Reconstruction from Substrings with Partial Overlap

Yonatan Yehezkeally[*], **Daniella Bar-Lev**[†], **Sagi Marcovich**[†], and **Eitan Yaakobi**[†]
[*]Institute for Communications Engineering, Technical University of Munich, 80333 Munich, Germany
[†]Department of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel

***Abstract*—This paper introduces a new family of reconstruction codes which is motivated by applications in DNA data storage and sequencing. In such applications, DNA strands are sequenced by reading some subset of their substrings. While previous works considered two extreme cases in which *all* substrings of some fixed length are read or substrings are read with no overlap, this work considers the setup in which consecutive substrings are read with some given minimum overlap. First, an upper bound is provided on the attainable rates of codes that guarantee unique reconstruction. Then, we present efficient constructions of asymptotically optimal codes that meet that upper bound.**

I. Introduction

String reconstruction refers to a large class of problems where information about a string can only be obtained in the form of multiple, incomplete and/or noisy observations. Examples of such problems are the *reconstruction problem* by Levenshtein [11], the *trace reconstruction problem* [3], and the k*-deck problem* [13].

Notably, when observations are comprised of unordered consecutive substrings, the *reconstruction from substring-compositions problem* [1], [4], [9], [14], [16], [23] and the *torn-paper problem* [2], [17], [18], [20] (closely related to the *shuffling channel* [8], [10], [22]) have received significant interest in the past decade due to applications in DNA- or polymer-based storage systems, resulting from contemporary sequencing technologies [4], [7], [16]. The former arises from an idealized assumption of full overlap (and uniform coverage) in read substrings, while the latter results from an assumption of no overlap; in applications, this models the question of whether the complete information string may be replicated and uniformly segmented for sequencing, or if segmentation occurs adversarially in the medium prior to sequencing.

Motivated by these two paradigms, we study in this paper a generalized (or intermediate) setting where an information string is observed through an arbitrary collection of its substrings, where the minimum length $L_{\min}$ of each retrieved substring, as well as the length L_{over} of overlap between consecutive substrings, are bounded from below. A similar setting was recently studied in [19], where both substrings' lengths and overlap were assumed to be random; we study the problem in the aforementioned worst-case regime. We ask what the minimum value of $L_{\min}$ is for which there exist codes of length-n strings allowing for unique reconstruction in this channel with asymptotically non-vanishing rates, and then what is the asymptotically optimal obtainable rate given the value of L_{over}. Having answered both questions, we demonstrate that in these regimes it is possible to efficiently encode and decode information for unique reconstruction attaining asymptotically optimal rates.

The rest of this paper is organized as follows. In Section II we present notation and definitions used throughout the paper. In Section III we bound from above the asymptotically attainable rate of codes for unique reconstruction as a function of $L_{\min}, L_{\text{over}}$, and then in Section IV we develop efficient encoding and decoding algorithms for such codes, asymptotically meeting this bound. Due to space limitations, some proofs have been delegated to an extended version in preprint on arXiv.org.

II. Definitions and Preliminaries

Let Σ be a finite alphabet of size q. Where advantageous, we assume Σ is equipped with a ring structure, and in particular identify elements $0, 1 \in \Sigma$. For a positive integer n, let $[n]$ denote the set $[n] \triangleq \{0, 1, \ldots, n-1\}$.

For two non-negative functions f, g of a common variable n, denoting $L \triangleq \limsup_{n\to\infty} \frac{f(n)}{g(n)}$ (in the wide sense) we say that $f = o_n(g)$ if $L = 0$, $f = \Omega_n(g)$ if $L > 0$, $f = O_n(g)$ if $L < \infty$, and $f = \omega_n(g)$ if $L = \infty$. We say that $f = \Theta_n(g)$ if $f = \Omega_n(g)$ and $f = O_n(g)$. If f is not positive, we say $f = O_n(g)$ ($f = o_n(g)$) if $|f| = O_n(g)$ (respectively, $|f| = o_n(g)$). If clear from context, we omit the subscript from aforementioned notations.

Let Σ^* denote the set of all finite strings over Σ. The length of a string $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1}) \in \Sigma^*$ is denoted by $|\boldsymbol{x}| = n$. For strings $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^*$, we denote their concatenation by $\boldsymbol{x} \circ \boldsymbol{y}$. We say that $\boldsymbol{v}$ is a *substring* of $\boldsymbol{x}$ if there exist strings $\boldsymbol{u}, \boldsymbol{w}$ such that $\boldsymbol{x} = \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}$. If $\boldsymbol{u}$ (respectively, $\boldsymbol{w}$) is empty, we say that $\boldsymbol{v}$ is a *prefix* (*suffix*) of $\boldsymbol{x}$. If the length of $\boldsymbol{v}$ is ℓ, we specifically say that $\boldsymbol{v}$ is an ℓ*-substring* of $\boldsymbol{x}$ (similarly, an ℓ-pre/suffix). For $I \subseteq [|\boldsymbol{x}|]$, we let $\boldsymbol{x}_I$ denote the subsequence of $\boldsymbol{x}$ obtained by restriction to the coordinates of I; specifically, for $i \in [|\boldsymbol{x}| - \ell + 1]$ we denote by $\boldsymbol{x}_{i+[\ell]}$ the ℓ-substring of $\boldsymbol{x}$ at *location* i (we reserve the term *index* for a different use), where $i + [\ell] = \{i + j : j \in [\ell]\}$.

