Optimal Reconstruction Codes for Deletion Channels

Johan Chrisnata*, Han Mao Kiah†, and Eitan Yaakobi*
*Department of Computer Science, Technion — Israel Institute of Technology, Haifa, 3200003 Israel
†School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371
Emails: johanchr001@ntu.edu.sg, hmkiah@ntu.edu.sg, yaakobi@cs.technion.ac.il

Abstract—**The sequence reconstruction problem, introduced by Levenshtein in 2001, considers a communication scenario where the sender transmits a codeword from some codebook and the receiver obtains multiple noisy reads of the codeword. Motivated by modern storage devices, we introduced a variant of the problem where the number of noisy reads N is fixed (Kiah *et al.* 2020). Of significance, for the single-deletion channel, using $\log_2 \log_2 n + O(1)$ redundant bits, we designed a reconstruction code of length n that reconstructs codewords from two distinct noisy reads.**

In this work, we show that $\log_2 \log_2 n - O(1)$ redundant bits are necessary for such reconstruction codes, thereby, demonstrating the optimality of our previous construction. Furthermore, we show that these reconstruction codes can be used in t-deletion channels (with $t \geqslant 2$) to uniquely reconstruct codewords from $n^{t-1} + O\left(n^{t-2}\right)$ distinct noisy reads.

I. Introduction

As our data needs surge, new technologies emerge to store these huge datasets. Interestingly, besides promising ultra-high storage density, certain emerging storage media, such as DNA based storage [1] and racetrack memories [2], [3], rely on technologies that provide users with multiple cheap, albeit noisy, reads. In our companion paper [4], we proposed a *coding solution* to leverage on these multiple reads to increase the information capacity, or equivalently, reduce the number of redundant bits.

Our code design problem is based on the *sequence reconstruction problem*, formulated by Levenshtein [5]. In Levenshtein's seminal work, he considers a communication scenario where the sender transmits a codeword from some codebook and the receiver obtains multiple noisy reads of the codeword. The common setup assumes the codebook to be the entire space and the problem is to determine the minimum number of distinct reads N that is required to reconstruct the transmitted codeword. In constrast, in our problem, the parameter N is fixed and our task is to design a *codebook* such that every codeword can be uniquely reconstructed from any N distinct noisy reads.

Hence, our fundamental problem is then: how large can this codebook be? Or equivalently, what is the *minimum redundancy*? Modifying a code construction in [3], we provided in [4] a number of reconstruction codes for the single-edit channel and its variants with $\log_2 \log_2 n + O(1)$ bits of redundancy. In this work, we focus on the converse of the problem and demonstrate that $\log_2 \log_2 n - O(1)$ redundant bits are *necessary*. To ease our exposition, we focus on channels with *deletions only*. Due to space constraints, we only provide an overview of the results and discuss their implications. Details can be found in [6] and the results have been presented in ISITA 2020.

II. Preliminaries

Consider a data storage scenario described by an error-ball function. Formally, given an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, an *error-ball* function B maps a *word* $\boldsymbol{x} \in \mathcal{X}$ to a subset of *noisy reads* $B(\boldsymbol{x}) \subseteq \mathcal{Y}$. Given a code $\mathcal{C} \subseteq \mathcal{X}$, we define the *read coverage* of $\mathcal{C}$, denoted by $\nu(\mathcal{C}; B)$, to be the quantity

$$\nu(\mathcal{C}; B) \triangleq \max\left\{|B(\boldsymbol{x}) \cap B(\boldsymbol{y})| : \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}, \boldsymbol{x} \neq \boldsymbol{y}\right\}.$$

In other words, $\nu(\mathcal{C}; B)$ is the maximum intersection between the error-balls of any two codewords in $\mathcal{C}$. The quantity $\nu(\mathcal{C}; B)$ was introduced by Levenshtein [5], where he showed that the number of reads[1] required to reconstruct a codeword from $\mathcal{C}$ is at least $\nu(\mathcal{C}; B) + 1$. The problem to determine $\nu(\mathcal{C}; B)$ is referred to as the *sequence reconstruction problem.*

The sequence reconstruction problem was studied in a variety of storage and communication scenarios [3], [7]–[10]. In these cases, $\mathcal{C}$ is usually assumed to be the entire space (all binary words of some fixed length) or a classical error-correcting code. However, in most storage scenarios, the number of noisy reads N is a fixed system parameter and when N is at most $\nu(\mathcal{C}; B)$, we are unable to uniquely reconstruct the codeword. In [4], we propose the study of *code design* when the read coverage is strictly less than $\nu(\mathcal{C}; B)$. Specifically, we say that $\mathcal{C}$ is an $(n, N; B)$*-reconstruction code* if $\mathcal{C} \subseteq \{0,1\}^n$ and $\nu(\mathcal{C}; B) < N$.

