Generalized Unique Reconstruction from Substrings

Yonatan Yehezkeally, *Member, IEEE*, Daniella Bar-Lev, *Student Member, IEEE*,
Sagi Marcovich, *Student Member, IEEE* and Eitan Yaakobi, *Senior Member, IEEE*

***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 pre-defined lengths are read or substrings are read with no overlap for the single string case, this work studies two extensions of this paradigm. The first extension 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, efficient constructions of codes that asymptotically meet that upper bound are presented. In the second extension, we study the setup where multiple strings are reconstructed together. Given the number of strings and their length, we first derive a lower bound on the read substrings' length ℓ that is necessary for the existence of multi-strand reconstruction codes with non-vanishing rates. We then present two constructions of such codes and show that their rates approach 1 for values of ℓ that asymptotically behave like the lower 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 [20], the *trace reconstruction problem* [3], [6], and the k*-deck problem* [8], [9], [23], [32].

Notably, when observations are comprised of unordered consecutive substrings, two distinct models have received significant interest in the past decade due to applications in DNA- or polymer-based storage systems, resulting from contemporary sequencing technologies [4], [13], [27]. The first is the *reconstruction from substring-compositions problem* [1], [4], [11], [15], [18], [24], [26], [27], [33], [35], [39] (including extensions for erroneous observations [5], [13], [24], [39]), which arises from an idealized assumption of full overlap (and uniform coverage) in read substrings; the second is the *torn-paper problem* [2], [28], [29], [36] (a problem closely related to the *shuffling channel* [16], [19], [34], [38]), which results from an assumption of no overlap. In applications, the distinction 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 of each retrieved substring, as well as the length of overlap between consecutive substrings, are bounded from below. A similar setting was recently studied in [30], where both substrings' lengths and overlap were assumed to be random; we study the problem in the aforementioned worst-case regime.

Further, in both sequencing and tandem-mass-spectrometry technologies, used for DNA and polymer-based storage systems respectively, it is typical that not a single string is read alone, but multiple strings simultaneously [7], [14], [17], [22], [31]. We therefore study a setting where retreived substrings are taken from a collection of information strings stored together, with no information on the string from which they originated. We remark that this extension was already studied by the authors for the torn-paper problem, in [2].

Our problem setting is therefore given as follows: a multiset of k length-n strings is transmitted, and substrings of all information strings are retrieved, such that the length of each substring is at least ℓ_{min}, and consecutive substrings of the same information string overlap in at least ℓ_{over} positions. We are interested in the minimum value of ℓ_{min}, as a function of k and n, for which there exist codes 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}. In these cases, we seek to develop efficient coding schemes which asymptotically attain optimal rates.

The rest of this paper is organized as follows. In Section II, we present notation and definitions which are used throughout the paper. In Section III, we overview and extend results in existing literature which already solve our problem setting in specific end-cases. In Sections IV and V, we present a solution to the aforementioned problem in the private case of a single string ($k = 1$), by respectively bounding from

Manuscript received October 10, 2022; revised February 23, 2023; accepted April 16, 2023. 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 and by the Israel Innovation Authority Grant 75855. 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. This article was presented in part at the 2021 IEEE Information Theory Workshop [DOI: 10.1109/ITW48936.2021.9611486], and at the 2022 International Symposium on Information Theory and Its Applications. *(Corresponding author: Yonatan Yehezkeally)*

Y. Yehezkeally is with the Institute for Communications Engineering, School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany (e-mail: yonatan.yehezkeally@tum.de). D. Bar-Lev, S. Marcovich, and E. Yaakobi are with the Department of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel (e-mails: {yaakobi,daniellalev,sagimar}@cs.technion.ac.il).

The first three authors contributed equally to this work.

above the asymptotically attainable rate of codes for unique reconstruction as a function of $\ell_{\min}, \ell_{\text{over}}$, and then developing efficient encoding and decoding algorithms for such codes, asymptotically meeting this bound. Then, in Section VI, we study solutions to the problem in a different private case, where $\ell_{\text{over}} = \ell_{\min} - 1$ (i.e., a multi-strand extension of the reconsturction from substrings problem); we likewise present bounds and two efficient constructions of multiset-codes for this case, whose rates asymptotically approach 1 for values of $\ell_{\min}$ asymptotically equivalent to the lower bound. We conclude in Section VII with a summary and closing remarks.

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 $[n] \triangleq \{0, 1, \dots, n-1\}$. We denote a *multiset* by $S = \{\{a, a, b, \dots\}\}$; i.e., elements are allowed to appear with multiplicity. For convenience we let $\|S\|$, for a multiset S, denote the number of unique elements in S.

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 $\boldsymbol{x} = (x_0, x_1, \dots, x_{n-1}) \in \Sigma^*$ 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-prefix/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 (i.e., when $\boldsymbol{x}$ is considered as a function from $[|\boldsymbol{x}|]$ into Σ); 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 define

$$\mathcal{X}_{n,k} \triangleq \{S = \{\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}\} : \forall i, \boldsymbol{x}_i \in \Sigma^n\},$$

and observe that $|\mathcal{X}_{n,k}| = \binom{k+q^n-1}{k}$. We consider in this paper the problem of multi-string reconstruction from substrings with partial overlap. That is, we assume that a message $S \in \mathcal{X}_{n,k}$ is observed only through a multiset of substrings of its elements, without order or information on the substring from which they originate, with the following restrictions: (i) all observed substrings are of length at least $\ell_{\min}$; and (ii) succeeding substrings of the same $\boldsymbol{x} \in S$ overlap with length at least ℓ_{over} (in particular, every symbol of $\boldsymbol{x}$ is observed in some substring).

More formally, a *substring-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset $\left\{\left\{\boldsymbol{x}_{i_j+[\ell_j]} : 1 \leqslant j \leqslant m\right\}\right\}$, 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 $i_m + \ell_m = n$. A complete substring-trace of $\boldsymbol{x} \in \Sigma^n$ is called an *$(\ell_{\min}, \ell_{\text{over}})$-trace* if $\ell_j \geqslant \ell_{\min} \geqslant \ell_{\text{over}}$ for all j, and $i_j + \ell_j - i_{j+1} \geqslant \ell_{\text{over}}$ 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 $i_1 > 0$).

See Figures 1 to 3 for an illustration if these substring-traces.

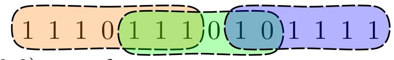

Figure 1. A $(6, 2)$-trace of $\boldsymbol{x}$.

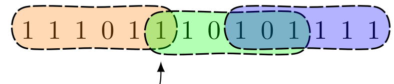

Figure 2. A complete substring-trace of $\boldsymbol{x}$ (not a $(6, 2)$-trace).

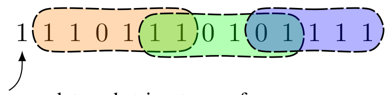

Figure 3. An incomplete substring-trace of $\boldsymbol{x}$.

The $(\ell_{\min}, \ell_{\text{over}})$-*trace spectrum* of $\boldsymbol{x} \in \Sigma^n$, denoted $\mathcal{T}_{\ell_{\min}}^{\ell_{\text{over}}}(\boldsymbol{x})$, is the set of all $(\ell_{\min}, \ell_{\text{over}})$-traces of $\boldsymbol{x}$. We extend the definition to $S \in \mathcal{X}_{n,k}$ by $\mathcal{T}_{\ell_{\min}}^{\ell_{\text{over}}}(S) \triangleq \bigcup_{\boldsymbol{x}\in S} \mathcal{T}_{\ell_{\min}}^{\ell_{\text{over}}}(\boldsymbol{x})$, where the union respects multiplicity (i.e., multiset union), and similarly extend the definitions of traces. Our channel accepts $S \in \mathcal{X}_{n,k}$ and outputs a single arbitrary $(\ell_{\min}, \ell_{\text{over}})$-trace of S.

For all $\mathcal{C} \subseteq \mathcal{X}_{n,k}$, we denote the *rate*, *redundancy* of $\mathcal{C}$ by $R(\mathcal{C}) \triangleq \frac{\log|\mathcal{C}|}{\log|\mathcal{X}_{n,k}|}$, $\operatorname{red}(\mathcal{C}) \triangleq \log|\mathcal{X}_{n,k}| - \log|\mathcal{C}|$, 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 *$(\ell_{\min}, \ell_{\text{over}})$-trace code* if for all $\boldsymbol{c}_1 \neq \boldsymbol{c}_2 \in \mathcal{C}$, $\mathcal{T}_{\ell_{\min}}^{\ell_{\text{over}}}(\boldsymbol{c}_1) \cap \mathcal{T}_{\ell_{\min}}^{\ell_{\text{over}}}(\boldsymbol{c}_2) = \emptyset$. We likewise define a *multi-strand $(\ell_{\min}, \ell_{\text{over}})$-trace code* $\mathcal{C} \subseteq \mathcal{X}_{n,k}$. The main goal of this work is to find, for $\ell_{\min}, \ell_{\text{over}}$ as functions of n, k, the maximum asymptotic rate of (multi-strand) $(\ell_{\min}, \ell_{\text{over}})$-trace codes. We will also be interested in efficient constructions of codes with rate asymptotically approaching that value.

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

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

(Here, if $\ell_{\min} - \ell_{\text{over}}$ does not divide $n - \ell_{\min}$ we allow the $\ell_{\min}$-suffix to contain a longer overlap with its preceding $\ell_{\min}$-substring.) We likewise let $\mathcal{L}_{\ell_{\min}}^{\ell_{\text{over}}}(S) \triangleq \bigcup_{\boldsymbol{x}\in S} \mathcal{L}_{\ell_{\min}}^{\ell_{\text{over}}}(\boldsymbol{x})$.

III. Repeat-free strings

In this this section, we discuss the special case of $(\ell, \ell-1)$-trace codes, which has been studied in literature in the context of reconstruction from substring compositions. To that end, we introduce the pertinent notion of *repeat-free* strings [11], which we denote herein for all $\ell \leqslant n$ by

$$\mathcal{RF}_\ell(n) \triangleq \left\{ \boldsymbol{x} \in \Sigma^n : \left\| \mathcal{L}_\ell^{\ell-1}(\boldsymbol{x}) \right\| = n - \ell + 1 \right\},$$

That is, the set of all length-n strings whose ℓ-substrings are all distinct. It was observed in [37] that if $\boldsymbol{x} \in \mathcal{RF}_\ell(n)$, then $\mathcal{L}_{\ell+1}^\ell(\boldsymbol{x}) \neq \mathcal{L}_{\ell+1}^\ell(\boldsymbol{y})$ for all $\boldsymbol{y} \in \Sigma^n$, $\boldsymbol{y} \neq \boldsymbol{x}$. A straightforward generalization of the arguments therein demonstrates the following lemma.

