Gilbert-Varshamov inspired lower bound on the maximal cardinality of indel and substitution correcting codes

Speé, W.J.P.; Weber, J.H.

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 Symposium on Information Theory and Signal Processing in the Benelux

Citation (APA)
Speé, W. J. P., & Weber, J. H. (2023). Gilbert-Varshamov inspired lower bound on the maximal cardinality of indel and substitution correcting codes. In *Proceedings of the 2023 Symposium on Information Theory and Signal Processing in the Benelux* (pp. 24-28)

Gilbert-Varshamov inspired lower bound on the maximal cardinality of indel and substitution correcting codes

Ward J. P. Spee
Delft University of Technology
Delft, The Netherlands
W.J.P.Spee@student.tudelft.nl

Jos H. Weber
Delft University of Technology
Delft, The Netherlands
J.H.Weber@tudelft.nl

***Abstract*—Recent advances in DNA data storage and racetrack memory have attracted renewed attention towards deletion, insertion and substitution correcting codes. Compared to codes aimed at correcting either substitution errors or deletion and insertion (indel) errors, the understanding of codes that correct combinations of substitution and indel errors lags behind. In this paper, we focus on the maximal size of q-ary t-indel s-substitution correcting codes. In particular, our main contribution is a Gilbert-Varshamov inspired lower bound on this size. Moreover, we study the asymptotic behaviour of this bound.**

***Index Terms*—Error correcting codes, Gilbert-Varshamov bound, indels, substitutions.**

I. Introduction

Coding techniques for correcting deletion, insertion and substitution errors have attracted increasing attention recently due to their applications in DNA data storage [1], [2] and racetrack memory [3], [4]. Codes that correct either substitution errors or deletion and insertion errors have been extensively studied in literature. In contrast, the simultaneous correction of combinations of these three error types is less understood. A central problem is to determine the maximal size of codes that correct combinations of deletion, insertion, and substitution errors.

Classical error correcting codes aimed at correcting substitution errors have been well-studied for over 75 years [5]. A fundamental result in this area is the well-known Gilbert-Varshamov bound [6], [7] which asserts the existence of a q-ary s-substitution correcting code with codewords of length n and with a code size of at least

$$\frac{q^n}{\sum_{i=0}^{2s}\binom{n}{i}(q-1)^i}.$$

This statement was initially proven by Gilbert [6] for binary codes, and later independently by Varshamov [7]. Subsequently, the bound has been improved and generalized in various settings. An overview of these improvements in the context of substitution correcting codes is given in [8].

In a seminal paper [9], Levenshtein initiated the study of deletion and insertion (indel) correcting codes. He showed that a code that is able to correct t deletions (or insertions) is able to correct any t' deletions and t'' insertions, whenever $t'+t''\leq t$. In other words, a t-deletion (insertion) correcting code is also a t-indel correcting code. This property shows the indifference between correcting deletions and insertions, which warrants the terminology of t-indel correcting codes. Inspired by the Gilbert-Varshamov bound and the work of Tolhuizen [10], a lower bound on the maximal size of t-indel correcting codes was given in [11]. Multiple bounds that improve upon this result were presented in [12] and [13].

In comparison with either substitution correcting codes or indel correcting codes, non-asymptotic lower bounds on the maximal cardinality of t-indel s-substitution correcting codes have been studied to a lesser degree in literature. Several t-indel s-substitution correcting codes have been constructed, e.g. in [14], [15], which naturally imply non-asymptotic lower bounds on the maximal size of these codes. In [9], Levenshtein also showed two asymptotic bounds which imply that a binary t-indel s-substitution correcting code of maximal size has an asymptotic redundancy between $(t+s)\log_2(n)$ and $2(t+s)\log_2(n)+o(\log_2(n))$. Moreover, note that each $(t+2s)$-indel correcting codes is also a t-indel s-substitution correcting code, because a substitution can be seen as a deletion followed by an insertion. Hence, lower bounds on the maximum size of $(t+2s)$-indel correcting codes imply lower bounds for t-indel s-substitution correcting codes as well.