We consider in this paper the problem of string reconstruction from substrings with partial overlap. That is, we assume that a message $\boldsymbol{x} \in \Sigma^n$ is observed only through a multiset of its substrings, without order, with the following restrictions: (i) all observed substrings are of length at least $L_{\min}$; and (ii) succeeding substrings overlap with length at least L_{over} (in particular, every symbol of $\boldsymbol{x}$ is observed in some substring).

This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant 801434. Y. Yehezkeally was supported by a Carl Friedrich von Siemens postdoctoral research fellowship of the Alexander von Humboldt Foundation. D. Bar-Lev, S. Marcovich, and E. Yaakobi were supported in part by the United States-Israel BSF grant no. 2018048.

The first three authors contributed equally to this work.

More formally, a *substring-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset $\big\{\{\boldsymbol{x}_{i_j+[\ell_j]} : 1 \leqslant j \leqslant m\}\big\}$, for some $m \in \mathbb{N}$, such that $i_1 < i_2 < \cdots < i_m$ and $\ell_j \in [n - i_j + 1]$. A substring-trace is *complete* if $i_1 = 0$, $i_{j+1} < i_j + \ell_j$ for all $j < m$, and $i_m + \ell_m - 1 = n$. A complete substring-trace of $\boldsymbol{x} \in \Sigma^n$ is called an $(L_{\min}, L_{\over})$-*trace* if $\ell_j \geqslant L_{\min} \geqslant L_{\over}$ for all j, and $i_j + \ell_j - i_{j+1} \geqslant L_{\over}$ for all $j < m$. For example, for $\boldsymbol{x} = 11101110101111$

- $\{\{1110111, 111010, 101111\}\}$ is a $(6, 2)$-trace of $\boldsymbol{x}$;
- $\{\{111011, 110101, 101111\}\}$ is a complete substring-trace of $\boldsymbol{x}$ which is not a $(6, 2)$-trace; and
- $\{\{110111, 110101, 01111\}\}$ is a substring-trace of $\boldsymbol{x}$ which is not complete (since $i_1 > 0$).

The $(L_{\min}, L_{\over})$-*trace spectrum* of $\boldsymbol{x} \in \Sigma^n$, denoted $\mathcal{T}^{L_{\over}}_{L_{\min}}(\boldsymbol{x})$, is the set of all $(L_{\min}, L_{\over})$-traces of $\boldsymbol{x}$. Our channel accepts $\boldsymbol{x} \in \Sigma^n$ and outputs some $(L_{\min}, L_{\over})$-trace of $\boldsymbol{x}$ error free.

For all $\mathcal{C} \subseteq \Sigma^n$ we denote the *rate*, *redundancy* of $\mathcal{C}$ by $R(\mathcal{C}) \triangleq \frac{1}{n}\log|\mathcal{C}|$, $\mathrm{red}(\mathcal{C}) \triangleq n - \log|\mathcal{C}|$, respectively. Throughout the paper, we use the base-q logarithms. Motivated by the above channel definition, a code $\mathcal{C} \subseteq \Sigma^n$ is called an $(L_{\min}, L_{\over})$-*trace code* if for all $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in \mathcal{C}$, $\mathcal{T}^{L_{\over}}_{L_{\min}}(\boldsymbol{x}_1) \cap \mathcal{T}^{L_{\over}}_{L_{\min}}(\boldsymbol{x}_2) = \emptyset$. The main goal of this work is to find, for $L_{\min}, L_{\over}$ as functions of n, the maximum asymptotic rate of $(L_{\min}, L_{\over})$-trace codes. We will also be interested in efficient constructions of $(L_{\min}, L_{\over})$-trace codes with rate asymptotically approaching that value.

For convenience of analysis we denote by $\mathcal{L}^{L_{\over}}_{L_{\min}}(\boldsymbol{x}) \in \mathcal{T}^{L_{\over}}_{L_{\min}}(\boldsymbol{x})$, for $\boldsymbol{x} \in \Sigma^n$, the $(L_{\min}, L_{\over})$-trace of $\boldsymbol{x}$ containing specifically its $L_{\min}$-prefix, and subsequent $L_{\min}$-substrings overlapping in precisely $L_{\over}$ coordinates. For example, if $\boldsymbol{x} = 11101110101111$ then

$$\mathcal{L}^2_4(\boldsymbol{x}) = \{\{1110, 1011, 1110, 1010, 1011, 1111\}\}.$$

(Here, if $L_{\min} - L_{\over}$ does not divide $n - L_{\min}$ we allow the $L_{\min}$-suffix to contain a longer overlap with its preceding $L_{\min}$-substring; in the sequel we assume for ease of presentation that this does not occur, a fact that again shall not affect our asymptotic analysis.) Since $\mathcal{L}^{L_{\over}}_{L_{\min}}(\boldsymbol{x}) \in \mathcal{T}^{L_{\over}}_{L_{\min}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Sigma^n$, observe that any $(L_{\min}, L_{\over})$-trace code $\mathcal{C} \subseteq \Sigma^n$ satisfies

$$|\mathcal{C}| \leqslant \left|\left\{\mathcal{L}^{L_{\over}}_{L_{\min}}(\boldsymbol{x}) : \boldsymbol{x} \in \Sigma^n\right\}\right|; \tag{1}$$

