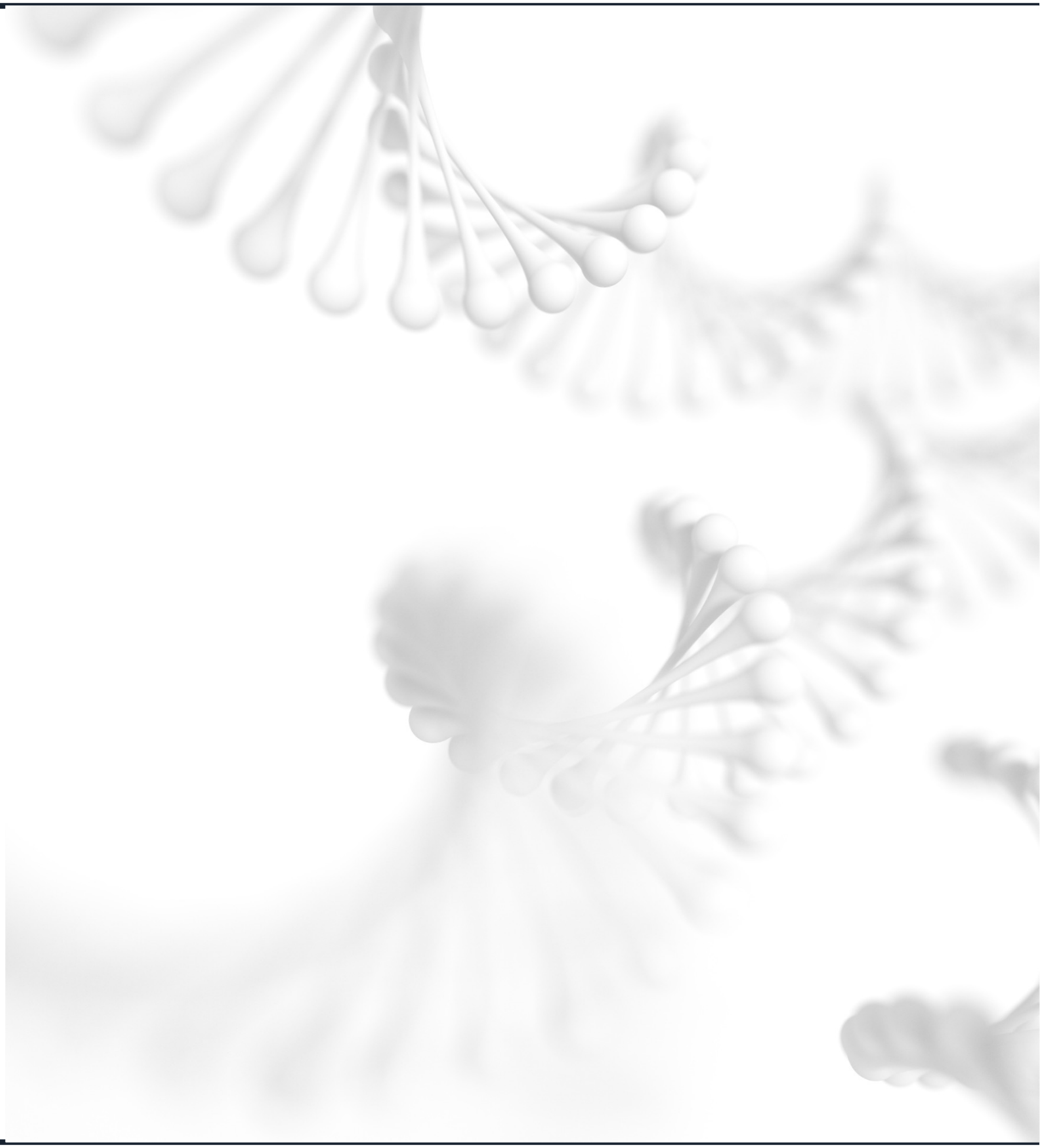
UNIQUE RECONSTRUCTION FROM SUBSTRINGS

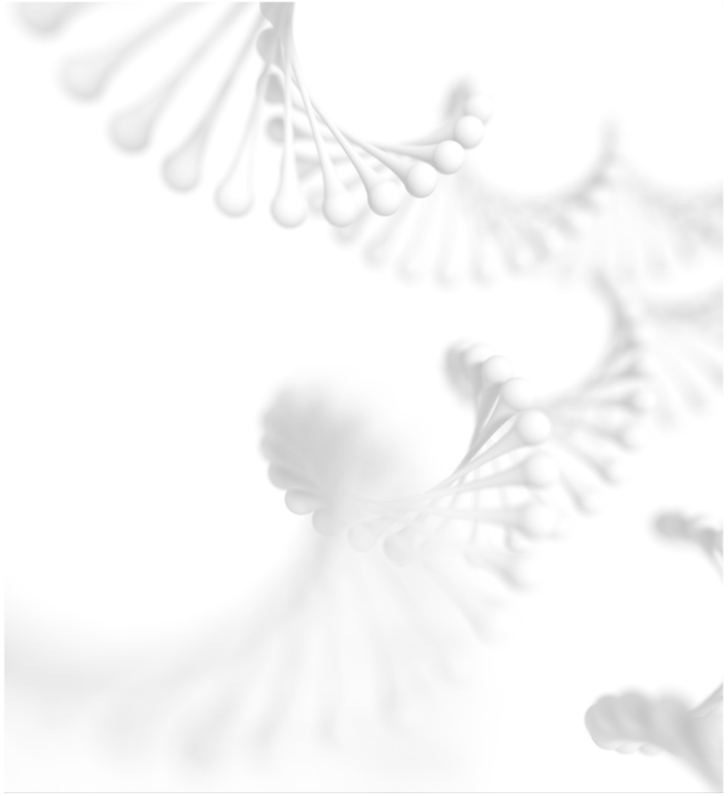
Daniella Bar-Lev | Technion-Israel Institute of Technology

A joint work with:

Yonatan Yehezkeally | Technical University of Munich
Sagi Marcovich | Technion-Israel Institute of Technology
Eitan Yaakobi | Technion-Israel Institute of Technology



DNA Storage



String Reconstruction

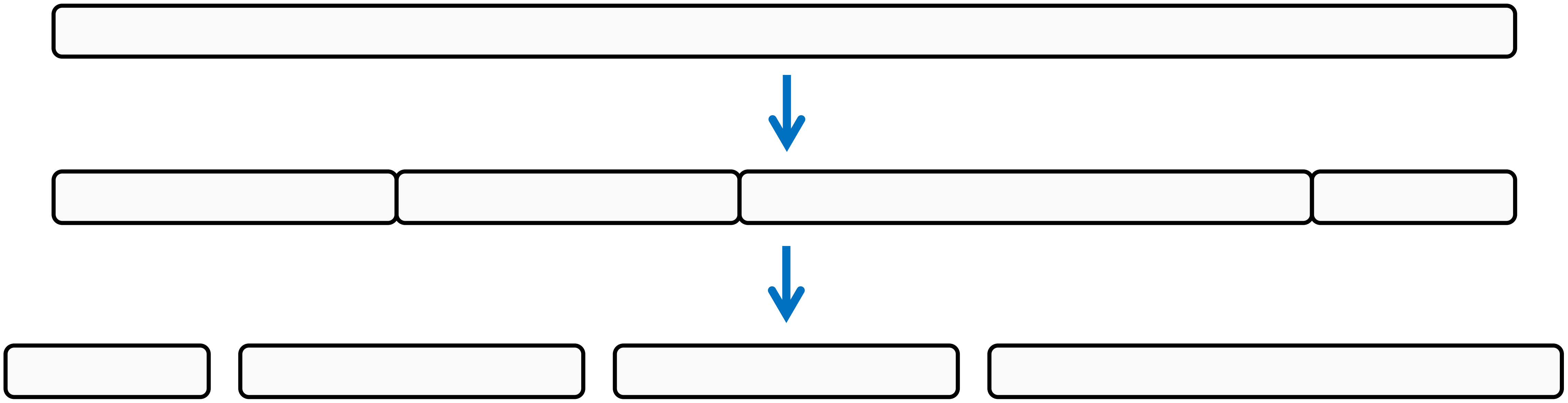
Reconstruct string from ***multiple***, ***incomplete*** *and/or* ***noisy*** *observations*

Examples:

- Levenshtein's *reconstruction problem* [1]
- *Trace reconstruction problem* [2]
- *k-deck problem* [3]

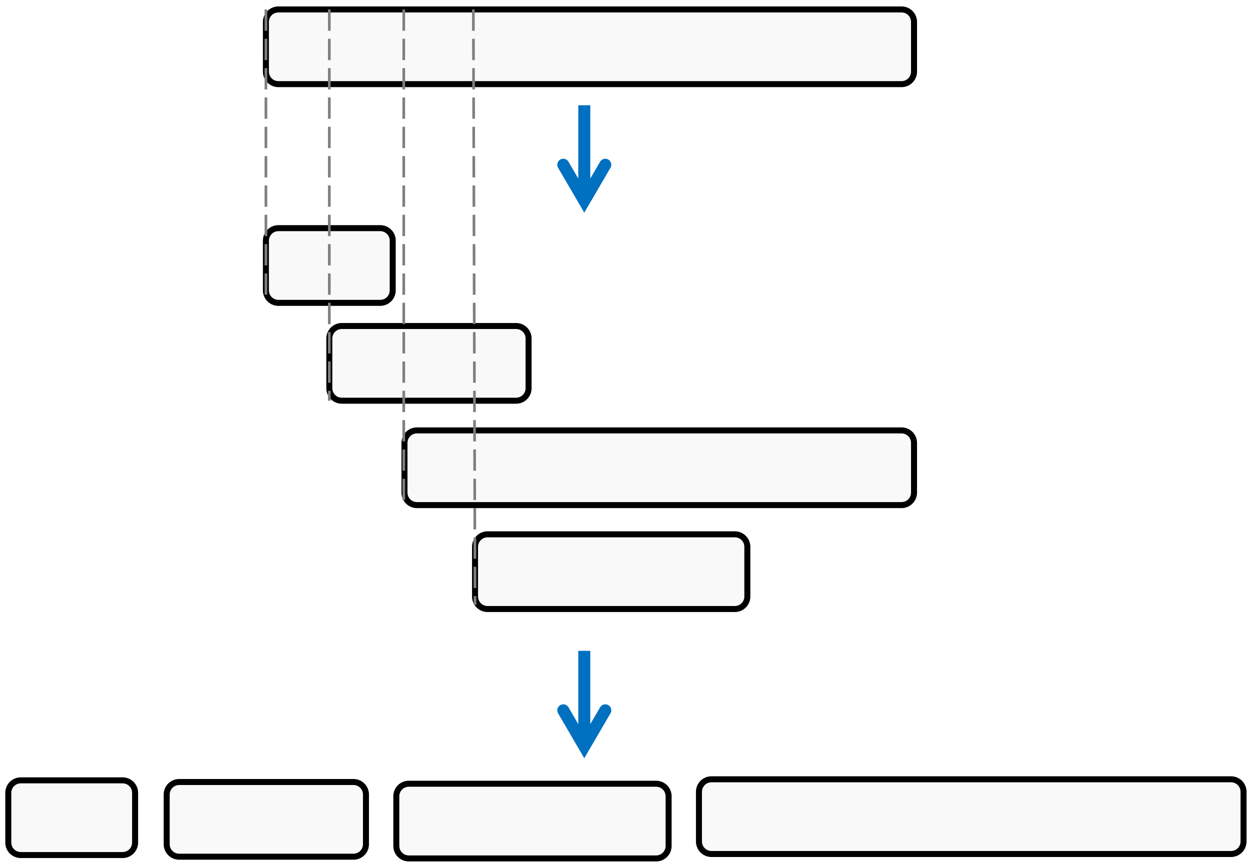
[1] V. I. Levenshtein, "Efficient reconstruction of sequences from subsequences or supersequences," J. Combin. Theory, Feb. 2001
[2] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in SODA, Society for Industrial and Applied Mathematics, Jan. 2004,.
[3] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, and P. Stockmeyer, "Reconstruction of sequences," Discrete Mathematics, 1991.

Torn-Paper Reconstruction



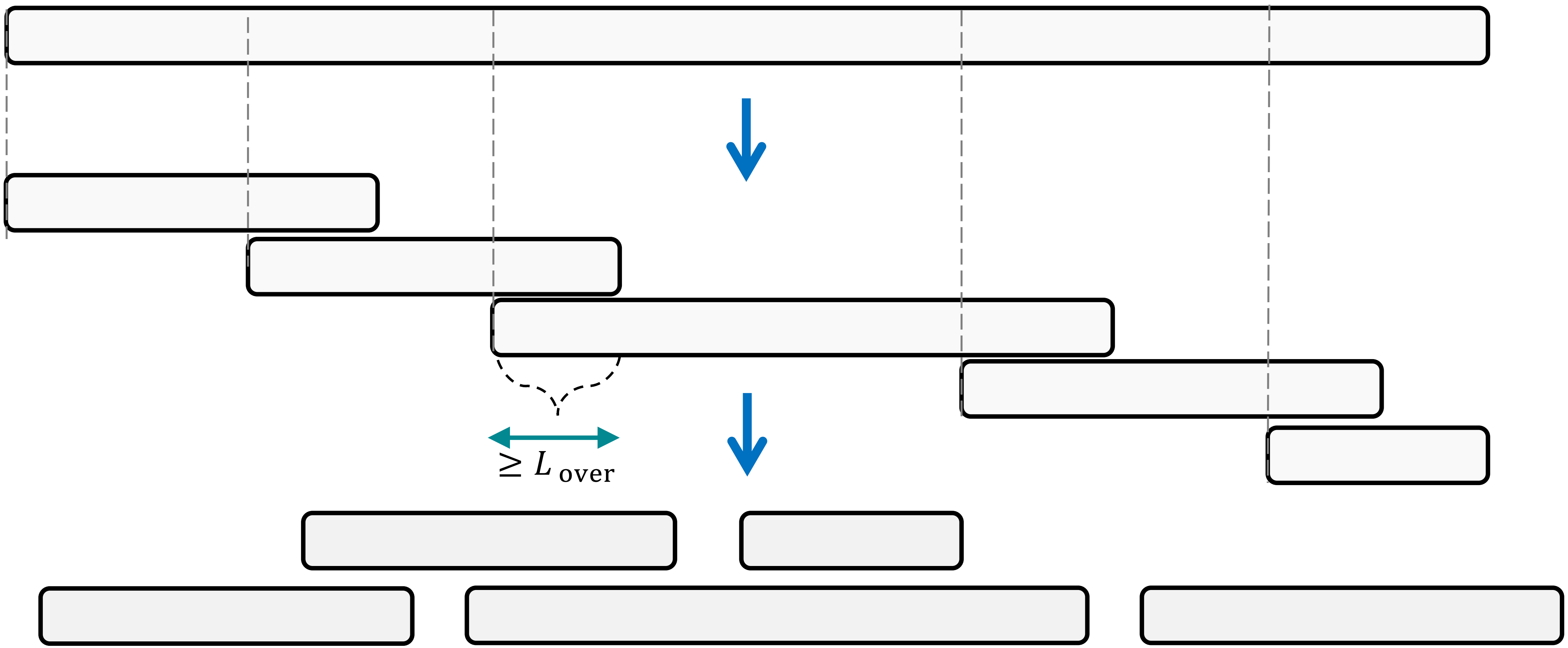
D. Bar-Lev, S. Marcovich, E. Yaakobi, and Y. Yehezkeally, “Adversarial torn-paper codes,” in *IEEE ISIT*, 2022.
S. Nassirpour, I. Shomorony, and A. Vahid, “Reassembly codes for the chop-and-shuffle channel,” in arXiv, 2022.
A. N. Ravi, A. Vahid, and I. Shomorony, “Capacity of the torn paper channel with lost pieces,” in *IEEE ISIT*, 2021.
I. Shomorony and A. Vahid, “Torn-paper coding,” *IEEE TIT*, Dec. 2021.