The last observation that a $(t+2s)$-indel correcting code is also a t-indel s-substitution correcting code might raise the preliminary question whether it is superfluous to consider the correction of substitutions separately. However, there are two arguments in favor of separating indel correction from substitution correcting. First, it was recognized by Song *et al.* [14] that $(t+2s)$-indel correcting codes are not necessarily optimal within the set of t-indel s-substitution correcting codes in terms of redundancy[1]. Secondly, in applications such as DNA data storage, the error rates of indels and substitutions differ [2]. Therefore, it is sensible to bound the number indels and substitutions by different parameters.

[1] For instance, the single-substitution correcting binary Hamming code with words of length 7 has size 16 [16]. In contrast, in [17, Thrm. 1] it was shown that a binary two-indel correcting code has a maximal size of at most 11.

In this paper, we study the maximal size of t-indel s-substitution correcting codes on a q-ary alphabet. In particular, our contribution is a Gilbert-Varshamov inspired lower bound on this size. Moreover, we will prove that this bound implies that a q-ary t-indel s-substitution correcting code of maximal size has an asymptotic redundancy of at most $2(t+s)\log_q(n)+o(\log(n))$. This extends Levenshtein's upper bound on the asymptotic redundancy to q-ary codes.

The organisation of this paper is as follows. In Section II, notation, terminology and several prior results are discussed. Next, a non-asymptotic lower bound inspired by the Gilbert-Varshamov bound is derived in Section III. Lastly, the asymptotic behaviour of this bound is studied in Section IV.

II. Definitions and preliminaries

For a finite set S, denote the cardinality of S by $|S|$. Consider the alphabet with $q \geq 2$ symbols given by $\mathcal{B}_q := \{0, 1, ..., q-1\}$. The set of q-ary words (i.e., vectors) of length n with symbols from $\mathcal{B}_q$ is denoted by $\mathcal{B}_q(n) := \{0, 1, ..., q-1\}^n$. A non-empty subset $\mathcal{C} \subseteq \mathcal{B}_q(n)$ is called a code and the elements of a code are called codewords. A code can be capable of correcting errors by ensuring that the codewords of $\mathcal{C}$ are 'sufficiently different', so that after several errors have occurred the resulting word still 'resembles' the original codeword, but not any of the other codewords. This idea forms the basis for the following definition of an indel and substitution correcting code.

For integers $0 \leq t \leq n$ and $0 \leq s \leq n$, a code $\mathcal{C} \subseteq \mathcal{B}_q(n)$ is said to be a *t-indel s-substitution correcting code* if any q-ary word (not necessarily of length n) can be obtained from no more than one codeword by exactly t' deletions, t'' insertions and s or fewer substitutions, whenever $t'+t'' \leq t$. A 0-indel s-substitution correcting code is simply called an *s-substitution correcting code* and analogously a t-indel 0-substitution correcting code is called a *t-indel correcting code*.

By only using codewords for communicating information, the code gains error-correcting capabilities at the cost of introducing redundancy. In order to maximize the amount of information that can be transmitted using a code, we are interested in the maximal size of a q-ary t-indel s-substitution correcting code with codewords of length n, which we denote by $M_q(n, t, s)$. The (information) rate of a code $\mathcal{C}$ is defined by $\frac{1}{n}\log_q(|\mathcal{C}|)$ and the redundancy by $n - \log(|\mathcal{C}|)$.

Denote by $\mathcal{V}_{t',t'',s}(\mathbf{x})$ the set of words that can be reached from $\mathbf{x} \in \mathcal{B}_q(n)$ by means of exactly t' deletions, t'' insertions and at most s substitutions. Clearly, the q-ary words in the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$ have length $n - t' + t''$. Moreover, we define $\mathcal{D}_t(\mathbf{x}) = \mathcal{V}_{t,0,0}(\mathbf{x})$, $\mathcal{I}_t(\mathbf{x}) = \mathcal{V}_{0,t,0}(\mathbf{x})$ and $\mathcal{S}_s(\mathbf{x}) = \mathcal{V}_{0,0,s}(\mathbf{x})$. These sets are highly related to t-indel s-substitution correcting codes, and allow for equivalent characterizations of these codes in terms of the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$. The following lemma collects various equivalent characterizations from e.g., [14, Sec. II], [18, Lem. 2] and [19, Lem. 2].