Applying the *profile-vectors* argument [5], we count the incidences of each possible $\boldsymbol{u} \in \Sigma^{L_{\min}}$ in $\mathcal{L}^{L_{\over}}_{L_{\min}}(\boldsymbol{x})$ and observe that the sum of incidences equals $1 + \left\lceil \frac{n-L_{\min}}{L_{\min}-L_{\over}} \right\rceil = \left\lceil \frac{n-L_{\over}}{L_{\min}-L_{\over}} \right\rceil$; thus, we have an embedding of $\left\{\mathcal{L}^{L_{\over}}_{L_{\min}}(\boldsymbol{x}) : \boldsymbol{x} \in \Sigma^n\right\}$ into

$$\left\{ f \colon \Sigma^{L_{\min}} \to \mathbb{N} : \sum_{\boldsymbol{u} \in \Sigma^{L_{\min}}} f(\boldsymbol{u}) = \left\lceil \frac{n - L_{\over}}{L_{\min} - L_{\over}} \right\rceil \right\},$$

and therefore

$$\begin{aligned} |\mathcal{C}| &\leqslant \binom{\left\lceil \frac{n-L_{\over}}{L_{\min}-L_{\over}} \right\rceil + q^{L_{\min}} - 1}{q^{L_{\min}} - 1} \\ &\leqslant \binom{\left\lceil \frac{n-L_{\over}}{L_{\min}-L_{\over}} \right\rceil + q^{L_{\min}}}{q^{L_{\min}}}. \end{aligned} \tag{2}$$

Before concluding this section we discuss the pertinent notion of *repeat-free* strings [6], which we denote herein for all $\ell < n$ by

$$\mathcal{RF}_\ell(n) \triangleq \big\{\boldsymbol{x} \in \Sigma^n : \boldsymbol{x}_{i+[\ell]} \neq \boldsymbol{x}_{j+[\ell]}, \forall i < j \in [n - \ell + 1]\big\}.$$

It was observed in [21] that for $\boldsymbol{x} \in \mathcal{RF}_\ell(n)$, $\mathcal{L}^\ell_{\ell+1}(\boldsymbol{x}) \neq \mathcal{L}^\ell_{\ell+1}(\boldsymbol{y})$ for all $\boldsymbol{y} \in \Sigma^n \backslash \{\boldsymbol{x}\}$. A straightforward generalization of the arguments therein demonstrates the following lemma.

Lemma 1 *Given $L_{\min} > L_{\over}$, for all $\boldsymbol{x} \in \mathcal{RF}_{L_{\over}}(n)$, there exists an efficient algorithm reconstructing $\boldsymbol{x}$ from any $(L_{\min}, L_{\over})$-trace of $\boldsymbol{x}$.*

Proof: Let T be any $(L_{\min}, L_{\over})$-trace of $\boldsymbol{x}$. For any $\boldsymbol{u} \in T$, suppose by negation that there exist $\boldsymbol{v}_1, \boldsymbol{v}_2 \in T$, $\boldsymbol{v}_1 \neq \boldsymbol{v}_2$, such that the ℓ_i-suffix of $\boldsymbol{v}_i$ equals the ℓ_i-prefix of $\boldsymbol{u}$, where $\ell_i \geqslant L_{\over}$, for $i \in \{1, 2\}$. Since $\boldsymbol{v}_1 \neq \boldsymbol{v}_2$, they occur in distinct locations in $\boldsymbol{x}$, and in particular their $\min\{\ell_1, \ell_2\}$-suffix occurs in distinct locations; this in contradiction to $\boldsymbol{x} \in \mathcal{RF}_{L_{\over}}(n)$. The same argument proves that there do not exist $\boldsymbol{v}_1, \boldsymbol{v}_2 \in T$, $\boldsymbol{v}_1 \neq \boldsymbol{v}_2$, such that the ℓ_i-prefix of $\boldsymbol{v}_i$ equals the ℓ_i-suffix of $\boldsymbol{u}$, where again $\ell_i \geqslant L_{\over}$, for $i \in \{1, 2\}$.

Hence, matching prefix to suffix, of lengths at least $L_{\over}$, one reconstructs $\boldsymbol{x}$ from T. Equivalently, for each $\boldsymbol{u} \in T$, finding the unique $\boldsymbol{v} \in T$ that contains the $L_{\over}$-prefix of $\boldsymbol{u}$ as a substring (which exists unless $\boldsymbol{u}$ is itself a prefix of $\boldsymbol{x}$) results with complete reconstruction. A naive implementation requires $O(n^2 L_{\over})$ run-time. ∎

Further, we note that if $\liminf L_{\over}/\log(n) > 1$, then [6] showed that $\mathcal{RF}_{L_{\over}}(n)$ forms a rate $1 - o_n(1)$ code in Σ^n with an efficient encoder/decoder pair; we summarize their result in the following lemma.

