Covering Codes for Insertions and Deletions

Andreas Lenz[*], **Cyrus Rashtchian**[†], **Paul H. Siegel**[‡], and **Eitan Yaakobi**[§]
[*]Institute for Communications Engineering, Technical University of Munich, Germany
[†]Computer Science and Engineering Department and the Qualcomm Institute, University of California, San Diego
[‡]Department of Electrical and Computer Engineering, CMRR, University of California, San Diego
[§]Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel
Emails: andreas.lenz@mytum.de, crashtchian@eng.ucsd.edu, psiegel@ucsd.edu, yaakobi@cs.technion.ac.il

***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 insertions and 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.**

I. Introduction

Covering codes are a core object of study in coding theory and discrete mathematics. While previous work has mostly studied covering codes with respect to substitutions, 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., [1], [2], [3], [4], [5], [6], [7]). In this paper, we study covering codes for insertions and deletions. Loosely speaking, we aim to cover a space of words by the union of balls around a minimum number of codewords. This means that the covering problem for R insertions deals with finding a small set of codewords of length n such that each word of length $n + R$ is a supersequence of some codeword. Similarly, for the case of deletions, each word of length $n - R$ must be a subsequence of some codeword. In both cases, the codewords can be viewed as the centers of balls with radius R under the Levenshtein distance. Notice, however, that the codewords and the covered words reside in different spaces as they have different lengths.

Although there is a rich literature on covering codes for the Hamming distance [8], as well as recent improvements for insertion/deletion error-correcting codes (e.g., [9], [10], [11], [12], [13]), 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, the number of words obtainable by inserting R symbols into $\boldsymbol{x}$ (cf. [14]) only depends on the length of $\boldsymbol{x}$. In contrast, deletion balls are irregular, and their sizes depend on many properties of their center $\boldsymbol{x}$, such as the number of runs. This irregularity means that, compared to the Hamming distance, it is inherently more challenging to derive bounds on the minimum covering code size.

Covering codes for insertions or deletions are similar in spirit to asymmetric covering codes for substitutions [19] since they share the irregularity of the balls. Afrati et al. have studied covering codes for insertions and deletions, motivated by designing MapReduce algorithms for similarity joins under the Levenshtein distance [17], [20]. Over an alphabet of size q, 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.

In this paper 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. Regarding R-insertions-covering codes, we provide nearly matching upper and lower bounds that differ only by a factor of seven for $R = 1$. This improves upon the upper bound result of Afrati et al. [17] 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 [16], and it seems beyond our current techniques to obtain a tighter bound.

For the case of R-deletion-covering codes, we first 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 a fixed number $R \geq 2$ of 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 $R > 1$.

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,

A. Lenz acknowledges support from the German-American Fulbright Commission for funding the visit to UCSD. E. Yaakobi acknowledges support from the Center for Memory and Recording Research at UCSD. This work is also funded by the European Research Council under the EU's Horizon 2020 research and innovation programme (grant No. 801434), by NSF grant CCF-BSF-1619053 and by the United States-Israel BSF grant 2018048.

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	$\frac{q^n}{(q-1)n+1}$	(∞)	$\frac{q^n}{(q-1)n+1}$		[15]
R-substitution	$\frac{cR\log R \cdot q^n}{V_{\mathsf{H}}^q(n,R)}$	(∞)	$\frac{q^n}{V_{\mathsf{H}}^q(n,R)}$		[16]
1-insertion	$\frac{7 \cdot q^{n+1}}{(n+1)(q-1)+1}$		$\frac{q^{n+1}}{(n+1)(q-1)+1}$		Theorem 4, 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 7, Theorem 1
1-deletion (binary)	$\frac{2^n}{n+1}$		$\frac{2^n}{n+1}$	(∞)	[17], [18]
1-deletion	$\frac{q^n}{(n+1)\lfloor q/2 \rfloor}$		$\frac{q^n(n-2)}{(q-1)n(n+1)}$		Theorem 3, Theorem 2
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 12, Theorem 2

$\rho(\boldsymbol{x}) := 1 + |\{1 \le i < n : x_i \ne x_{i+1}\}|$. 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. [14], that insertion balls are regular, i.e., only depend on $\mathsf{len}(x)$, q, and t. Thus, we denote by $V_{\mathsf{I}}^q(n,t)$ the insertion ball size of length-n words over Σ_q. We will consider two subproblems, namely covering words with only insertions or only deletions. Formally, we have the following definitions.

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, $\bigcup_{\boldsymbol{c}\in\mathcal{C}} \mathsf{Ball}_{\mathsf{I}}^q(\boldsymbol{c},R) = \Sigma_q^{n+R}$. Analogously, 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.

We now turn to establishing 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 on the number of words a codeword can cover. For the case of insertions, the ball size is known and is independent of the center and we directly obtain the following theorem.

Theorem 1. *For all n and R, it holds that*

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

Furthermore, for fixed R and large n,

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

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. [21] that enables the computation of a bound even though the ball sizes are irregular by bounding the solution of a linear program. We state the resulting lower bound on deletion-covering codes in the following theorem.