Lemma 1. *Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers, and let $\mathcal{C} \subseteq \mathcal{B}_q(n)$ be a code. Then, the following five statements are equivalent:*

1) *$\mathcal{C}$ is a t-indel s-substitution correcting code.*
2) *$\mathcal{V}_{t',t'',s}(\mathbf{c}_1) \cap \mathcal{V}_{t',t'',s}(\mathbf{c}_2) = \emptyset$ for all distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$, and for all integers $t', t'' \geq 0$ such that $t' + t'' \leq t$.*
3) *$\mathcal{V}_{t,0,s}(\mathbf{c}_1) \cap \mathcal{V}_{t,0,s}(\mathbf{c}_2) = \emptyset$ for all distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$.*
4) *$\mathcal{V}_{0,t,s}(\mathbf{c}_1) \cap \mathcal{V}_{0,t,s}(\mathbf{c}_2) = \emptyset$ for all distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$.*
5) *$\mathbf{c}_2 \notin \mathcal{V}_{t,t,2s}(\mathbf{c}_1)$ for all distinct $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$.*

For general parameters t', t'' and s, and words $\mathbf{x} \in \mathcal{B}_q(n)$ determining the cardinality of $\mathcal{V}_{t',t'',s}(\mathbf{x})$ is a non-trivial task [20]. In the highly specific case that $t' = t'' = 0$ it holds for each $\mathbf{x} \in \mathcal{B}_q(n)$ [5] that

$$|\mathcal{S}_s(\mathbf{x})| = \sum_{i=0}^{s} \binom{n}{i} (q-1)^i. \qquad (1)$$

The quantity $S^s_{n,q} := \sum_{i=0}^{s} \binom{n}{i}(q-1)^i$ will be referred to as the size of the q-ary Hamming sphere of radius s. Moreover, it has been established [21] that

$$|\mathcal{I}_t(\mathbf{x})| = S^t_{n+t,q} = \sum_{i=0}^{t} \binom{n+t}{i} (q-1)^i. \qquad (2)$$

Interestingly, the cardinalities of $\mathcal{S}_s(\mathbf{x})$ and $\mathcal{I}_t(\mathbf{x})$ depend on $\mathbf{x}$ only via the parameters n and q. In contrast, $|\mathcal{D}_t(\mathbf{x})|$ depends on the structure of the word $\mathbf{x}$ as well as the parameters n and q. To the best of authors' knowledge, an analytic formula of $|\mathcal{D}_t(\mathbf{x})|$ is not known for general t and therefore we must rely on bounds (see e.g., [9], [22], [23]). For $t \leq 5$, an analytic formula of $|\mathcal{D}_t(\mathbf{x})|$ has been provided in [24], but these expressions are rather involved for $t \geq 2$. Lastly, we mention that using the observation that $\mathbf{x} \in \mathcal{I}_t(\mathbf{y})$ if and only if $\mathbf{y} \in \mathcal{D}_t(\mathbf{x})$, it was shown in [11] that the average cardinality of $\mathcal{D}_t(\mathbf{x})$ is given by

$$\frac{1}{q^n} \sum_{\mathbf{x} \in \mathcal{B}_q(n)} |\mathcal{D}_t(\mathbf{x})| = \frac{1}{q^n} \sum_{\mathbf{y} \in \mathcal{B}_q(n-t)} |\mathcal{I}_t(\mathbf{y})| \overset{(2)}{=} \frac{1}{q^t} \sum_{i=0}^{t} \binom{n}{i} (q-1)^i. \qquad (3)$$

III. Gilbert-Varshamov inspired lower bound

The well-known Gilbert-Varshamov lower bound for s-substitution correcting codes [6], [7] is given by

$$M_q(n, 0, s) \geq \frac{q^n}{\sum_{i=0}^{2s} \binom{n}{i} (q-1)^i}. \qquad (4)$$

This bound is commonly proven using a sphere-covering argument where the spheres are given by $\mathcal{S}_{2s}(\mathbf{c})$ centered around the codewords $\mathbf{c} \in \mathcal{C}$ (see e.g., [5, Thrm. 4.3]). In the case of substitutions, this proof is facilitated by the fact that these spheres are of equal size.