Lemma 2 *[6, Sec. V] For $k = \lceil 2\log\log(n) \rceil$ there exists an efficient encoder into $\mathcal{RF}_{\lceil \log(n) \rceil + 5(k+2)}(n)$, which in addition produces strings containing no $(2k+2)$-length run of zeros. Redundancy is introduced only in the initialization phase encoding into*

$$Z(n, k) \triangleq \{\boldsymbol{u} \in \Sigma^n : \boldsymbol{u} \textit{ has no length-}k \textit{ run of zeros}\}$$

(this is the well-understood Run-length-limited (RLL) *constraint; see, e.g., [15, Sec. 1.2]).*
Further, an efficient decoder exists for the provided encoder.

Analysis of the asymptotic rate achieved by the encoder of Lemma 2 is given in the following lemma.

Lemma 3 *There exist efficient encoders into $Z(n, k)$ requiring $\left\lceil \frac{q}{q-2} n/q^k \right\rceil$ redundant symbols for $q > 2$ [23, Lem. 5], or $2\left\lceil 2n/2^k \right\rceil$ for $q = 2$ [12, Sec. III].*

For our purposes, however, it will be beneficial to observe that the arguments used in [6, Sec. V] apply without change to any other choice of k, resulting in the following corollary. It demonstrates that $\mathcal{RF}_\ell(n)$ forms a rate $1-o_n(1)$ code in Σ^n as long as $\ell-\log(n)=\omega(1)$.

Corollary 4 *For $\ell(n)=\lceil\log(n)\rceil+\omega_n(1)$ and any*

$$t\leqslant 2\lfloor(\ell(n)-\lceil\log(n)\rceil)/5\rfloor-3$$

there exists an efficient encoder/decoder pair into $\mathcal{RF}_{\ell(n)}(n)$, producing strings containing no t-length run of zeros, and requiring at most $\left\lceil\frac{q^2}{q-2}q^{-\lfloor t/2\rfloor}n\right\rceil$ redundant symbols for $q>2$ or $2\left\lceil 4\cdot 2^{-\lfloor t/2\rfloor}n\right\rceil$ symbols for $q=2$, i.e., rate $1-O_n(q^{-t/2})$.

Proof: For any $s\leqslant\lfloor(\ell(n)-\lceil\log(n)\rceil)/5\rfloor-2$, letting $k=s$ in Lemma 2 produces ℓ'-repeat-free strings, for some $\ell'\leqslant\ell(n)$, hence in particular also $\ell(n)$-repeat-free strings, containing no $(2s+2)$-length run of zeros.

Letting $s\triangleq\lfloor t/2\rfloor-1$ we encode $\ell(n)$-repeat-free strings containing no t-length run of zeros. Finally, the encoders of Lemma 3 then require the claimed redundancy. ∎

Given Corollary 4 and the preceding discussion, we focus in the sequel on the complement, unsolved case of $\limsup L_{\text{over}}/\log(n)\leqslant 1$.

III. Bounds

In this section we demonstrate an upper bound on the achievable asymptotic rate of $(L_{\min},L_{\text{over}})$-trace codes.

Lemma 5 *If $L_{\min}=a\log(n)+O_n(1)$ and $L_{\text{over}}=\gamma L_{\min}+O_n(1)$, for some $a>1$ and $0<\gamma\leqslant\frac{1}{a}$, then any $(L_{\min},L_{\text{over}})$-trace code $\mathcal{C}\subseteq\Sigma^n$ satisfies*

$$R(\mathcal{C})\leqslant\frac{1-1/a}{1-\gamma}+O\left(\frac{\log\log(n)}{\log(n)}\right).$$

(Note that $\frac{1-1/a}{1-\gamma}\leqslant 1$ if and only if $\lim L_{\text{over}}/\log(n)=\gamma a\leqslant 1$.)

Proof: Observe for $v\geqslant u\geqslant 0$ that

$$\begin{aligned}\log\binom{u+v}{u}&\leqslant\log\frac{(u+v)^u}{u!}<\log\left(\left(\tfrac{e}{u}(u+v)\right)^u\right)\\&=u\left(\log(e)+\log\left(1+\tfrac{v}{u}\right)\right)\\&=u\left(\log(e)+\left(\log\left(\tfrac{u}{v}+1\right)+\log\left(\tfrac{v}{u}\right)\right)\right)\\&\leqslant u\left(\left(1+\tfrac{u}{v}\right)\log(e)+\log\left(\tfrac{v}{u}\right)\right)\\&\leqslant u\left(2\log(e)+\log\left(\tfrac{v}{u}\right)\right).\end{aligned}$$

Letting $v\triangleq q^{L_{\min}}$ and $u\triangleq\left\lceil\frac{n}{L_{\min}-L_{\text{over}}}\right\rceil<v$, we observe that $\log(\frac{v}{u})=O(\log(n))$ and by Eq. (2), $|\mathcal{C}|\leqslant\binom{u+v}{u}$, hence we have

$$\begin{aligned}\log|\mathcal{C}|&\leqslant n\frac{L_{\min}-\log(n)+\log(L_{\min}-L_{\text{over}})+2\log(e)}{L_{\min}-L_{\text{over}}}+\\&\quad+O(\log(n))\\&=n\left(\frac{L_{\min}-\log(n)}{L_{\min}-L_{\text{over}}}+O\left(\frac{\log\log(n)}{\log(n)}\right)\right)\\&=n\left(\frac{1-1/a}{1-\gamma}+O\left(\frac{\log\log(n)}{\log(n)}\right)\right).\end{aligned}$$

