Systematic Single-Deletion Multiple-Substitution Correcting Codes

Wentu Song*, Nikita Polyanskii†, Kui Cai*, and Xuan He*
*Science, Mathematics and Technology Cluster, Singapore University of Technology and Design, Singapore
†Technical University of Munich, Germany, and Skolkovo Institute of Science and Technology, Russia
Email: {wentu_song, cai_kui}@sutd.edu.sg, nikitapolyansky@gmail.com, helaoxuan@126.com

I. Introduction

The problem of constructing deletion/insertion correcting codes was introduced by Levenshtein [1] and recently has attracted an increasing attention due to their relevance to the DNA-based data storage [2]. In [1], Levenshtein proved that for any t-deletion correcting code $\mathscr{C}$ of length n, the redundancy of $\mathscr{C}$ (defined as $n-\log|\mathscr{C}|$) is asymptotically at least $t\log n+o(\log n)$, and an optimal t-deletion correcting code has redundancy at most $2t\log n+o(\log n)$.

The first class of single-deletion correcting codes with redundancy $\log n+O(1)$ are the well-known Varshamov-Tenengolts (VT) codes [3]. A decoding algorithm of the VT codes was proposed in [1], and a systematic encoding algorithm of the VT codes was proposed in [4], both have linear-time complexity. Multiple-deletion correcting codes with small asymptotical redundancy were studied in [5]−[10].

However, in many application scenarios, such as DNA data storage and file synchronization, it is necessary to correct the edit errors (i.e., the combination of insertions, deletions and substitutions), which motivates the problem of constructing codes that can correct insertions, deletions and/or substitutions. A modified VT construction with redundancy $\log n+O(1)$ was presented in [1] to correct a single edit. Quaternary codes correctting a single edit for DNA data storage were considered in [11]. In [12], a family of single-deletion single-substitution correcting codes with redundancy $6\log n+8$ was constructed using four VT-like parity check equations. The codes constructed in [10] are capable of correcting combination of insertions, deletions and substitutions such that the total number of insertions, deletions and substitutions is upper bounded by k. Such codes have redundancy $4k\log n+o(\log n)$, which is the best known construction with respect to redundancy.

In this paper, we study the problem of constructing single-deletion s-substitution correcting codes, i.e., codes that can correct any combination of a single deletion and up to s substitutions. It was shown by Smagloy *et.al.* in [12] that the redundancy r of such codes satisfies $r\geq(s+1)\log n+o(\log n)$. The main result of this paper is a construction of a family of single-deletion s-substitution correcting codes with a systematic encoding function and with redundancy r satisfying $r\leq(3s+4)\log n+o(\log n)$. The encoding and decoding complexity of the proposed codes are $O\left(n^{s+3}\right)$ and $O\left(n^{s+2}\right)$, respectively.

We use a set of higher order weight vectors, denoted by $\mathbf{a}^{(j)}$, $j=1,\ldots,2s+1$, to construct parity checks, where $\mathbf{a}^{(j)}=\left(1^{j-1},1^{j-1}+2^{j-1},\ldots,\sum_{i=1}^{n}i^{j-1}\right)$. Similar weight vectors are used in [9], [10] and [12]. According to the construction in [10], there exist codes correcting a single deletion and s substitutions with redundancy $4(s+1)\log n+o(\log n)$. In this work, by using a pre-coding function of BCH code and the syndrome compression technique [13], our construction achieves the redundancy of $(3s+4)\log n+o(\log n)$, decreasing by $s\log n$ compared to [10].

II. Deletion and Substitution Correcting Codes

Let t, s and L be positive integers such that $t+s<L$. For any $\mathbf{x}\in\{0,1\}^L$, $\mathscr{B}_{t,s}(\mathbf{x})$ denotes the set of all sequences that can be obtained from $\mathbf{x}$ by deletion of t symbols and substitution of at most s symbols. Let n be a positive integer. A set $\mathscr{C}\subseteq\{0,1\}^n$ is called a code of length n. The code $\mathscr{C}$ is called a *t-deletion s-substitution correcting code* if for any $\mathbf{c}\in\mathscr{C}$, $\mathbf{c}$ can be correctly recovered from any $\mathbf{y}\in\mathscr{B}_{t,s}(\mathbf{c})$.