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

Proof: Let T be any $(\ell_{\min}, \ell_{\text{over}})$-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 $\ell_i \geqslant \ell_{\text{over}}$, 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 $\boldsymbol{x} \in \mathcal{RF}_{\ell_{\text{over}}}(n)$. 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 $\ell_i \geqslant \ell_{\text{over}}$, 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 $O(n^2 \ell_{\text{over}})$ run-time. ■

We also denote *multi-strand ℓ-repeat-free strings*

$$\mathcal{RF}_\ell(n,k) \triangleq \left\{ S \in \mathcal{X}_{n,k} : \left\| \mathcal{L}_\ell^{\ell-1}(S) \right\| = k(n-\ell+1) \right\},$$

and observe the following corollary of Lemma 1.

Corollary 2 *For all $S \in \mathcal{RF}_{\ell_{\text{over}}}(n,k)$ there exists an efficient algorithm reconstructing S from any $(\ell_{\min}, \ell_{\text{over}})$-trace of S.*

Proof: Observe that S is a set, $S \subseteq \mathcal{RF}_{\ell_{\text{over}}}(n)$, and that $\left\{ \mathcal{L}_{\ell_{\text{over}}}^{\ell_{\text{over}}-1}(\boldsymbol{x}) : \boldsymbol{x} \in S \right\}$ are pairwise-disjoint, hence the reconstruction algorithm of Lemma 1 may operate on all elements of S in parallel without interference. ■

As a consequence of Corollary 2, $\mathcal{RF}_{\ell_{\text{over}}}(n,k)$ forms a multi-strand $(\ell_{\min}, \ell_{\text{over}})$-trace code in $\mathcal{X}_{n,k}$ (likewise, $\mathcal{RF}_{\ell_{\text{over}}}(n)$ in Σ^n).

Further, we note for $k=1$ that if $\liminf \ell_{\text{over}}/\log(n) > 1$, then [11] showed that $\mathcal{RF}_{\ell_{\text{over}}}(n)$ forms a rate $1 - o_n(1)$ code in Σ^n with an efficient encoder/decoder pair. Before summarizing their results, we will require the following notation; let

$$\mathcal{RLL}_s(n) \triangleq \{ \boldsymbol{u} \in \Sigma^n : \boldsymbol{u} \text{ has no length-}s \text{ run of zeros} \}.$$

This is the well-understood *run-length-limited* constraint (see, e.g., [25, Sec. 1.2]).

Then, from [11] we have the following lemma.

Lemma 3 *1) [11, Sec. IV] There exists an efficient encoder/decoder pair into $\mathcal{RF}_{2\lceil \log(n) \rceil + 2}(n)$, requiring a single redundant symbol.*

2) [11, Sec. V] There exists an efficient encoder/decoder pair into $\mathcal{RF}_{s'}(n) \cap \mathcal{RLL}_{s''}(n)$, where

$$s' \triangleq \lceil \log(n) \rceil + 10 \lceil \log\log(n) \rceil + 10;$$
$$s'' \triangleq 4 \lceil \log\log(n) \rceil + 2.$$

The required redundancy is

$$s'' + 1 + \operatorname{red}\left(\mathcal{RLL}_{2\lceil \log\log(n) \rceil}(n - s'' - 1) \right).$$

Analysis of the asymptotic rate achieved by the encoder of Lemma 3, part 2 is given in the following lemma.

Lemma 4 *There exist efficient encoders into $\mathcal{RLL}_s(n)$ requiring $2\lceil 2n/2^s \rceil$ redundant symbols for $q=2$ [21, Sec. III], or $\left\lceil \frac{q}{q-2} n/q^s \right\rceil$ for $q > 2$.*

Proof: The claim for $q=2$ is proven in [21, Sec. III]. Hence, we need only extend it when $q > 2$, and to do so we rely on the concept of the encoder in [21, Alg. 1]. First, the information string $\boldsymbol{x} \in \Sigma^m$ is divided into blocks of length N (where the last block is permitted to be shorter), to be determined later. Then, in each block:

1) Append a 1 to the block.
2) From left to right, search for zero-runs of length s; if one is encountered, remove it, and append the index of its incidence to the block using s symbols, such that the last symbol is restricted not to be either $\{0,1\}$.
3) Continue, until no further zero-runs of length s exist.

Note that this process concludes in finite time (since in each iteration of part 2 it advances by at least s locations of the original block, and appended symbols contain no zero-run of length s). Further, with the given restriction, s symbols may index a total of $q^{s-1}(q-2)$ locations for the beginning of the zero s-substring. It is therefore required to set $N \triangleq q^{s-1}(q-2) + s - 1$.

Also observe that a possible decoder can use the last symbol to indicate whether a zero-run of length s was removed and indexed (which it can then inject in the correct place, discarding the index), or if the process is concluded (in which case the suffix '1' should also be discarded).

Next, since every encoded block ends with a nonzero symbol, these blocks can be concatenated without violating the constraint. Observe, then, that a single redundant symbol is added per block, hence the claimed overall redundancy.

Finally, note that both encoder and decoder operate in polynomial time in the input length. ■

Lemma 4 provides efficient encoders/decoders into $\mathcal{RLL}_s(n)$; to complete the picture, we observe that their redundancy has asymptotically optimal order of magnitude; indeed, by [21, Lem. 3] we have that $\operatorname{red}(\mathcal{RLL}_s(n)) \geqslant \frac{\log(e)}{2} \left(1 - \frac{1}{q}\right)^2 \frac{n-2s}{q^s}$.

Next, although the encoder of part 2 of Lemma 3 asymptotically achieves rate 1, it is of interest to encode into $\mathcal{RF}_\ell(n)$ using less redundancy, for any $\ell < 2\log(n)$. We will show that the approach of [11, Sec. V] can be generalized to this end.

Theorem 5 *For integers $\ell(n), t$ satisfying*

$$\lceil \log\log(n) \rceil + 4 \leqslant t \leqslant \lfloor (\ell(n) - \lceil \log(n) \rceil)/3 \rfloor$$

(for $q = 2$, require $\lceil\log\log(n)\rceil + 5 \leqslant t \leqslant \lfloor(\ell(n) - \lceil\log(n)\rceil)/3\rfloor$) there exists an efficient encoder/decoder pair into $\mathcal{RF}_{\ell(n)}(n) \cap \mathcal{RLL}_t(n)$, requiring at most $t + 1 + \lceil\frac{q^4}{q-2}n/q^t\rceil$ redundant symbols (for $q = 2$, this is $t + 1 + 2\lceil 16n/2^t\rceil$), i.e., rate $1 - O_n\big(\frac{t}{n} + q^{-t}\big)$.

Proof: The proof follows the steps of [11, Sec. V], with some amendments; where their arguments hold without change, we shall clearly cite the relevant proposition while reproducing its proof (when possible, we prioritize intuition over formality in our proof, without sacrificing rigour). The rest of the proof is organized in stages, to improve readability.