∎

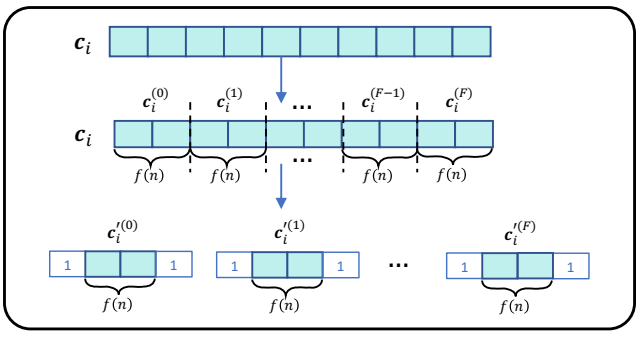

Figure 1. Index generation. Each index $\boldsymbol{c}_i$ is first partitioned into $F+1$ non-overlapping segments of length $f(n)$. Then, each of the segments is concatenated with a single 1 in each edge.

In particular, Lemma 5 implies the following lower bound on $L_{\min}$ for the existence of codes with asymptotically non-vanishing rates.

Corollary 6 *If $\limsup_n L_{\min}/\log(n)\leqslant 1$, then $R(\mathcal{C})=o_n(1)$ for any $(L_{\min},L_{\text{over}})$-trace code $\mathcal{C}\subseteq\Sigma^n$.*

IV. Construction

In this section we present an efficient encoder for $(L_{\min},L_{\text{over}})$-trace codes, achieving asymptotically optimal rate, for the case $\limsup L_{\text{over}}/\log(n)\leqslant 1$. Throughout the section, we let $L_{\min}\triangleq\lceil a\log(n)\rceil$ and $L_{\text{over}}\triangleq\lceil\gamma L_{\min}\rceil$, for some $a>1$ and $0<\gamma\leqslant 1/a$. Further, we let f be any integer function satisfying $f(n)=\omega(1)$ and $f(n)=o(\log(n))$, and finally $I\triangleq\left\lceil\frac{1-\gamma a}{1-\gamma}\log(n)+(\log(n))^{0.5+\epsilon}\right\rceil$ for some small $\epsilon>0$. In the sequel we tacitly assume that $q^I L_{\min}$ divides n.

Construction A The construction is based on the following two ingredients:

- *Index generation*. Let $(\boldsymbol{c}_i)_{i\in[q^I]}$, $\boldsymbol{c}_i\in\Sigma^I$ be indices in ascending lexicographic order. We encode each $\boldsymbol{c}_i$ independently as follows (see Figure 1). Denoting $F\triangleq\lceil I/f(n)\rceil-1$, we partition $\boldsymbol{c}_i$ into $F+1$ non-overlapping segments; more formally, define $\left\{\boldsymbol{c}_i^{(k)}\right\}_{k\in[F]}\subseteq\Sigma^{f(n)}$, and $\boldsymbol{c}_i^{(F)}\in\Sigma^{I\bmod f(n)}$, by

$$\boldsymbol{c}_i^{(0)}\circ\boldsymbol{c}_i^{(1)}\circ\cdots\circ\boldsymbol{c}_i^{(F)}=\boldsymbol{c}_i.$$

  Lastly, denote $\boldsymbol{c}_i'^{(k)}\triangleq 1\circ\boldsymbol{c}_i^{(k)}\circ 1\in\Sigma^{f(n)+2}$ (similarly, $\boldsymbol{c}_i'^{(F)}$). We refer to $\boldsymbol{c}_i$ (or simply i) as an *index* in the construction, and to $\left\{\boldsymbol{c}_i'^{(k)}\right\}_{k\in[F+1]}$ as segments of an *encoded index*.
- *Repeat-free (RF) encoding*. For $\ell\in\mathbb{N}$, we use an RF encoder which receives strings of length m and returns strings of length $N_{n,\ell}(m)$ which contain no repeated substrings of length ℓ, and no zero-runs of length $f(n)$ (see Corollary 4). We denote such an encoder by $E_{m,\ell}^{\mathcal{RF}}$.

The constructed code, denoted by $\mathcal{C}_A(n)$, is carried as follows. Denote

$$r \triangleq f(n) + 3 + (F+1)(f(n)+2),$$
$$\ell \triangleq \left\lceil \frac{L_{\text{over}} - 2f(n) - 5}{1 + (f(n)+2)/\left\lfloor \frac{L_{\min}-r}{F+1} \right\rfloor} \right\rceil.$$