Tolhuizen [10] recognized that the Gilbert-Varshamov bound is also implied by Turán's theorem [25] from extremal graph theory. A particular consequence of the latter approach is that it easily generalizes to the case in which the spheres are not of equal size. For instance, this is the case for t-indel correcting codes when dealing with the spheres $\mathcal{V}_{t,t,0}(\mathbf{c})$. The approach from Tolhuizen was used by Levenshtein [11] to bound the maximal size of a t-indel correcting code from below. In particular, it was shown that

$$M_q(n,t,0) \geq \frac{q^{n+t}}{\left(\sum_{i=0}^{t}\binom{n}{i}(q-1)^i\right)^2}. \tag{5}$$

For completeness, we mention that other Gilbert-Varshamov related lower bounds on $M_q(n,t,0)$ are given in [12], [13].

Next, it is a natural step to generalize the argument from Tolhuizen to t-indel s-substitution correcting codes.

Lemma 2. *Let $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$ be integers. The following gives a lower bound on $M_q(n,t,s)$,*

$$M_q(n,t,s) \geq \frac{q^n}{\mathcal{V}^{avr}_{t,t,2s}}, \tag{6}$$

where $\mathcal{V}^{avr}_{t,t,2s} := q^{-n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{V}_{t,t,2s}(\mathbf{x})|$.

Proof. The idea of this proof is to translate the problem of finding a large code to the problem of finding a large clique[2]. This allows us to apply the argument from [10, Sec. II] to derive the desired lower bound on $M_q(n,t,s)$.

Define the undirected graph $G = (V,E)$ without loops or double edges as follows. Let $V = \mathcal{B}_q(n)$ be the set of nodes of G. Two distinct nodes $\mathbf{x}$ and $\mathbf{y}$ from V are joined by an edge in E if $\mathbf{x} \notin \mathcal{V}_{t,t,2s}(\mathbf{y})$. This is well-defined because it holds that $\mathbf{x} \notin \mathcal{V}_{t,t,2s}(\mathbf{y})$ if and only if $\mathbf{y} \notin \mathcal{V}_{t,t,2s}(\mathbf{x})$. Intuitively, the pairs of nodes that are connected by an edge can both be codewords in a t-indel s-substitution correcting code. The number of nodes equals $|V| = q^n$ and the number of edges is given by

$$\begin{aligned}
|E| &= \frac{1}{2}\sum_{\mathbf{x}\in V}(|V \setminus \mathcal{V}_{t,t,2s}(\mathbf{x})|) \\
&= \frac{1}{2}\sum_{\mathbf{x}\in V}(|V| - |\mathcal{V}_{t,t,2s}(\mathbf{x})|) \\
&= \frac{1}{2}q^{2n} - \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{V}_{t,t,2s}(\mathbf{x})| \\
&= \frac{1}{2}q^n(q^n - \mathcal{V}^{avr}_{t,t,2s}),
\end{aligned}$$

where the first equality follows from the fact that each $\mathbf{x} \in V$ has $|V \setminus \mathcal{V}_{t,t,2s}(\mathbf{x})|$ incident edges. Therefore, summing $|V \setminus \mathcal{V}_{t,t,2s}(\mathbf{x})|$ over all nodes in $\mathbf{x} \in V$ equals $2|E|$ since each edge is counted twice. Observe that from the definition of the edges in G and Lemma 1 it follows that a clique of size k in G corresponds to a t-indel s-substitution correcting code $\mathcal{C}$ of size k.