Reconstruction from Substring-Composition



R. Gabrys and O. Milenkovic, “Unique reconstruction of coded sequences from multiset substring spectra,” *IEEE TIT*, Jun. 2019.
G. Bresler, M. Bresler, and D. Tse, “Optimal assembly for high throughput shotgun sequencing,” *BMC Bioinformatics*, Jul. 2013.
H. M. Kiah, G. J. Puleo, and O. Milenkovic, “Codes for DNA sequence profiles,” *IEEE TIT*, Jun. 2016.
S. Marcovich and E. Yaakobi, “Reconstruction of strings from their substrings spectrum,” *IEEE TIT*, Jul. 2021.
A. S. Motahari, G. Bresler, and D. N. C. Tse, “Information theory of DNA shotgun sequencing,” *IEEE TIT*, Oct. 2013.
Y. Yehezkeally, S. Marcovich, and E. Yaakobi, “Multi-strand reconstruction from substrings,” in *IEEE ITW*, 2021

Partial-Overlap Channel



Probabelistic version of this channel:

Ravi, A. N., Vahid, A., and Shomorony, I. (2022). Coded Shotgun Sequencing. *ACM JSAIT*, 3(1), 147-159.

Agenda

01

Torn-Paper Channel

02

Substring-Composition Channel

03

Partial Overlap Channel

04

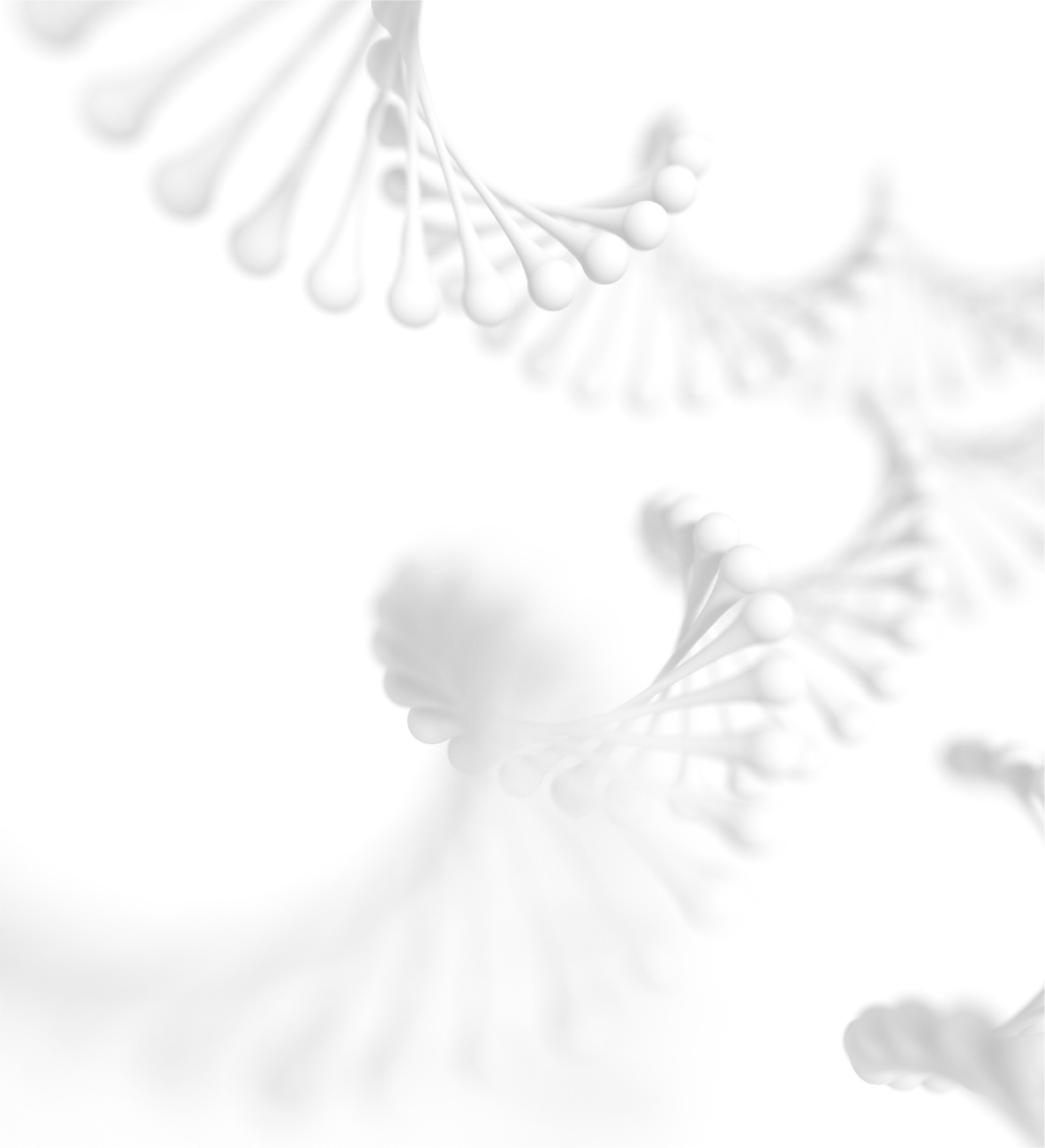
Future Directions

Notations

$[n] = \{0,1,\ldots,n-1\}$.

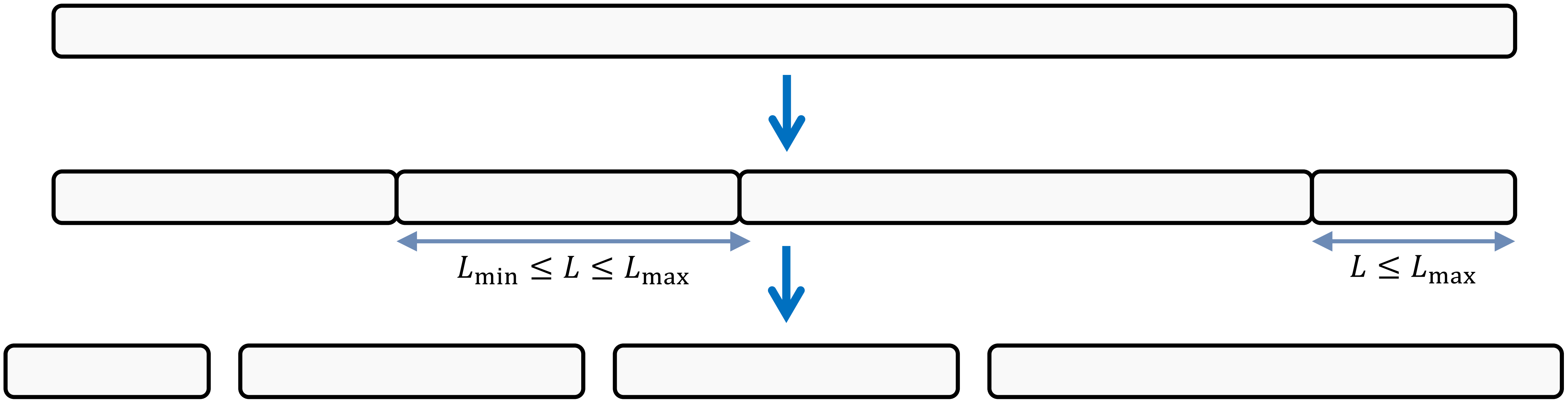
$\boldsymbol{x} \circ \boldsymbol{y}$: concatenation of $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^*$.

ℓ-substring of $\boldsymbol{x}$: substring of length ℓ.



Torn-Paper Channel
Single-Strand

Single-Strand Torn-Paper Reconstruction



D. Bar-Lev, S. Marcovich, E. Yaakobi, and Y. Yehezkeally, “Adversarial torn-paper codes,” in *IEEE ISIT*, 2022.
S. Nassirpour, I. Shomorony, and A. Vahid, “Reassembly codes for the chop-and-shuffle channel,” in arXiv, 2022.
A. N. Ravi, A. Vahid, and I. Shomorony, “Capacity of the torn paper channel with lost pieces,” in *IEEE ISIT*, 2021.
I. Shomorony and A. Vahid, “Torn-paper coding,” *IEEE TIT*, Dec. 2021.

Definitions

An *$(L_{\min}, L_{\max})$-segmentation* of $\boldsymbol{x} \in \Sigma^n$ is a multiset $\{\{\boldsymbol{u}_0, \boldsymbol{u}_1, \dots, \boldsymbol{u}_{m-1}\}\}$ of substrings such that:

(1) $\boldsymbol{x} = \boldsymbol{u}_0 \circ \boldsymbol{u}_1 \circ \cdots \circ \boldsymbol{u}_{m-1}$.

(2) $L_{min} \le |\boldsymbol{u}_i| \le L_{max}$ for $0 \le i < m-1$ and $|\boldsymbol{u}_{m-1}| \le L_{max}$

$\mathcal{T}_{L_{\min}}^{L_{\max}}(\boldsymbol{x})$: the multiset of all $(L_{\min}, L_{\max})$-segmentations of $\boldsymbol{x}$.

Example: $\boldsymbol{x}$=0001011001

$\{\{0001, 01, 10, 01\}\}$ is a (2,4)-segmentation of $\boldsymbol{x}$.

Definitions

An *$(L_{\min}, L_{\max})$-segmentation* of $\boldsymbol{x} \in \Sigma^n$ is a multiset $\{\{\boldsymbol{u}_0, \boldsymbol{u}_1, \dots, \boldsymbol{u}_{m-1}\}\}$ of substrings such that:

(1) $\boldsymbol{x} = \boldsymbol{u}_0 \circ \boldsymbol{u}_1 \circ \cdots \circ \boldsymbol{u}_{m-1}$.

(2) $L_{min} \leq |\boldsymbol{u}_i| \leq L_{max}$ for $0 \leq i < m-1$ and $|\boldsymbol{u}_{m-1}| \leq L_{max}$

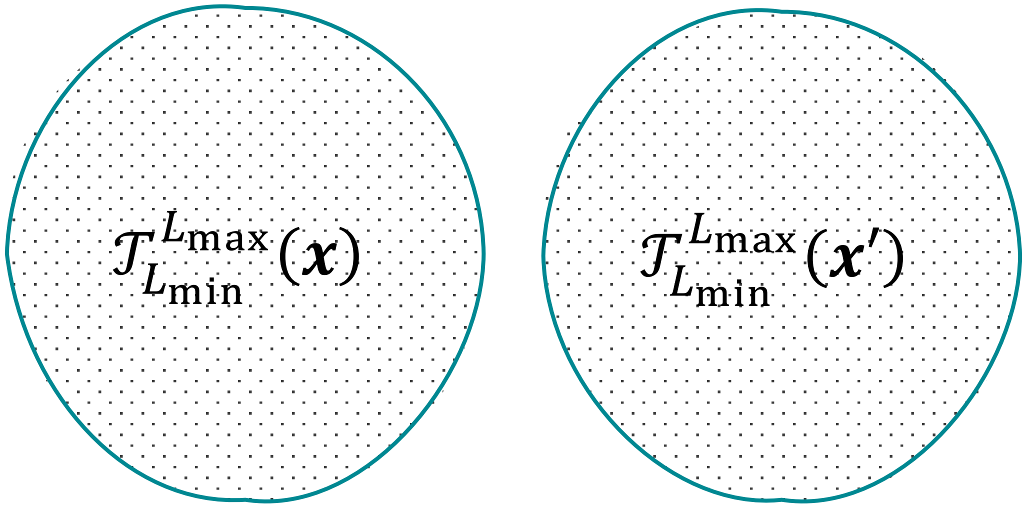
$\mathcal{T}_{L_{\min}}^{L_{\max}}(\boldsymbol{x})$: the multiset of all $(L_{\min}, L_{\max})$-segmentations of $\boldsymbol{x}$.

Channel input: $\boldsymbol{x} \in \Sigma^n$

Channel output: an $(L_{\min}, L_{\max})$-segmentation of $\boldsymbol{x}$.

Definitions

$(L_{\text{min}}, L_{\text{over}})$- single strand torn-paper code:
code $\mathcal{C} \subseteq \Sigma^n$ such that for any two distinct strings $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}$, it holds that $\mathcal{T}_{L_{\text{min}}}^{L_{\text{max}}}(\boldsymbol{x}) \cap \mathcal{T}_{L_{\text{min}}}^{L_{\text{max}}}(\boldsymbol{x}') = \emptyset$.



Single-Strand Torn-Paper Codes: Rate

Theorem: If $L_{\min} = a \log n + O_n(1)$, for some value $a > 1$, then for any $(L_{\min}, L_{\max})$-single-strand torn-paper code $\mathcal{C} \subseteq \Sigma^n$,

$$\mathrm{R}(\mathcal{C}) \leq 1 - \frac{1}{a} + o(1).$$

Preliminaries: Gray Code

A *Gray Code* is an ordering of Σ^n such that any two adjacent strings differ in only one symbol position.

Example:

$\Sigma = \{0,1\}$ and $n = 3$

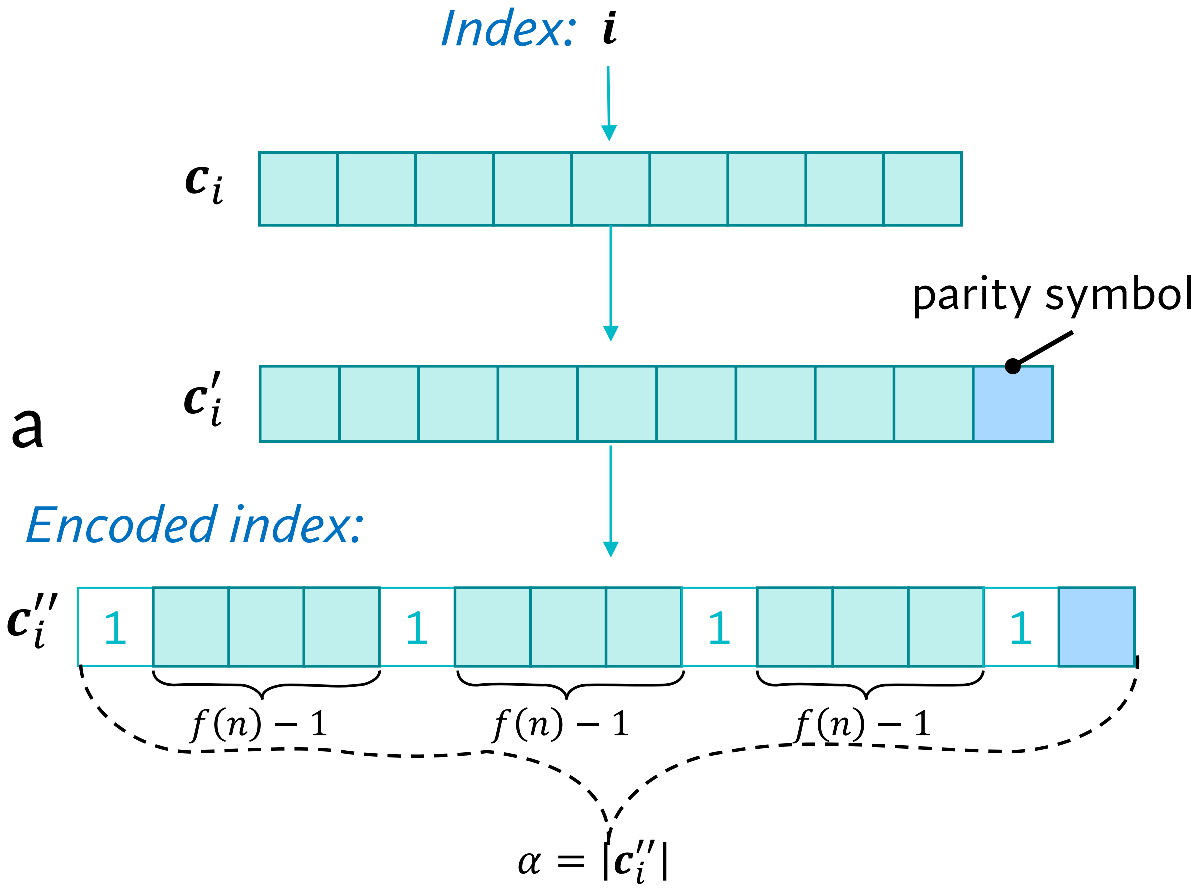
$|\Sigma^3| = 8$

0	000	4	110
1	001	5	100
2	011	6	101
3	010	7	111

Preliminaries: Index Generation

Consider:

- Index length: $I = \lceil \log_q(n/L_{\min}) \rceil$
- Block size: $f(n) = o(\log n)$
- $\boldsymbol{c}_0, \boldsymbol{c}_1, \dots, \boldsymbol{c}_{q^I-1}$: the codewords of a q-ary Gray code, in order.