Let m be an integer such that

$$N_{n,\ell}(m) = q^{-I} n (1 - r/L_{\min}).$$

(See the proof of Theorem 8 for an explanation of why such m exists for the choices of $n, L_{\min}, f(n), I, r, \ell$, and the observation that $m \geqslant N_{n,\ell}(m)\big(1 - O\big(q^{-f(n)/2}\big)\big)$.) An information string $\boldsymbol{x} \in \Sigma^{q^I m}$ is first partitioned into q^I non-overlapping segments $\boldsymbol{x} = \boldsymbol{x}_0 \circ \cdots \circ \boldsymbol{x}_{q^I-1}$, where $\boldsymbol{x}_i \in \Sigma^m$ for each $i \in [q^I]$. Each $\boldsymbol{x}_i$ is then independently encoded into a string $\boldsymbol{z}_i$ of length $q^{-I}n$ as explained below; the motivation for this step, which we later describe more formally, is to satisfy two properties: (i) the index i can be decoded from any $L_{\min}$-substring of $\boldsymbol{z}_i$; and (ii) the string $\boldsymbol{z}_i$ can be uniquely reconstructed from an $(L_{\min}, L_{\text{over}})$-trace of $\boldsymbol{z}_i$. Lastly, we let

$$\text{Enc}_A(\boldsymbol{x}) \triangleq \boldsymbol{z} = \boldsymbol{z}_0 \circ \cdots \circ \boldsymbol{z}_{q^I-1}.$$

The encoding $\boldsymbol{x}_i \mapsto \boldsymbol{z}_i$, for all $i \in [q^I]$, is performed as follows (see Figure 2).

1) Let $\boldsymbol{y}_i \triangleq E^{\mathcal{RF}}_{m,\ell}(\boldsymbol{x}_i)$, where $\boldsymbol{y}_i \in \Sigma^{N_{n,\ell}(m)}$.
2) Partition $\boldsymbol{y}_i$ into $n/(q^I L_{\min})$ non-overlapping segments of length $L_{\min} - r$ (recall $(L_{\min}-r)n/(q^I L_{\min}) = N_{n,\ell}(m)$) by denoting
$$\boldsymbol{y}_i = \boldsymbol{y}_{i,0} \circ \boldsymbol{y}_{i,1} \circ \cdots \circ \boldsymbol{y}_{i,n/(q^I L_{\min})-1}.$$
3) For all $j \in [n/(q^I L_{\min})]$:
   a) Partition each $\boldsymbol{y}_{i,j}$ into $F+1$ non-overlapping segments of equal lengths (up to ± 1, if necessary)
   $$\boldsymbol{y}_{i,j} = \boldsymbol{y}^{(0)}_{i,j} \circ \boldsymbol{y}^{(1)}_{i,j} \circ \cdots \circ \boldsymbol{y}^{(F)}_{i,j}.$$
   b) Combine $\big\{\boldsymbol{y}^{(k)}_{i,j} : k \in [F+1]\big\}$ with segments of the encoded index i, as follows. Define for all $k \in [F+1]$
   $$\boldsymbol{z}^{(k)}_{i,j} \triangleq \boldsymbol{y}^{(k)}_{i,j} \circ \boldsymbol{c}'^{(k)}_i,$$
   then
   $$\boldsymbol{z}_{i,j} \triangleq \begin{cases} 10^{f(n)}11 \circ \boldsymbol{z}^{(0)}_{i,j} \circ \cdots \circ \boldsymbol{z}^{(F)}_{i,j}, & j = 0; \\ 10^{f(n)}01 \circ \boldsymbol{z}^{(0)}_{i,j} \circ \cdots \circ \boldsymbol{z}^{(F)}_{i,j}, & j > 0 \end{cases}$$
   (where we refer to the substrings $10^{f(n)}01, 10^{f(n)}11$ as *synchronization markers*).
4) Concatenate
$$\boldsymbol{z}_i \triangleq \boldsymbol{z}_{i,0} \circ \cdots \circ \boldsymbol{z}_{i,n/(q^I L_{\min})-1}. \qquad \square$$

Before we prove the correctness of Construction A, we first analyze $R(\mathcal{C}_A)$.

Lemma 7 *1)* $r = I + \frac{2I}{f(n)} + 2f(n) + O(1)$.
2) Denoting $\lambda \triangleq 1 - \frac{I}{L_{\min}}$, *we have* $\ell = \lambda L_{\text{over}} - O(f(n))$.

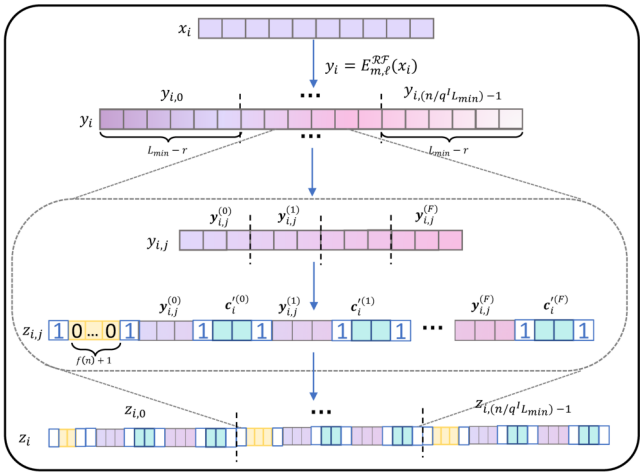

Figure 2. Encoding $\boldsymbol{x}_i$ into $\boldsymbol{z}_i$, as detailed in Construction A.

Based on these properties, we may show that Construction A asymptotically meets the bound of Lemma 5.

