On the Asymptotic Rate of Optimal Codes That Correct Tandem Duplications for Nanopore Sequencing

Wenjun Yu, Zuo Ye, *Member, IEEE*, and Moshe Schwartz, *Fellow, IEEE*

Abstract—We study codes that can correct backtracking errors during nanopore sequencing. In this channel, a sequence of length n over an alphabet of size q is being read by a sliding window of length ℓ, where from each window we obtain only its composition. Backtracking errors cause some windows to repeat, hence manifesting as tandem-duplication errors of fixed length k in the ℓ-read vector of window compositions. While existing constructions for duplication-correcting codes can be straightforwardly adapted to this model, even resulting in optimal codes, their asymptotic rate is hard to find. In the regime of unbounded number of duplication errors, we either give the exact asymptotic rate of optimal codes, or bounds on it, depending on the values of k, ℓ and q. In the regime of a constant number of duplication errors, t, we find the redundancy of optimal codes to be $t \log_q n + O(1)$ when $\ell | k$, and only upper bounded by this quantity otherwise.

Index Terms—Nanopore sequencing, tandem duplication, error-correcting codes.

I. INTRODUCTION

WITH the exponentially increasing amount of data generated and the need for long-term archival data storage, DNA storage is emerging as a contender technology, capable of breaking existing bottlenecks of conventional electronic storage systems by offering high data density, longevity and ease of copying information [6]. Among the many challenges facing us in the process of enabling DNA storage, we focus on the sequencing technologies used to read DNA.

Several technologies have been developed to improve the reading performance of DNA. In particular, nanopore sequencing has attracted the attention of researchers and practitioners alike due to better portability, ability to read longer strands, and low initial cost. A DNA strand is simply a long string over an alphabet of size 4, the elements of which are called nucleotides (or bases). The main nanopore sequencing process involves pushing a fragment of a DNA strand through a microscopic pore. While moving through it, a window of length ℓ nucleotides is observed by the machine. As the DNA molecule continues moving through the pore, ideally, we should obtain a sliding-window reading of the entire string with a window of size ℓ.

Unfortunately, certain physical aspects of the nanopore sequencer lead to various distortions in the final readout such as random dwell times, fading, inter-symbol interference (ISI), and random noise. Additionally, the passage of the DNA fragment through the pore is often irregular and may involve backtracking (jumping back a few positions) or skipping (jumping forward a few positions), creating duplicated parts in the sliding-window reading, or missing windows. Finally, it is not always easy to get the window reading, forcing us to resort to weaker forms of reading, for example, only obtaining the (unordered) composition of the bases within each window.

Prior work in this area focused on developing various mathematical models for the sequencer and designing codes to efficiently correct errors in the final readouts. For example, [15] studied the model that incorporates ISI, deletions and measurement noise. In [3], the authors considered codes that correct substitution errors, and provided the optimal redundancy for single substitution with window of length $\ell \geqslant 3$. For the window of length $\ell = 2$, the work in [16] generalized previous works to propose the optimal redundancy for two substitution errors. Additionally, [2] studied the single deletion nanopore channel and proposed a code with optimal redundancy up to an additive constant. In [5], the authors considered codes that correct backtracking or skipping errors and presented numerical results for the asymptotic rate of backtracking error-correcting codes.

In this paper, we adopt the nanopore-channel model studied in [2], [3], and [16]: the DNA sequence is read by sliding windows of length ℓ, where we only obtain the composition of each scanned window. However, unlike [3], [16] which studied substitution errors, and [2] which studied skipping errors (causing deletions), we focus on backtracking errors. In more detail, each DNA sequence of length n is first transformed into an $(n + \ell - 1)$-length vector called an ℓ-read vector through the ISI channel, where the i-th coordinate of the ℓ-read vector is obtained by collecting the consecutive symbols between the $(i - \ell + 1)$-st and i-th entries of the original DNA sequence and outputting their (unordered) multiset composition. This

process can be regarded as a detecting window of length ℓ passing through the whole sequence by shifting one position at each step. However, as the DNA molecule moves through the pore, a backtracking error causes it to move back k bases. Thus, a few of the sliding windows are repeated, in essence, causing a tandem duplication of length k to occur in the ℓ-read vector. Reference [5] focused exclusively on the specific case of moving back $k = 2$ bases. In contrast, this paper extends the study to the general case for any integer $k \geqslant 1$, providing a broader and more comprehensive analysis of the problem.