1) In the first stage, an information string $\boldsymbol{x} \in \Sigma^m$ is encoded into $\boldsymbol{y} \in \mathcal{RLL}_{t-3}(n - t - 1)$, where m is determined by, e.g., Lemma 4.
2) Next, we wish to eliminate from $\boldsymbol{y}$ repeated substrings of length $s \triangleq \lceil\log(n)\rceil + t + 2$. The elimination stage requires an indexing function $\boldsymbol{h}\colon [n] \to \mathcal{RLL}_{t-3}(N)$ (i.e., an integer N satisfying $|\mathcal{RLL}_{t-3}(N)| \geqslant n$). By Lemma 4 an explicit function exists if $\lceil\frac{q}{q-2}N/q^{t-3}\rceil \leqslant N - \log(n)$ (for $q = 2$ that is $2\lceil 2N/2^{t-3}\rceil \leqslant N - \log(n)$), or equivalently $\big(1 - \frac{q^{4-t}}{q-2}\big)N \geqslant \lceil\log(n)\rceil$. With the assumed lower bound on t, this requirement is satisfied by $N \triangleq \lceil\log(n)\rceil + 1$, for sufficiently large n.
3) In the elimination stage (based on [11, Alg. 3]), $\boldsymbol{y}$ is processed from left to right; whenever $j > i$ are found such that $\boldsymbol{y}_{j+[s]} = \boldsymbol{y}_{i+[s]}$ (and again, j is minimal satisfying this requirement), the segment $\boldsymbol{y}_{j+[s]}$ is deleted and replaced with
$$1\ 0^{t-3}\ 1 \circ \boldsymbol{h}(i) \circ 1,$$
where we consider $10^{t-3}1$ to be a marker, indicating the replaced segment (based on the first step, this marker does not appear elsewhere in $\boldsymbol{y}$). Based on the fact that any elimination reduces the string length by 1, this stage is concluded in $O(n^2)$ steps. We denote the resulting string by $\boldsymbol{w}$, of length n' (for some $n' \leqslant n - t - 1$, depending on how many eliminations were performed). Trivially, the only instances of 0^{t-3} in $\boldsymbol{w}$ are the markers used in replaced substrings (and $\boldsymbol{w} \in \mathcal{RLL}_{t-2}(n')$). By following the same approach as in [11, Lem. 19] (which in turn was based on [13, Cla. 10]) one observes that $\boldsymbol{w} \in \mathcal{RF}_s(n')$ and that the process can be reversed; the former is trivial since the process only terminates when no repeated sequences remain. The latter is done by decoding from right to left, where replaced substrings are identified by the presence of markers, and the eliminated substrings are restored based on $\boldsymbol{h}(i)$. To prove this is possible, one needs only show that after any iteration of the process, the right-most instance of a marker is the one injected in that iteration. Indeed, since the process scans for j from left to right, if j is the location identified (i.e., $\boldsymbol{y}_{j+[s]}$ was replaced) in the last iteration, and j' in the iteration before that, then by necessity $j \geqslant j' - s + 1$; clearly, then, if $j < j'$ then the marker injected at location j' was overwritten in the last iteration. I.e., the marker injected at any iteration either overwrites the last injected marker, or appears in the replaced string to its right.
4) The process is concluded in an expansion stage, meant to output strings of length n from which $\boldsymbol{w}$ (hence also $\boldsymbol{x}$) can be decoded. For that purpose, an arbitrary string $\boldsymbol{v} \in \mathcal{RF}_s(n) \cap \mathcal{RLL}_{t-2}(n)$ is generated in a fashion to be described below, and interleaved with $1\ 0^{t-2}\ 1$ marker-segments after every s positions; i.e., if $\boldsymbol{v} = \boldsymbol{v}_0 \circ \boldsymbol{v}_1 \circ \cdots \circ \boldsymbol{v}_{\lceil n/s\rceil - 1}$, where $|\boldsymbol{v}_i| = s$ for all $i \in [\lfloor n/s\rfloor]$ and $\big|\boldsymbol{v}_{\lceil n/s\rceil-1}\big| \leqslant s$, then
$$\boldsymbol{w}' \triangleq \boldsymbol{v}_0 \circ 1\ 0^{t-2}\ 1 \circ \boldsymbol{v}_1 \circ 1\ 0^{t-2}\ 1 \circ \cdots \circ \boldsymbol{v}_{\lceil n/s\rceil-1}.$$
Clearly, $\boldsymbol{w}' \in \mathcal{RF}_{s+t}(n'') \cap \mathcal{RLL}_{t-1}(n'')$, where $n'' \triangleq |\boldsymbol{w}'| \geqslant n$. It is straightforward that the only instances of 0^{t-2} in $\boldsymbol{w}'$ are the markers interleaved into it. Based on these observations, it is proven similarly to [11, Lem. 23] that
$$\hat{\boldsymbol{w}} \triangleq \boldsymbol{w} \circ 1\ 0^{t-1}\ 1 \circ \boldsymbol{w}'$$
is $(s + 2t - 2) = (\lceil\log(n)\rceil + 3t)$-repeat-free and t-run-length-limited; this is done by observing that any substring of this length of $\hat{\boldsymbol{w}}$ contains markers (potentially unless it is a substring of $\boldsymbol{w}$, in which case the absence of markers indicates that fact), their length ($(t - i)$ for $i \in \{1, 2, 3\}$) indicates which portion of of $\hat{\boldsymbol{w}}$ it is taken from, and if it does not cover the unique instance of 0^{t-1} in $\hat{\boldsymbol{w}}$ then it contains s consecutive symbols of either $\boldsymbol{w}$ or $\boldsymbol{v}$, hence is unique.
Observe that by the upper bound on t, $\hat{\boldsymbol{w}}$ is $\ell(n)$-repeat-free. Also, the n-prefix of $\hat{\boldsymbol{w}}$ contains $\boldsymbol{w} \circ 1\ 0^{t-1}\ 1$, hence $\boldsymbol{w}$ can uniquely be extracted from it. That prefix is therefore output as the encoded information.
5) Finally, it remains to describe how any arbitrary $\boldsymbol{v} \in \mathcal{RF}_s(n) \cap \mathcal{RLL}_{t-2}(n)$ might be generated (a single example suffices). To achieve this, any total order on Σ is chosen where 0 is the minimum, and 1 the maximum. For a string $\boldsymbol{u} \in \Sigma^s$, let its *necklace* be the lexicographic least cyclic rotation of $\boldsymbol{u}$, and its *periodic-reduction* be the minimum period of $\boldsymbol{u}$. It was shown in [12] that concatenating in lexicographic order periodic-reductions of necklaces of length s produces a de Bruijn sequence $\boldsymbol{b} \in \Sigma^{q^s}$ (in fact, this is the lexicographically least de Bruijn sequence of that length).
The last instance of 0^t in $\boldsymbol{b}$ is in the necklace $0^t\ 1^{s-t}$ (since only the "0" necklace ends with 0, and the longest zero-run in any necklace appears at its beginning). Hence, letting $i \in [q^s]$ be the unique location such that $\boldsymbol{b}_{i+[s]} = 0^t\ 1^{s-t}$, we let $\boldsymbol{v} \triangleq \boldsymbol{b}_{q^s-n+[n]}$ and to conclude we need only show that $q^s - n > i$.
This was done in [11, Lem. 20] in case that s is prime, and we generalize for all s; we do so by counting $\big|\big\{\boldsymbol{b}_{j+[s]} : i < j \leqslant q^s - s\big\}\big| = q^s - s - i$. By the proof of [12, Th. 4] every $\boldsymbol{u} \in \Sigma^s$ appears in $\boldsymbol{b}$ in a location intersecting the appearance of the periodic-reduction of its necklace, potentially unless $u_0 = 1$.
Observe that it is sufficient that $\boldsymbol{u} \in \mathcal{RLL}_t(s)$ satisfies $u_{s-1} \neq 0$ for its necklace to be greater than $0^t\ 1^{s-t}$;

therefore for all $\boldsymbol{u} \in \mathcal{RLL}_t(s-2)$ and $u', u'' \in \Sigma$ satisfying $u' \neq 1$, $u'' \neq 0$ there exists $j > i$ such that $\boldsymbol{b}_{j+[s]} = u' \circ \boldsymbol{u} \circ u''$. It follows that

$$\begin{aligned} q^s - s - i &= \left|\left\{\boldsymbol{b}_{j+[s]} : i < j \leqslant q^s - s\right\}\right| \\ &\geqslant (q-1)^2 |\mathcal{RLL}_t(s-2)| \\ &\overset{(*)}{\geqslant} (q-1)^2 q^{(s-2)\left(1-2q^{-t}\right)} \\ &\geqslant (q-1)^2 n^{(1+t/\log(n))\left(1-2q^{-t}\right)}, \end{aligned}$$

where $(*)$ is justified by [21, Lem. 3] for sufficiently large t. Since $t > \lceil \log\log(n) \rceil$, for sufficiently large n we have $(1 + t/\log(n))(1 - 2q^{-t}) > 1$, as required.

Finally, redundancy of this construction is $t + 1$, plus the redundancy of encoding into $\mathcal{RLL}_{t-3}(n-t-1)$; Lemma 4 now concludes the proof. ■

In summary, we have the following corollary:

Corollary 6 *By Lemma 1, $\mathcal{RF}_{\ell_{\mathrm{over}}}(n)$ forms an $(\ell_{\min}, \ell_{\mathrm{over}})$-trace code in Σ^n, which by Theorem 5 has $1 - o_n(1)$ rate whenever $\ell_{\mathrm{over}} \geqslant \lceil \log(n) \rceil + 3\lceil \log\log(n) \rceil + 12$.*

In the sequel, we therefore focus on the complement, unsolved case of $\limsup \ell_{\mathrm{over}}/\log(n) \leqslant 1$.

IV. Bounds

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

Lemma 7 *Any multi-strand $(\ell_{\min}, \ell_{\mathrm{over}})$-trace code $\mathcal{C} \subseteq \mathcal{X}_{n,k}$ satisfies*

$$|\mathcal{C}| \leqslant \binom{k\left\lceil \frac{n-\ell_{\mathrm{over}}}{\ell_{\min}-\ell_{\mathrm{over}}} \right\rceil + q^{\ell_{\min}}}{q^{\ell_{\min}}}.$$

Proof: Since $\mathcal{L}_{\ell_{\min}}^{\ell_{\mathrm{over}}}(\boldsymbol{x}) \in \mathcal{T}_{\ell_{\min}}^{\ell_{\mathrm{over}}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Sigma^n$, we have

$$|\mathcal{C}| \leqslant \left|\left\{\mathcal{L}_{\ell_{\min}}^{\ell_{\mathrm{over}}}(S) : S \in \mathcal{X}_{n,k}\right\}\right|.$$

Similarly to the argument used in [5], we count the incidences of each possible $\boldsymbol{u} \in \Sigma^{\ell_{\min}}$ in $\mathcal{L}_{\ell_{\min}}^{\ell_{\mathrm{over}}}(S)$, resulting in $f_S \colon \Sigma^{\ell_{\min}} \to \mathbb{N}$ (dubbed a *profile-vector* in [5]). Observe that $\sum_{\boldsymbol{u} \in \Sigma^{\ell_{\min}}} f_S(\boldsymbol{u}) = k\left(1 + \left\lceil \frac{n-\ell_{\min}}{\ell_{\min}-\ell_{\mathrm{over}}} \right\rceil\right) = k\left\lceil \frac{n-\ell_{\mathrm{over}}}{\ell_{\min}-\ell_{\mathrm{over}}} \right\rceil$; thus, we have an embedding of $\left\{\mathcal{L}_{\ell_{\min}}^{\ell_{\mathrm{over}}}(S) : S \in \mathcal{X}_{n,k}\right\}$ into

$$\left\{\boldsymbol{f} \in \mathbb{N}^{q^{\ell_{\min}}} : \sum_{i \in [q^{\ell_{\min}}]} \boldsymbol{f}_i = k\left\lceil \frac{n-\ell_{\mathrm{over}}}{\ell_{\min}-\ell_{\mathrm{over}}} \right\rceil \right\},$$

and therefore

$$|\mathcal{C}| \leqslant \binom{k\left\lceil \frac{n-\ell_{\mathrm{over}}}{\ell_{\min}-\ell_{\mathrm{over}}} \right\rceil + q^{\ell_{\min}} - 1}{q^{\ell_{\min}} - 1},$$

which concludes the proof. ■

Lemma 8 *For $k = 1$, if $\ell_{\min} = a\log(n) + O_n(1)$ and $\ell_{\mathrm{over}} = \gamma\ell_{\min} + O_n(1)$, for some $a > 1$ and $0 \leqslant \gamma \leqslant \frac{1}{a}$, then any $(\ell_{\min}, \ell_{\mathrm{over}})$-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 γ is a linear scaling of the required overlap between consecutive segments, in proportion to their required minimum length. We scale that minimum length linearly with $\log(n)$ (where the constant a indicates the ratio), a decision informed by the statement of this lemma, and the succeeding corollary. Finally, observe that in this notation, $\frac{1-1/a}{1-\gamma} \leqslant 1$ if and only if $\lim \ell_{\mathrm{over}}/\log(n) = \gamma a \leqslant 1$.)

Proof: From the known bound $u! > (u/e)^u$ we observe for all $v \geqslant u > 0$ that

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

where the last inequality holds since $\frac{u}{v} + 1 \leqslant 2 < e$.

Letting $v \triangleq q^{\ell_{\min}}$ and $u \triangleq \left\lceil \frac{n-\ell_{\mathrm{over}}}{\ell_{\min}-\ell_{\mathrm{over}}} \right\rceil < v$, we observe that $\log(\frac{v}{u}) = O(\log(n))$ and $\log(u) \geqslant \log(n - \ell_{\mathrm{over}}) - \log\log(n) + O(1)$; observing

$$\begin{aligned} \log(n - \ell_{\mathrm{over}}) &= \log(n) + \log\left(1 - \tfrac{\ell_{\mathrm{over}}}{n}\right) \\ &\geqslant \log(n) - \frac{\log(e)\ell_{\mathrm{over}}}{n - \ell_{\mathrm{over}}} \\ &= \log(n) - O\left(\frac{\log(n)}{n}\right), \end{aligned}$$

where we used $\ln(1-x) \geqslant \frac{-x}{1-x}$, we summarize $\log(u) \geqslant \log(n) - \log\log(n) + O(1)$.