Theorem 8 *Letting* $f(n) \triangleq \left\lceil \sqrt{\log(n)} \right\rceil$ *we have*

$$R(\mathcal{C}_A(n)) \geqslant \frac{1-1/a}{1-\gamma} - \frac{1/a}{(\log(n))^{0.5-\epsilon}} - O\left(\frac{1}{\sqrt{\log(n)}}\right).$$

Proof: Recalling $q^I N_{n,\ell}(m) = \left(1 - \frac{r}{L_{\min}}\right)n$, we observe

$$\begin{aligned}\log(N_{n,\ell}(m)) &= \log(n) - I + \log\left(1 - \frac{r}{L_{\min}}\right) \\ &= \frac{(a-1)\gamma}{1-\gamma}\log(n) - (\log(n))^{0.5+\epsilon} + O(1)\end{aligned}$$

whereas by part 2 of Lemma 7

$$\begin{aligned}\ell &= \left(1 - \frac{1/a-\gamma}{1-\gamma} - \frac{(\log(n))^{\epsilon}}{a\sqrt{\log(n)}}\right)\gamma a \log(n) - O\left(\sqrt{\log(n)}\right) \\ &= \frac{(a-1)\gamma}{1-\gamma}\log(n) - \gamma(\log(n))^{0.5+\epsilon} - O\left(\sqrt{\log(n)}\right).\end{aligned}$$

Hence

$$\ell - \log(N_{n,\ell}(m)) = (1-\gamma)(\log(n))^{0.5+\epsilon} - O\left(\sqrt{\log(n)}\right)$$

and for sufficiently large n,

$$f(n) \leqslant 2\lfloor(\ell - \lceil\log(N_{n,\ell}(m))\rceil)/5\rfloor - 3,$$

satisfying the condition of Corollary 4. We can therefore efficiently encode $\boldsymbol{y}_i = E^{\mathcal{RF}}_{m,\ell}(\boldsymbol{x}_i)$ (and vice versa, decode $\boldsymbol{x}_i$) while attaining

$$N_{n,\ell}(m) - m \leqslant O\left(q^{-f(n)/2} N_{n,\ell}(m)\right).$$

In particular,

$$\begin{aligned}q^I m &= q^I N_{n,\ell} - q^I(N_{n,\ell} - m) \\ &\geqslant n\left(1 - \frac{r}{L_{\min}} - O\left(q^{-f(n)/2}\left(1 - \frac{r}{L_{\min}}\right)\right)\right).\end{aligned}$$

By part 1 of Lemma 7,

$$\frac{r}{L_{\min}} = \frac{1/a-\gamma}{1-\gamma} + \frac{1/a}{(\log(n))^{0.5-\epsilon}} + O\left(\frac{1}{\sqrt{\log(n)}}\right),$$

hence noting $R(\mathcal{C}_A(n)) = q^I m/n$ concludes the proof. ∎

We note that the choice $f(n) = \left\lceil \sqrt{\log(n)} \right\rceil$ in Theorem 8 is optimal, since ϵ in Construction A must satisfy $\epsilon \geqslant \max\left\{ \frac{\log(f(n))}{\log\log(n)}, 1 - \frac{\log(f(n))}{\log\log(n)} \right\} - 0.5$.

Finally, we prove the correctness of Construction A. We begin with two technical lemmas.

Lemma 9 *Every $L_{\min}$-substring $\boldsymbol{u}$ of $\boldsymbol{z}$ contains as subsequences at least an $(I-\mu)$-suffix of $\boldsymbol{c}_i$, and an μ-prefix of either $\boldsymbol{c}_i$ or $\boldsymbol{c}_{i+1}$, for some $i \in [q^I]$ and $\mu \in [I]$, in identifiable locations.*

Lemma 10 *Every L_{over}-substring $\boldsymbol{v}$ of $\boldsymbol{z}$ contains at least ℓ consecutive symbols of $\boldsymbol{y} \triangleq \boldsymbol{y}_0 \circ \cdots \circ \boldsymbol{y}_{q^I-1}$.*

Combining both lemmas, we have the following theorem.

Theorem 11 *For all admissible values of n, the code $\mathcal{C}_A(n)$ is an $(L_{\min}, L_{\text{over}})$-trace code.*

Proof: Take $\boldsymbol{z} \in \mathcal{C}_A(n)$ and let $T \in \mathcal{T}_{L_{\min}}^{L_{\text{over}}}(\boldsymbol{z})$ be any $(L_{\min}, L_{\text{over}})$-trace of $\boldsymbol{z}$.

For $\boldsymbol{u} \in T$, we extract the $(I-\mu)$-suffix of $\boldsymbol{c}_i$, and an μ-prefix of either $\boldsymbol{c}_i$ or $\boldsymbol{c}_{i+1}$, for some i, guaranteed by Lemma 9. Observe that if this prefix belongs to $\boldsymbol{c}_{i+1}$, then $\boldsymbol{u}$ also contains a complete synchronization marker $10^f(n)11$ (the instance appearing as prefix of $\boldsymbol{z}_{i+1}$), hence these two cases may be distinguished. Further, note that the μ-prefix of $\boldsymbol{c}_{i+1}$ equals the μ-prefix of $\boldsymbol{c}_i$, unless the $(I-\mu)$-suffix of $\boldsymbol{c}_i$ is the greatest element of $\Sigma^{I-\mu}$ (in lexicographic order). In both cases, one can correctly deduce that the location of $\boldsymbol{u}$ in $\boldsymbol{z}$ begins in the segment $\boldsymbol{z}_i$. It is therefore possible to partition T by index i (corresponding to the starting location of each substring).