Theorem 2. *For all n and $0 < R < n$, it holds that*

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

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

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

Furthermore, for fixed R and large n, we have

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

The proof of this theorem is omitted for brevity and can be found in the full version of this paper [22].

III. 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 deletions and insertions for both binary and non-binary alphabets. 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

The well-known Varshamov-Tenengolts (VT) [18] codes are perfect binary single-deletion-correcting codes and thus also single-deletion-covering codes. VT codes have a non-binary extension, presented by Tenengolts in [23], which can correct a single deletion in the non-binary case. However, this family of codes is no longer perfect and therefore not single-deletion-covering. Our main result in this section is another non-binary extension of the binary VT codes, which does satisfy the single-deletion covering property.

Definition 2. *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*

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

where for $m \in \mathbb{N}_0$, we denote by $(m)_2$ the value $(m \bmod 2)$.

For $q = 2$ and $b = 0$ this definition is equivalent to that of the VT code and we abbreviate $\mathcal{C}_{\mathsf{VT}}(n; a) \stackrel{\text{def}}{=} \mathcal{C}^2_{\mathsf{NBVT}}(n; a, 0)$. The following theorem proves that $\mathcal{C}^q_{\mathsf{NBVT}}(n; a, b)$ is indeed a non-binary single-deletion-covering code.

Theorem 3. *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,*

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

The theorem is based on the fact that the VT-code is a single-deletion covering code and can be found in full detail in [22]. Lastly, we note that this construction improves upon the construction in [17][1], 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. Interestingly, in contrast to the case of a single deletion, the VT code is not a perfect code for a single insertion. In fact, this can be verified by simple counting arguments using that the VT code has size only roughly half the size required by Theorem 1. This missing factor of two can intuitively be explained from the point of view of the covered space Σ_2^{n+1}. Let $\boldsymbol{y} \in \Sigma_2^{n+1}$ be a word with checksum $b \equiv \sum_{i=1}^{n+1} i(y_i)_2 \bmod (n+1)$ that shall be covered by the VT-code $\mathcal{C}_{\mathsf{VT}}(n; a)$. Then, $\boldsymbol{y}$ is covered by $\mathcal{C}_{\mathsf{VT}}(n; a)$ if there exists a deletion position j such that the resulting checksum, after deleting bit j, is equal to a. Now, it is known that most words $\boldsymbol{y}$ have only roughly $\frac{n}{2}$ possible results after a single deletion. However, there are $n+1$ possible checksums b for $\boldsymbol{y}$, which means that roughly half the words $\boldsymbol{y}$ will not be covered by $\mathcal{C}_{\mathsf{VT}}(n; a)$.

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 [24], 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 the following. Our main result is stated in the following theorem

Theorem 4. *For all $n \geq 1$ and $q \geq 2$ 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, Ellis, and Kahng [19]. 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 [19].) 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 efficiently.

Lemma 5. *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 $A \otimes B = \{\boldsymbol{ab} \mid \boldsymbol{a} \in A, \boldsymbol{b} \in B\}$ is the tensor product of two sets and $\boldsymbol{ab}$ is the concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$*

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 reduces the overall code size by a constant factor compared to uniform sampling. The 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 $\mu_{\mathsf{I}}/V^q_{\mathsf{I}}(n_2, 1)$ because this is the factor saved by the use of induction later in the construction.

Lemma 6. *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, $T = \Sigma_q^{n_1+1} \setminus \bigcup_{\boldsymbol{s} \in S} \mathsf{Ball}^q_{\mathsf{I}}(\boldsymbol{s}, 1)$, while the sizes of S and T satisfy*

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

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 $q_{\boldsymbol{x}} \stackrel{\text{def}}{=} cV^q_{\mathsf{D}}(\boldsymbol{x}, 1)^{-1}$ for a constant $c > 0$ to be set later. Let T be all remaining words that are not covered by S, i.e., $T = \Sigma_q^{n_1+1} \setminus \bigcup_{\boldsymbol{s} \in S} \mathsf{Ball}^q_{\mathsf{I}}(\boldsymbol{s}, 1)$.

[1]The result is stated in Corollary 5.5 in [17]. However, note that the authors of this paper refer to deletion-covering codes as insertion-covering codes and also note that their result is stated over length-$(n+1)$ codes.

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 $V_{\mathsf{D}}^q(\boldsymbol{y},1) = \rho(\boldsymbol{y})$ [25], 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:

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

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 [25] 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 $1-z \leq \mathrm{e}^{-z}$ for all $z \in \mathbb{R}$, we have $(1-q_{\boldsymbol{y}})^{V_{\mathsf{D}}^q(\boldsymbol{y},1)} \leq \mathrm{e}^{-c}$ and we bound W by

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

Finally, we use $\sum_{k=0}^{n}\binom{n}{k}x^k = (1+x)^n$ 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

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

where we used in equality (b) 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 exist 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 $\mu_{\mathsf{I}}q^{n_1+1}/V_{\mathsf{I}}^q(n,1)$, it follows that there exists an (S,T) pair satisfying the desired bound. □

Putting everything together, we prove Theorem 4 for single-insertion-covering codes.

