large n, we have the asymptotic relation

$$q\sum_{r=1}^{n-R}\frac{(q-1)^{r-1}\binom{n-R-1}{r-1}}{\binom{r+3R-1}{R}} \geq \frac{R!q^n}{n^R(q-1)^R}\left(1-o(1)\right).$$

Proof: We first note that

$$\binom{n-R-1}{r-1} = \frac{r(r+1)\cdots(r+R-1)}{(n-1)(n-2)\cdots(n-R)}\binom{n-1}{r+R-1}.$$

Hence,

$$\begin{aligned} q\sum_{r=1}^{n-R}\frac{(q-1)^{r-1}\binom{n-R-1}{r-1}}{\binom{r+3R-1}{R}} &= \frac{q}{(n-1)(n-2)\cdots(n-R)} \\ &\cdot \sum_{r=1}^{n-R}\frac{(q-1)^{r-1}r(r+1)\cdots(r+R-1)}{\binom{r+3R-1}{R}}\binom{n-1}{r+R-1} \\ &\geq \frac{qR!}{n^R}\sum_{r=1}^{n-R}\frac{(q-1)^{r-1}r(r+1)\cdots(r+R-1)}{(r+2R)\cdots(r+3R-1)}\binom{n-1}{r+R-1}. \end{aligned}$$

For $0 \leq i \leq R-1$, it holds that $\frac{r+i}{r+2R+i} \geq \frac{r}{r+2R}$ and so

$$\begin{aligned} \frac{r(r+1)\cdots(r+R-1)}{(r+2R)\cdots(r+3R-1)} &\geq \left(\frac{r}{r+2R}\right)^R \\ &= \left(1-\frac{2R}{r+2R}\right)^R \\ &\geq 1-\frac{2R^2}{r+2R}, \end{aligned}$$

where the last step holds by the the inequality $(1-x)^d \geq 1-xd$ for $d \geq 1$. Furthermore, for $r \geq 2R(\sqrt{n}R-1) \stackrel{\text{def}}{=} b$, we have that $1-\frac{2R^2}{r+2R} \geq 1-\frac{1}{\sqrt{n}}$, and thus we deduce that

$$\begin{aligned} &q\sum_{r=1}^{n-R}\frac{(q-1)^{r-1}\binom{n-R-1}{r-1}}{\binom{r+3R-1}{R}} \\ &\geq q(1-\frac{1}{\sqrt{n}})\frac{R!}{n^R}\sum_{r=b}^{n-R}(q-1)^{r-1}\binom{n-1}{r+R-1} \\ &= q(1-\frac{1}{\sqrt{n}})\frac{R!}{n^R}\sum_{r=b+R-1}^{n-1}(q-1)^{r-R}\binom{n-1}{r} \\ &= q(1-\frac{1}{\sqrt{n}})\frac{R!}{n^R(q-1)^R}\sum_{r=b+R-1}^{n-1}(q-1)^{r}\binom{n-1}{r} \\ &= (1-\frac{1}{\sqrt{n}})\frac{R!}{n^R(q-1)^R}\left(q^n-\sum_{r=1}^{b+R-2}\binom{n-1}{r}\right) \\ &\geq \frac{R!q^n}{n^R(q-1)^R}\left(1-o(1)\right). \end{aligned}$$

□

REFERENCES

[1] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes*. Amsterdam, The Netherlands: Elsevier, 1997, vol. 54.

[2] H. Hamalainen, I. Honkala, S. Litsyn, and P. Ostergard, "Football Pools–A game for mathematicians," *Amer. Math. Monthly*, vol. 102, no. 7, p. 579, Aug. 1995.

[3] R. Smolensky, "On representations by low-degree polynomials," in *Proc. IEEE 34th Annu. Found. Comput. Sci.*, Nov. 1993, pp. 130–138.

[4] D. Micciancio, "Almost perfect lattices, the covering radius problem, and applications to Ajtai's connection factor," *SIAM J. Comput.*, vol. 34, no. 1, pp. 118–169, Jan. 2004.

[5] R. Pagh, "Locality-sensitive hashing without false negatives," in *Proc. 27th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2016, pp. 1–9.

[6] M. Braverman, R. Gelles, J. Mao, and R. Ostrovsky, "Coding for interactive communication correcting insertions and deletions," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6256–6270, Oct. 2017.

[7] D. Chakraborty, D. Das, E. Goldenberg, M. Koucky, and M. Saks, "Approximating edit distance within constant factor in truly sub-quadratic time," in *Proc. IEEE 59th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2018, pp. 979–990.

[8] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," 2019, *arXiv:1910.07199*. [Online]. Available: http://arxiv.org/abs/1910.07199

[9] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probab. Surv.*, vol. 6, pp. 1–33, 2009.

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

[11] Y. Liron and M. Langberg, "A characterization of the number of sub-sequences obtained via the deletion channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2300–2312, May 2015.

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

[13] L. Organick *et al.*, "Random access in large-scale DNA data storage," *Nat. Biotechnol.*, vol. 36, no. 3, p. 242, 2018.

[14] S. Chandak *et al.*, "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Sep. 2019, pp. 147–156.

[15] B. Bukh, V. Guruswami, and J. Hastad, "An improved bound on the fraction of correctable deletions," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 93–103, Jan. 2017.

[16] V. Guruswami, B. Haeupler, and A. Shahrasbi, "Optimally resilient codes for list-decoding from insertions and deletions," 2019, *arXiv:1909.10683*. [Online]. Available: http://arxiv.org/abs/1909.10683

[17] V. Guruswami and R. Li, "Polynomial time decodable codes for the binary deletion channel," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2171–2178, Apr. 2019.

[18] V. Guruswami and C. Wang, "Deletion codes in the high-noise and high-rate regimes," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 1961–1970, Apr. 2017.