Preliminaries: Run-Length Limited Encoding

For $t < N$ the set of *run-length limited (RLL)* strings is

$$\mathcal{RLL}_t(N) \triangleq \{\boldsymbol{x} \in \Sigma^N: \quad \textit{no } t\textit{-length runs of zeros}\}.$$

RLL encoder: input length m, output length $N = N_n(m)$, and $t = f(n)$

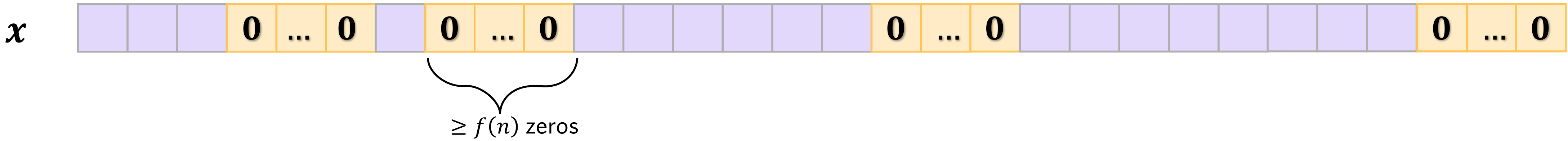
$$E_m^{RLL}: \Sigma^m \to \mathcal{RLL}_{f(n)}(N_n(m)).$$

[1] 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.
[2, Lem. 5] Y. Yehezkeally, S. Marcovich, and E. Yaakobi, "Multi-strand reconstruction from substrings," in *IEEE ITW*, 2021.

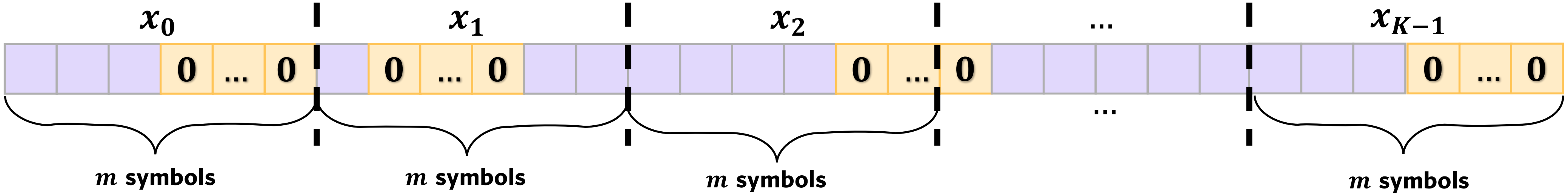
Construction A

Encoder for $(L_{\min}, L_{\max})$-single strand Torn-Paper code $\mathcal{C}_A(n)$:

Input: a sequence $\boldsymbol{x} \in \Sigma^{Km}$



Devide $\boldsymbol{x}$ into q^I non-overlapping substrings of length m.

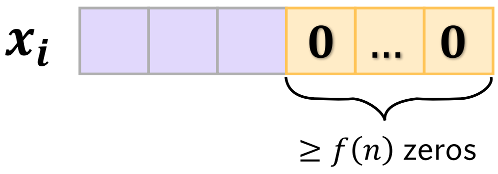


Next, we encode each substring $\boldsymbol{x_i}$ into $\boldsymbol{z_i} \in \Sigma^{L_{\min}}$ independently.

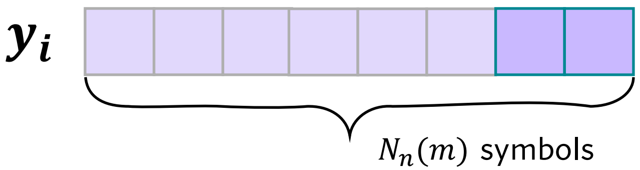
Construction A

Encoder for $(L_{\min}, L_{\max})$-single strand Torn-Paper code $\mathcal{C}_A(n)$:

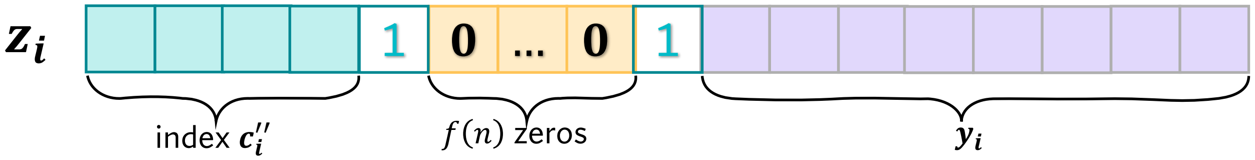
For any $0 \le i \le K-1$:



(1) Encode $\boldsymbol{x_i}$ using the RLL encoder to obtain $\boldsymbol{y_i} = E_m^{RLL}(\boldsymbol{x_i})$.



(2) Let $\boldsymbol{c}_i''$ be the i-th encoded index and let $\boldsymbol{z}_{\mathrm{i}} = \boldsymbol{c}_{\mathrm{i}}'' \circ \mathbf{10}^{f(n)}\mathbf{1} \circ \boldsymbol{y_i}$.

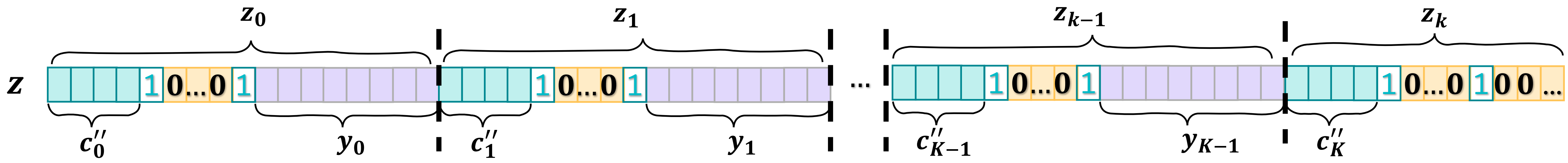


$|\boldsymbol{z_i}| = L_{\min}$.

Construction A

Encoder for $(L_{\min}, L_{\max})$-single strand Torn-Paper code $\mathcal{C}_A(n)$:

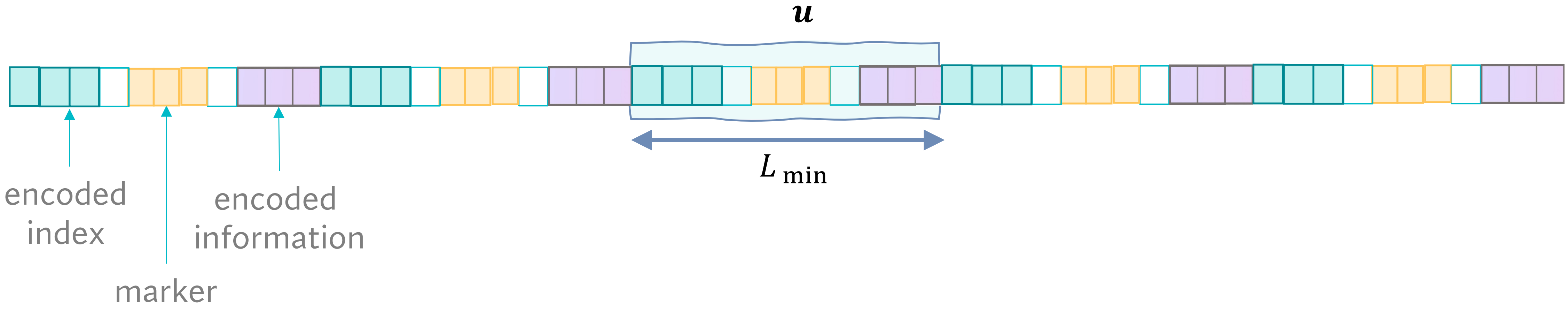
Finally, let $\mathrm{Enc}_A(\boldsymbol{x}) \triangleq \boldsymbol{z} = \boldsymbol{z}_0 \circ \boldsymbol{z}_1 \circ \cdots \circ \boldsymbol{z}_{K-1} \circ \boldsymbol{z}_K$



Construction A – Encoder Correctness

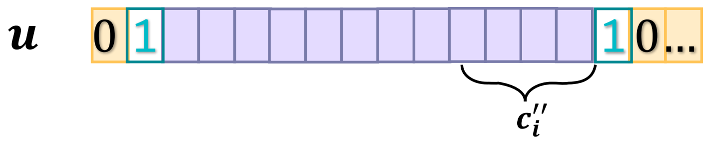
Every $L_{\min}$-substring $\boldsymbol{u}$ of $\boldsymbol{z}$:

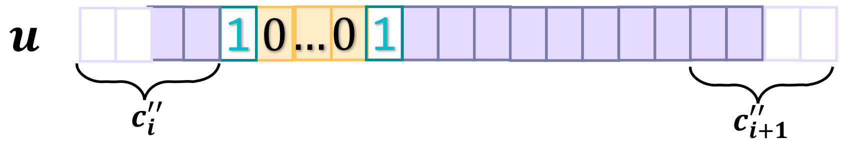
- does not contain any occurrences of the marker $\boxed{1}\boxed{0}\boxed{0}\boxed{0}\boxed{1}$ except those explicitly added after each encoded index.
- either $\boldsymbol{u}$ contains an occurrence of the marker or it has a suffix-prefix pair whose concatenation is the marker.



Construction A – Encoder Correctness

Either $\boldsymbol{u}$ contains an occurrence of $\boxed{1}\,0\,0\,\boxed{}\,\boxed{1}$ or it has a suffix-prefix pair whose concatenation is $\boxed{1}\,0\,0\,\boxed{}\,\boxed{1}$.





$\boldsymbol{c}_i''$ and $\boldsymbol{c}_{i+1}''$ differ only at the parity symbol and one more coordinate.

$\boldsymbol{c}''$: the concatenation of the $(\alpha - j)$-suffix of $\boldsymbol{u}$ with the j-prefix of $\boldsymbol{u}$.

Case 1:

$\boldsymbol{c}_i''$	0	1	0	1
$\boldsymbol{c}_{i+1}''$	1	1	0	0
$\boldsymbol{c}''$	1	1	0	1

$\boldsymbol{c}''$ is a copy of $\boldsymbol{c}_{i+1}''$ with an erroneous parity symbol.

Case 2:

$\boldsymbol{c}_i''$	1	1	1	1
$\boldsymbol{c}_{i+1}''$	1	1	0	0
$\boldsymbol{c}''$	1	1	1	1

$\boldsymbol{c}''$ is equal to $\boldsymbol{c}_i''$.

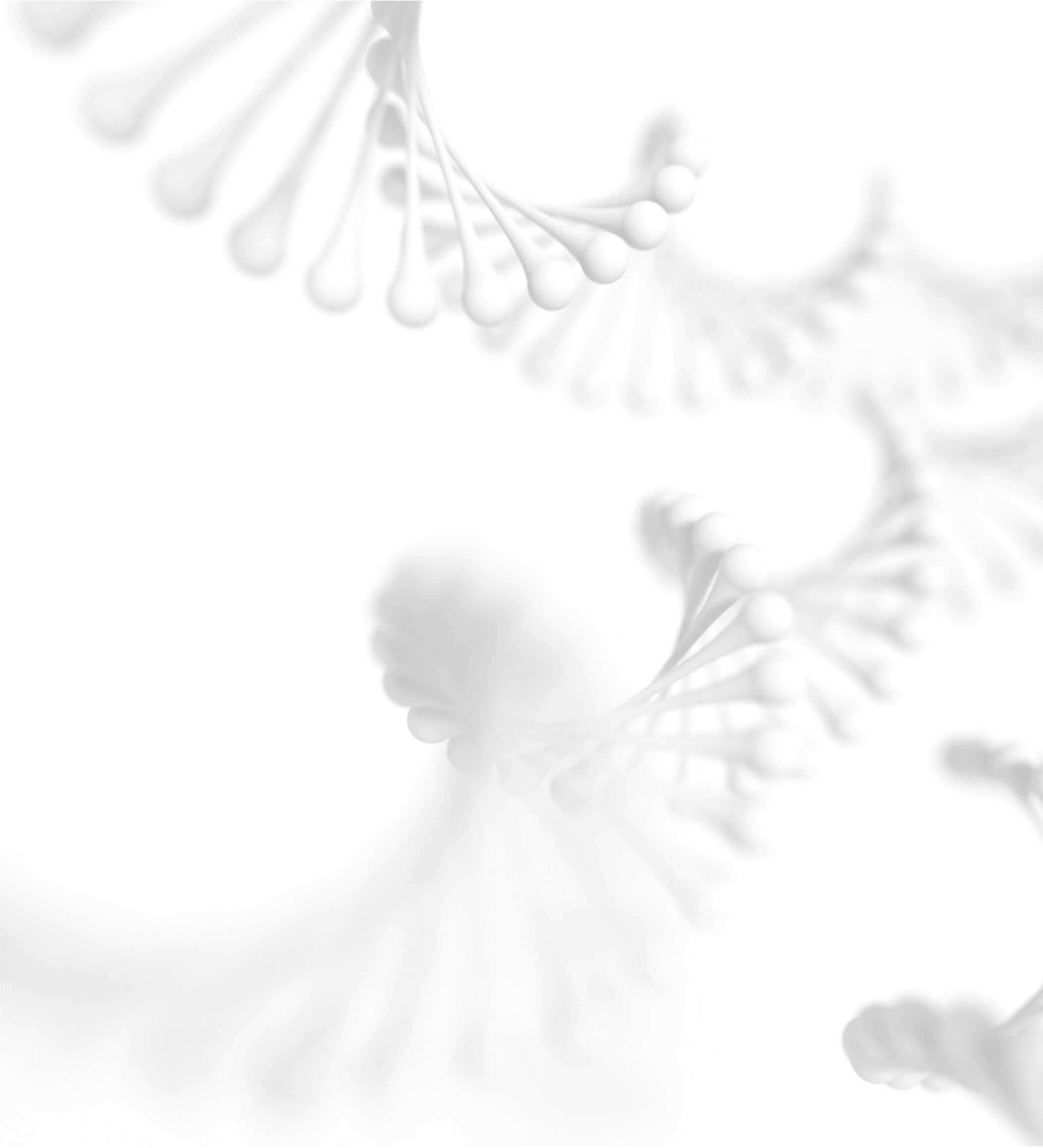
Hence, if the parity symbol is correct then i is the decoding of $\boldsymbol{c}''$ and otherwise i is the decoding of $\boldsymbol{c}''$ minus one.