Duplication errors were introduced in [7], in the context of DNA storage in living organisms, and the first duplication-correcting codes were developed in [9]. These allow for arbitrary alphabets, parametric duplication length, and various duplication types. Following this, codes correcting any number of tandem duplications were studied in [9], [21], and [22]. Codes that correct only a fixed number of tandem duplications (sometimes just one) were studied in [8], [10], and [13]. Additionally, codes that correct mixtures of duplications, substitutions, as well as insertions and deletions, have been the focus of [17], [18], and [19].

In this paper, we are interested in codes over an alphabet of size q and length n, that are capable of correcting tandem duplications of fixed length k in their ℓ-read vectors. Unlike skipping errors, backtracking errors (in which the DNA strand moves back in the pore) may potentially happen an unbounded number of times. Thus, we are interested both in the regime of an unbounded number of duplications, or a constant number of duplications. Since codes capable of correcting tandem duplications exist in both regimes, a simple and effective strategy for constructing the desired codes is the following: Consider the set of all q^n possible sequences of length n over the alphabet. After applying the sliding window, these result in q^n ℓ-read vectors that reside in a much larger space of size $\sim q^{\ell n}$. Note that most of the sequences in that larger space are *not* ℓ-read vectors, since they do not represent a sliding-window reading of a sequence over the original alphabet. However, in that large space we can construct a known duplication-correcting code, and from its set of codewords, keep only those that are actually ℓ-read vectors. Thus, we can harness the power of known optimal constructions, but we are left with a difficult problem: what are the rates of the resulting codes?

The main contribution of this paper is finding the exact asymptotic rate, or bounds on it, for optimal codes over an alphabet of size q, that are capable of correcting tandem duplications of fixed length k in their ℓ-read vectors, both in the regime of unbounded number of errors, and in the regime of a constant number of errors. For the case of unbounded errors, we study the optimal code that parallels the construction given in [9]. Depending on the values of k and ℓ, sometimes depending on whether $\ell | k$, we can compute the exact asymptotic rate of the optimal code, and in others, we provide bounds. These results are summarized in Table I. For the case of a constant number of errors, t, we study the code that parallels the construction given in [11]. We show that the redundancy of the optimal code is $t \log_q n + O(1)$ when $\ell | k$, and only upper bounded by this quantity otherwise. All results

generalize the results of [9] and [11], which are the special case of $\ell = 1$, namely, when sequences are read letter by letter.

The paper is organized as follows. We begin in Section II by introducing the relevant notation and definitions. Section III focuses on the regime of unbounded duplication errors, whereas Section IV studies a constant number of errors. We conclude in Section V with a short discussion of the results and open problems.

II. PRELIMINARIES

For any $i, j \in \mathbb{Z}$, $i \leqslant j$, we define $[i, j] \triangleq \{i, i + 1, \ldots, j\}$. For $n \in \mathbb{N}$ we then define $[n] \triangleq [1, n]$. Consider some finite alphabet Σ. A sequence (or, a vector) of length n over Σ is an n-tuple $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, where $x_i \in \Sigma$ for all i. We shall generally use bold lower-case letters for sequences, and the same letter in regular font and an index, to denote a letter of the sequence. Indexing of sequence positions will start from 1. The length of $\boldsymbol{x}$ is denoted by $|\boldsymbol{x}| \triangleq n$, and whenever it makes sense, the norm of $\boldsymbol{x}$ is defined as $\|\boldsymbol{x}\|_1 \triangleq \sum_{i=1}^{n} |x_i|$. The set of all sequences of length n over Σ is denoted by Σ^n. The unique empty sequence (the sequence of length 0) is denoted by $\boldsymbol{\varepsilon}$, and we define $\Sigma^0 \triangleq \{\boldsymbol{\varepsilon}\}$. We then also define $\Sigma^{\leqslant n} \triangleq \bigcup_{i=0}^{n} \Sigma^i$, $\Sigma^* \triangleq \bigcup_{i \geqslant 0} \Sigma^i$, and $\Sigma^+ \triangleq \bigcup_{i \geqslant 1} \Sigma^i$. Given two sequences, $\boldsymbol{x}, \boldsymbol{y} \in \Sigma^*$, we use either $\boldsymbol{xy}$ or $(\boldsymbol{x}, \boldsymbol{y})$ to denote their concatenation. For any $\ell \in \mathbb{N}$, we also define $\boldsymbol{x}^\ell$ to be the concatenation of ℓ copies of $\boldsymbol{x}$. Naturally, we define $\boldsymbol{x}^0 \triangleq \boldsymbol{\varepsilon}$.