Let $\mathscr{C}\subseteq\{0,1\}^n$ be a code and $k\in[n]=\{1,\cdots,n\}$. A set $I\subseteq[n]$ of size $|I|=k$ is said to be an *information set* of $\mathscr{C}$ if for every $\mathbf{x}\in\{0,1\}^k$, there is at least one codeword $\mathbf{c}\in\mathscr{C}$ such that $\mathbf{c}_I=\mathbf{x}$, where $\mathbf{c}_I$ is the subsequence of $\mathbf{x}$ by deleting all symbols $x_j, j\in[n]\backslash I$. Clearly, $k\leq\log|\mathscr{C}|$.

Clearly, if $\mathscr{C}$ has an encoding function $\mathcal{E}:\{0,1\}^k\rightarrow\{0,1\}^n$, then the redundancy of $\mathscr{C}$ equals to $n-k$.

Lemma 1: For any positive integer k, there exists a positive integer n_0 and a function $h:\{0,1\}^k\rightarrow\{0,1\}^{n_0}$ such that $n_0-k\leq s(\log(n_0)+2)$ and $\mathscr{C}_0\triangleq\{h(\mathbf{x}):\mathbf{x}\in\{0,1\}^k\}$ is a systematic linear code of minimum distance at least $2s+1$.

III. Construction of Single-Deletion s-Substitution Correcting Codes

In this section, we propose a family of systematic single-deletion s-substitution correcting codes. The length of the information sequences is denoted by k, where k is a positive integer. Let $h:\{0,1\}^k\rightarrow\{0,1\}^{n_0}$ be the function constructed in Lemma 1. Then the encoding function of the proposed code,

W. Song, K. Cai, and X. He's work was supported by Singapore Ministry of Education Academic Research Fund Tier 2 MOE2019-T2-2-123.

N. Polyanskii's work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under Grant No. WA3907/1-1.

This abstract is condensed from [15].

denoted by $\mathcal{E}$, is defined as

$$\mathcal{E}(\mathbf{x}) = \Big(h(\mathbf{x}), g(h(\mathbf{x})), \mathrm{Rep}_{2s+2}\big(f(g(h(\mathbf{x})))\big)\Big),\ \forall\ \mathbf{x} \in \{0,1\}^k, \tag{1}$$

where $\mathrm{Rep}_{2s+2}(\cdot)$ is the encoding function of the $(2s+2)$-fold repetition code. The functions $g : \{0,1\}^{n_0} \to \{0,1\}^{n_1}$ and $f : \{0,1\}^{n_1} \to \{0,1\}^{n_2}$ satisfy the following three conditions:

(C1) $n_1 = 2(s+2)\log(n_0) + o(\log(n_0))$ and $n_2 = o(\log(n_0))$;

(C2) For every $\mathbf{x} \in \{0,1\}^k$, $h(\mathbf{x})$ can be recovered from $g\big(h(\mathbf{x})\big)$ and any given sequence in $\mathscr{B}_{1,s}(h(\mathbf{x}))$;

(C3) For every $\mathbf{x} \in \{0,1\}^k$, $g(h(\mathbf{x}))$ can be recovered from $f(g(h(\mathbf{x})))$ and any given sequence in $\mathscr{B}_{1,s}(g(h(\mathbf{x})))$.

From any $\mathbf{y} \in \mathscr{B}_{1,s}(\mathcal{E}(\mathbf{x}))$, $\mathbf{x}$ can be recovered as follows: First, we can obtain three sequences $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}$ such that $\mathbf{y}^{(1)} \in \mathscr{B}_{1,s}(h(\mathbf{x}))$, $\mathbf{y}^{(2)} \in \mathscr{B}_{1,s}(g(h(\mathbf{x})))$ and $\mathbf{y}^{(3)} \in \mathscr{B}_{1,s}\big(\mathrm{Rep}_{2s+2}(f(g(h(\mathbf{x}))))\big)$. Note that $f\big(g(h(\mathbf{x}))\big)$ can always be recovered from $\mathbf{y}^{(3)}$. Then by condition (C3), $g\big(h(\mathbf{x})\big)$ can be recovered from $f\big(g(h(\mathbf{x}))\big)$ and $\mathbf{y}^{(2)}$. Further, by condition (C2), $h(\mathbf{x})$ can be recovered from $g\big(h(\mathbf{x})\big)$ and $\mathbf{y}^{(1)}$. Finally, $\mathbf{x}$ can be recovered from $h(\mathbf{x})$ by the inverse function h^{-1} of h. Hence, the encoding function $\mathcal{E}$ gives a single-deletion s-substitution correcting code. Since by Lemma 1, h is a systematic encoding function, so $\mathcal{E}$ is also a systematic encoding function. Moreover, by the construction, the proposed code has length $n = n_0 + n_1 + n_2 \geq n_0$, so by Lemma 1 and condition (C1), its redundancy r satisfies $r \leq n_0 - k + n_1 + n_2 \leq s(\log(n_0)+2) + 2(s+2)\log(n_0) + o(\log(n_0)) \leq (3s+4)\log n + o(\log n)$.