Construction A – Rate

Theorem: Letting $f(n) \triangleq (1 + o(1))\sqrt{\log(n)}$ we have that

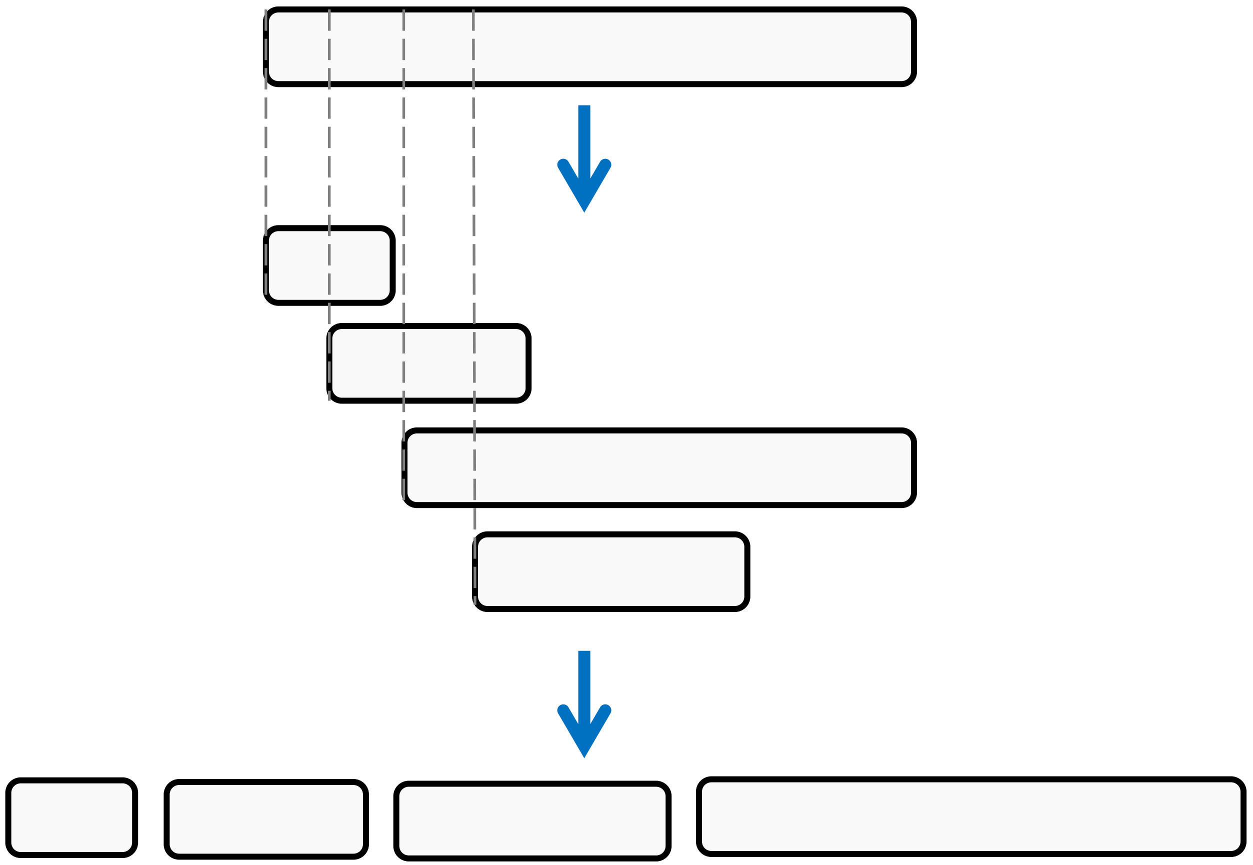
$$\mathrm{red}\big(\mathcal{C}_A(n)\big) \leq \frac{n}{a}\left(1 + \frac{2 + o(1)}{\sqrt{\log(n)}}\right).$$

Hence, Construction A asymptotically meets the upper bound.



Multi-Strand Reconstruction from Substring-Composition

Reconstruction from Substring-Composition



R. Gabrys and O. Milenkovic, “Unique reconstruction of coded sequences from multiset substring spectra,” *IEEE TIT*, Jun. 2019.
G. Bresler, M. Bresler, and D. Tse, “Optimal assembly for high throughput shotgun sequencing,” *BMC Bioinformatics*, Jul. 2013.
H. M. Kiah, G. J. Puleo, and O. Milenkovic, “Codes for DNA sequence profiles,” *IEEE TIT*, Jun. 2016.
S. Marcovich and E. Yaakobi, “Reconstruction of strings from their substrings spectrum,” *IEEE TIT*, Jul. 2021.
A. S. Motahari, G. Bresler, and D. N. C. Tse, “Information theory of DNA shotgun sequencing,” *IEEE TIT*, Oct. 2013.
Y. Yehezkeally, S. Marcovich, and E. Yaakobi, “Multi-strand reconstruction from substrings,” in *IEEE ITW*, 2021

Notations

$$\mathcal{X}_{n,k} \triangleq \left\{ \boldsymbol{S} = \{\{\boldsymbol{x_1}, \dots, \boldsymbol{x_k}\}\} : \quad \forall i, \boldsymbol{x_i} \in \Sigma^n \right\}.$$

$||S||$: number of unique elements in S.

Definitions

An *ℓ-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset of substrings such that:
(1) All substrings are of length at least ℓ.
(2) Succeeding substrings overlap is at least $\ell - 1$.
(3) $\boldsymbol{x}$ is covered by the substrings.

ℓ-trace spectrum of $\boldsymbol{x}$, denoted by $\mathcal{T}_\ell(\boldsymbol{x})$: set of all ℓ-traces of $\boldsymbol{x}$.

Example: $\boldsymbol{x}$=1110111

$\{\{11101,1101,10111,0111\}\}$ is a 4-trace of $\boldsymbol{x}$.

Definitions

An *ℓ-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset of substrings such that:
(1) All substrings are of length at least ℓ.
(2) Succeeding substrings overlap is at least $\ell - 1$.
(3) $\boldsymbol{x}$ is covered by the substrings.

ℓ-trace spectrum of $\boldsymbol{x}$, denoted by $\mathcal{T}_\ell(\boldsymbol{x})$: set of all ℓ-traces of $\boldsymbol{x}$.

Example: $\boldsymbol{x}=1110111$

$\{\{11101, 1101, 10111, 0111\}\}$ is a 4-trace of $\boldsymbol{x}$.

$\mathcal{T}_\ell(\boldsymbol{S}) \triangleq \bigcup_{\boldsymbol{x}\in\boldsymbol{S}} \mathcal{T}_\ell(\boldsymbol{x})$.

Definitions

An *ℓ-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset of substrings such that:
(1) All substrings are of length at least ℓ.
(2) Succeeding substrings overlap is at least $\ell - 1$.
(3) $\boldsymbol{x}$ is covered by the substrings.

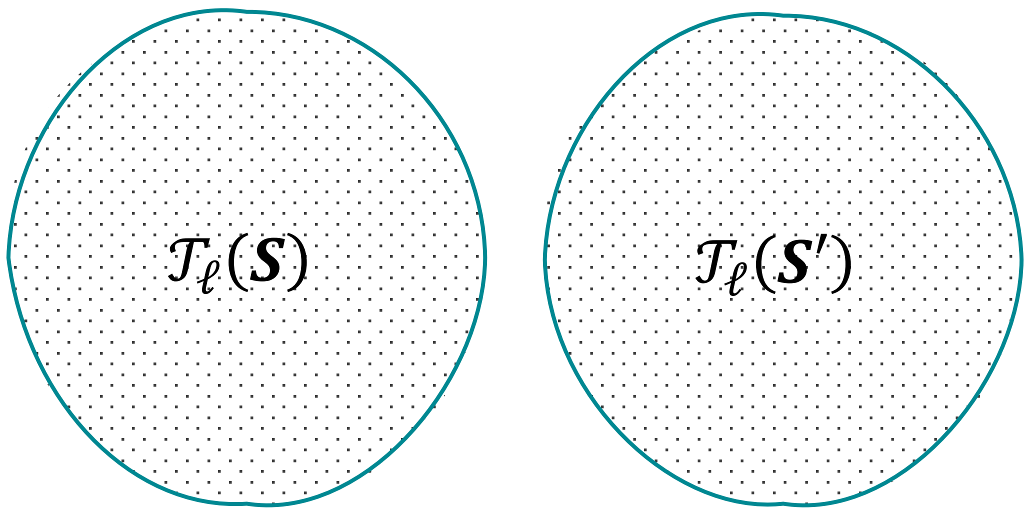
ℓ-trace spectrum of $\boldsymbol{x}$, denoted by $\mathcal{T}_\ell(\boldsymbol{x})$: set of all ℓ-traces of $\boldsymbol{x}$.

Channel input: $\boldsymbol{S} \in \mathcal{X}_{n,k}$
Channel output: an ℓ-trace of $\boldsymbol{S}$.

$\mathcal{T}_\ell(\boldsymbol{S}) \triangleq \cup_{\boldsymbol{x} \in \boldsymbol{S}} \mathcal{T}_\ell(\boldsymbol{x})$.

Definitions

Multi-strand ℓ-trace code: code $\mathcal{C} \subseteq \mathcal{X}_{n,k}$ such that for any two distinct multisets $\boldsymbol{S}, \boldsymbol{S}' \in \mathcal{C}$, it holds that $\mathcal{T}_\ell(\boldsymbol{S}) \cap \mathcal{T}_\ell(\boldsymbol{S}') = \emptyset$.



Preliminaries: Repeat-Free Encoding

$\mathcal{L}_\ell(\boldsymbol{x}) \in \mathcal{T}_\ell(\boldsymbol{x})$ is the ℓ-trace that consists of all the ℓ-substrings of $\boldsymbol{x}$.

Example: $\boldsymbol{x} = 11101110101111$

$$\mathcal{L}_{12}(\boldsymbol{x}) = \{\{111011101011,\ 110111010111,\ 101110101111\}\}$$

For $\ell < n$ the set of *ℓ-repeat-free (RF)* strings is

$$\begin{aligned}\mathcal{RF}_\ell(n) &\triangleq \{\boldsymbol{x} \in \Sigma^n:\ \ no\ \ell - substring\ repeats\} \\ &= \left\{\boldsymbol{x} \in \Sigma^n:\ \left|\left|\mathcal{L}_\ell(\boldsymbol{x})\right|\right| = n - \ell + 1\right\}.\end{aligned}$$

A *multi-strand ℓ-repeat-free* strings:

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

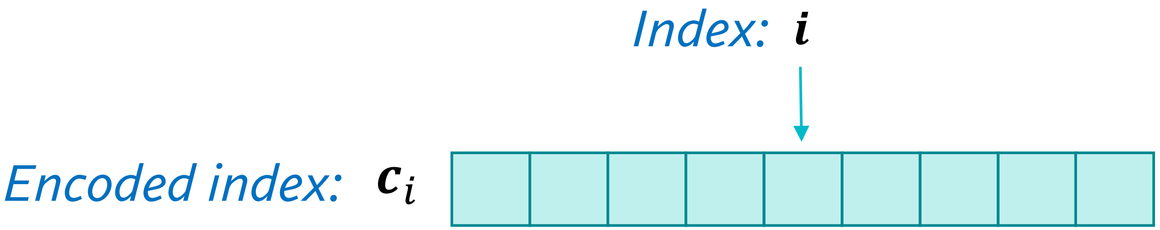
Preliminaries: Repeat-Free Encoding

Lemma: For all $\boldsymbol{S} \in \mathcal{RF}_\ell(n,k)$ there exists an efficient algorithm reconstructing $\boldsymbol{S}$ from any ℓ-trace of $\boldsymbol{S}$.

Preliminaries: Index Generation

Consider:

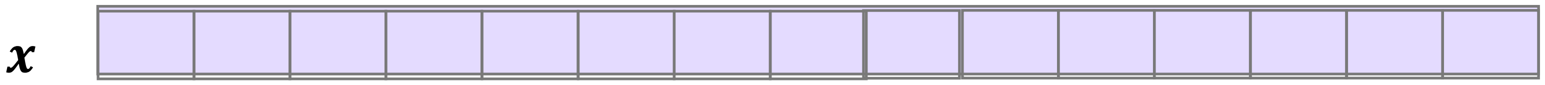
- Index length: $I = \lceil \log_q(k) \rceil$



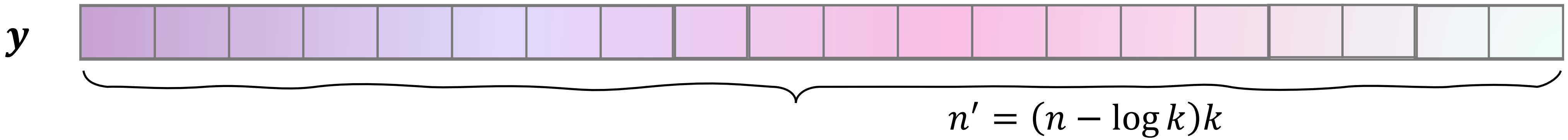
Construction B

Encoder for multi-strand ℓ-trace code $\mathcal{C}_B$:

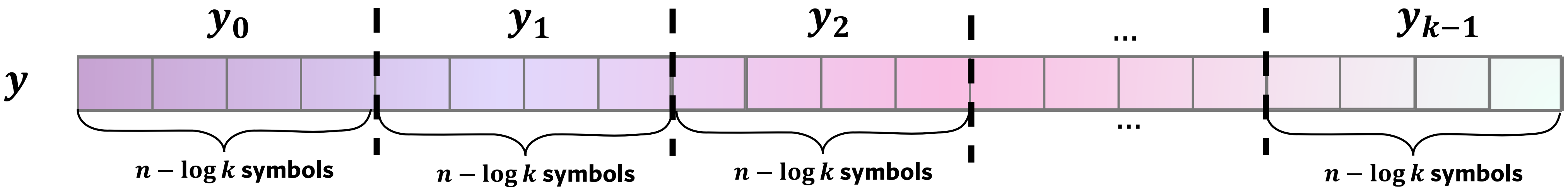
Input: a sequence $\boldsymbol{x} \in \Sigma^m$



Encode $\boldsymbol{x}$ using the RF encoder for $\ell' = \ell - \log k$ to obtain $\boldsymbol{y} = E^{\mathcal{RF}}_{m,\ell'}(\boldsymbol{x})$.

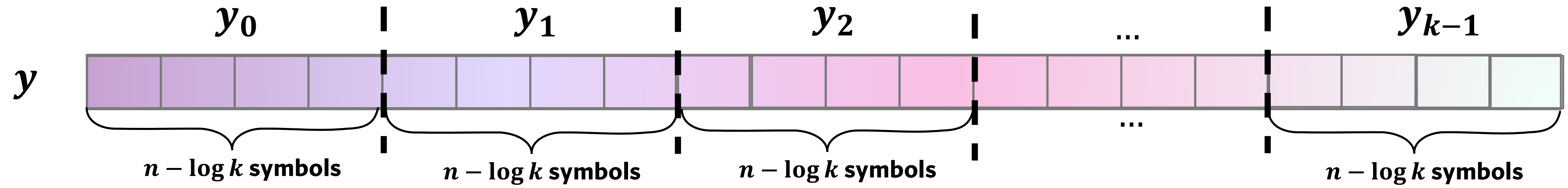


Divide $\boldsymbol{y}$ into k non-overlapping substrings of length $n - \log k$.

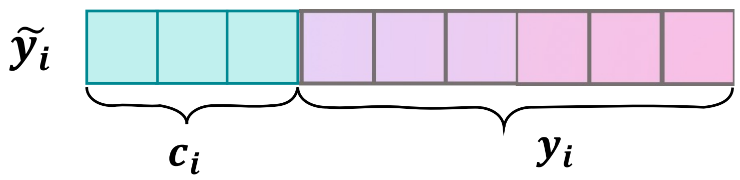


Construction B

Encoder for multi-strand ℓ-trace code $\mathcal{C}_B$:



For each $\boldsymbol{y}_i$ define $\tilde{\boldsymbol{y}}_i = \boldsymbol{c}_i \circ \boldsymbol{y}_i$

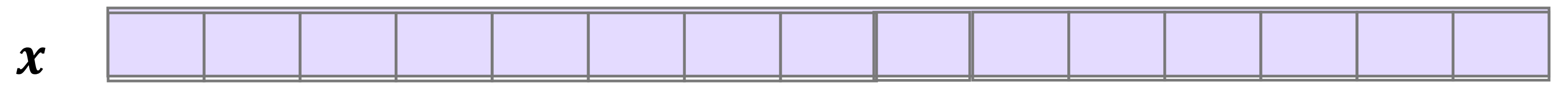


Finally, let $\mathrm{Enc}_B(\boldsymbol{x}) \triangleq \{\{\tilde{\boldsymbol{y}}_i : i \in [k]\}\} \in \mathcal{X}_{n,k}$.