This gives rise to a *new quantity of interest* that measures the *trade-off between codebook redundancy and read coverage*. Specifically, given N and an error-ball B, we study the quantity

$$\rho(n, N; B) \triangleq \min\left\{n - \log|\mathcal{C}| : \mathcal{C} \subseteq \{0,1\}^n,\ \nu(\mathcal{C}; B) < N\right\}.$$

A. The Sequence Reconstruction Problem for Deletion Channels

In this work, we focus on channels that introduce *deletions only*. Specifically, let $\mathcal{D}_t(\boldsymbol{x})$ denote the deletion ball of $\boldsymbol{x}$ with exactly t deletions. Let $D_t(n)$ denote the maximum deletion ball size of words of length n, that is, $D_t(n) = \max\{|\mathcal{D}_t(\boldsymbol{x})| : \boldsymbol{x} \in \{0,1\}^n\}$. It is well known (see for example, [11]) that

$$D_t(n) = \sum_{i=0}^{t} \binom{n-t}{i} = n^t + O(n^{t-1}), \text{ for } 0 \leqslant t \leqslant n.$$

For convenience, we assign $D_t(n) = 0$ when $t < 0$ or $t > n$.

For purposes of brevity, we let $\nu_t(n)$ denote $\nu(\{0,1\}^n; \mathcal{D}_t)$, the read coverage of $\{0,1\}^n$. We have the following landmark result of Levenshtein.

Theorem 1 (Levenshtein [11]).

$$\nu_t(n) = 2D_{t-1}(n-2) = 2n^{t-1} + O(n^{t-2}).$$

Recently, the authors of [8] studied the sequence reconstruction problem when $\mathcal{C}$ is a single-deletion-correcting code or an $(n, 1; \mathcal{D}_1)$-reconstruction code. Namely, they showed that $\mathcal{C}$ allows unique reconstruction with significantly less reads (as compared to $\nu_t(n)$) for deletions with $t \geqslant 2$.

B. Reconstruction Codes with $N = 2$ for Single Deletions

When we use the whole space $\{0,1\}^n$ as our codebook, we require $\nu_1(n) + 1 = 3$ noisy reads to uniquely reconstruct any codeword. Hence, we have $\rho(n, N; \mathcal{D}_1) = 0$ for $N \geqslant 3$.

In contrast, when $N = 1$, or, when we have only one noisy read, we recover the usual notion of error-correcting codes and

[1] In the original paper, Levenshtein used the term "channels", instead of reads. Here, we used the term "reads" to reflect the data storage scenario.

the classical Varshamov-Tenengolts (VT) code is an $(n,1;\mathcal{D}_1)$-reconstruction code whose redundancy is at most $\log_2(n+1)$ [12]. Hence, we have $\rho(n,1;\mathcal{D}_1)=\log_2 n+\Theta(1)$. Therefore, it remains to ask: how should we design the codebook when we have only two noisy reads? Or, what is the value of $\rho\left(n,2;\mathcal{D}_1\right)$?

Now, the first construction of a $(n,2;\mathcal{D}_1)$-reconstruction code was proposed in [3] for the design of codes in racetrack memory. The codebook uses $\log_2\log_2 n+O(1)$ redundant bits and in [4], we modified the construction to obtain codebooks that uniquely reconstruct codewords for the single-edit channel and its variants. The construction can be seen as a generalization of the classical VT code proposed by Levenshtein [12] and the shifted VT codes proposed by Schoeny *et al.* [13].

Definition 2 (Constrained Shifted VT Codes [3], [4]). For $n\geqslant P>0$ and P even, let $c\in\mathbb{Z}_{1+P/2}$ and $d\in\mathbb{Z}_2$. The *constrained shifted VT code* $\mathcal{C}_{\rm CSVT}(n,P;c,d)$ is defined to be the set of all words $\boldsymbol{x}=x_1x_2\cdots x_n$ such that the following holds.