Using the cardinalities of V and E it follows from the argument in [10, Sec. II] that there exists a clique in G of size $\lceil \frac{q^n}{\mathcal{V}^{avr}_{t,t,2s}} \rceil$. For brevity, we do not repeat this argument here. In turn, this implies that there exists an equally large t-indel s-substitution correcting code, which concludes the proof. □

In order to evaluate the lower bound in Lemma 2 the size of $\mathcal{V}_{t,t,2s}(\mathbf{x})$ averaged over all $\mathbf{x} \in \mathcal{B}_q(n)$ needs to be determined. To the best of the authors' knowledge, an analytic formula for $|\mathcal{V}_{t,t,2s}(\mathbf{x})|$ or $\mathcal{V}^{avr}_{t,t,2s}$ is not known for general parameters n, q, t and s. For this reason, we employ an upper bound on $\mathcal{V}^{avr}_{t,t,2s}$ to obtain an explicit result.

Theorem 3. *For integers $n \geq 1$, $q \geq 2$, $0 \leq t \leq n$ and $0 \leq s \leq n$, the following gives a lower bound on $M_q(n,t,s)$,*

$$M_q(n,t,s) \geq \frac{q^{n+t}}{\left(\sum\limits_{i=0}^{t}\binom{n}{i}(q-1)^i\right)^2 \sum\limits_{i=0}^{2s}\binom{n-t}{i}(q-1)^i}. \tag{7}$$

Proof. We claim that $\mathcal{V}^{avg}_{t,t,2s}$ can be upper bounded by

$$\frac{1}{q^t}\left(\sum_{i=0}^{t}\binom{n}{i}(q-1)^i\right)^2 \sum_{i=0}^{2s}\binom{n-t}{i}(q-1)^i. \tag{8}$$

In this case, the result of the theorem follows immediately from applying the upper bound to Lemma 2. Therefore, this proof is limited to proving this claim. In what follows, a superscript $^-$ will be used to denote a word in $\mathcal{B}_q(n-t)$, whereas an omission thereof is meant for words in $\mathcal{B}_q(n)$.

To this end, observe that each element in $\mathcal{V}_{t,t,2s}(\mathbf{x})$ can be reached from $\mathbf{x} \in \mathcal{B}_q(n)$ by first deleting precisely t symbols, followed by substituting at most $2s$ symbols and lastly inserting exactly t symbols. Hence, it follows that

$$|\mathcal{V}_{t,t,2s}(\mathbf{x})| \leq \sum_{\mathbf{y}^-\in\mathcal{D}_t(\mathbf{x})}\ \sum_{\mathbf{z}^-\in\mathcal{S}_{2s}(\mathbf{y}^-)}|\mathcal{I}_t(\mathbf{z}^-)|. \tag{9}$$

In order to evaluate the right-hand side of this expression, recall from (1) and (2) that the cardinalities of the sets $\mathcal{I}_t(\mathbf{x}^-)$ and $\mathcal{S}_{2s}(\mathbf{x}^-)$ do not depend on the choice of $\mathbf{x}^- \in \mathcal{B}_q(n-t)$. Moreover, the cardinality of $\mathcal{D}_t(\mathbf{x})$ averaged over all $\mathbf{x} \in \mathcal{B}_q(n)$ was given in (3). By combining these results and carefully taking into account the lengths of the words, it follows that