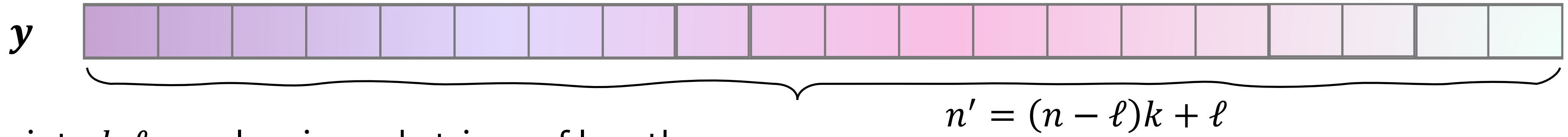
Construction C

Encoder for multi-strand ℓ-trace code $\mathcal{C}_C$:

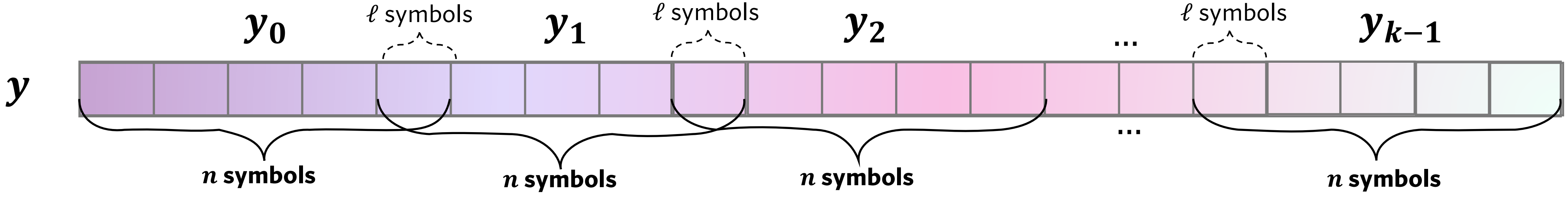
Input: a sequence $\boldsymbol{x} \in \Sigma^m$



Encode $\boldsymbol{x}$ using the RF encoder to obtain $\boldsymbol{y} = E^{\mathcal{RF}}_{m,\ell}(\boldsymbol{x})$.

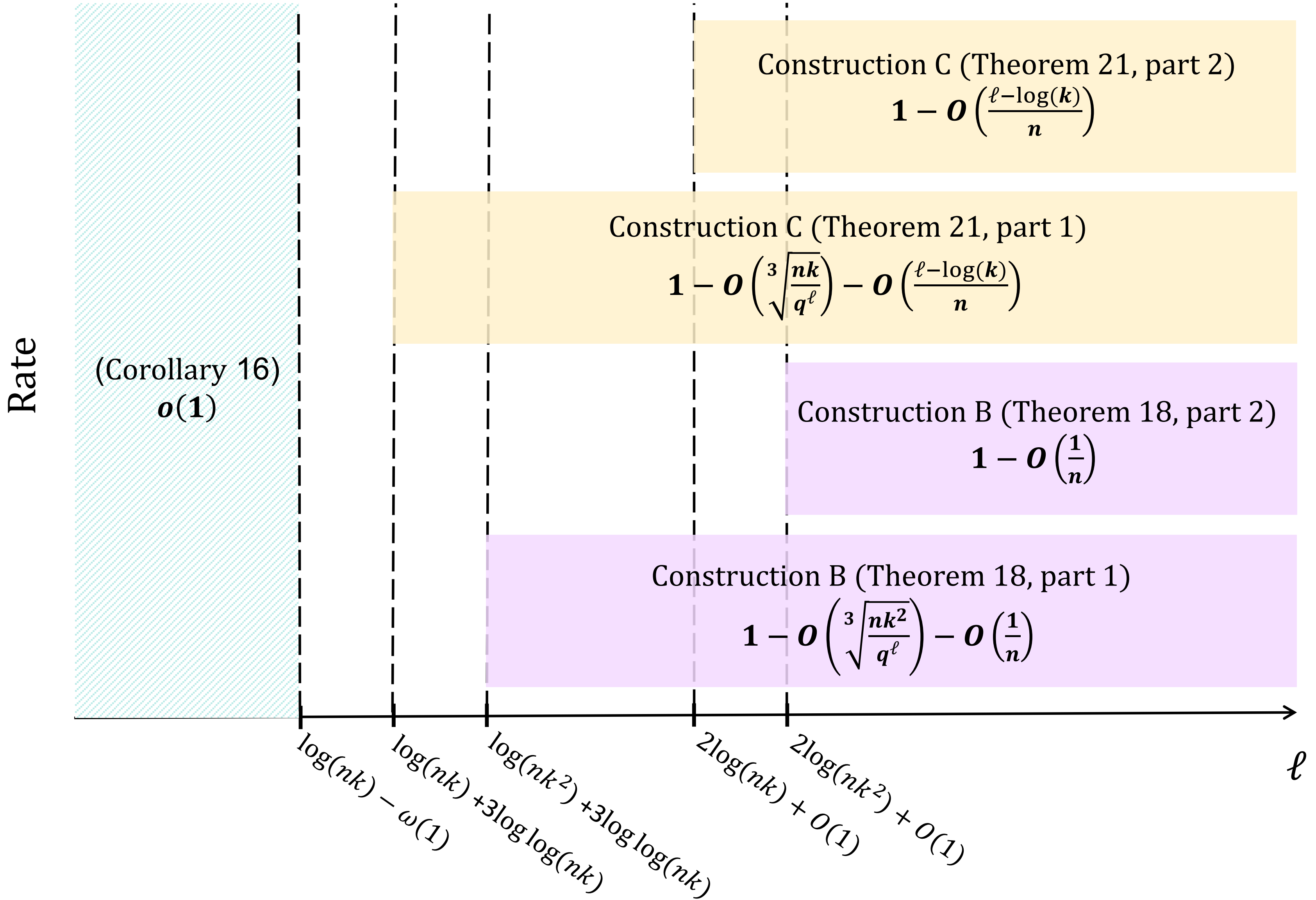


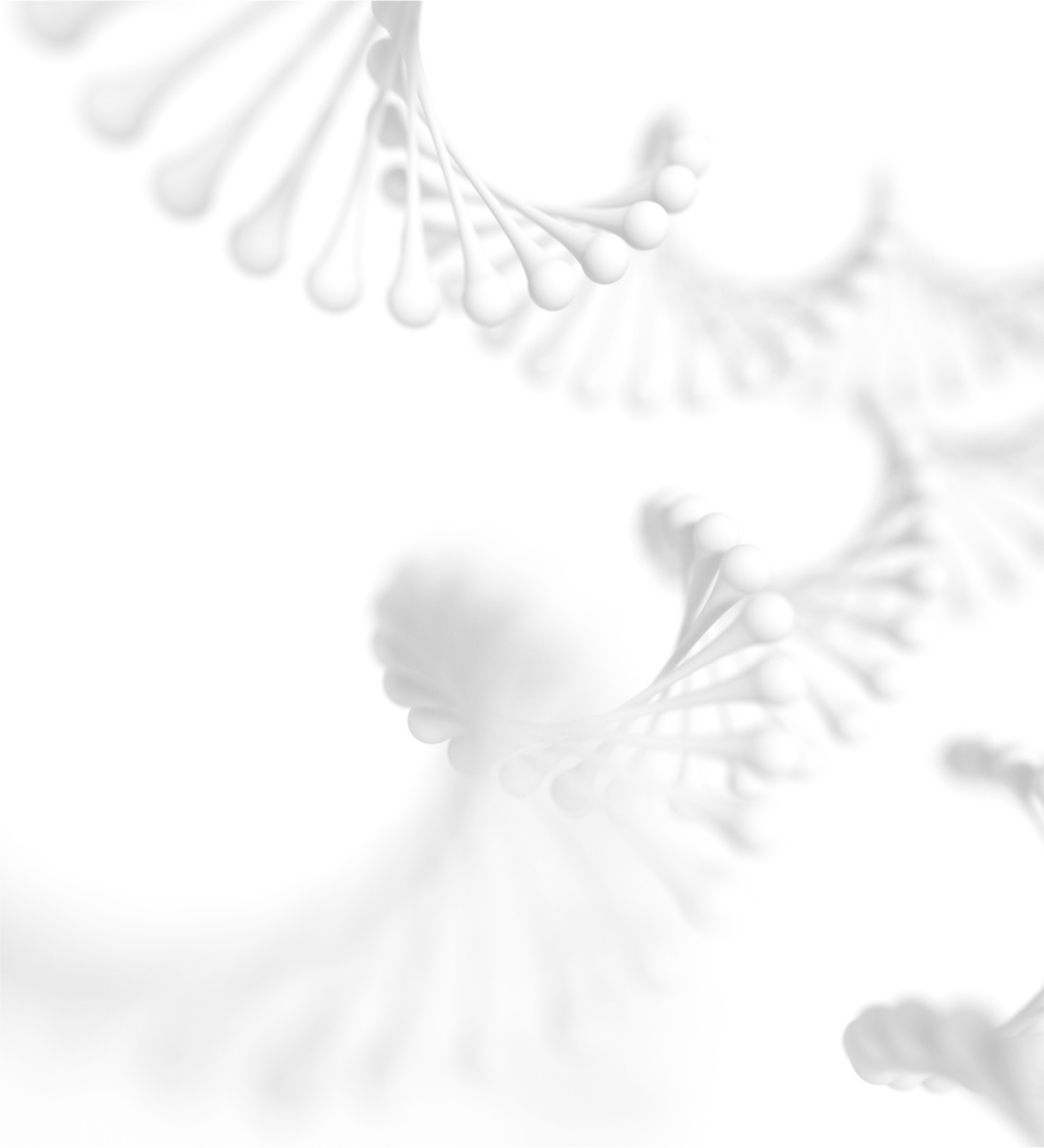
Divide $\boldsymbol{y}$ into k ℓ-overlapping substrings of length n.



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

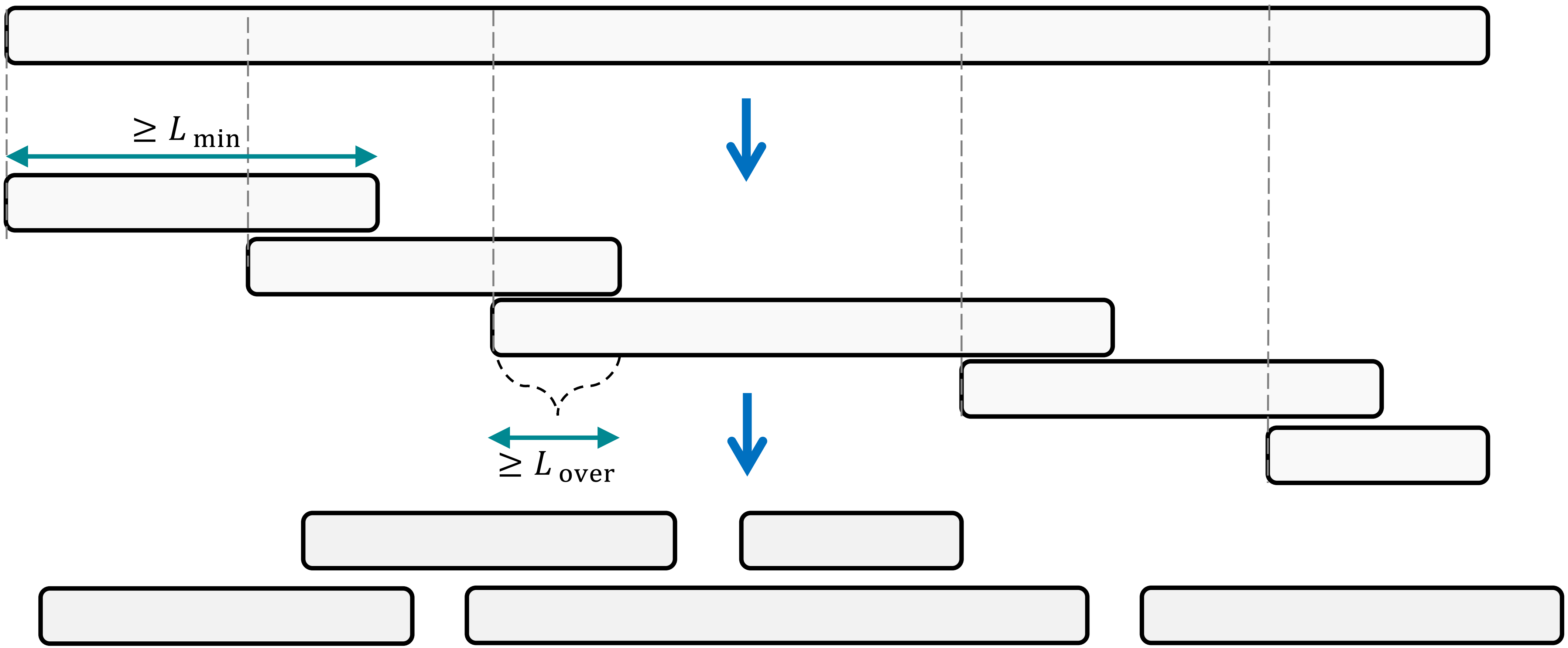
Constructions B&C – Rate





Reconstruction from Substrings with Partial Overlap

Partial-Overlap Channel



Probabelistic version of this channel:

Ravi, A. N., Vahid, A., and Shomorony, I. (2022). Coded Shotgun Sequencing. *ACM JSAIT*, 3(1), 147-159.

Definitions

An $(L_{\min}, L_{\text{over}})$*-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset of substrings such that:
(1) All substrings are of length at least $L_{\min}$.
(2) Succeeding substrings overlap at least L_{over}.
(3) $\boldsymbol{x}$ is covered by the substrings.

$(L_{\min}, L_{\text{over}})$- *trace spectrum* of $\boldsymbol{x}$, denoted by $\mathcal{T}_{L_{\min}}^{L_{\text{over}}}(\boldsymbol{x})$:
set of all $(L_{\min}, L_{\text{over}})$-traces of $\boldsymbol{x}$.

Example: $\boldsymbol{x}=11101110101111$
$\{\{1110111, 1110101, 011111\}\}$ is a (6,2)-trace of $\boldsymbol{x}$.

Definitions

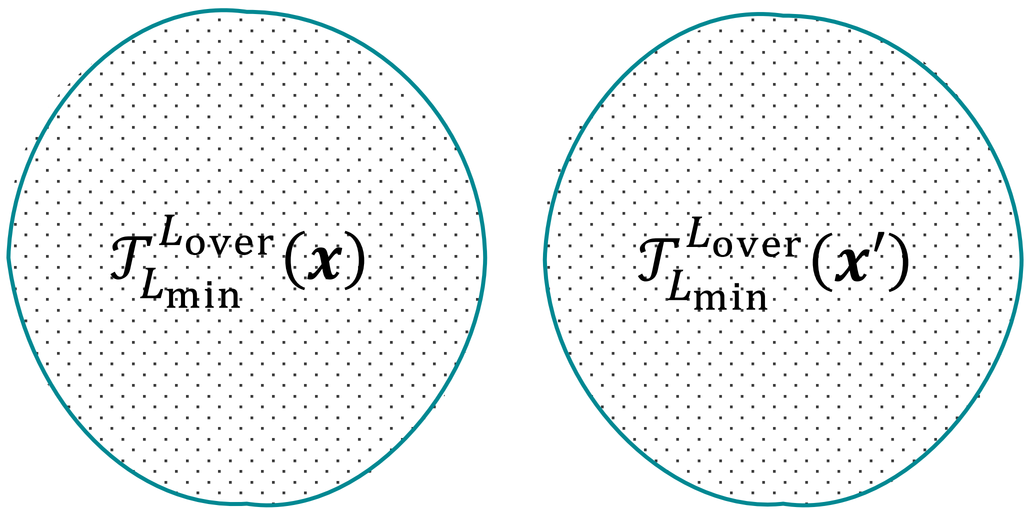
An $(L_{\min}, L_{\text{over}})$*-trace* of $\boldsymbol{x} \in \Sigma^n$ is a multiset of substrings such that:
(1) All substrings are of length at least $L_{\min}$.
(2) Succeeding substrings overlap at least L_{over}.
(3) $\boldsymbol{x}$ is covered by the substrings.