[19] B. Haeupler, A. Rubinstein, and A. Shahrasbi, "Near-linear time insertion-deletion codes and (1+ε)-approximating edit distance via indexing," in *Proc. 51st Annu. ACM SIGACT Symp. Theory Comput. STOC*, 2019, pp. 697–708.

[20] J. Sima and J. Bruck, "Optimal k-deletion correcting codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 847–851.

[21] V. I. Levenshtein, *Elements Coding Theory*. Moscow, Russia: Nauka, 1974.

[22] V. I. Levenshtein, "On perfect codes in deletion and insertion metric," *Discrete Math. Appl.*, vol. 2, no. 3, pp. 241–258, 1992.

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

[24] N. J. A. Sloane, "On single-deletion-correcting codes," in *Codes and Designs*, K. T. Arasu and A. Seress, Eds. Columbus, OH, USA: Ohio State Univ., 2002, pp. 273–291.

[25] G. A. Kabatyanski and V. I. Panchenko, "Packings and coverings of the Hamming space by unit balls," *Dokl. Akad. Nauk SSSR.*, vol. 303, no. 3, pp. 550–552, 1988.

[26] M. Krivelevich, B. Sudakov, and V. H. Vu, "Covering codes with improved density," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1812–1815, Jul. 2003.

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.

[27] F. N. Afrati, A. D. Sarma, A. Rajaraman, P. Rule, S. Salihoglu, and J. D. Ullman, "Anchor-points algorithms for Hamming and edit distances using MapReduce," in *Proc. Int. Conf. Database Theory*, 2014, pp. 24–28.
[28] R. R. Varshamov and G. M. Tenengolts, "Codes which correct single asymmetric errors," *Autom. Remote Control*, vol. 26, no. 2, pp. 286–290, 1965.
[29] 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.
[30] F. N. Afrati, A. D. Sarma, D. Menestrina, A. Parameswaran, and J. D. Ullman, "Fuzzy joins using MapReduce," in *Proc. IEEE 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 498–509.
[31] J. N. Cooper, R. B. Ellis, and A. B. Kahng, "Asymmetric binary covering codes," *J. Combinat. Theory A*, vol. 100, no. 2, pp. 232–249, Nov. 2002.
[32] P. Erdó and H. Hanani, "On a limit theorem in combinatorial analysis," *Publ. Math. Debrecen*, vol. 10, pp. 10–13, 1963.
[33] V. Rödl, "On a packing and covering problem," *Eur. J. Combinatorics*, vol. 6, no. 1, pp. 69–78, Mar. 1985.
[34] B. Bukh and C. Cox, "Periodic words, common subsequences and frogs," 2019, *arXiv:1912.03510*. [Online]. Available: http://arxiv.org/abs/1912.03510
[35] S. Ganguly, E. Mossel, and M. Z. Racz, "Sequence assembly from corrupted shotgun reads," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 265–269.
[36] M. Kiwi, M. Loebl, and J. Matoušek, "Expected length of the longest common subsequence for large alphabets," *Adv. Math.*, vol. 197, no. 2, pp. 480–498, Nov. 2005.
[37] G. S. Lueker, "Improved bounds on the average length of longest common subsequences," *J. ACM*, vol. 56, no. 3, pp. 1–38, May 2009.
[38] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, Mar. 2001.
[39] M. Schimd and G. Bilardi, "Bounds and estimates on the average edit distance," in *String Processing and Information Retrieval*, N. R. Brisaboa and S. J. Puglisi, Eds. Cham, Switzerland: Springer, 2019, pp. 91–106.
[40] D. Applegate, E. M. Rains, and N. J. A. Sloane, "On asymmetric coverings and covering numbers," *J. Combinat. Des.*, vol. 11, no. 3, pp. 218–228, 2003.
[41] 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.
[42] G. Tenengolts, "Nonbinary codes, correcting single deletion or insertion (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 5, pp. 766–769, Sep. 1984.
[43] B. Ginsburg, "A number-theoretic function with an application in the theory of coding," *Probl. Kibernetiki*, vol. 19, pp. 249–252, 1967.
[44] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Doklady*, vol. 28, pp. 707–710, 1966.

Andreas Lenz (Student Member, IEEE) 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. He is currently pursuing the Ph.D. degree with the Coding for Communications and Data Storage (COD) Group, TUM, where he is involved in research about coding theory for insertion and deletion errors and modern data storage systems. During his studies, his research interests included 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.

Cyrus Rashtchian received the Ph.D. degree in computer science and engineering from the University of Washington, Seattle, in 2018. He is currently a Data Science Fellow with the University of California at San Diego, La Jolla, CA, USA, affiliated with the Computer Science and Engineering Department and the Qualcomm Institute. His research interests include trace reconstruction, DNA data storage, robust classification, clustering, and in general, a geometric perspective on machine learning and algorithms.

Paul H. Siegel (Life Fellow, IEEE) received the B.S. and Ph.D. degrees in mathematics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1975 and 1979, respectively. He held the Chaim Weizmann Postdoctoral 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. In 1995, he joined the University of California at San Diego, La Jolla, CA, USA, as a Faculty Member, 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 an 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 a co-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 was the 2015 Padovani Lecturer of the IEEE Information Theory Society. He has 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 for Coding Techniques of 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 of 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.

Eitan Yaakobi (Senior Member, IEEE) 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 at San Diego, La Jolla, CA, USA, in 2011. From 2011 to 2013, he was a Post-Doctoral Researcher with the Department of Electrical Engineering, the California Institute of Technology, and the Center for Memory and Recording Research, University of California at 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.

Authorized licensed use limited to: Eitan Yaakobi. Downloaded on July 10,2021 at 14:49:11 UTC from IEEE Xplore. Restrictions apply.