(i) $\mathrm{Syn}(\boldsymbol{x})=c \pmod{1+P/2}$.
(ii) $\sum_{i=1}^n x_i=d \pmod 2$.
(iii) The longest 2-periodic run in $\boldsymbol{x}$ is at most P.

Here, $\mathrm{Syn}(\boldsymbol{x})$ denotes the *VT syndrome* $\mathrm{Syn}(\boldsymbol{x})\triangleq\sum_{i=1}^n ix_i$ and a *2-periodic* run refers to a continguous substring $x_ix_{i+1}\cdots x_j$ where $x_k=x_{k+2}$ for all $i\leqslant k\leqslant j-2$.

When $P=2n$ and we remove Condition (ii)[2] we recover the classical VT code that corrects a single deletion. On the other hand, when we remove the Condition (iii), we recover the shifted VT code that is used in the correction of a single burst of deletions [13]. It was recently demonstrated that the CSVT code enables unique reconstruction whenever we have two distinct noisy reads.

Theorem 3 ([3], [4]). *For all choices of c and d, we have that $\mathcal{C}_{\rm CSVT}(n,P;c,d)$ is an $(n,2;\mathcal{D}_1)$-reconstruction code. Furthermore, if we set $P=\lceil\log_2 n\rceil+2$, the code $\mathcal{C}_{\rm CSVT}(n,P;c,d)$ has redundancy $1+\log_2(\lceil\log+2n\rceil+4)=\log_2\log_2 n+O(1)$ for some choice of c and d. Thus, $\rho(n,2;\mathcal{D}_1)\leqslant\log_2\log_2 n+O(1)$.*

III. Main Contributions

Our first contribution is to show that the codes in Theorem 3 are asymptotically *optimal*. Specifically, we show that an $(n,2;\mathcal{D}_1)$-reconstruction code requires at least $\log_2\log_2 n-O(1)$ redundant bits. In our proof, we cast the code construction as an independent set problem in graph theory and borrow graph theoretic tools to provide the necessary bounds.

Theorem 4. *Let $\mathcal{C}$ be an $(n,2;\mathcal{D}_1)$-reconstruction code. For $\epsilon>0$, we have that*

$$\log_2|\mathcal{C}|\leqslant n-\log_2\log_2 n+\log_2(1-\epsilon)+o(1).$$

Therefore, $\rho(n,2;\mathcal{D}_1)=\log_2\log_2 n-O(1)$. Combining with Theorem 3, we have that $\rho(n,2;\mathcal{D}_1)=\log_2\log_2 n+\Theta(1)$.

Therefore, the CSVT code constructed in Theorem 3 is asymptotically optimal and we have that $\rho(n,2;\mathcal{D}_1)=\log\log n+\Theta(1)$. Hence, we have the complete solution for ρ in the case for $t=1$.

Theorem 5. *The value $\rho(n,N;\mathcal{D}_1)$ satisfies*

$$\rho(n,N;\mathcal{D}_1)=\begin{cases}\log_2 n+\Theta(1), & \textit{when } N=1,\\ \log_2\log_2 n+\Theta(1), & \textit{when } N=2,\\ 0, & \textit{when } N\geqslant 3.\end{cases}$$

Theorem 5 shows that as the number of noisy reads increases, the optimal number of redundant bits is gracefully reduced from $\log_2 n+\Theta(1)$ to $\log_2\log_2 n+\Theta(1)$, and then to zero.

For our second contribution, in the spirit of [8], we look at an $(n,2;\mathcal{D}_1)$-reconstruction code $\mathcal{C}$ and study its performance when the channel introduces more deletions. To this end, we derived the following combinatorial result.

Theorem 6. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be binary words of length $n\geqslant 7$ and $t\geqslant 2$. If $|\mathcal{D}_1(\boldsymbol{x})\cap\mathcal{D}_1(\boldsymbol{y})|=1$, then we have that*

$$\begin{aligned}|\mathcal{D}_t(\boldsymbol{x})\cap\mathcal{D}_t(\boldsymbol{y})|&\leqslant D_{t-1}(n-3)+D_{t-2}(n-3)+2D_{t-2}(n-5)\\&\leq D_{t-1}(n-1)+\nu_{t-1}(n-3)\\&=n^{t-1}+O(n^{t-2})\ \textit{for fixed values of } t.\end{aligned}$$