Let $L \geq 3$ be an arbitrarily fixed integer and $\xi(L) = (s+1)(2s+1)\log L + (2s+1)\log(2s+1)$. Then we can construct a function $f : \{0,1\}^L \to \{0,1\}^{\lceil \xi(L) \rceil}$ such that

$$f(\mathbf{x})_j = \mathbf{x} \cdot \mathbf{a}^{(j)} \mod (2s+1)L^j,\ j \in \{1, \cdots, 2s+1\}.$$

The construction of f is similar to the Sima-Bruck-Gabrys construction in [10]. For the case of $t = 1$, the construction in [10] consists of $2(s+1)+1 = 2s+3$ components, that is, $f(\mathbf{x}) = \big(f(\mathbf{x})_1, f(\mathbf{x})_2, \ldots, f(\mathbf{x})_{2s+3}\big)$, while in this paper, we prove that $2s+1$ components are sufficient, that is, we only need $f(\mathbf{x}) = \big(f(\mathbf{x})_1, f(\mathbf{x})_2, \ldots, f(\mathbf{x})_{2s+1}\big)$.

Let $L = n$ and let $\mathscr{C}_{\mathbf{r}} = \big\{\mathbf{c} \in \{0,1\}^n : f(\mathbf{c}) = \mathbf{r}\big\}$ for any fixed $\mathbf{r} = (r_1, r_2, \ldots, r_{2s+1}) \in \prod_{j=1}^{2s+1} \big[0, (2s+1)n^j - 1\big]$. Then we can prove that $\mathscr{C}_{\mathbf{r}}$ is a single-deletion s-substitution correcting code with redundancy $r(\mathscr{C}_{\mathbf{r}})$ satisfies $r(\mathscr{C}_{\mathbf{r}}) \leq (s+1)(2s+1)\log n + (2s+1)\log(2s+1)$. For $s = 1$, $\mathscr{C}_{\mathbf{r}}$ is a single-deletion single-substitution correcting code with redundancy $r(\mathscr{C}_{\mathbf{r}}) \leq 6\log n + 3$, which is similar to the Construction 11 of [12], where an additional component $f(\mathbf{x})_0 = \sum_{i=1}^{n} x_i \bmod 5$ is used and the redundancy of the corresponding code is at most $6\log n + 8$.

For $s \geq 2$, by utilizing the compression method, we can further reduce the redundancy (see the following lemma).

Lemma 2: Let h be constructed as in Lemma 1 and $n_1 = 2(s+2)\log(n_0) + o(\log(n_0))$. There exists a function $g : \{0,1\}^{n_0} \to \{0,1\}^{n_1}$ such that for any $\mathbf{x} \in \{0,1\}^k$, $h(\mathbf{x})$ can be recovered from $g(h(\mathbf{x}))$ and any $\mathbf{y} \in \mathscr{B}_{1,s}(h(\mathbf{x}))$. Moreover, $g(h(\mathbf{x}))$ can be computed in time $O\big((n_0)^{s+3}\big)$, and $h(\mathbf{x})$ can be computed from $g(h(\mathbf{x}))$ and $\mathbf{y}$ in time $O\big((n_0)^{s+2}\big)$.

Now, we can present our main result of this paper.