For each substring $\boldsymbol{u}$ of index i, intersecting both $\boldsymbol{y}_i, \boldsymbol{y}_{i+1}$, $\boldsymbol{u}$ must contain a complete synchronization marker $10^f(n)11$ (the instance appearing as prefix of $\boldsymbol{z}_{i+1}$); hence its location in $\boldsymbol{u}$ implies the exact location of $\boldsymbol{u}$ in $\boldsymbol{z}$. For all other substrings of index i, it holds by Lemma 10, and since each $\boldsymbol{y}_i$ is ℓ-repeat-free, that there exist a unique way to concatenate these substrings (excluding overlap) as shown in Lemma 1.

Finally, once $\boldsymbol{z}$ is reconstructed we may extract $\{\boldsymbol{y}_i\}_{i\in[q^I]}$, then decode $\{\boldsymbol{x}_i\}_{i\in[q^I]}$ with the decoder of $E_{m,\ell}^{\mathcal{RF}}$. ∎

REFERENCES

[1] J. Acharya, H. Das, O. Milenkovic, A. Orlitsky, and S. Pan, "String reconstruction from substring compositions," *SIAM J. Discrete Math.*, vol. 29, no. 3, pp. 1340–1371, 2015.

[2] D. Bar-Lev, S. Marcovich, E. Yaakobi, and Y. Yehezkeally, "Adversarial torn-paper codes," in *Proceedings of the 2022 IEEE International Symposium on Information Theory (ISIT), Espoo, Finland*, Jun. 2022, pp. 2934–2939.

[3] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'04), New Orleans, LA, USA*. Society for Industrial and Applied Mathematics, Jan. 2004, pp. 910–918.

[4] G. Bresler, M. Bresler, and D. Tse, "Optimal assembly for high throughput shotgun sequencing," *BMC Bioinformatics*, vol. 14, no. 5, p. S18, Jul. 2013.

[5] Z. Chang, J. Chrisnata, M. F. Ezerman, and H. M. Kiah, "Rates of DNA sequence profiles for practical values of read lengths," *IEEE Trans. on Inform. Theory*, vol. 63, no. 11, pp. 7166–7177, Nov. 2017.

[6] O. Elishco, R. Gabrys, M. Médard, and E. Yaakobi, "Repeat-free codes," *IEEE Trans. on Inform. Theory*, vol. 67, no. 9, pp. 5749–5764, Sep. 2021.

[7] R. Gabrys and O. Milenkovic, "Unique reconstruction of coded strings from multiset substring spectra," *IEEE Trans. on Inform. Theory*, vol. 65, no. 12, pp. 7682–7696, Dec. 2019.

[8] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany*, Jun. 2017, pp. 3130–3134.

[9] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. on Inform. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[10] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *Proceedings of the 2019 IEEE Information Theory Workshop (ITW), Visby, Sweden*, Aug. 2019.

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

[12] M. Levy and E. Yaakobi, "Mutually uncorrelated codes for DNA storage," *IEEE Trans. on Inform. Theory*, vol. 65, no. 6, pp. 3671–3691, Jun. 2019.

[13] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, and P. Stockmeyer, "Reconstruction of sequences," *Discrete Mathematics*, vol. 94, no. 3, pp. 209–219, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0012365X77900449

[14] S. Marcovich and E. Yaakobi, "Reconstruction of strings from their substrings spectrum," *IEEE Trans. on Inform. Theory*, vol. 67, no. 7, pp. 4369–4384, Jul. 2021.

[15] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," Oct. 2001, unpublished Lecture Notes. [Online]. Available: www.math.ubc.ca/~marcus/Handbook

[16] A. S. Motahari, G. Bresler, and D. N. C. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. on Inform. Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.

[17] S. Nassirpour, I. Shomorony, and A. Vahid, "Reassembly codes for the chop-and-shuffle channel," *arXiv preprint arXiv:2201.03590*, 2022.

[18] A. N. Ravi, A. Vahid, and I. Shomorony, "Capacity of the torn paper channel with lost pieces," in *Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Victoria, Australia*, Jul. 2021, pp. 1937–1942.

[19] ——, "Coded shotgun sequencing," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 1, pp. 147–159, Mar. 2022.

[20] I. Shomorony and A. Vahid, "Torn-paper coding," *IEEE Trans. on Inform. Theory*, vol. 67, no. 12, pp. 7904–7913, Dec. 2021.

[21] E. Ukkonen, "Approximate string-matching with q-grams and maximal matches," *Theoretical Computer Science*, vol. 92, no. 1, pp. 191–211, 1992.

[22] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability," *arXiv preprint arXiv:2109.12549*, Sep. 2021. [Online]. Available: https://arxiv.org/abs/2109.12549

[23] Y. Yehezkeally, S. Marcovich, and E. Yaakobi, "Multi-strand reconstruction from substrings," in *Proceedings of the 2021 IEEE Information Theory Workshop (ITW), Kanazawa, Japan*, Oct. 2021.