$$\begin{aligned}
\mathcal{V}^{avg}_{t,t,2s} &= q^{-n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{V}_{t,t,2s}(\mathbf{x})| \\
&\overset{(9)}{\leq} \frac{1}{q^n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}\ \sum_{\mathbf{y}^-\in\mathcal{D}_t(\mathbf{x})}\ \sum_{\mathbf{z}^-\in\mathcal{S}_{2s}(\mathbf{y}^-)}|\mathcal{I}_t(\mathbf{z}^-)| \\
&\overset{(2)}{=} \frac{1}{q^n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}\ \sum_{\mathbf{y}^-\in\mathcal{D}_t(\mathbf{x})}\ \sum_{\mathbf{z}^-\in\mathcal{S}_{2s}(\mathbf{y}^-)}S^t_{n,q} \\
&\overset{(1)}{=} \frac{1}{q^n}\sum_{\mathbf{x}\in\mathcal{B}_q(n)}\ \sum_{\mathbf{y}^-\in\mathcal{D}_t(\mathbf{x})}S^{2s}_{n-t,q}\cdot S^t_{n,q} \\
&= \frac{1}{q^n}\cdot S^t_{n,q}\cdot S^{2s}_{n-t,q}\cdot\sum_{\mathbf{x}\in\mathcal{B}_q(n)}|\mathcal{D}_t(\mathbf{x})| \\
&\overset{(3)}{=} \frac{1}{q^t}\cdot(S^t_{n,q})^2\cdot S^{2s}_{n-t,q}.
\end{aligned}$$

[2] A clique of a graph G is an induced subgraph that is complete, i.e., all pairs of vertices are connected by an edge.

Note that the last expression is equivalent to (8), which proves the claim. □

Observe that the lower bounds (4) and (5) are special cases of the latter theorem, since they are recovered by setting $t = 0$ and $s = 0$, respectively. Obviously, the bound from Theorem 3 can be improved with the availability of exact expressions, or tighter bounds on $\mathcal{V}^{avg}_{t,t,2s}$.

IV. Asymptotic behaviour

In this section we discuss the asymptotic behaviour of Theorem 3 in two settings based on the dependency of t and s with respect to n.

First, consider the setting in which the parameters q, t and s are fixed, and we let n tend to infinity. In this setting, Levenshtein [9] showed two asymptotic bounds on $M_2(n, t, s)$ which imply that the asymptotic redundancy of a binary t-indel s-substitution correcting code of maximal size lies between $(t + s)\log_2(n)$ and $(2t + 2s)\log_2(n) + o(\log_2(n))$. Here, we provide an alternative proof for the asymptotic upper bound and extend the result from binary to q-ary codes, by showing that it is implied by the non-asymptotic lower bound on $M_q(n, t, s)$ of Theorem 3.

Lemma 4. *Let $q \geq 2$ be an integer. For non-negative integers s and t such that $s + t \geq 1$, the following holds*

$$\limsup_{n\to\infty} \frac{n - \log_q(M_q(n, t, s))}{(2t + 2s)\log_q(n)} \leq 1.$$

Proof. Theorem 3 states that

$$M_q(n, t, s) \geq \frac{q^{n+t}}{(S^t_{n,q})^2 \cdot S^{2s}_{n-t,q}}.$$

This implies that the redundancy of an optimal t-indel s-substitution correcting code is bounded by

$$n - \log_q(M_q(n, t, s)) \leq -t + 2\log_q(S^t_{n,q}) + \log_q(S^{2s}_{n-t,q}).$$

Note that for a fixed integer $k \geq 1$ it holds that $\binom{n}{k} = \frac{1}{k!}n^k + o(n^k)$. In turn, it follows that $S^s_{n,q} = \frac{(q-1)^s}{s!}n^s + o(n^s)$, and $\log_q(S^s_{n,q}) = s\log_q(n) + o(\log_q(n))$. By combining these observations we obtain

$$\limsup_{n\to\infty} \frac{n - \log_q(M_q(n, t, s))}{(2t + 2s)\log_q(n)} \leq \limsup_{n\to\infty} \frac{-t + 2\log_q(S^t_{n,q}) + \log_q(S^{2s}_{n-t,q})}{(2t + 2s)\log_q(n)} = 1,$$

as desired. □

The following statement is immediate from the previous lemma.