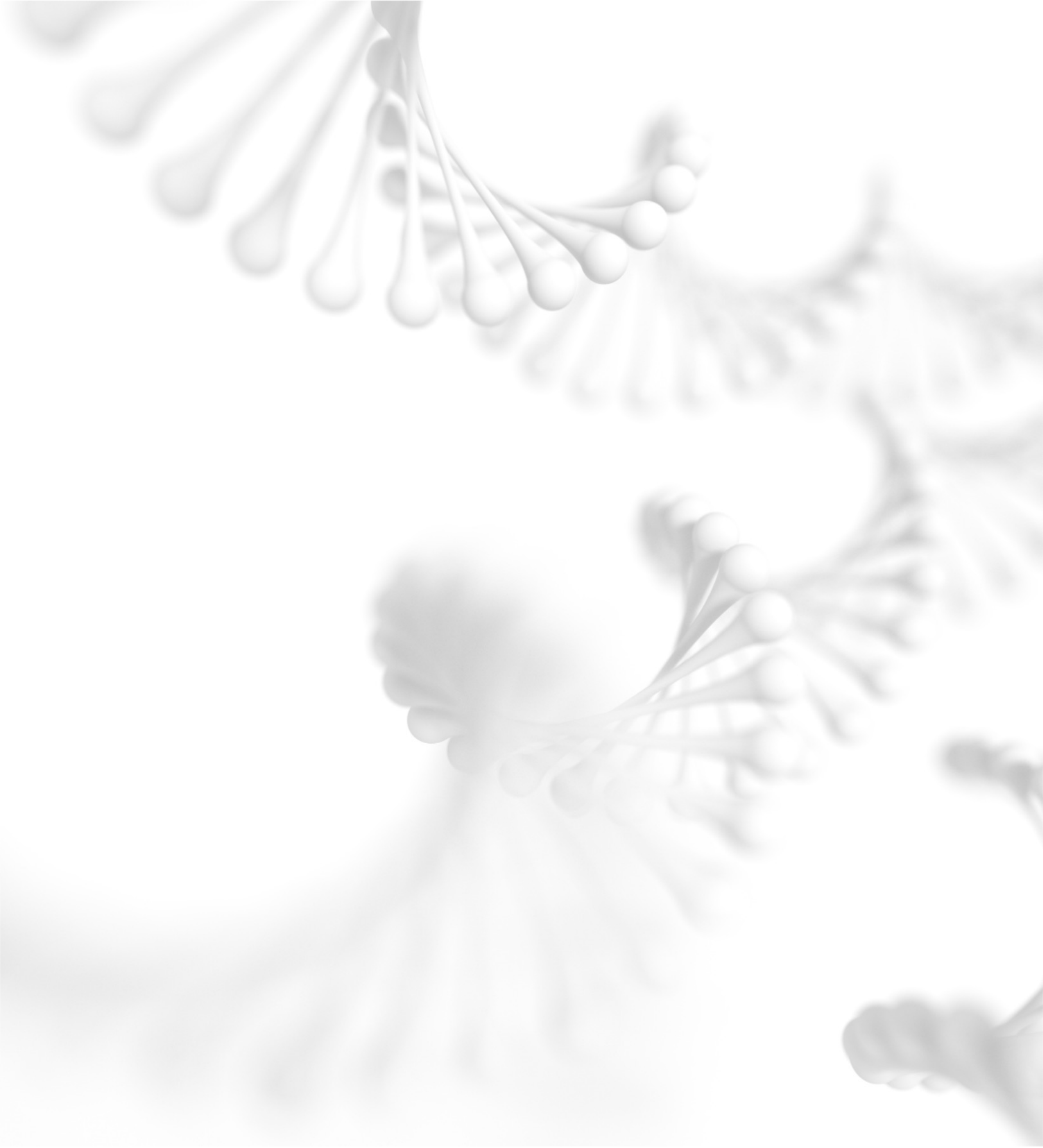
$(L_{\min}, L_{\text{over}})$*- trace spectrum* of $\boldsymbol{x}$, denoted by $\mathcal{T}_{L_{\min}}^{L_{\text{over}}}(\boldsymbol{x})$:
set of all $(L_{\min}, L_{\text{over}})$-traces of $\boldsymbol{x}$.

Channel input: $\boldsymbol{x} \in \Sigma^n$
Channel output: an $(L_{\min}, L_{\text{over}})$-trace of $\boldsymbol{x}$.

Definitions

$(L_{\mathrm{min}}, L_{\mathrm{over}})$-trace code: code $\mathcal{C} \subseteq \Sigma^n$ such that for any two distinct strings $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}$, it holds that $\mathcal{T}_{L_{\mathrm{min}}}^{L_{\mathrm{over}}}(\boldsymbol{x}) \cap \mathcal{T}_{L_{\mathrm{min}}}^{L_{\mathrm{over}}}(\boldsymbol{x}') = \emptyset$.





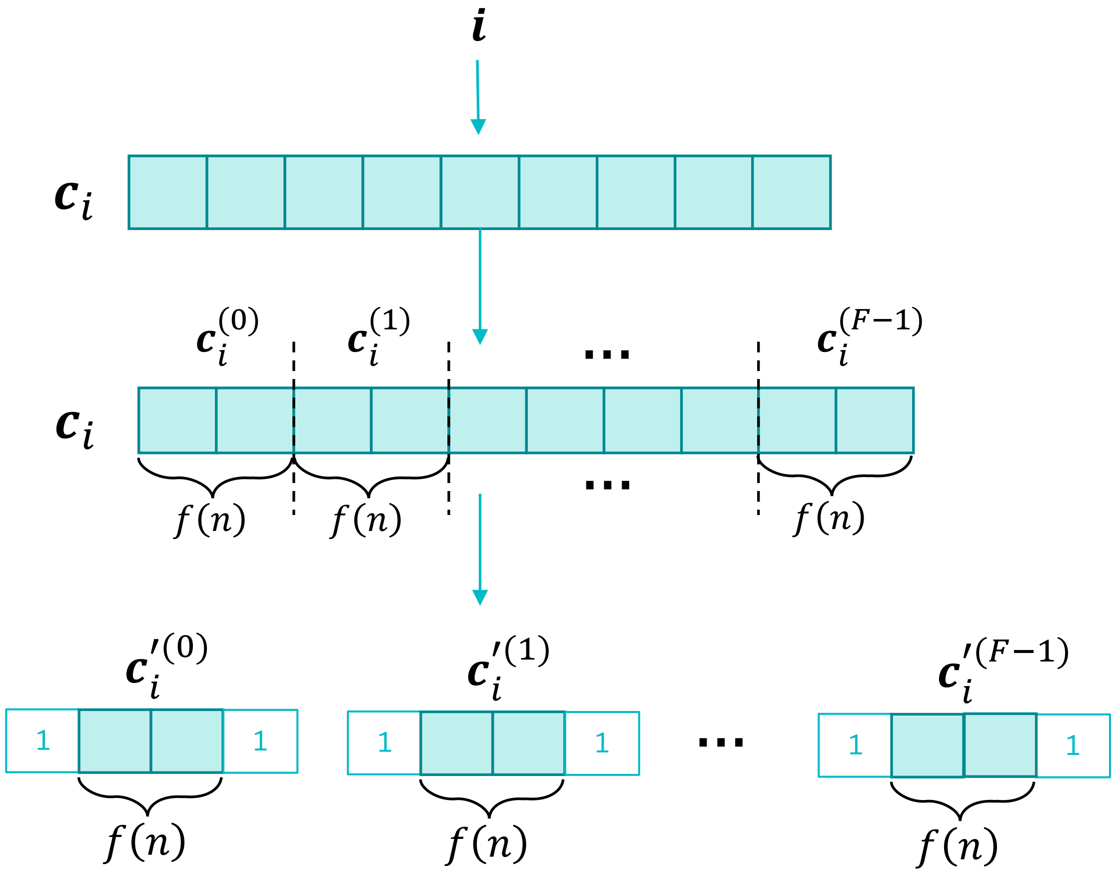
Construction for $(L_{\mathrm{min}}, L_{\mathrm{over}})$-Trace Codes

$$L_{\mathrm{min}} = a\log(n) + O_n(1), \qquad a > 1$$
$$L_{\mathrm{over}} = \gamma L_{\mathrm{min}} + O_n(1), \qquad 0 < \gamma \leq \frac{1}{a}$$

Preliminaries: Index Generation

Consider:

- Index length: $I \approx \left(1 - \frac{1-\gamma a}{1-\gamma}\right) L_{\min}$
- Block size: $f(n) = o(\log n)$
- \# blocks: $F = \frac{I}{f(n)}$



Segments of the encoded index

Preliminaries: Repeat-Free Encoding

For $\ell < N$ the set of *repeat-free (RF)* strings is

$$\mathcal{RF}_\ell(N) \triangleq \{\boldsymbol{x} \in \Sigma^N: \ \ no\ \ell - substring\ repeats\}.$$

Lemma: For $q > 2$ and integers $\ell(N) > \lceil\log(N)\rceil + 3\lceil\log\log(N)\rceil$ and $\lceil\log\log(N)\rceil < t \leq \left\lfloor\frac{\ell(N)-\lceil\log(N)\rceil}{3}\right\rfloor$, there exists an efficient **encoder/decoder** pair of $\boldsymbol{\mathcal{RF}_{\ell(N)}(N)}$ that also does **not contain any $\boldsymbol{t}$-length runs of zeros**, **with rate $\boldsymbol{1 - O_n\left(\frac{t}{n} + q^{-t}\right)}$**.

Define encoder $\boldsymbol{E_{m,\ell}^{\mathcal{RF}}}$: input length m, output length $N = N_{n,\ell}(m)$, $t = f(n)$

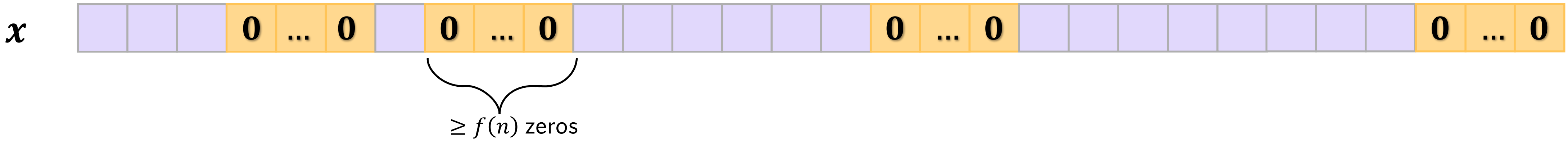
Similar result was proven in [1] for specific values of $\ell(n)$ and t

[1] 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.

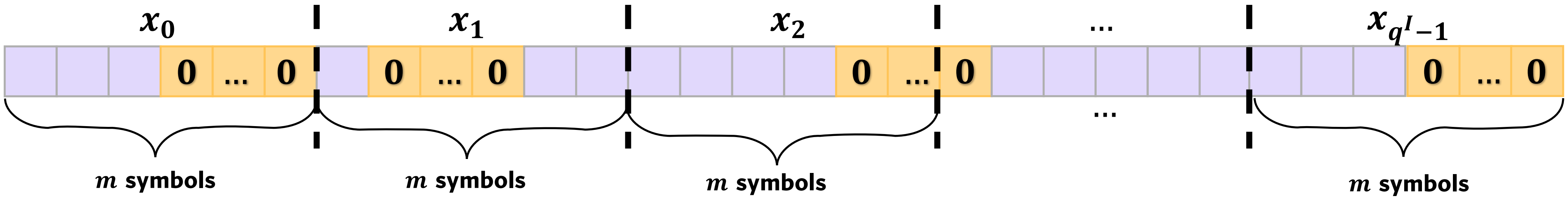
Construction D

Encoder for $(L_{\min}, L_{\text{over}})$-trace code $\mathcal{C}_D$:

Input: a sequence $\boldsymbol{x} \in \Sigma^{q^I m}$



Devide $\boldsymbol{x}$ into q^I non-overlapping substrings of length m.



Next, we encode each substring $\boldsymbol{x}_i$ into $\boldsymbol{z}_i \in \Sigma^{n/q^I}$ independently.

Construction D

Encoder for $(L_{\min}, L_{\text{over}})$-trace code $\mathcal{C}_D$:

For any $0 \leq i \leq q^I - 1$ we want $\boldsymbol{z_i}$ to satisfy two properties:

(1) the index i can be decoded from any $L_{\min}$-substring of $\boldsymbol{z_i}$, and
(2) the string $\boldsymbol{z_i}$ can be uniquely reconstructed from any $(L_{\min}, L_{\text{over}})$-trace of $\boldsymbol{z_i}$.

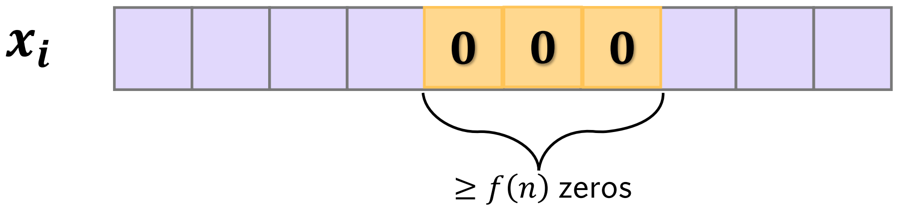
Then, we let

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

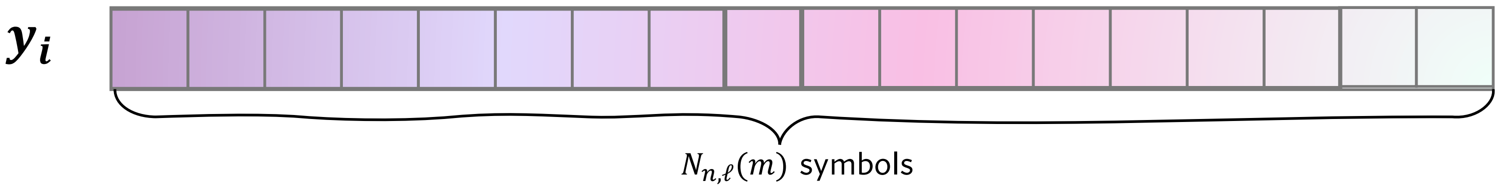
Construction D

Encoder for $(L_{\min}, L_{\text{over}})$-trace code $\mathcal{C}_D$:

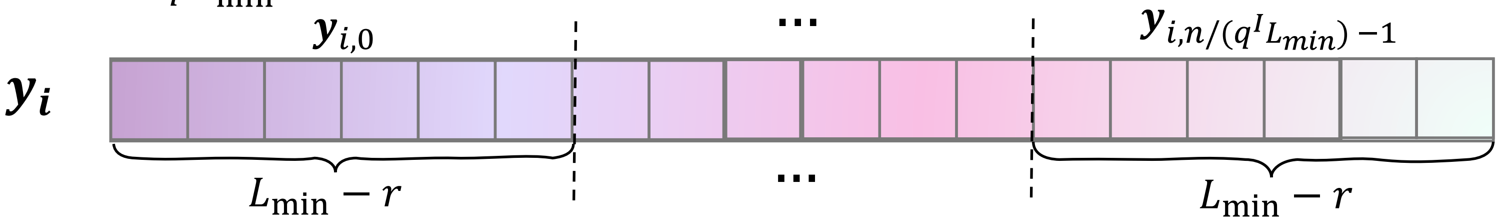
For any $0 \le i \le q^I - 1$:



(1) Encode $\boldsymbol{x_i}$ using the RF encoder to obtain $\boldsymbol{y_i} = E_{m,\ell}^{\mathcal{RF}}(\boldsymbol{x_i})$.