Hence, consider any two codewords $\boldsymbol{x}$ and $\boldsymbol{y}$ from an $(n,2;\mathcal{D}_1)$-reconstruction code $\mathcal{C}$. Since $|\mathcal{D}_1(\boldsymbol{x})\cap\mathcal{D}_1(\boldsymbol{y})|\leq 1$, from Theorem 6, we have that the read coverage $\nu(\mathcal{C};\mathcal{D}_t)$ is at most $N_t^{(2)}(n)$ where $N_t^{(2)}(n)=D_{t-1}(n-1)+\nu_{t-1}(n-3)$. Hence, $\mathcal{C}$ is an $(n,N_t^{(2)}(n)+1;\mathcal{D}_t)$-reconstruction code. Therefore, using the codes from Theorem 3, we have the following result.

Theorem 7. *Let $n\geqslant 6$ and $t\geqslant 2$. Set $N_t^{(1)}(n)=2D_{t-2}(n-4)+2D_{t-2}(n-5)+2D_{t-2}(n-7)+D_{t-3}(n-6)+D_{t-3}(n-7)$, $N_t^{(2)}(n)=D_{t-1}(n-1)+\nu_{t-1}(n-3)$ and $N_t'(n)=\max\left\{N_t^{(1)}(n),N_t^{(2)}(n)\right\}$. If $\mathcal{C}$ is an $(n,2;\mathcal{D}_1)$-reconstruction code, then $\mathcal{C}$ is also an $(n,N_t'(n)+1;\mathcal{D}_t)$-reconstruction code. Furthermore, when the value of t is fixed, this implies that $\rho\left(n,N_t^{(2)}(n)+1;\mathcal{D}_t\right)\leqslant\log_2\log_2 n+O(1)$.*

References

[1] S. Yazdi, H. M. Kiah, E. R. Garcia, J. Ma, H. Zhao, and O. Milenkovic. DNA-based storage: Trends and methods. *IEEE Trans. Molecular, Biological, Multi-Scale Commun.*, 1(3):230–248, 2015.

[2] S. S. Parkin, M. Hayashi, and L. Thomas, "Magnetic domain-wall racetrack memory," *Science*, vol. 320, pp. 190–194, 2008.

[3] Y. M. Chee, H. M. Kiah, A. Vardy, E. Yaakobi, and V. K. Vu. "Coding for racetrack memories," *IEEE Trans. on Information Theory*, 2018.

[4] H. M. Kiah, T. T. Nguyen and E. Yaakobi, "Coding for Sequence Reconstruction for Single Edits," In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, Jun. 2020. (*arXiv preprint arxiv:2001.01376)*

[5] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. on Information Theory*, 47(1), pp. 2–22, 2001.

[6] J. Chrisnata, H. M. Kiah, and E. Yaakobi, "Optimal Reconstruction Codes for Deletion Channels," In *Proc. Intl. Symp. Inform. Theory and Its Appl. (ISITA)*, Hawaii, Oct. 2020. (*arXiv preprint arxiv:2004.06032*)

[7] M. Cheraghchi, R. Gabrys, O. Milenkovic and J. Ribeiro, "Coded trace reconstruction," *arXiv preprint arxiv:1903.09992*, 2019

[8] R. Gabrys, and E. Yaakobi. "Sequence reconstruction over the deletion channel," *IEEE Trans. on Information Theory*, 64(4), pp.2924-2931, 2018.

[9] Y. Yehezkeally and M. Schwartz. "Reconstruction codes for DNA sequences with uniform tandem-duplication errors," In *Information Theory (ISIT), 2018 IEEE International Symposium on*, pages 2535–2539. IEEE, 2018.

[10] M. Abu Sini, and E. Yaakobi, "Reconstruction of Sequences in DNA Storage". In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 2019.

[11] V. I. Levenshtein, "Efficient Reconstruction of Sequences from Their Subsequences or Supersequences," *Journal of Combinatorial Theory, Series A*, 93, pp. 310–332, 2001.

[12] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[13] C. Schoeny, A. Wachter-Zeh, R. Gabrys, and E. Yaakobi. "Codes correcting a burst of deletions or insertions." *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 1971–1985, 2017.

[2] When $P=2n$, then any 2-periodic run is at most $n<P$. Hence, Condition (iii) is always true.