Corollary 5. *A maximal size t-indel s-substitution correcting code has an asymptotic redundancy of at most $(2t + 2s)\log_q(n) + o(\log_q(n))$.*

Secondly, we consider the asymptotic regime in which $q \geq 2$ and $\tau, \sigma \in [0, 1]$ are fixed and n tends to infinity. We set[3] $t = \tau n$, $s = \sigma n$. Define the asymptotic rate by

$$R_q(\tau, \sigma) := \liminf_{n\to\infty} \frac{1}{n} \log_q(M_q(n, \tau n, \sigma n)). \qquad (10)$$

For $\sigma = 0$ and $\tau > 0$, bounds on $M_q(n, t, 0)$ have been used to derive results on $R_q(\tau, 0)$ in e.g., [11], [17], [26]. On the other hand, for $\tau = 0$ and $\sigma > 0$ a summary of several results on $R_q(0, \sigma)$ can be found in [5]. Here, we use Theorem 3 to derive a lower bound on $R_q(\tau, \sigma)$.

To this end, let $H_q(x) = x\log_q(q-1) - x\log_q(x) - (1-x)\log_q(1-x)$ on $[0, 1 - \frac{1}{q}]$ with $H_q(0) = 0$ denote the q-ary entropy function. The extended q-ary entropy function is given by $H^*_q(x) = H_q(\min\{x, 1 - \frac{1}{q}\})$ on $[0, \infty)$. Recall the following useful property of the extended q-ary entropy function [17], for each $\lambda \in (0, 1)$ it holds that

$$\lim_{n\to\infty} \frac{1}{n} \log_q \left(\sum_{i=0}^{\lambda n} \binom{n}{i} (q-1)^i \right) = H^*_q(\lambda). \qquad (11)$$

This property enables us to derive the following lower bound on $R_q(\tau, \sigma)$.

Lemma 6. *Let $q \geq 2$ be an integer and $\tau, \sigma \in (0, 1)$. Then, it holds that*

$$R_q(\tau, \sigma) \geq 1 + \tau - 2H^*_q(\tau) - (1 - \tau)H^*_q(\frac{2\sigma}{1-\tau}).$$

Proof. Theorem 3 states for $n \geq 1$ that

$$M_q(n, \tau n, \sigma n) \geq \frac{q^{n+\tau n}}{(S^{\tau n}_{n,q})^2 \cdot S^{2\sigma n}_{n-\tau n,q}}.$$

By applying this bound to the rate function $R_q(\tau, \sigma)$, it readily follows that

$$\begin{aligned} R_q(\tau, \sigma) \geq& \liminf_{n\to\infty} \frac{1}{n} \log_q \left(\frac{q^{n+\tau n}}{(S^{\tau n}_{n,q})^2 \cdot S^{2\sigma n}_{n-\tau n,q}} \right) \\ =& 1 + \tau - 2\liminf_{n\to\infty} \frac{1}{n} \log_q(S^{\tau n}_{n,q}) \\ &- \liminf_{n\to\infty} \frac{1}{n} \log_q(S^{2\sigma n}_{n-\tau n,q}) \\ =& 1 + \tau - 2H^*_q(\tau) \\ &- \liminf_{n'\to\infty} \frac{1-\tau}{n'} \log_q(S^{\frac{2\sigma}{1-\tau}n'}_{n',q}) \qquad (12) \\ =& 1 + \tau - 2H^*_q(\tau) \\ &- (1 - \tau)H^*_q(\frac{2\sigma}{1-\tau}), \end{aligned}$$

where we applied the change of variables $n' = n - \tau n$ in (12), and used (11) to evaluate the limit inferiori. □

[3] In what follows, we will be slightly imprecise by setting $t = \tau n$, $s = \sigma n$ which may not be integer-valued. However, in the asymptotic regime this does not change the over-all results.

V. Concluding remarks

In this paper, we have presented a non-asymptotic lower bound on the maximal cardinality of a t-indel s-substitution correcting code. In order to improve this lower bound, an interesting research challenge is to find an expression or tighter upper bound for the size of the set $\mathcal{V}_{t',t'',s}(\mathbf{x})$.

More generally, it could also be investigated whether the numerous existing lower and upper bounds on the maximum cardinality of either t-indel correcting codes or s-substitution correcting codes can be generalized to bounds on $M_q(n,t,s)$.