Next, by Lemma 7 $|\mathcal{C}| \leqslant \binom{u+v}{u}$, hence we have

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

■

In particular, Lemma 8 implies the following lower bound on $\ell_{\min}$ for the existence of codes with asymptotically non-vanishing rates.

Corollary 9 *Take $(\ell_{\min}^{(n)})_{n>0}, (\ell_{\mathrm{over}}^{(n)})_{n>0}$, and let $\mathcal{C}^n \subseteq \Sigma^n$ be $(\ell_{\min}^{(n)}, \ell_{\mathrm{over}}^{(n)})$-trace codes. If $\limsup_n \ell_{\min}^n/\log(n) \leqslant 1$, then $R(\mathcal{C}^n) = o_n(1)$.*

Proof: Since $\mathcal{C}^n$ are also $({\ell'_{\min}}^{(n)}, 0)$-trace codes for ${\ell'_{\min}}^{(n)} \geqslant \ell_{\min}^{(n)}$, it follows from Lemma 8 that $R(\mathcal{C}^n) \leqslant \frac{1-1/a}{1-0} + o(1)$ for all $a > 1$, hence the claim follows. ■

V. A Construction of Trace Codes

In this section we present an efficient encoder for $(\ell_{\min}, \ell_{\text{over}})$-trace codes (i.e., in the case $k = 1$), achieving asymptotically optimal rate, for the case $\limsup \ell_{\text{over}} / \log(n) \leqslant 1$ (complementing the results of Section II). Throughout the section, we let

$$\begin{aligned} \ell_{\min} &\triangleq \lceil a \log(n) \rceil; \\ \ell_{\text{over}} &\triangleq \lceil \gamma \ell_{\min} \rceil, \end{aligned} \tag{1}$$

for some $a > 1$ and $0 < \gamma \leqslant 1/a$. Further, we let f be any integer function satisfying $f(n) = o(\log(n))$ and $f(n) \geqslant \log\log(n) + 4$, and finally

$$I \triangleq \left\lceil \frac{1-\gamma a}{1-\gamma} \log(n) + (\log(n))^{0.5+\epsilon} \right\rceil, \tag{2}$$

for some small $\epsilon > 0$. In our construction, I is the number of symbols dedicated to (unencoded-)indices, which are then partitioned into length-$f(n)$ fragments, as described below. When analyzing the redundancy of our construction, we shall optimize it by a proper choice of $f(n)$, in Theorem 15.

The main idea of the construction presented below of an $(\ell_{\min}, \ell_{\text{over}})$-trace code $\mathcal{C}_A(n)$ is to encode an information string $\boldsymbol{x}$ into $(\boldsymbol{z}_i)_{i \in [q^I]}$ so that the following two properties are satisfied: (i) the index i can be decoded from any $\ell_{\min}$-substring of $\boldsymbol{z}_i$; and (ii) the string $\boldsymbol{z}_i$ can be uniquely reconstructed from an $(\ell_{\min}, \ell_{\text{over}})$-trace of $\boldsymbol{z}_i$. This is performed by interleaving segments of indices in appropriate locations in the encoded strings. Then, we let

$$\mathrm{Enc}_A(\boldsymbol{x}) \triangleq \boldsymbol{z} = \boldsymbol{z}_0 \circ \cdots \circ \boldsymbol{z}_{q^I - 1} \in \mathcal{C}_A(n).$$

Before presenting the construction, we describe the method of index-generation.

Definition 10 *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 4). Denoting $F \triangleq \lceil I / f(n) \rceil$, we partition $\boldsymbol{c}_i$ into F non-overlapping segments of equal lengths $\{\boldsymbol{c}_i^{(h)}\}_{h \in [F]}$; here and in the sequel, we say a string is partitioned into non-overlapping segments of equal lengths if $\boldsymbol{c}_i^{(0)} \circ \boldsymbol{c}_i^{(1)} \circ \cdots \circ \boldsymbol{c}_i^{(F-1)} = \boldsymbol{c}_i$ and*

$$\left|\boldsymbol{c}_i^{(h)}\right| = \begin{cases} \lceil I/F \rceil, & h < I \bmod F; \\ \lfloor I/F \rfloor, & \textit{otherwise}. \end{cases}$$

Observe that $\left|\boldsymbol{c}_i^{(h)}\right| \leqslant f(n)$ for all $h \in [F]$. We then denote ${\boldsymbol{c}'_i}^{(h)} \triangleq 1 \circ \boldsymbol{c}_i^{(h)} \circ 1$. We refer to $\boldsymbol{c}_i$ (or simply i) as an index *in the construction, and to $\{{\boldsymbol{c}'_i}^{(h)}\}_{h \in [F]}$ as segments of an* encoded index.

Further, for $N \leqslant n$ to be defined later, and $\ell > \lceil \log(N) \rceil + 3f(n)$, we denote the encoder of Theorem 5

$$E_{N,\ell}^{\mathcal{RF}} : \Sigma^{m(N)} \to \mathcal{RF}_\ell(N) \cap \mathcal{RLL}_{f(n)+1}(N).$$

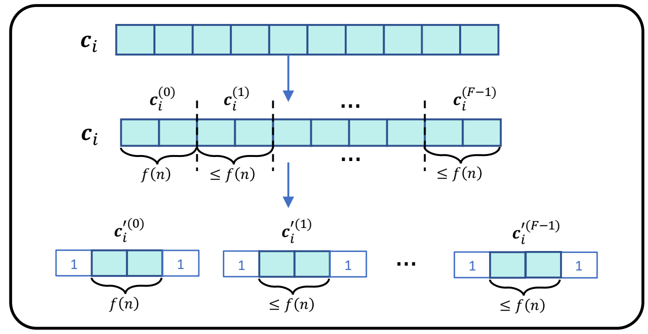

Figure 4. Index generation. Each index $\boldsymbol{c}_i$ is first partitioned into $F+1$ non-overlapping segments of length $f(n)$. Then, each of the segments is concatenated with a single 1 in each edge.

Here,

$$\begin{aligned} m(N) &\triangleq N - \mathrm{red}(E_{N,\ell}^{\mathcal{RF}}) \\ &\geqslant N\left(1 - \frac{f(n)+2}{N} - \frac{q^3}{q-2} q^{-f(n)}\right). \end{aligned} \tag{3}$$

(For $q = 2$, $m(N) \geqslant N\left(1 - \frac{f(n)+4}{N} - 2^{4-t}\right)$.)

Construction A The encoding into $\boldsymbol{z}_i$, for all $i \in [q^I]$, is performed as follows (see Figure 5). We denote

$$\begin{aligned} r &\triangleq I + 2F + f(n) + 4 \\ &= I + \frac{2I}{f(n)} + f(n) + O(1), \end{aligned} \tag{4}$$

then define

$$\ell \triangleq \left\lceil \frac{\ell_{\text{over}} - 2f(n) - 6}{1 + (f(n)+2) / \left\lfloor \frac{\ell_{\min} - r}{F} \right\rfloor} \right\rceil. \tag{5}$$

(see Lemmas 11 and 12, respectively, for the reason for these definitions). Also, for all $i \in [q^I]$

$$N_i \triangleq \begin{cases} \lceil q^{-I} n \rceil - \lceil n/(q^I \ell_{\min}) \rceil r, & i < n \bmod q^I; \\ \lfloor q^{-I} n \rfloor - \lceil n/(q^I \ell_{\min}) \rceil r, & \text{otherwise}. \end{cases} \tag{6}$$

Now, for all $i \in [q^I]$ define $\boldsymbol{y}_i \triangleq E_{N_i,\ell}^{\mathcal{RF}}(\boldsymbol{x}_i) \in \Sigma^{N_i}$, where $\boldsymbol{x}_i \in \Sigma^{m(N_i)}$ and

$$\boldsymbol{x} = \boldsymbol{x}_0 \circ \boldsymbol{x}_1 \circ \cdots \circ \boldsymbol{x}_{q^I - 1}$$

is an arbitrary information string (see the proof of Theorem 15 for a choice of $f(n)$ satisfying the conditions of Theorem 5, hence assuring the existence of $E_{N_i,\ell}^{\mathcal{RF}}$).

Next, for all $i \in [q^I]$

1) Partition $\boldsymbol{y}_i$ into $\lceil n/(q^I \ell_{\min}) \rceil$ non-overlapping segments of equal length

$$\boldsymbol{y}_i = \boldsymbol{y}_{i,0} \circ \boldsymbol{y}_{i,1} \circ \cdots \circ \boldsymbol{y}_{i, \lceil n/(q^I \ell_{\min}) \rceil - 1}.$$

2) For all $j \in [\lceil n/(q^I \ell_{\min}) \rceil]$:
   a) Partition each $\boldsymbol{y}_{i,j}$ into F non-overlapping segments of equal lengths

$$\boldsymbol{y}_{i,j} = \boldsymbol{y}_{i,j}^{(0)} \circ \boldsymbol{y}_{i,j}^{(1)} \circ \cdots \circ \boldsymbol{y}_{i,j}^{(F-1)}.$$

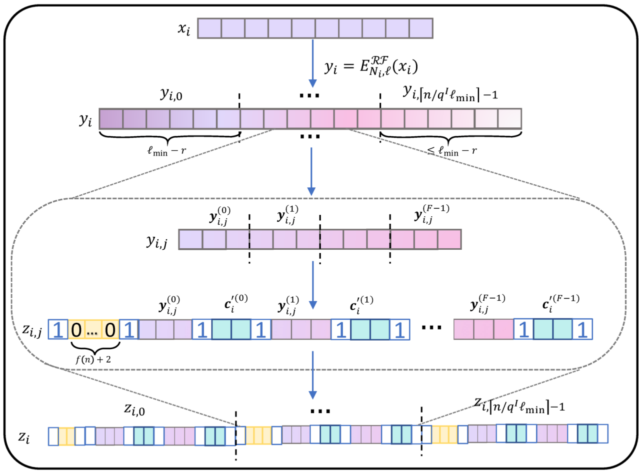

Figure 5. Encoding $\boldsymbol{x}_i$ into $\boldsymbol{z}_i$, as detailed in Construction A.

b) Combine $\left\{\boldsymbol{y}^{(h)}_{i,j} : h \in [F]\right\}$ with segments of the encoded index i, as follows. Define for all $h \in [F]$
$$\boldsymbol{z}^{(h)}_{i,j} \triangleq \boldsymbol{y}^{(h)}_{i,j} \circ \boldsymbol{c}'^{(h)}_i,$$
then
$$\boldsymbol{z}_{i,j} \triangleq \begin{cases} 1\,0^{f(n)+1}\,1\,1 \circ \boldsymbol{z}^{(0)}_{i,j} \circ \cdots \circ \boldsymbol{z}^{(F-1)}_{i,j}, & j=0; \\ 1\,0^{f(n)+1}\,0\,1 \circ \boldsymbol{z}^{(0)}_{i,j} \circ \cdots \circ \boldsymbol{z}^{(F-1)}_{i,j}, & j>0 \end{cases}$$
(we refer to the substrings $10^{f(n)+1}11, 10^{f(n)+1}01$ as *synchronization markers*).