Theorem 1: Let f be constructed with $L = n_1$. The encoding function $\mathcal{E}$ defined by (1) gives a systematic single-deletion s-substitution correcting code $\mathscr{C}$ of length $n = n_0 + n_1 + n_2$. The redundancy $r(\mathscr{C})$ of $\mathscr{C}$ satisfies $r(\mathscr{C}) \leq (3s+4)\log n + o(\log n)$, and the encoding and decoding complexity of $\mathscr{C}$ are $O\big(n^{s+3}\big)$ and $O\big(n^{s+2}\big)$, respectively.

IV. Discussions and Future Work

The key improvement of our construction is a pre-coding process using the BCH codes, i.e., the function h constructed by Lemma 1. This technique can also be generalized to the construction of t-deletion s-substitution correcting codes for $t > 1$ such that the redundancy decreases by $s\log n$ compared to the construction in [10].

Using a similar approach of pre-coding we can obtain an explicit construction of t-deletion correcting codes whose redundancy is $(4t-1)\log n$ (improved by $\log n$ compared to the construction in [10]). Another possible line of research is nonbinary t-deletion s-substitution correcting codes. For nonbinary case, we can use codes from [14] for pre-coding.

References

[1] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals (in Russian)," *Doklady Akademii Nauk SSR*, vol. 163, no. 4, pp. 845-848, 1965.

[2] R. Heckel, G. Mikutis, and R. N. Grass, "A Characterization of the DNA Data Storage Channel," *Scientific Reports*, vol. 9, no. 1, pp. 9663, 2019. Available online at: https://doi.org/10.1038/s41598-019-45832-6

[3] R. R. Varshamov and G. M. Tenengolts, "Codes which correct single asymmetric errors (in Russian)," *Automatika i Telemkhanika*, vol. 161, no. 3, pp. 288-292, 1965.

[4] K. A. S. Abdel-Ghaffar and H. C. Ferreira, "Systematic encoding of the Varshamov-Tenengolts codes and the Constantin-Rao codes," *IEEE Trans Inf. Theory*, vol. 44, no. 1, pp. 340-345, 1998.

[5] J. Brakensiek, V. Guruswami, and S. Zbarsky, "Efficient low-redundancy codes for correcting multiple deletions," *IEEE Trans. on Inf. Th.*, vol. 64, no. 5, pp. 3403-3410, 2018.

[6] R. Gabrys and F. Sala, "Codes correcting two deletions," *IEEE Trans. Inform. Theory*, vol. 65, no. 2, pp. 965-974, Feb 2019.

[7] J. Sima, N. Raviv, and J. Bruck, "Two deletion correcting codes from indicator vectors," *IEEE Trans. Inform. Theory*, pp. 1-1, 2019.

[8] V. Guruswami and Johan Håstad, "Explicit two-deletion codes with redundancy matching the existential bound," arXiv preprint arXiv:2007.10592 (2020).

[9] J. Sima and J. Bruck, "Optimal k-deletion correcting codes," 2019, Available online at: https://arxiv.org/abs/1910.12247

[10] J. Sima, R. Gabrys, and J. Bruck, "Optimal Systematic t-Deletion Correcting Codes," in *Proc. ISIT*, 2020.

[11] K. Cai, Y. M. Chee, R. Gabrys, H. M. Kiah, and T. T. Nguyen, "Optimal Codes Correcting a Single Indel/Edit for DNA-Based Data Storage," accepted by *IEEE Trans. Inform. Theory*, Special Issue Dedicated to the Memory of Vladimir I. Levenshtein, Nov. 2020. Available online at arXiv:1910.06501.

[12] I. Smagloy, L. Welter, A. Wachter-Zeh, and E. Yaakobi, "Single-Deletion Single-Substitution Correcting Codes," 2020, Available online at: https://arxiv.org/abs/2005.09352

[13] J. Sima, R. Gabrys, and J. Bruck, "Syndrome Compression for Optimal Redundancy Codes," in *Proc. ISIT*, 2020.

[14] S. Yekhanin and I. Dumer, "Long nonbinary codes exceeding the Gilbert-Varshamov bound for any fixed distance," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2357-2362, Oct. 2004.

[15] W. Song, N. Polyanskii, K. Cai, and X. He, "Systematic Single-Deletion Multiple-Substitution Correcting Codes," 2020, Available online at: https://arxiv.org/abs/2006.11516