Proof of Theorem 4. 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 $n_1+n_2+1 = n$ and $S \subseteq \Sigma_q^{n_1}$ and $T \subseteq \Sigma_q^{n_1+1}$ denote sets guaranteed by Lemma 6. Note that clearly $n_2 < n$ in Lemma 6, which will be useful later. As these sets S and T fulfill the requirement of Lemma 5, we define $\mathcal{C}_n = (S \otimes \Sigma_q^{n_2+1}) \cup (T \otimes \mathcal{C}_{n_2})$, which is a single-insertion-covering code $\mathcal{C}_n \subseteq \Sigma_q^n$ of size

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

where in (c) 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 6, 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 4, 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.

IV. Multiple-Insertion/Deletion-Covering Codes

We now turn to 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 3. *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 since the deletion balls are non-uniform and the density is thus defined with respect to the lower bound obtained in Theorem 2 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. 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 and are omitted for brevity.

A. Multiple-Insertion-Covering Codes

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

Theorem 7. *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 4, 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 8 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 9 and 10, 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 7. The following lemma gives an upper bound on the sizes of the almost-covering code S and the complement T of its coverage.

Lemma 8. *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 $\lim_{n\to\infty} f_{n,R} = 1$.

The proof of Lemma 8 is, similarly to that of Lemma 6, based on a random choice of S and T, where in our random selection of codewords, we favor codewords that cover words with small $V_{\mathsf{D}}(\boldsymbol{y}, R)$ to ensure that each word is covered with high enough probability. For a detailed outline the reader is referred to the full version of the paper [22]. Analogous to Lemma 9, we define the operation that allows to assemble covering codes given shorter covering codes.

Lemma 9. *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 $(S \otimes \mathcal{C}_1) \cup (T \otimes \mathcal{C}_2)$ is an $R = R_1 + R_2$-insertion-covering code of length $n = n_1 + n_2$ with size at most $|S| \cdot |\mathcal{C}_1| + |T| \cdot |\mathcal{C}_2|$.*

Based on Lemmas 8 and 9 we can now derive a recursive bound on the optimal density $\mu_{\mathsf{I}}(n, R)$.

Lemma 10. *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 9, 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 8 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

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

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 11 (cf. [16])**.** *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 7.

Proof of Theorem 7. Using Lemma 10 and 11, 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 12. *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 $q = 2$,

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

The outline of the proof is similar to that of insertions and is omitted for brevity. Details can be found in [22]. Note that compared to the case of insertions, we could use in Theorem 12 that for the binary case $\mu_{\mathsf{D}}^*(1) = 1$, which results in a tighter bound also for the case of $R > 1$.

REFERENCES

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

[2] D. Chakraborty, D. Das, E. Goldenberg, M. Koucky, and M. Saks, "Approximating edit distance within constant factor in truly sub-quadratic time," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018, pp. 979–990.

[3] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *arXiv:1910.07199 [cs, math]*, Nov. 2019.

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

[5] V. I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences," *Journal of Combinatorial Theory, Series A*, vol. 93, no. 2, pp. 310–332, 2001.

[6] Y. Liron and M. Langberg, "A characterization of the number of subsequences obtained via the deletion channel," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2300–2312, 2015.

[7] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Random access in large-scale dna data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

[8] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes*. Elsevier, 1997, vol. 54.

[9] B. Bukh, V. Guruswami, and J. Håstad, "An improved bound on the fraction of correctable deletions," *IEEE Transactions on Information Theory*, vol. 63, no. 1, pp. 93–103, 2016.

[10] V. Guruswami, B. Haeupler, and A. Shahrasbi, "Optimally resilient codes for list-decoding from insertions and deletions," *arXiv preprint arXiv:1909.10683*, 2019.

[11] 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.

[12] 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.

[13] B. Haeupler, A. Rubinstein, and A. Shahrasbi, "Near-linear time insertion-deletion codes and (1+ ε)-approximating edit distance via indexing," in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2019, pp. 697–708.

[14] V. I. Levenshtein, "Elements of coding theory," *Discrete Mathematics and Mathematical Problems of Cybernetics*, pp. 207–305, 1974.

[15] 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, 1988.

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

[17] 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 *International Conference on Database Theory*, 2014.

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

[19] J. N. Cooper, R. B. Ellis, and A. B. Kahng, "Asymmetric binary covering codes," *Journal of Combinatorial Theory, Series A*, vol. 100, no. 2, pp. 232–249, 2002.

[20] F. N. Afrati, A. D. Sarma, D. Menestrina, A. Parameswaran, and J. D. Ullman, "Fuzzy joins using MapReduce," in *IEEE International Conference on Data Engineering (ICDE)*, 2012.

[21] D. Applegate, E. M. Reins, and N. J. A. Sloane, "On asymmetric coverings and covering numbers," *Journal of Combinatorial Designs*, vol. 11, no. 3, pp. 218–228, Apr. 2003.

[22] A. Lenz, C. Rashtchian, P. H. Siegel, and E. Yaakobi, "Covering codes for insertions and deletions," *arXiv:1911.09944 [cs, math]*, Nov. 2019.

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

[24] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 28, pp. 707–710, 1966.

[25] 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.