Consider a sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \Sigma^n$. We say $\boldsymbol{v} \in \Sigma^k$, $k \geqslant 1$, is a *k-factor* of $\boldsymbol{x}$, if there exist $\boldsymbol{u}, \boldsymbol{w} \in \Sigma^*$ such that $\boldsymbol{x} = \boldsymbol{uvw}$. If the length of $\boldsymbol{v}$ is unknown or unimportant, we just say that $\boldsymbol{v}$ is a factor of $\boldsymbol{x}$. Given $i, j \in \mathbb{Z}$, $i \leqslant j$, the window from position i to j is defined as

$$\boldsymbol{x}[i, j] \triangleq (x_i, x_{i+1}, \ldots, x_j).$$

To allow for windows that extend outside of $[n]$, we define $x_i \triangleq \boldsymbol{\varepsilon}$ for all $i \notin [n]$. Thus, the factors of $\boldsymbol{x}$ are of the form $\boldsymbol{x}[i, j]$ for all $1 \leqslant i \leqslant j \leqslant n$. We also define the removal of the $[i, j]$ window from $\boldsymbol{x}$ by

$$M_{[i,j]}(\boldsymbol{x}) \triangleq (x_1, x_2, \ldots, x_{i-1}, x_{j+1}, x_{j+2}, \ldots, x_n).$$

We shall need to count the number of times a specific letter appears in a sequence. Let $a \in \Sigma$ be some letter. Then $\boldsymbol{x}|_a$ counts the number of times a appears in $\boldsymbol{x}$, namely,

$$\boldsymbol{x}|_a \triangleq |\{i \in [n] \ : \ x_i = a\}|.$$

Example 1: Consider the binary sequence

$$\boldsymbol{x} = (0, 1, 1, 0, 0, 0, 1).$$

Then $\boldsymbol{x}[2, 4] = (1, 1, 0)$, and $M_{[2,4]}(\boldsymbol{x}) = (0, 0, 0, 1)$. Also notice that $\boldsymbol{x}[-3, 2] = (0, 1)$ and $\boldsymbol{x}[5, 12] = (0, 0, 1)$. Of course, $\boldsymbol{x}|_0 = 4$ since there are four occurrences of 0 in $\boldsymbol{x}$.

Let $\boldsymbol{x} \in \Sigma^n$ be a sequence of length n, and assume $1 \leqslant m \leqslant n - 1$. We say $\boldsymbol{x}$ is *m-periodic*, if $x_{i+m} = x_i$ for all $i \in [1, n - m]$. We note that by saying $\boldsymbol{x}$ is m-periodic we do not contend that m is the minimal positive integer with this property, but rather merely that it is a period of $\boldsymbol{x}$. Thus, we have

$$\boldsymbol{x} = \left(\boldsymbol{x}[1, m]^{\lceil n/m \rceil}\right)[1, n].$$

If no such m exists we say that x is aperiodic.

In what follows, we shall be examining sequences over various alphabets, which will all be derived from an original sequence. We shall assume, w.l.o.g., that the original sequence is over the alphabet $\mathbb{Z}_q \triangleq \{0, 1, \ldots, q-1\}$, for some integer $q \geqslant 2$. Assume, therefore, that $x = (x_1, x_2, \ldots, x_n) \in \mathbb{Z}_q^n$ is such a sequence. The *composition* of x is represented by the polynomial in the variables $z_0, z_1, \ldots, z_{q-1}$,

$$c(x) \triangleq x|_0 \cdot z_0 + x|_1 \cdot z_1 + \ldots + x|_{q-1} \cdot z_{q-1}.$$

We note that $c(\varepsilon) = 0$. Additionally, if x and y are sequences, then $c(xy) = c(x) + c(y)$.