(2) Partition $\boldsymbol{y_i}$ into $\frac{n}{q^I L_{\min}}$ non-overlapping segments of length $L_{\min} - r$.

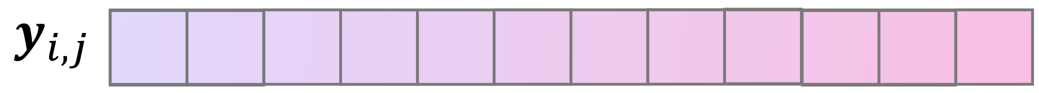


$$N_{n,\ell}(m) \approx \frac{1-\gamma a}{1-\gamma} n q^{-I}$$

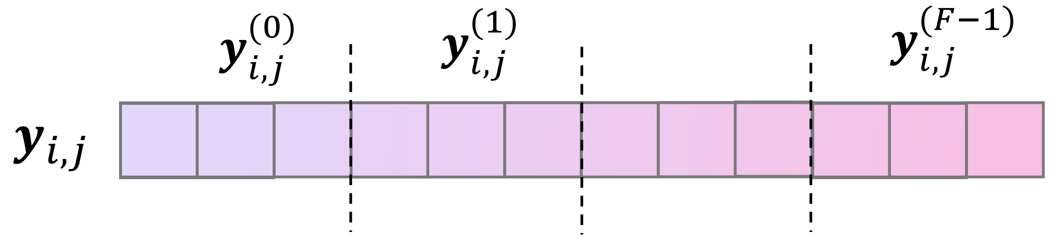
Construction D

Encoder for $(L_{\min}, L_{\mathrm{over}})$-trace code $\mathcal{C}_D$:

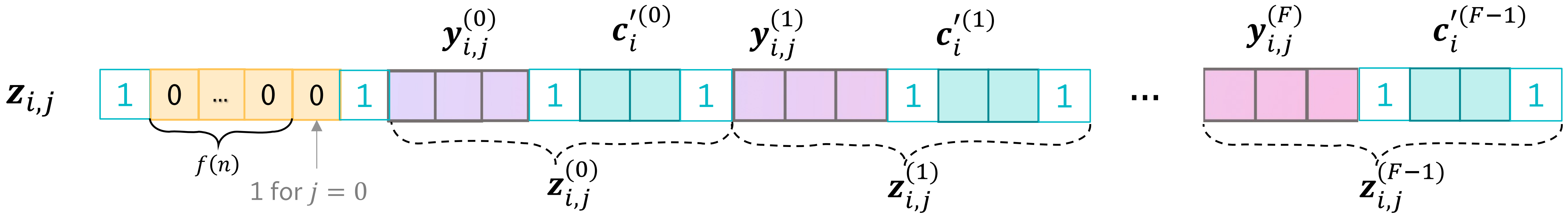
(3) For all $j \in [n/(q^I L_{\min})]$:



(3.1) Partition $\boldsymbol{y}_{i,j}$ into F non-overlapping segments of equal lengths (up to ± 1, if necessary).



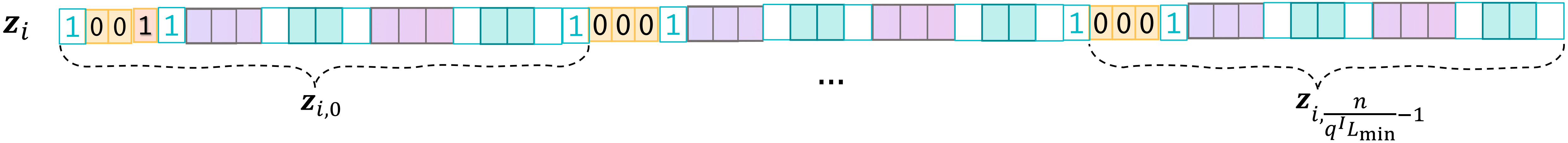
(3.2) For $k \in [F]$, let $\boldsymbol{z}_{i,j}^{(k)} \triangleq \boldsymbol{y}_{i,j}^{(k)} \circ \boldsymbol{c}_{\boldsymbol{i}}^{\prime(k)}$



Construction D

Encoder for $(L_{\min}, L_{\text{over}})$-trace code $\mathcal{C}_D$:

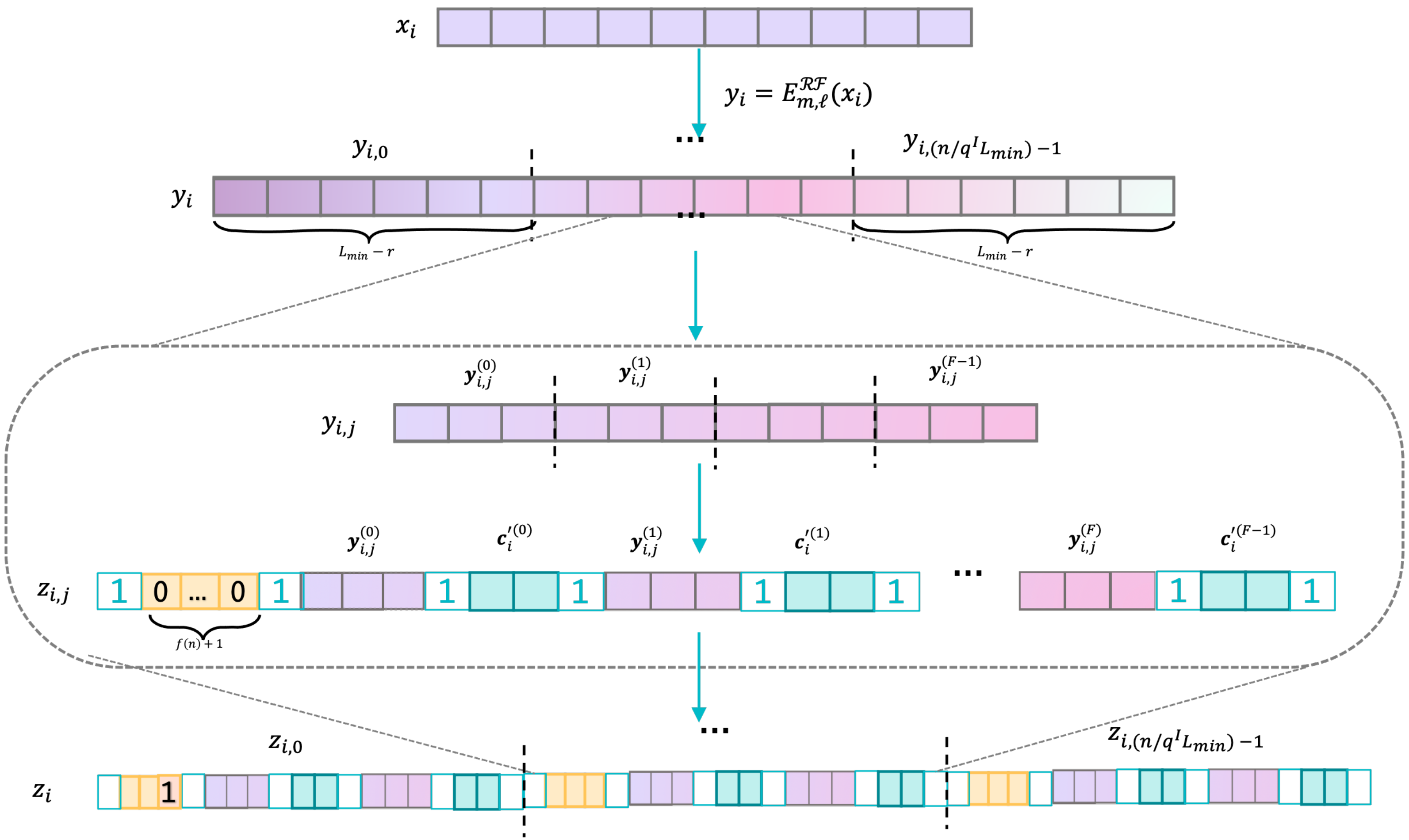
(4) Concatenate $\boldsymbol{z}_i = \boldsymbol{z}_{i,0} \circ \boldsymbol{z}_{i,1} \circ \cdots \circ \boldsymbol{z}_{i,\frac{n}{q^I L_{\min}}-1}$



Lastly, we concatenate all the $\boldsymbol{z}_i$ strings, in order, to obtain the code word

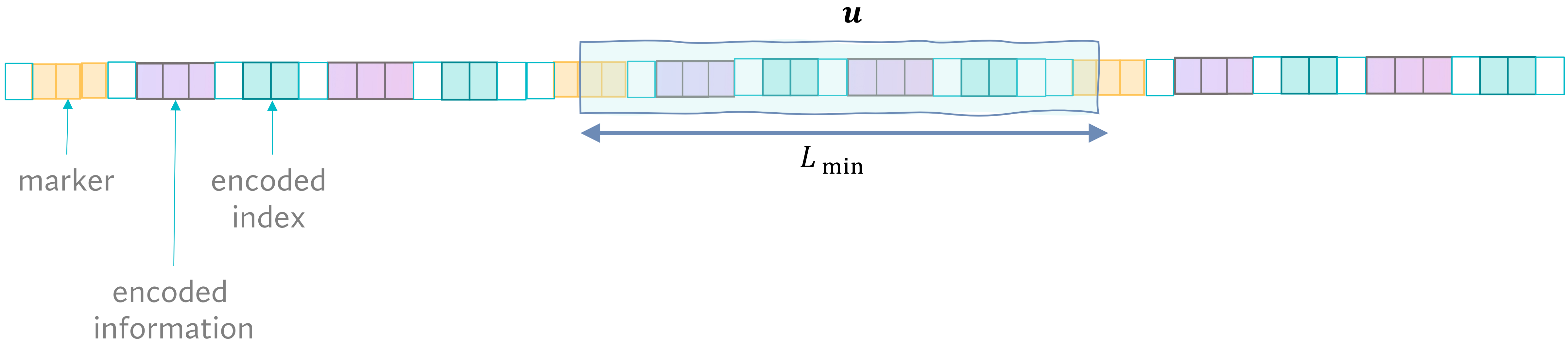
$$\text{Enc}_D(\boldsymbol{x}) \triangleq \boldsymbol{z} = \boldsymbol{z}_0 \circ \cdots \circ \boldsymbol{z}_{q^I-1}.$$

Construction D – Overview



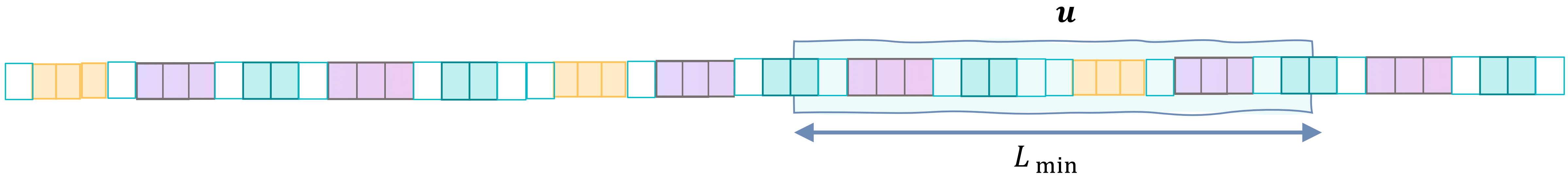
Construction D – Encoder Correctness

Lemma: Every $L_{\min}$-substring $\boldsymbol{u}$ of $\boldsymbol{z}$ contains as subsequences at least an $(I-\mu)$-suffix of $\boldsymbol{c}_i$ and 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.

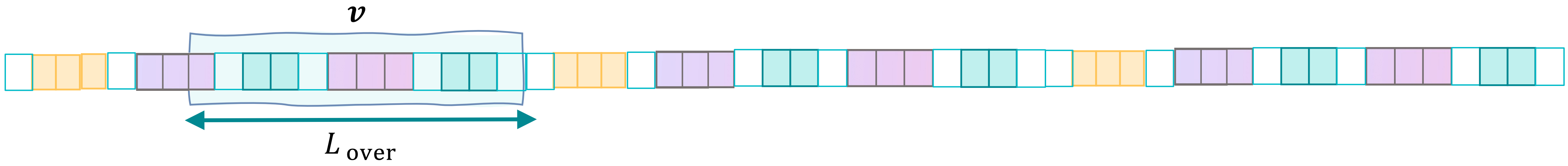


Construction D – Encoder Correctness

Lemma: Every $L_{\min}$-substring $\boldsymbol{u}$ of $\boldsymbol{z}$ contains as subsequences at least an $(I-\mu)$-suffix of $\boldsymbol{c}_i$ and 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.



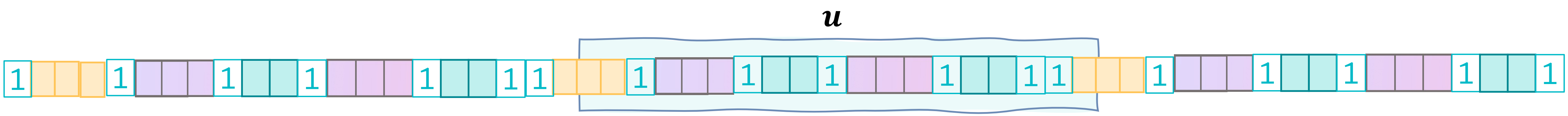
Lemma: Every L_{over}-substring $\boldsymbol{v}$ of $\boldsymbol{z}$ contains at least ℓ consecutive symbols of $\boldsymbol{y} \triangleq \boldsymbol{y}_0 \circ \boldsymbol{y}_1 \circ \cdots \circ \boldsymbol{y}_{q^I-1}$.



Construction D – Encoder Correctness

Theorem : for all admissible values of n, the code $\mathcal{C}_D$ is an $(L_{\min}, L_{\text{over}})$-trace code.

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



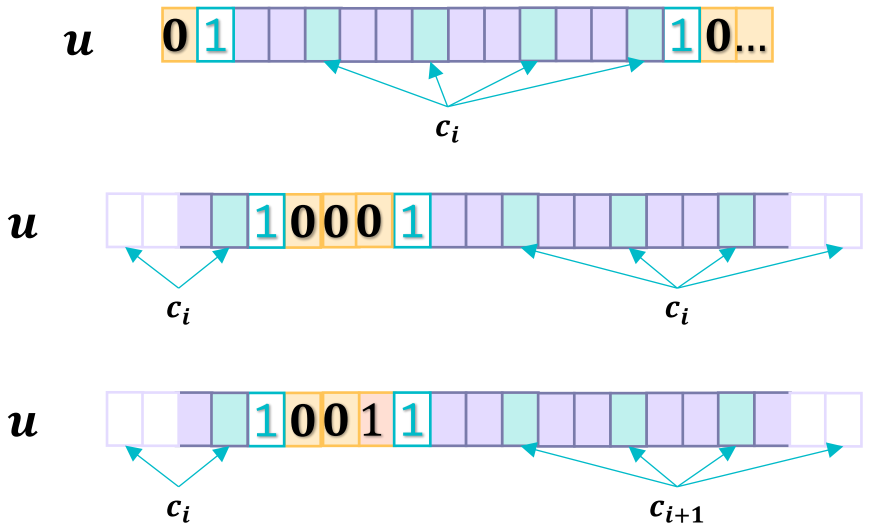
Let $\boldsymbol{u} \in T$, we can assume w.l.o.g. that $|\boldsymbol{u}| = L = L_{\min}$.