3) Concatenate
$$\boldsymbol{z}_i \triangleq \boldsymbol{z}_{i,0} \circ \cdots \circ \boldsymbol{z}_{i,\lceil n/(q^I\ell_{\min})\rceil-1}. \qquad \square$$

First, we prove the correctness of Construction A. We begin with two technical lemmas which are key to the proof of correctness in Theorem 13.

Lemma 11 *Every $\ell_{\min}$-substring $\boldsymbol{u}$ of $\boldsymbol{z} \in \mathcal{C}_A(n)$ contains as subsequences at least an $(I-\mu)$-suffix of an index $\boldsymbol{c}_i$ (see Definition 10), 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.*

Proof: Note that
$$\begin{aligned}\ell_{\min} - q^I\ell_{\min}^2/n &\leqslant \frac{\ell_{\min}}{1+q^I\ell_{\min}/n} = \frac{n/q^I}{n/(q^I\ell_{\min})+1} \\ &\leqslant \frac{n/q^I}{\lceil n/(q^I\ell_{\min})\rceil} \leqslant \frac{\lceil n/q^I\rceil}{\lceil n/(q^I\ell_{\min})\rceil} \leqslant \ell_{\min}\end{aligned}$$
Observing from Eq. (2) that $q^I\ell_{\min}^2 = o(n)$, and by subtracting r from the above inequality, it holds from Eq. (6) for sufficiently large n and all $i \in [q^I]$ that $\ell_{\min} - r - 1 \leqslant N_i/\lceil n/(q^I\ell_{\min})\rceil \leqslant \ell_{\min} - r$. Hence also for all $j \in [\lceil n/(q^I\ell_{\min})\rceil]$ it holds that
$$\ell_{\min} - r - 1 \leqslant \left|\boldsymbol{y}_{i,j}\right| \leqslant \ell_{\min} - r. \tag{7}$$
By part 2 of Construction A it follows that $|\boldsymbol{z}_{i,j}| = |\boldsymbol{y}_{i,j}| + r \in \{\ell_{\min}, \ell_{\min}-1\}$.

Next, observe that instances of synchronization markers only appear in $\boldsymbol{z}$ at the beginning of $\{\boldsymbol{z}_{i,j}\}_{i,j}$. From the last paragraph, either $\boldsymbol{u}$ contains a complete synchronization marker as substring, or it contains a suffix-prefix pair whose concatenation is an instance of a synchronization marker; in both cases, the exact locations in which symbols of the indices $\left\{\boldsymbol{c}'^{(h)}_i\right\}$ appear can be determined. Extracting $\left\{\boldsymbol{c}_i{}^{(h)}\right\}$, these contain a suffix of $\boldsymbol{c}_i$ and a prefix of either $\boldsymbol{c}_i, \boldsymbol{c}_{i+1}$ (depending on whether $\boldsymbol{u}$ is a substring of $\boldsymbol{z}_i$ for some i) whose combined lengths is I, again since for all i,j, $|\boldsymbol{z}_{i,j}| \leqslant \ell_{\min}$ and $\boldsymbol{z}_{i,j}$ contains all symbols of $\boldsymbol{c}_i$. Taking $\mu \in [I]$ to be the length of the prefix ($\mu = 0$ indicates the possibility that all symbols of the same index appear in $\boldsymbol{u}$) concludes the proof. ∎

Lemma 12 *Every ℓ_{over}-substring $\boldsymbol{v}$ of $\boldsymbol{z} \in \mathcal{C}_A(n)$ contains at least ℓ consecutive symbols of $\boldsymbol{y} \triangleq \boldsymbol{y}_0 \circ \cdots \circ \boldsymbol{y}_{q^I-1}$ (see Eq. (5)).*

Proof: At worst, $\boldsymbol{v}$ either begins or ends with a complete instance of a synchronization marker; hence the remaining $\ell_{\text{over}} - f(n) - 4$ symbols are sampled from $\left\{\boldsymbol{z}^{(h)}_{i,j}\right\}$, and again, at worst end with a complete segment of an encoded index. Since from Definition 10 $\left|\boldsymbol{c}'^{(h)}_i\right| \leqslant f(n)+2$ and by Eq. (7) $\left|\boldsymbol{y}^{(h)}_{i,j}\right| \geqslant \left\lfloor \frac{\ell_{\min}-r}{F} \right\rfloor$ for all i,j,h, $\boldsymbol{v}$ contains at least
$$\left\lceil \frac{\ell_{\text{over}} - 2f(n) - 6}{1 + (f(n)+2)/\left\lfloor \frac{\ell_{\min}-r}{F} \right\rfloor} \right\rceil = \ell$$
consecutive symbols of $\boldsymbol{y}$. ∎

Combining both lemmas, we have the following theorem.

Theorem 13 *For all admissible values of n, the code $\mathcal{C}_A(n)$ is an $(\ell_{\min}, \ell_{\text{over}})$-trace code.*

Proof: Take $\boldsymbol{z} \in \mathcal{C}_A(n)$ and let $T \in \mathcal{T}^{\ell_{\text{over}}}_{\ell_{\min}}(\boldsymbol{z})$, i.e., any $(\ell_{\min}, \ell_{\text{over}})$-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 11. 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 every symbol of the $(I-\mu)$-suffix of $\boldsymbol{c}_i$ is $(q-1)$, in which case it is the q-ary expansion of the successor natural number to that prefix. 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 12, 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 $E^{\mathcal{RF}}_{N,\ell}$. ∎

Next, we analyze $R(\mathcal{C}_A(n))$. First, we require a simplified (asymptotic) expression for ℓ, used in Construction A for repeat-free encoding, which we derive in the next lemma.

Lemma 14 *Denoting $\lambda \triangleq 1 - \frac{I}{\ell_{\min}}$, we have*

$$\ell = \lambda \ell_{\text{over}} - O\left(f(n) + \frac{\log(n)}{f(n)}\right).$$

Proof: Recall the definition $\ell = \left\lceil \frac{\ell_{\text{over}} - 2f(n) - 6}{1 + (f(n)+2) \big/ \left\lfloor \frac{\ell_{\min} - r}{F} \right\rfloor} \right\rceil$ in Eq. (5), where $F = \lceil I/f(n) \rceil$ and r is defined in Eq. (4). We begin by observing

$$\begin{aligned}
\frac{f(n)+2}{\left\lfloor \frac{\ell_{\min}-r}{F} \right\rfloor} &= \frac{F(f(n)+2)}{\ell_{\min} - r - O(F)} \\
&= \frac{I + O(f(n))}{\ell_{\min} - I - \frac{2I}{f(n)} - O\left(f(n) + \frac{\log(n)}{f(n)}\right)} \\
&= \frac{I}{\ell_{\min} - I} \cdot \frac{1 + O\left(\frac{f(n)}{\log(n)}\right)}{1 - O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right)} \\
&= \frac{I}{\ell_{\min} - I} + O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right) \\
&= \frac{1-\lambda}{\lambda} + O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right),
\end{aligned}$$

where the second to last equality is justified by $\frac{1}{1-x} = 1 + x + \frac{x^2}{1-x}$ for $x \neq 1$, and since from Equations (1) and (2) $\frac{I}{\ell_{\min} - I} = O(1)$. Finally,

$$\begin{aligned}
\ell &= \frac{\ell_{\text{over}} - 2f(n) - 6}{1 + \frac{1-\lambda}{\lambda} + O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right)} + O(1) \\
&= \frac{\lambda \ell_{\text{over}} - O(f(n))}{1 + O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right)} + O(1) \\
&= (\lambda \ell_{\text{over}} - O(f(n)))\left(1 - O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right)\right) \\
&= \lambda \ell_{\text{over}} - O\left(\frac{\log(n)}{f(n)} + f(n)\right),
\end{aligned}$$

where again the second to last equality is based on $\frac{1}{1-x} = 1 + x + \frac{x^2}{1-x}$. ■

Based on this property, we show that Construction A asymptotically meets the bound of Lemma 8.

Theorem 15 *Letting $f(n) \triangleq \left\lceil \sqrt{\log(n)} \right\rceil$, the use of Theorem 5 in Construction A is justified, and we have*

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

Proof: We start by noting from Equations (1), (2) and (4)

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

and

$$\begin{aligned}
\frac{r}{\ell_{\min}} &= \frac{I\left(1 + O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right)\right)}{\ell_{\min}} \\
&= \frac{I}{\ell_{\min}} + O\left(\frac{1}{f(n)} + \frac{f(n)}{\log(n)}\right) \\
&= \frac{1/a - \gamma}{1-\gamma} + \frac{(\log(n))^{\epsilon}}{a\sqrt{\log(n)}} + O\left(\frac{1}{\sqrt{\log(n)}}\right).
\end{aligned}$$

Now, from Eq. (6)

$$\begin{aligned}
N_i &\geqslant \lfloor n/q^I \rfloor - \lceil n/(q^I \ell_{\min}) \rceil r \\
&\geqslant q^{-I} n(1 - r/\ell_{\min}) - (r+1) \\
&= q^{-I} n\left(1 - \frac{r}{\ell_{\min}} - \frac{q^I(r+1)}{n}\right) = \Omega(q^{-I} n),
\end{aligned}$$

hence

$$\begin{aligned}
\log(N_i) &\geqslant \log(n) - I + O(1) \\
&= \frac{(a-1)\gamma}{1-\gamma}\log(n) - (\log(n))^{0.5+\epsilon} + O(1).
\end{aligned}$$

In particular for sufficiently large n we have

$$f(n) \geqslant \lceil \log\log(N_i) \rceil + 5. \tag{8}$$

Next, by Lemma 14,

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

Hence

$$\begin{aligned}
\ell - \lceil \log(N_{n,\ell}(m)) \rceil &= (1-\gamma)(\log(n))^{0.5+\epsilon} - O\left(\sqrt{\log(n)}\right) \\
&> 3f(n)
\end{aligned} \tag{9}$$

for sufficiently large n. Together, Equations (8) and (9) satisfy the conditions of Theorem 5, allowing us to efficiently encode $\boldsymbol{y}_i = E^{\mathcal{RF}}_{N_i,\ell}(\boldsymbol{x}_i) \in \mathcal{RF}_\ell(N_i)$ (and vice versa, decode $\boldsymbol{x}_i$) while attaining from Eq. (3) $\frac{m(N_i)}{N_i} \geqslant 1 - \frac{f(n)+2}{N_i} - \frac{q^3}{q-2} q^{-f(n)} = 1 - O(q^{-f(n)})$, where the coefficient of the asymptotic notation does not depend on i. Hence,

$$\begin{aligned}
\sum_{i \in [q^I]} m(N_i) &\geqslant \left(1 - O(q^{-f(n)})\right) \sum_{i \in [q^I]} N_i \\
&= \left(1 - O(q^{-f(n)})\right)\left(n - q^I \lceil n/(q^I \ell_{\min}) \rceil r\right) \\
&\geqslant n\left(1 - O(q^{-f(n)})\right)\left(1 - \frac{r}{\ell_{\min}} - \frac{q^I r}{n}\right),
\end{aligned}$$

where the equality on the second line follows from Eq. (6), which concludes the proof. ■

From the proof of Theorem 15 we note that ϵ in Construction A must satisfy $\epsilon \geqslant \max\left\{\frac{\log(f(n))}{\log\log(n)}, 1 - \frac{\log(f(n))}{\log\log(n)}\right\} - 0.5$; it follows that the choice $f(n) = \left\lceil \sqrt{\log(n)} \right\rceil$ is optimal, in the sense that $\frac{\log(f(n))}{\log\log(n)} = \frac{1}{2} + o(1)$.