In the nanopore-sequencing channel setting, a DNA sequence is read by a sliding window of length ℓ, however, from each such window only its composition is obtained. To that end, we define the following alphabet

$$\Psi_{\ell,q} \triangleq \left\{ c(y) \ : \ y \in \mathbb{Z}_q^{\leqslant \ell} \right\}. \tag{1}$$

In other words, $\Psi_{\ell,q}$ contains all possible compositions of sequences of length at most ℓ over $\mathbb{Z}_q$. By the definition, $|\Psi_{\ell,q}|$ is the number of vectors $(x_0, \ldots, x_{q-1}) \in \mathbb{Z}^q$ satisfying $\sum_{i=0}^{q-1} x_i \leqslant \ell$ and $x_i \geqslant 0$ for all $i \in [0, q-1]$. Evidently, $|\Psi_{\ell,q}| = \binom{\ell+q}{\ell}$. We then define the composition of the window of length ℓ that ends at position i by

$$R_{i,\ell}(x) \triangleq c(x[i - \ell + 1, i]),$$

for all $i \in [n - \ell + 1]$. We also observe that whenever $i \notin [n + \ell - 1]$ we have $x[i - \ell + 1, i] = \varepsilon$, and so $R_{i,\ell}(x) = 0$. We can now define the *ℓ-read sequence* of x as

$$\begin{aligned}
\mathcal{R}_\ell(x) &\triangleq (c(x[-\ell + 2, 1]), c(x[-\ell + 3, 2]), \\
&\qquad \ldots, c(x[n, n + \ell - 1])) \\
&= (R_{1,\ell}(x), R_{2,\ell}(x), \ldots, R_{n+\ell-1,\ell}(x)) \in \Psi_{\ell,q}^{n+\ell-1}.
\end{aligned}$$

Example 2: Assume that $q = 4$, namely, we are working over $\mathbb{Z}_4$. Let

$$x = (1, 2, 0, 1, 3, 1, 2, 2, 0, 0) \in \mathbb{Z}_4^{10}.$$

Take window size $\ell = 2$, so then

$$\begin{aligned}
\mathcal{R}_2(x) &= (z_1, z_1 + z_2, z_0 + z_2, z_0 + z_1, z_1 + z_3, \\
&\qquad z_1 + z_3, z_1 + z_2, 2z_2, z_0 + z_2, 2z_0, z_0) \in \Psi_{2,4}^{11}.
\end{aligned}$$

In the nanopore-channel setting, a sequence x is transmitted, while $\mathcal{R}_\ell(x)$ is received, assuming a perfect noiseless reading. However, as described in [12], backtracking errors may occur. When such an error occurs, the nanopore reading mechanism shifts back the molecule in the pore by k positions, therefore repeating the last k entries read, before continuing on the reading of the remainder of the molecule. These backtracking errors are easily described using the tandem-duplication error mechanism already used in [9] in a different context.

Generally speaking, assume $y \in \Sigma^*$ is some sequence over the finite alphabet $\Sigma = \Psi_{\ell,q}$. If a tandem-duplication error of length k occurs at position i, we denote the resulting erroneous sequence by

$$T_{i,k}(y) = uv^2w,$$

where $u, v, w \in \Sigma^*$, $y = uvw$, $|u| = i$, and $|v| = k$. Thus, this error repeats a k-factor of y located right after an i-prefix of y. We say t such errors occurred, resulting in the sequence z, if there exist $i_1, i_2, \ldots, i_t$ such that

$$z = T_{i_t,k}\left(T_{i_{t-1},k}\left(\ldots T_{i_1,k}(y)\right)\right).$$

We denote this relation between y and z by

$$y \Longrightarrow_k^t z,$$

and say that z is a *t-descendant* of y. If the number of errors is finite but unknown or unimportant, we write $y \Longrightarrow^* z$ and say that z is a *descendant* of y. For convenience, y is a 0-descendant of itself. The set of all *t-descendants* of y is denoted by

$$D_k^t(y) \triangleq \left\{ z \in \Sigma^* \ : \ y \Longrightarrow_k^t z \right\},$$

and similarly,

$$D_k^*(y) \triangleq \left\{ z \in \Sigma^* \ : \ y \Longrightarrow_k^* z \right\} = \bigcup_{j=0}^{\infty} D_k^j(y).$$