<u>Note that:</u>

1. The segment $\boldsymbol{u}$ does not contain any occurrences of the marker $\boxed{1}\boxed{0\ 0\ \ }\boxed{1}$ except those explicitly added as a prefix of $\boldsymbol{z}_{i,j}$ (for some i, j).
2. For any i, j, $|\boldsymbol{z}_{i,j}| \leq L_{\min}$.
3. Either $\boldsymbol{u}$ contains an occurrence of $\boxed{1}\boxed{0\ 0\ \ }\boxed{1}$ or it has a suffix-prefix pair whose concatenation is $\boxed{1}\boxed{0\ 0\ \ }\boxed{1}$.

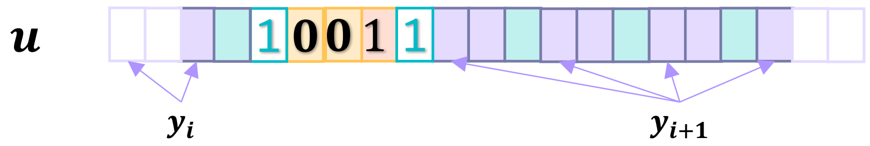
Construction D – Encoder Correctness

Either $\boldsymbol{u}$ contains an occurrence of $1\,0\,0\,\square\,1$ or it has a suffix-prefix pair whose concatenation is $1\,0\,0\,\square\,1$.



We can always correctly deduce i from $\boldsymbol{u}$.
It is therefore possible to partition T by the index i (corresponding to the substring $\boldsymbol{z_i}$).

Construction D – Encoder Correctness



If $\boldsymbol{u}$ is intersecting both $\boldsymbol{y}_i, \boldsymbol{y}_{i+1}$, then $\boldsymbol{u}$ contains the complete synchronization marker $\boxed{1}\boxed{0}\boxed{0}\boxed{1}\boxed{1}$ hence its location in $\boldsymbol{u}$ implies the exact location of $\boldsymbol{u}$ in $\boldsymbol{z}$.

Construction D – Encoder Correctness

Lemma: Every L_{over}-substring $\boldsymbol{v}$ of $\boldsymbol{z}$ contains at least ℓ consecutive symbols of $\boldsymbol{y} \triangleq \boldsymbol{y}_0 \circ \boldsymbol{y}_1 \circ \cdots \circ \boldsymbol{y}_{q^I-1}$.

Since each $\boldsymbol{y}_i$ is ℓ-repeat-free, there exist a unique way to match the overlaps of these substrings.

Construction D – Encoder Correctness

Lemma: Every L_{over}-substring $\boldsymbol{v}$ of $\boldsymbol{z}$ contains at least ℓ consecutive symbols of $\boldsymbol{y} \triangleq \boldsymbol{y_0} \circ \boldsymbol{y_1} \circ \cdots \circ \boldsymbol{y_{q^I-1}}$.

Since each $\boldsymbol{y_i}$ is ℓ-repeat-free, there exist a unique way to match the overlaps of these substrings.

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 $E_{m,\ell}^{\mathcal{RF}}$.

Construction D – Rate

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

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

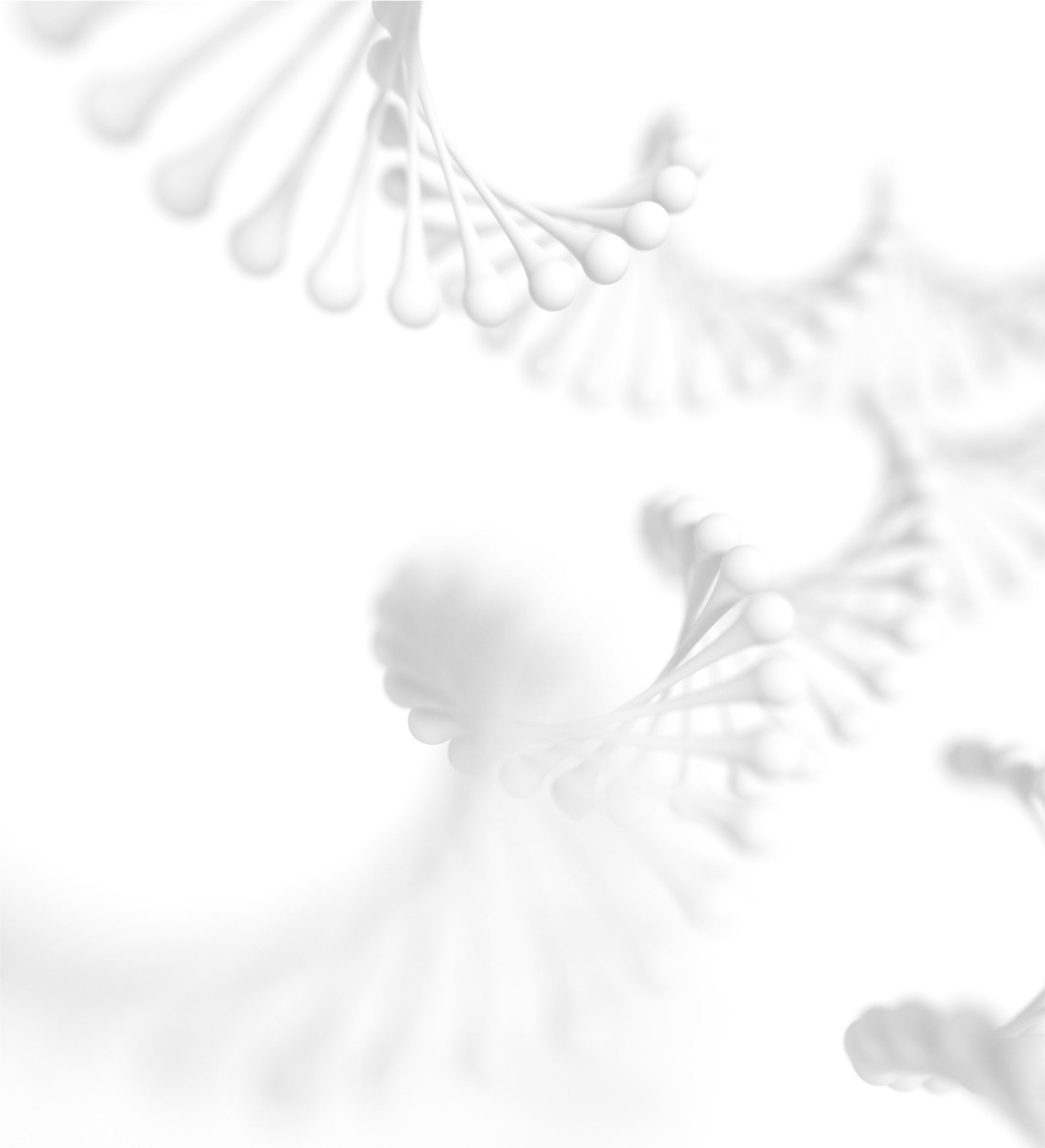
Maximum Asymptotic Rate of $(L_{\min}, L_{\mathrm{over}})$-Trace Codes

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

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

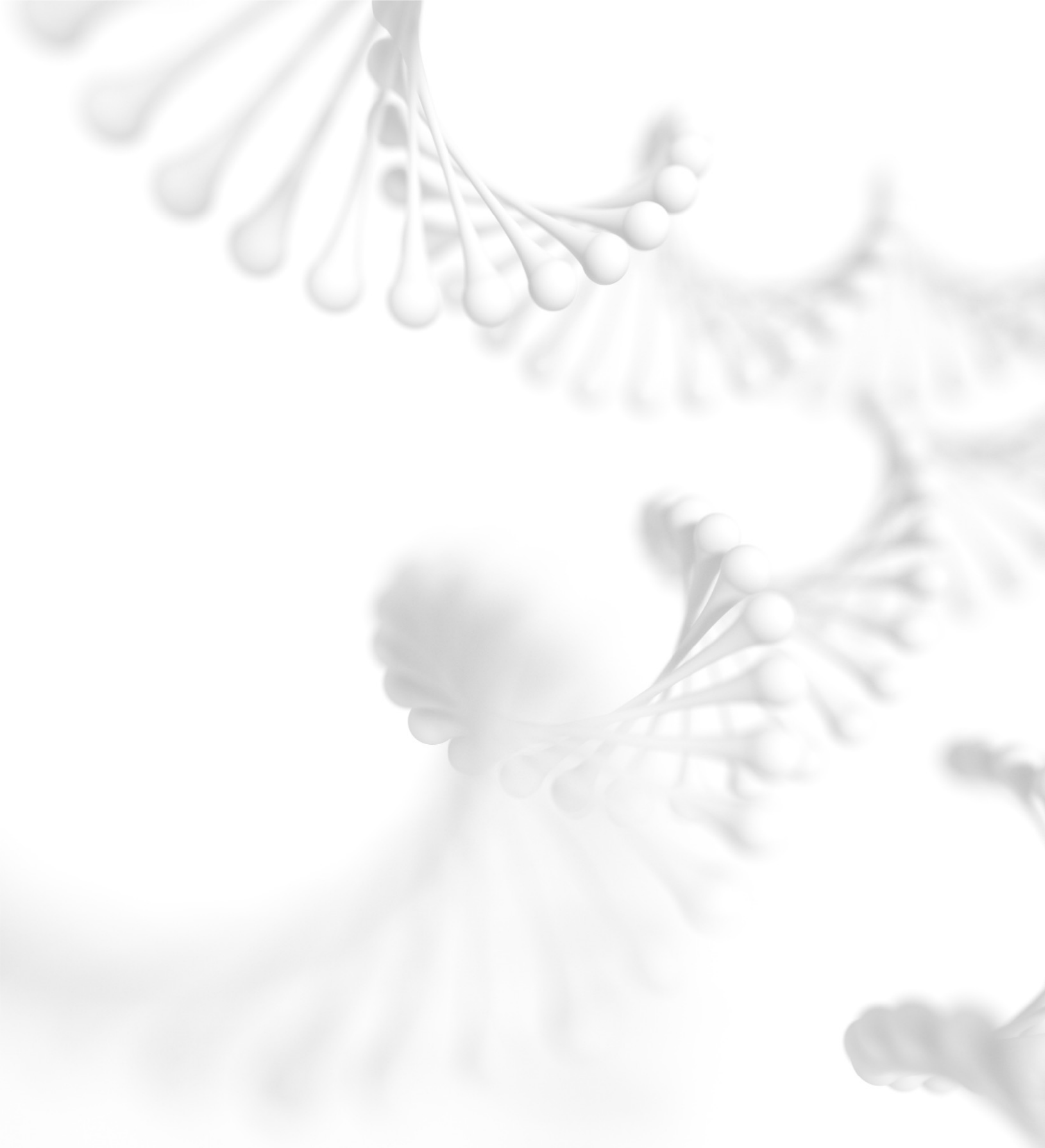
Hence, Construction D asymptotically meets the upper bound.

Corollary: If $\limsup_{n\to\infty} \frac{L_{\min}}{\log(n)} \leq 1$, then $R(\mathcal{C}) = o_n(1)$ for any $(L_{\min}, L_{\mathrm{over}})$-trace code $\mathcal{C} \subseteq \Sigma^n$.



Future Work

- Erroneous versions of the channels.
- Multistrand setup for the partial overlap channel.
- Analyze the worst-case for a more realistic setup.



References

[1] **D. Bar-Lev**, S. Marcovich, E. Yaakobi, and Y. Yehezkeally, “Adversarial Torn-Paper Codes”, submitted to *IEEE Transactions on Information Theory*.

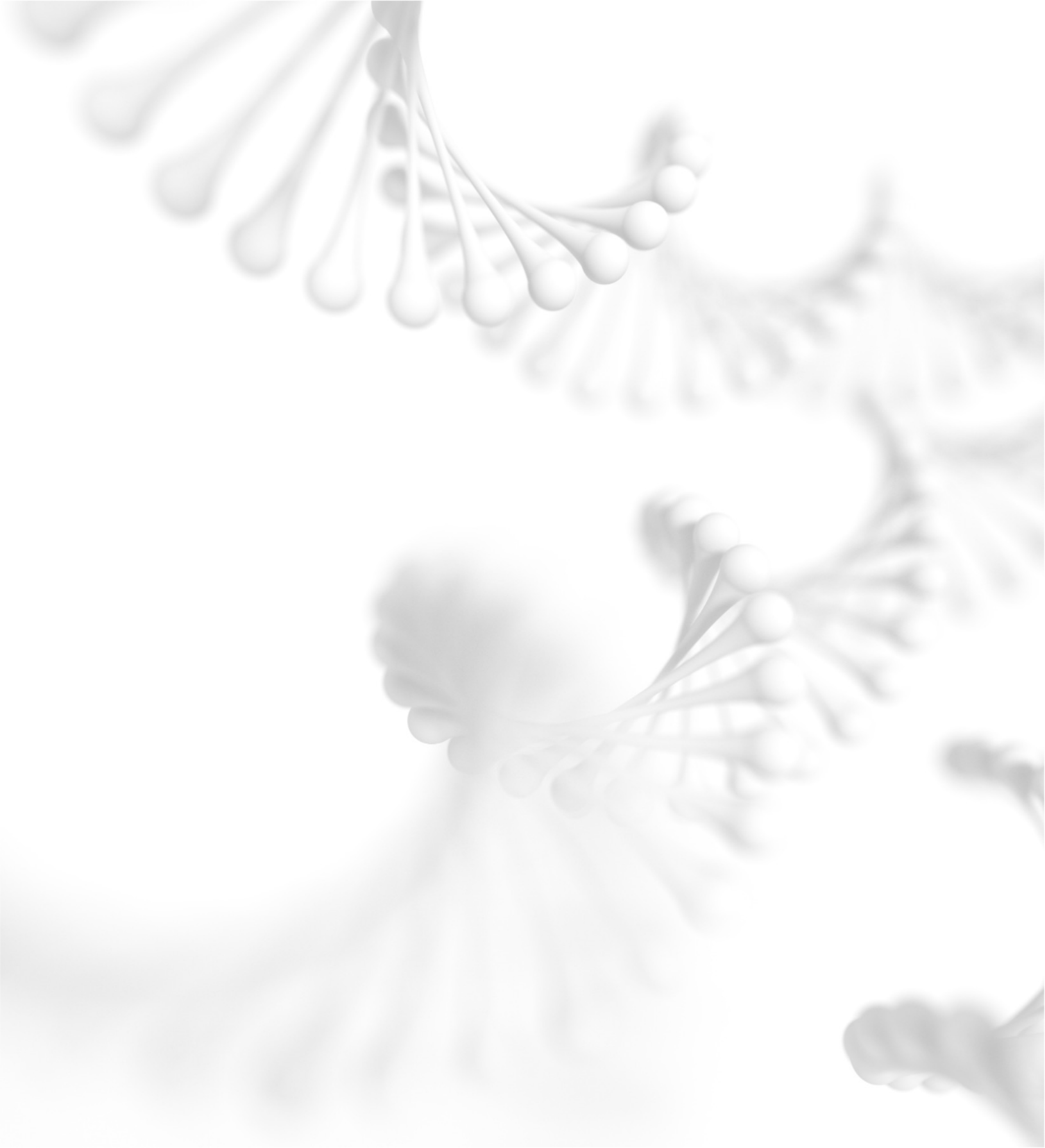
[2] Y. Yehezkeally, **D. Bar-Lev**, S. Marcovich, and E. Yaakobi, “Generalized Unique Reconstruction from Substrings”, submitted to *IEEE Transactions on Information Theory*.

and also

[3] Y. Yehezkeally, S. Marcovich, and E. Yaakobi, “Multi-strand reconstruction from substrings,” in *IEEE ITW*, 2021

[4] **D. Bar-Lev**, S. Marcovich, E. Yaakobi, Y. Yehezkeally, “Adversarial Torn-Paper Codes,” in *IEEE ISIT*, 2022.

[5] Y. Yehezkeally**, D. Bar-Lev**, S. Marcovich, and E. Yaakobi, “Reconstruction from Substrings with Partial Overlap”, in *IEEE ISITA.*, 2022.



Thank You!