VI. Multi-strand reconstruction from substring-compositions

In this section, we study an extension of the reconstruction from substring-compositions problem, i.e., $(\ell, \ell-1)$-trace codes, to multisets of strings, i.e, to codes over $\mathcal{X}_{n,k}$ for $k > 1$. For a string $\boldsymbol{x} \in \Sigma^n$ we denote for brevity an $(\ell, \ell-1)$-trace of $\boldsymbol{x}$, $\mathcal{T}_\ell^{\ell-1}(\boldsymbol{x})$ and $\mathcal{L}_\ell^{\ell-1}(\boldsymbol{x})$ by an ℓ-trace, $\mathcal{T}_\ell(\boldsymbol{x})$ and $\mathcal{L}_\ell(\boldsymbol{x})$, respectively. We say $\mathcal{L}_\ell(\boldsymbol{x})$ in particular is the *ℓ-profile* of $\boldsymbol{x}$, the multiset of all of its ℓ-substrings.

We shall assume throughout in asymptotic analysis that as n grows, $\limsup \frac{\log(k)}{n} < 1$, which is most relevant in applications (see, e.g., [10] for an overview of typical string-lengths in applications); the complement case is of independent theoretical interest, and is left for future work. Hence, we have the following lemma.

Lemma 16 $\log|\mathcal{X}_{n,k}| = k(n - \log(k/e)) + o(k) = \Theta(nk)$.

Proof: From Stirling's approximation we have $(k/e)^k \leqslant k! \leqslant e\sqrt{k}(k/e)^k$, implying

$$\frac{1}{e\sqrt{k}}\left(\frac{q^n}{k/e}\right)^k \leqslant \frac{q^{nk}}{k!} \leqslant |\mathcal{X}_{n,k}| \leqslant \frac{(q^n+k)^k}{k!} \leqslant \left(\frac{q^n(1+k/q^n)}{k/e}\right)^k. \qquad \blacksquare$$

For a *multi-strand ℓ-trace code* $\mathcal{C}$ we have from Lemma 7 that $|\mathcal{C}| \leqslant \binom{nk+q^\ell}{q^\ell}$. A corollary of Lemma 8 is therefore stated:

Corollary 17 *Assume* $\limsup \frac{\log(k)}{n} < 1$. *If* $\log(nk) - \ell = \omega_{nk}(1)$ *then for any multi-strand ℓ-trace code* $\mathcal{C} \subseteq \mathcal{X}_{n,k}$ *it holds that*

$$R(\mathcal{C}) = o_{nk}(1).$$

Proof: By the observation in the proof of Lemma 8

$$R(\mathcal{C}) \leqslant \frac{1}{|\mathcal{X}_{n,k}|}\log\binom{nk+q^\ell}{q^\ell} = O\left(\frac{q^\ell}{nk}(2\log(e) + \log(nk) - \ell)\right),$$

where the equality follows from Lemma 16. $\blacksquare$

On the other hand, recall from Corollary 2 that $\mathcal{RF}_\ell(n,k) \subseteq \mathcal{X}_{n,k}$ is a multi-strand $(\ell+1)$-trace code. Next, we show in contrast to Corollary 17 that if $\ell - \log(nk) = \omega_{nk}(1)$, then $R(\mathcal{RF}_\ell(n,k)) = 1 - o_{nk}(1)$. We shall do so by presenting two explicit constructions of multi-strand ℓ-repeat-free codes with efficient encoders and decoders. For convenience, we assume all quantities to have integer values; a straightforward adjustment of the described methods applies for all values.

A. Index-based construction

Construction B Denote $n' \triangleq (n - \log(k))k$, and take m such that $E\colon \Sigma^m \to \mathcal{RF}_{\ell'}(n')$ is any repeat-free encoder, for a given ℓ'. Let $\boldsymbol{x} \in \Sigma^m$ be an arbitrary information string, and encode it into $\boldsymbol{y} \triangleq E(\boldsymbol{x})$. Take $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{k-1} \in \Sigma^{n-\log(k)}$ such that $\boldsymbol{y} = \boldsymbol{y}_0 \circ \boldsymbol{y}_1 \circ \cdots \circ \boldsymbol{y}_{k-1}$. Let $\boldsymbol{c}_i \in \Sigma^{\log(k)}$ be a q-ary expansion of $i \in [k]$. Denote $\widetilde{\boldsymbol{y}}_i \triangleq \boldsymbol{c}_i \circ \boldsymbol{y}_i$; then,

$$\mathrm{Enc}_B(\boldsymbol{x}) \triangleq \{\{\widetilde{\boldsymbol{y}}_i : i \in [k]\}\} \in \mathcal{X}_{n,k}. \qquad \square$$

We denote $\mathcal{C}_B(n,k) \triangleq \mathrm{Enc}_B(\Sigma^m)$. The decoding success of Construction B follows from the next lemma.

Lemma 18 $\mathcal{C}_B(n,k) \subseteq \mathcal{RF}_\ell(n,k)$, *where* $\ell = \ell' + \log(k)$.

Proof: For $\boldsymbol{x} \in \Sigma^m$, note that $\boldsymbol{y} \triangleq \mathrm{Enc}_B(\boldsymbol{x}) = \boldsymbol{y}_0 \circ \boldsymbol{y}_1 \circ \cdots \circ \boldsymbol{y}_{k-1} \in \mathcal{RF}_{\ell'}(n')$ and thus $\|\mathcal{L}_{\ell'}(\boldsymbol{y})\| = n' - \ell' + 1$. It follows that $\|\mathcal{L}_{\ell'}(\{\{\boldsymbol{y}_i : i \in [k]\}\})\| = (n' - \ell' + 1) - (k-1)(\ell' - 1) = k(n - \ell + 1)$.

Now, let $\boldsymbol{u}, \boldsymbol{v}$ be ℓ-substrings of $\widetilde{\boldsymbol{y}}_i, \widetilde{\boldsymbol{y}}_j$ respectively; note that the ℓ'-suffixes of $\boldsymbol{u}, \boldsymbol{v}$ are ℓ'-substrings of $\boldsymbol{y}_i, \boldsymbol{y}_j$ respectively, and hence if $\boldsymbol{u} = \boldsymbol{v}$ then $i = j$ and their locations in $\boldsymbol{y}_i$ agree. It follows that the locations of $\boldsymbol{u}, \boldsymbol{v}$ in $\widetilde{\boldsymbol{y}}_i$ agree as well, and the claim follows. $\blacksquare$

Recall, then, that given $\mathcal{L}_{\ell+1}(\mathrm{Enc}_B(\boldsymbol{x}))$, an efficient algorithm produces the set of strings $\{\widetilde{\boldsymbol{y}}_i : i \in [k]\}$. Then, by ordering and subsequent removal of the length-$\log(k)$ indices from these strings, one obtains the string $\boldsymbol{y} = E(\boldsymbol{x})$, and consequently, $\boldsymbol{x}$. Note that the role of the indices in this construction is crucial to deduce $\boldsymbol{y}$ from its ℓ-profile; without indices the order of these k substrings could not have been derived, hence one would only obtain $\boldsymbol{y}$ up to a permutation of its non-overlapping (n'/k)-substrings. The next theorem analyzes the parameters of codes that can be constructed using Construction B based upon Lemma 3 and Theorem 5.

Theorem 19 *Given* $\ell(n,k)$, *denote* $f(n,k) \triangleq \ell(n,k) - \log(nk) - \log(k)$. *Further, let* $\ell' \triangleq \ell(n,k) - \log(k)$. *Here, we assume Construction B is operated with* n, k, ℓ'. *Observe*

$$\ell' - \log(n') = f(n,k) - \log\left(1 - \frac{\log(k)}{n}\right) \geqslant f(n,k).$$

1) *If* $f(n,k) \geqslant 3\log\log(nk) + 12$ *then utilizing Theorem 5 in Construction B we obtain*

$$R(\mathcal{C}_B(n,k)) \geqslant 1 - \frac{q^{4-\lfloor f(n,k)/3 \rfloor}}{q-2} - \frac{\log(e)}{n - \log(k)} - o\left(\tfrac{1}{n}\right).$$

(For $q = 2$ *that is* $R(\mathcal{C}_B(n,k)) \geqslant 1 - 2^{5-\lfloor f(n,k)/3 \rfloor} - \frac{\log(e)}{n-\log(k)} - o\left(\frac{1}{n}\right)$.*)*

2) *If* $f(n,k) \geqslant \log(nk) + 2 + 2\log\left(1 - \frac{\log(k)}{n}\right)$ *then utilizing Lemma 3 in Construction B we have*

$$R(\mathcal{C}_B(n,k)) \geqslant 1 - \frac{\log(e)}{n - \log(k)} - o\left(\tfrac{1}{n}\right).$$

Proof:

1) Note that by the assumption, Theorem 5 may be applied for some choice of t. Since Construction B does not require $\boldsymbol{y}$ to be run-length constrained, we let $t \triangleq \lfloor (\ell' - \log(n'))/3 \rfloor$ and observe

$$\begin{aligned} m &\geqslant n' - t - 1 - \left\lceil \tfrac{q^4}{q-2} n'/q^t \right\rceil \\ &= n' - \left\lceil \tfrac{q^4}{q-2} n'/q^{\lfloor (\ell' - \log(n'))/3 \rfloor} \right\rceil \\ &\geqslant \left(1 - \tfrac{q^4}{q-2} q^{-\lfloor f(n,k)/3 \rfloor}\right) k(n - \log(k)) - 1. \end{aligned}$$

(For $q = 2$, that is $m \geqslant \left(1 - 2^{5-\lfloor f(n,k)/3 \rfloor}\right)k(n - \log(k)) - 3$.)