As is common in coding theory, we define the ball of radius t around y by

$$B_{k,t}(y) \triangleq \bigcup_{j=0}^{t} D_k^j(y),$$

and for an unbounded radius, we define

$$B_{k,\infty}(y) \triangleq D_k^*(y).$$

We now introduce the channel we shall focus on in this paper. In this channel model, a sequence $x \in \mathbb{Z}_q^n$ is transmitted by being stored as a DNA sequence. The reading is performed using nanopore sequencing, which ideally, should result in the ℓ-read vector $\mathcal{R}_\ell(x)$. However, since backtracking errors may occur, tandem duplications occur in $\mathcal{R}_\ell(x)$. We assume these errors are of length k. Assuming no more than t such errors occur, the receiver obtains $z \in B_{k,t}(\mathcal{R}_\ell(x))$. If the number of backtracking errors is unbounded, $z \in B_{k,\infty}(\mathcal{R}_\ell(x))$. In summary,

$$x \in \mathbb{Z}_q^n \xrightarrow{\ell\text{-read}} \mathcal{R}_\ell(x) \xrightarrow{\text{tandem duplication}} z \in B_{k,\infty}(\mathcal{R}_\ell(x)).$$

Example 3: We continue the setting of Example 2. Assume a single backtracking error occurred with length $k = 3$, resulting in a tandem duplication of length 3 in $\mathcal{R}_2(x)$, following a prefix of length 5,

$$\begin{aligned}
T_{5,3}(\mathcal{R}_2(x)) &= (z_1, z_1 + z_2, z_0 + z_2, z_0 + z_1, z_1 + z_3, z_1 + z_3, \\
&\qquad z_1 + z_2, 2z_2, \underline{z_1 + z_3, z_1 + z_2, 2z_2}, z_0 + z_2, \\
&\qquad 2z_0, z_0) \in B_{3,1}(\mathcal{R}_2(x)),
\end{aligned}$$

where the underline shows the duplicated 3-factor.

We emphasize that after a tandem duplication, the resulting vector of polynomials is not necessarily an ℓ-read vector of some sequence. Example 3 illustrates this point since there is no sequence x' such that $\mathcal{R}_2(x') = T_{5,3}(\mathcal{R}_2(x))$. In this particular case, this can be seen by the two adjacent entries in $T_{5,3}(\mathcal{R}_2(x))$ containing $2z_2, z_1 + z_3$ which can never be adjacent in a 2-read vector.

We shall design error-correcting codes for this channel, which we now define.

Definition 1: Let $\ell \in \mathbb{N}$ be the window read size, $k \in \mathbb{N}$ the tandem-duplication length, $t \in \mathbb{N}$ the maximum number of tandem-duplication errors (or $t = \infty$ if this number is unbounded), $n \in \mathbb{N}$ the code length, and $q \in \mathbb{N}$ the alphabet size. A code of length n is a subset $\mathcal{C} \subseteq \mathbb{Z}_q^n$. We say it can correct t tandem-duplication errors of length k in the ℓ-read nanopore channel if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, $\boldsymbol{x} \neq \boldsymbol{y}$, we have

$$B_{k,t}(\mathcal{R}_\ell(\boldsymbol{x})) \cap B_{k,t}(\mathcal{R}_\ell(\boldsymbol{y})) = \emptyset.$$

The *rate* of the code $\mathcal{C}$ is defined as

$$R(\mathcal{C}) \triangleq \frac{1}{n} \log_q |\mathcal{C}|,$$

and its *redundancy* is defined as $n(1 - R(\mathcal{C}))$.