VI. Acknowledgments

We would like to thank Khaled Abdel-Ghaffar and Ludo Tolhuizen for their helpful comments which have greatly improved the quality of this paper.

References

[1] G. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science (New York, N.Y.)*, vol. 337, p. 1628, Sep. 2012.

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

[3] S. S. Parkin, M. Hayashi, and L. Thomas, "Magnetic domain-wall racetrack memory," *Science*, vol. 320, no. 5873, pp. 190–194, Apr. 2008.

[4] Y. M. Chee, H. M. Kiah, A. Vardy, V. K. Vu, and E. Yaakobi, "Coding for racetrack memories," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 7094–7112, 2018.

[5] R. Roth, *Introduction to Coding Theory*. Cambridge: Cambridge University Press, 2006.

[6] E. N. Gilbert, "A comparison of signalling alphabets," *The Bell System Technical Journal*, vol. 31, no. 3, pp. 504–522, May 1952.

[7] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," *Doklady Akademii Nauk SSSR*, vol. 117, no. 5, pp. 739–741, Jun. 1957.

[8] J. Tao and A. Vardy, "Asymptotic improvement of the gilbert-varshamov bound on the size of binary codes," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1655–1664, 2004.

[9] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966, doklady Akademii Nauk SSSR, vol. 163, no. 4, pp. 845-848, Aug. 1965.

[10] L. Tolhuizen, "The generalized Gilbert-Varshamov bound is implied by Turan's theorem," *IEEE Transactions on Information Theory*, vol. 43, no. 5, pp. 1605–1606, Sep. 1997.

[11] V. I. Levenshtein, "Bounds for deletion/insertion correcting codes," *Proceedings IEEE International Symposium on Information Theory*, Jul. 2002.

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

[13] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Three novel combinatorial theorems for the insertion/deletion channel," *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2702–2706, Jun. 2015.

[14] W. Song, N. Polyanskii, K. Cai, and X. He, "On multiple-deletion multiple-substitution correcting codes," *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 2655–2660, Sep. 2021.

[15] W. Song, N. Polyanskii, K. Cai, and X. He, "Systematic codes correcting multiple-deletion and multiple-substitution errors," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6402–6416, 2022.

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

[17] L. Tolhuizen, "Upper bounds on the size of insertion/deletion correcting codes," *Proceedings 8-th International Workshop on Algebraic and Combinatorial Coding Theory, Russia*, pp. 242–246, Sept. 2002.

[18] D. Cullina and N. Kiyavash, "An improvement to Levenshtein's upper bound on the cardinality of deletion correcting codes," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3862–3870, 2014.

[19] I. Smagloy, L. Welter, A. Wachter-Zeh, and E. Yaakobi, "Single-deletion single-substitution correcting codes," *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 775–780, Aug. 2020.

[20] M. Abu-Sini and E. Yaakobi, "On Levenshtein's reconstruction problem under insertions, deletions, and substitutions," *IEEE Transactions on Information Theory*, vol. 67, no. 11, pp. 7132–7158, Sep. 2021.

[21] V. I. Levenshtein, "Elements of the coding theory (in Russian)," *Discrete mathematics and mathematics problems of cybernetics Nauka, Moscow*, pp. 207–235, 1974.

[22] D. S. Hirschberg and M. Regnier, "Tight bounds on the number of string subsequences," *Journal of Discrete Algorithms*, vol. 1, Jun. 2001.

[23] 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, Mar. 2015.

[24] H. Mercier, M. Khabbazian, and V. K. Bhargava, "On the number of subsequences when deleting symbols from a string," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3279–3285, Jun. 2008.

[25] J. H. van Lint and R. M. Wilson, *A course in combinatorics*, 2nd ed. Cambridge university press, 2001.

[26] A. A. Kulkarni and N. Kiyavash, "Non-asymptotic upper bounds for deletion correcting codes," *IEEE Transactions on Information Theory*, vol. 59, no. 8, pp. 5115–5130, Apr. 2013.