It then follows from Lemma 16 that

$$\begin{aligned} R(\mathcal{C}_B(n,k)) &= \frac{m}{\log|\mathcal{X}_{n,k}|} \\ &\geqslant \frac{\left(1 - \frac{q^{4-\lfloor f(n,k)/3 \rfloor}}{q-2}\right)k(n - \log(k)) - 1}{k(n - \log(k/e)) + o(k)} \\ &= \frac{1 - \frac{q^{4-\lfloor f(n,k)/3 \rfloor}}{q-2} - \frac{1}{(n-\log(k))k}}{1 + \frac{\log(e)+o(1)}{n-\log(k)}} \\ &= 1 - \frac{q^{4-\lfloor f(n,k)/3 \rfloor}}{q-2} - \frac{\log(e)}{n - \log(k)} - o\left(\tfrac{1}{n}\right), \end{aligned}$$

where again the last equality follows from $\frac{1}{1-x} = 1 + x + \frac{x^2}{1-x}$, and from the observation $n - \log(k) = \Theta(n)$. (Similarly for $q = 2$.)

2) Equivalently, $\ell' \geqslant 2\log(n') + 2$, hence by Lemma 3 we have $m = n' - 1$. Following the same steps as in the last part,

$$\begin{aligned} R(\mathcal{C}_B(n,k)) &= \frac{m}{\log|\mathcal{X}_{n,k}|} \\ &\geqslant 1 - \frac{\log(e)}{n - \log(k)} - o\left(\tfrac{1}{n}\right). \end{aligned}$$ ■

B. Overlap-based construction

While in Construction B we added indices in order to overcome the lack of ordering when the string $\boldsymbol{y} = E(\boldsymbol{x})$ is partitioned into k substrings, in Construction C we tackle this constraint differently. To wit, we again partition $\boldsymbol{y}$, but include overlapping segments between consecutive substrings. The overlapping segments will guarantee in decoding that, given the set of k substrings, there will be a unique way to concatenate them into one long string. As opposed to Construction B, this approach eliminates the need to decrease the length used for repeat-free encoders with respect to that of the read substrings, i.e., ℓ.

Construction C For a given ℓ, denote $n' \triangleq nk - (k-1)\ell = (n-\ell)k + \ell$, and take m such that $E\colon \Sigma^m \to \mathcal{RF}_\ell(n')$ is any repeat-free encoder. Let $\boldsymbol{x} \in \Sigma^m$ be an arbitrary information string, and encode it into $\boldsymbol{y} \triangleq E(\boldsymbol{x})$. Define k length-n strings $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{k-1} \in \Sigma^n$ by segmenting $\boldsymbol{y}$ with an overlap of ℓ symbols between consecutive segments; more precisely, let $\boldsymbol{y}_i \triangleq (y_{i,1}, \dots, y_{i,n})$ for $i \in [k]$, where

$$y_{i,j} \triangleq y_{i(n-\ell)+j}; \quad j \in [n].$$

Then,

$$\mathrm{Enc}_C(\boldsymbol{x}) \triangleq \{\{\boldsymbol{y}_i : i \in [k]\}\} \in \mathcal{X}_{n,k}.$$ □

We denote $\mathcal{C}_C(n,k) \triangleq \mathrm{Enc}_C(\Sigma^m)$. The decoding success of Construction C follows from the following simple observation.

Lemma 20 *For all $\boldsymbol{x} \in \Sigma^m$ it holds that $\mathcal{L}_{\ell+1}(\boldsymbol{y}) = \mathcal{L}_{\ell+1}(\mathrm{Enc}_C(\boldsymbol{x}))$.*

Proof: Since $\boldsymbol{y}_i$ is a substring of $\boldsymbol{y}$ for all i, it follows that $\mathcal{L}_{\ell+1}(\mathrm{Enc}_C(\boldsymbol{x})) \subseteq \mathcal{L}_{\ell+1}(\boldsymbol{y})$. For the other direction, note that $\boldsymbol{y}_i, \boldsymbol{y}_{i+1}$ are overlapping substrings of $\boldsymbol{y}$ for all $1 \leqslant i < k$, with a common substring of length ℓ; thus all $(\ell+1)$-substrings of $\boldsymbol{y}$ are also substrings of some $\boldsymbol{y}_i$. ■

Lemma 20 immediately implies the next corollary.

Corollary 21 $\mathcal{C}_C(n,k) \subseteq \mathcal{RF}_\ell(n,k)$.

Proof: By Lemma 20 and since $\boldsymbol{y} \in \mathcal{RF}_\ell(n')$. ■

We are now ready to analyze the code parameters that Construction C can achieve, again based on Lemma 3 and Theorem 5.

Theorem 22 *Given $\ell(n,k)$, denote $f(n,k) \triangleq \ell(n,k) - \log(nk)$.*

1) If $f(n,k) \geqslant 3\log\log(nk) + 12$ then utilizing Theorem 5 in Construction C we obtain

$$\begin{aligned} R(\mathcal{C}_C(n,k)) \geqslant 1 - \frac{q^{4-\lfloor f(n,k)/3 \rfloor}}{q-2} - \\ - (1 + o(1))\frac{\log(n) + f(n,k)}{n - \log(k)}. \end{aligned}$$

(for $q = 2$, that is $R(\mathcal{C}_C(n,k)) \geqslant 1 - 2^{5-\lfloor f(n,k)/3 \rfloor} - (1 + o(1))\frac{\log(n)+f(n,k)}{n-\log(k)}$).

2) If $f(n,k) \geqslant \log(nk)+2+2\log\left(1 - \left(1 - \frac{1}{k}\right)\frac{\ell(n,k)}{n}\right)$ then utilizing Lemma 3 in Construction C we have

$$R(\mathcal{C}_C(n,k)) \geqslant 1 - (1 + o(1))\frac{\log(n) + f(n,k)}{n - \log(k)}.$$

Proof: Recalling from Construction C that $n' = nk - (k-1)\ell(n,k)$, we begin by observing

$$\begin{aligned} \frac{n'}{nk} &= 1 - \left(1 - \frac{1}{k}\right)\frac{\ell(n,k)}{n} \\ &= 1 - \frac{\log(k)}{n} - (1 + o(1))\frac{\log(n) + f(n,k)}{n}, \end{aligned}$$

hence $\ell(n,k) - \log(n') = \ell(n,k) - \log(nk) + O(1) = f(n,k) + O(1)$. Also, by multiplying the above equality with $\frac{n}{n-\log(k)}$ we have have

$$\frac{n'}{(n - \log(k))k} = 1 - (1 + o(1))\frac{\log(n) + f(n,k)}{n - \log(k)}.$$

Next,

1) As in the proof of Theorem 19,

$$m \geqslant \left(1 - \tfrac{q^4}{q-2}q^{-\lfloor f(n,k)/3 \rfloor}\right)n' - 1,$$

and following the same steps

$$\begin{aligned} R(\mathcal{C}_C(n,k)) &= \frac{m}{\log|\mathcal{X}_{n,k}|} \\ &\geqslant 1 - \frac{q^{4-\lfloor f(n,k)/3 \rfloor}}{q-2} - \\ &\quad - (1 + o(1))\frac{\log(n) + f(n,k)}{n - \log(k)}. \end{aligned}$$

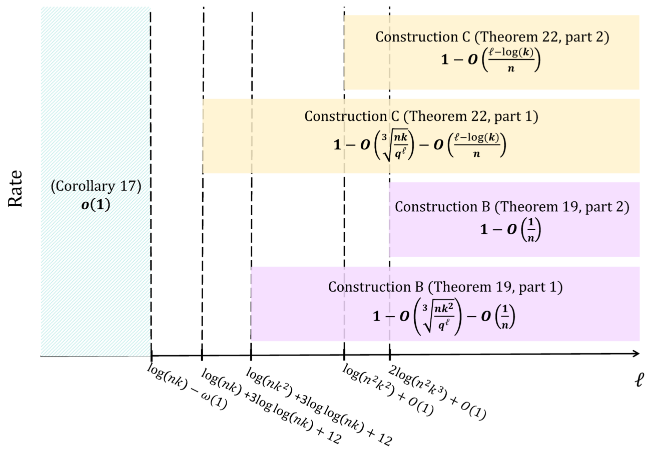

Figure 6. Trade-off of window-length to constructions' rates.

2) Again, we have $\ell(n,k) \geqslant 2\log(n')+2$, hence $m = n'-1$. It follows that

$$R(\mathcal{C}_C(n,k)) = 1-(1+o(1))\frac{\log(n)+f(n,k)}{n-\log(k)}. \qquad \blacksquare$$

We note that inherent to Construction C is that the last step might introduce more redundancy than is required for repeat-free encoding. Indeed, for $f(n,k) \geqslant 3\log(n)$ the latter term in Theorem 22 becomes significant, and the construction's rate is then correspondingly decreasing in $f(n,k)$; this is an oddity since $\mathcal{RF}_{\ell_1}(n,k) \subseteq \mathcal{RF}_{\ell_2}(n,k)$ for all $\ell_1 \leqslant \ell_2$.

C. Constructions' rates

In this section we study the performance of the two proposed constructions. We first seek to give a converse to Corollary 17 and establish the result on the minimum value of ℓ which guarantees that the asymptotic rate of multi-strand ℓ-reconstruction codes (in fact, $\mathcal{RF}_{\ell-1}(n,k)$) is 1. This result is established in the next corollary using Construction C.

Corollary 23 *For n,k satisfying $\limsup \frac{\log(k)}{n} < 1$ and for $\ell \geqslant \log(nk)+3\log\log(nk)+12$, it holds that $R(\mathcal{RF}_\ell(n,k)) = 1-o_{nk}(1)$.*

Note that if one aims to achieve rate $1-o(1)$ using Construction B, then the minimum value of $\ell(n,k)$ should be $\log(nk^2)+3\log\log(nk)+12$, i.e., there exists a gap of $\log(k)$ with respect to the result in Corollary 23. However, for comparable values of $\ell(n,k)$, Construction B may offer better code rate; a comparison of the rates of both constructions, based on Theorems 19 and 22, for applicable values of $\ell(n,k)$ is illustrated in Figure 6, in context of the lower bound of Corollary 17. The following observation follows from these results.

Lemma 24 *$R(\mathcal{C}_B(n)) > R(\mathcal{C}_C(n))$ for sufficiently large n if*

1) $\ell(n,k) \geqslant \log(n^2k^3)+2+2\log\left(1-\frac{\log(k)}{n}\right)$; or

2) if $k=\Omega(n^2)$, for $\ell-\log(n^4k^2)+3\log\log(n^4k) = \omega(1)$; or

3) if $\log(k)=\omega(\sqrt{n})$, for $\ell \geqslant \log(nk^2)+3\log\log(nk)+12$.

Proof: Clearly the claim holds if $\ell(n,k) \geqslant \log(n^2k^3)+2+2\log\left(1-\frac{\log(k)}{n}\right)$ by part 2 of Theorem 19, satisfying part 1.

For lower values of $\ell = \ell(n,k)$, suffice that $\frac{\ell-\log(k)}{n} = \omega\left(\sqrt[3]{\frac{nk^2}{q^\ell}}\right)$. Reorganizing $q^{(\ell-\log(k))/3}(\ell-\log(k))/3 = \omega\left(\sqrt[3]{n^4k}\right)$, we equivalently have $\frac{\ln(q)}{3}(\ell-\log(k)) = W_0\left(\sqrt[3]{n^4k}\right)+\omega(1)$, where $W_0(x) = \ln(x)-\ln\ln(x)+o(1)$ is the principal brunch of the Lambert W function. Hence, a sufficient condition is that

$$\ell-\log(n^4k^2)+3\log\log(n^4k) = \omega(1). \tag{10}$$

For part 2, observe that

$$\begin{aligned}&\left(\log(n^2k^3)+2+2\log\left(1-\frac{\log(k)}{n}\right)\right)\\ &\qquad -\log(n^4k^2)+3\log\log(n^4k)\\ &=\log(k/n^2)+3\log\log(n^4k)+O(1),\end{aligned}$$

hence $k=\Omega(n^2)$ ensures that there exist values of ℓ satisfying both Eq. (10) and $\ell < \log(n^2k^3)+2+2\log\left(1-\frac{\log(k)}{n}\right)$, i.e., not already covered by part 1.

Finally, part 3 is justified by

$$\begin{aligned}&\left(\log(nk^2)+3\log\log(nk)+12\right)\\ &\qquad -\log(n^4k^2)+3\log\log(n^4k)\\ &=3\log\left(\frac{\log(nk)\log(n^4k)}{n}\right)+12,\end{aligned}$$

and the observation that $\log(nk)\log(n^4k)=\omega(n)$ if and only if $\log(k)=\omega(\sqrt{n})$. $\blacksquare$

VII. Conclusion

In this work, we generalized both the reconstruction from substring-composition problem, and the torn-paper problem, by studying an intermediate setting of partial overlap between read substrings. Our analysis is done in worst-case (i.e., adversarial) regime, as opposed to the probabilistic treatment of this problem in [30]. For the case of a single string ($k=1$), we proved an upper bound on achievable code rates (implying in particular a lower bound on the length of read substrings, required for asymptotically non-vanishing codes' rates), and developed an efficient construction asymptotically achieving optimal rate. Pleasingly, at the two extreme points, Construction A essentially degenerates to known constructions for either the torn-paper channel [2] (for $\gamma = 0$) or for reconstruction from substring-composition [11] (for $\gamma \to 1/a$). Finally, we demonstrate that like in the torn-paper extreme, one may also extend solutions to the reconstruction from substring-composition problem to multiset-codes. It is left for future work to extend the intermediate setting in this fashion.

Before concluding, we suggest that one might consider a slightly different channel definition to that of Section VI, where the k strands are required to be distinct from one another, i.e., when information is stored in the space

$$\mathcal{X}^*_{n,k} \triangleq \{S \subseteq \Sigma^n : |S| = k\}.$$

A priori, it seems feasible that the added restriction might allow for lower redundancy (when measured in $\mathcal{X}^*_{n,k}$). However, we note that $\left|\mathcal{X}^*_{n,k}\right| = \binom{q^n}{k}$, thus a similar development to Lemma 16 yields

$$\frac{(q^n-k)^k}{k!} \leqslant \left|\mathcal{X}^*_{n,k}\right| \leqslant \frac{q^{nk}}{k!}.$$

It follows that $\log\left|\mathcal{X}^*_{n,k}\right| = k(n-\log(k/e))+o(k)$ as well. A careful examination reveals that Constructions B and C actually encode into $\mathcal{X}^*_{n,k}\cap\mathcal{RF}_\ell(n,k)$, and hence the results of this work also hold for that setup of the problem.

Acknowledgments

The authors gratefully acknowledge the two anonymous reviewers, and associate editor, whose insight and suggestions helped shape this paper and greatly improve its presentation.

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] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Trans. on Inform. Theory*, vol. 66, no. 10, pp. 6084–6103, Oct. 2020.

[7] C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, and J. Korlach, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nature Methods*, vol. 10, no. 6, pp. 563–569, Jun. 2013.

[8] J. Chrisnata, H. M. Kiah, S. Rao, A. Vardy, E. Yaakobi, and H. Yao, "On the number of distinct k-decks: Enumeration and bounds," in *Proceedings of the 2019 19th International Symposium on Communications and Information Technologies (ISCIT), Ho Chi Minh City, Vietnam, Vietnam*, Sep. 2019, pp. 519–524.

[9] M. Dudık and L. J. Schulman, "Reconstruction from subsequences," *J. Combin. Theory Ser. A*, vol. 103, no. 2, pp. 337–348, Aug. 2003.

[10] M. Eisenstein, "Enzymatic DNA synthesis enters new phase," *Nature Biotechnology*, vol. 38, no. 10, pp. 1113–1115, Oct. 2020.

[11] 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.

[12] H. Fredricksen and J. Maiorana, "Necklaces of beads in k colors and k-ary de Bruijn sequences," *Discrete Mathematics*, vol. 23, no. 3, pp. 207–210, 1978.

[13] 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.

[14] R. Gabrys, S. Pattabiraman, and O. Milenkovic, "Reconstruction of sets of strings from prefix/suffix compositions," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 3–12, Jan. 2023.

[15] S. Ganguly, E. Mossel, and M. Racz, "Sequence assembly from corrupted shotgun reads," in *Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain*, Jul. 2016, pp. 265–269.

[16] 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.

[17] A. R. Khan, M. T. Pervez, M. E. Babar, N. Naveed, and M. Shoaib, "A comprehensive study of de novo genome assemblers: Current challenges and future prospective," *Evolutionary Bioinformatics*, vol. 14, Jan. 2018, PMID: 29511353.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] N. J. Loman, J. Quick, and J. T. Simpson, "A complete bacterial genome assembled de novo using only nanopore sequencing data," *Nature Methods*, vol. 12, no. 8, pp. 733–735, Aug. 2015.

[23] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, and P. Stockmeyer, "Reconstruction of sequences," *Discrete Mathematics*, vol. 94, no. 3, pp. 209–219, 1991.

[24] 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.

[25] 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

[26] A. Motahari, K. Ramchandran, D. Tse, and N. Ma, "Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads," in *Proceedings of the 2013 IEEE International Symposium on Information Theory (ISIT), Istanbul, Turkey*, Jul. 2013, pp. 1640–1644.

[27] 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.

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

[29] 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.

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

[31] S. L. Salzberg, "Mind the gaps," *Nature Methods*, vol. 7, no. 2, pp. 105–106, Feb. 2010.

[32] A. D. Scott, "Reconstruction of sequences," *Discrete Mathematics*, vol. 175, no. 1, pp. 231–238, 1997.

[33] I. Shomorony, T. Courtade, and D. Tse, "Do read errors matter for genome assembly?" in *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China*, Jun. 2015, pp. 919–923.

[34] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," in *Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France*, Jul. 2019, pp. 762–766.

[35] I. Shomorony, G. M. Kamath, F. Xia, T. A. Courtade, and D. N. Tse, "Partial DNA assembly: A rate-distortion perspective," in *Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain*, Jul. 2016, pp. 1799–1803.

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

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

[38] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability bounds," *IEEE Trans. on Inform. Theory*, vol. 68, no. 9, pp. 5657–5700, Sep. 2022.

[39] Y. Yehezkeally and N. Polyanskii, "On codes for the noisy substring channel," in *Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Victoria, Australia*, Jul. 2021, pp. 1700–1705.

Yonatan Yehezkeally (S'12–M'20) is the Carl Friedrich von Siemens post-doctoral research fellow of the Alexander von Humboldt Foundation, in the Associate Professorship of Coding and Cryptography (Prof. Wachter-Zeh), School of Computation, Information and Technology, Technical University of Munich. His research interests include coding for novel storage media, with a focus on DNA-based storage and nascent sequencing technologies, as well as combinatorial structures and finite group theory.

Yonatan received the B.Sc. (*cum laude*) degree in Mathematics, and the M.Sc. (*summa cum laude*) and Ph.D. degrees in Electrical and Computer Engineering, in 2013, 2017 and 2020 respectively, all from Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Daniella Bar-Lev (S'20) is a Ph.D. student in the Computer Science Department at the Technion — Israel Institute of Technology. She received the B.Sc. degrees in computer science and mathematics, and the M.Sc. degree in computer science from the Technion — Israel Institute of Technology, Haifa, Israel, in 2019 and 2021, respectively. Her research interests include algorithms, discrete mathematics, coding theory, and DNA storage.

Sagi Marcovich (S'20) is a Ph.D. student in the Computer Science Department at the Technion — Israel Institute of Technology. He received the B.Sc. degree in software engineering and his M.Sc. degree in computer science from the Technion — Israel Institute of Technology, Haifa, Israel in 2016 and 2021, respectively. His research interests include algorithms, information theory, and coding theory with applications to DNA based storage.

Eitan Yaakobi (S'07–M'12–SM'17) is an Associate Professor at the Computer Science Department at the Technion — Israel Institute of Technology. He also holds a courtesy appointment in the Technion's Electrical and Computer Engineering (ECE) Department. He received the B.A. degrees in computer science and mathematics, and the M.Sc. degree in computer science from the Technion — Israel Institute of Technology, Haifa, Israel, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California, San Diego, in 2011. Between 2011-2013, he was a postdoctoral researcher in the department of Electrical Engineering at the California Institute of Technology and at the Center for Memory and Recording Research at the University of California, San Diego. His research interests include information and coding theory with applications to non-volatile memories, associative memories, DNA storage, data storage and retrieval, and private information retrieval. He received the Marconi Society Young Scholar in 2009 and the Intel Ph.D. Fellowship in 2010-2011. Since 2020, he serves as an Associate Editor for Coding snd Decoding for the IEEE TRANSACTIONS ON INFORMATION THEORY. Since 2016, he is affiliated with the Center for Memory and Recording Research at the University of California, San Diego, and since 2018, he is affiliated with the Institute of Advanced Studies, Technical University of Munich, where he holds a four-year Hans Fischer Fellowship, funded by the German Excellence Initiative and the EU 7th Framework Program. He is a recipient os several grants, including the ERC Consolidator Grant.