III. CORRECTING AN UNBOUNDED NUMBER OF ERRORS

In this section we describe an optimal construction for codes that are capable of correcting an unbounded number of tandem duplications in the nanopore channel. The construction parallels the construction of [9], in the sense discussed in the introduction. While the construction is straightforward, and even though we can show it is optimal, it is difficult to find its asymptotic rate. The main challenge we face is the fact that our domain is complicated – the set of ℓ-read vectors. We manage to find the exact rate for some values of ℓ and k (the tandem-duplication length), and bound it for the other cases.

The main tool used in the construction is the k-step derivative. It was first employed by [9] to construct codes correcting any number of tandem duplications, and was later used in [10], [11], [13], [17], [19], [20], and [21].

Definition 2: Assuming G is some (additive) Abelian group, let $\boldsymbol{y} = (y_1, \ldots, y_m) \in G^m$ be a sequence of group elements of length m. For $k \geq 1$, the *k-step derivative of $\boldsymbol{y}$* is defined by

$$\Delta_k(\boldsymbol{y}) \triangleq (y_1 - y_{1-k}, y_2 - y_{2-k}, \ldots, y_m - y_{m-k}) \in G^m,$$

where for $i \leq 0$ we define $y_i = 0$.

Building on the definition of $\Psi_{\ell,q}$ from (1), we now define

$$\Xi_{\ell,q} \triangleq \left\{ c(\boldsymbol{x}) - c(\boldsymbol{y}) \ : \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_q^{\leq \ell} \right\},$$

where we emphasize that since $c(\cdot)$ has integer coefficients, the subtraction is over the integers. Namely, $\Xi_{\ell,q}$ contains the differences of any two polynomials from $\Psi_{\ell,q}$. We shall mostly be interested in the k-step derivative of the ℓ-read vector of a sequence[1]. Assume $\boldsymbol{x} \in \mathbb{Z}_q^n$ be a sequence, and let $\mathcal{R}_\ell(\boldsymbol{x}) = (R_{1,\ell}(\boldsymbol{x}), R_{2,\ell}(\boldsymbol{x}), \ldots, R_{n+\ell-1,\ell}(\boldsymbol{x})) \in \Psi_{\ell,q}^{n+\ell-1}$ be its ℓ-read vector. Then the k-step derivative of $\mathcal{R}_\ell(\boldsymbol{x})$ is simply

$$\begin{aligned}
\Delta_k(\mathcal{R}_\ell(\boldsymbol{x})) \triangleq &(R_{1,\ell}(\boldsymbol{x}) - R_{1-k,\ell}(\boldsymbol{x}), R_{2,\ell}(\boldsymbol{x}) - R_{2-k,\ell}(\boldsymbol{x}), \\
&\ldots, R_{n+\ell-1,\ell}(\boldsymbol{x}) - R_{n+\ell-k-1,\ell}(\boldsymbol{x})) \\
&\in \Xi_{\ell,q}^{n+\ell-1}.
\end{aligned}$$

[1] Here, the group G from the Definition 2 contains all the linear polynomials in the variables $z_0, \ldots, z_{q-1}$.

Example 4: Let us continue Example 2 and Example 3. Taking the 3-step derivative of $\mathcal{R}_2(\boldsymbol{x})$ and $T_{5,3}(\mathcal{R}_2(\boldsymbol{x}))$ we obtain

$$\begin{aligned}
\Delta_3(\mathcal{R}_2(\boldsymbol{x})) = (&z_1, z_1 + z_2, z_0 + z_2, z_0, z_3 - z_2, \\
&z_1 + z_3 - z_0 - z_2, z_2 - z_0, 2z_2 - z_1 - z_3, \\
&z_0 + z_2 - z_1 - z_3, 2z_0 - z_1 - z_2, \\
&z_0 - 2z_2), \\
\Delta_3(T_{5,3}(\mathcal{R}_2(\boldsymbol{x}))) = (&z_1, z_1 + z_2, z_0 + z_2, z_0, z_3 - z_2, \\
&z_1 + z_3 - z_0 - z_2, z_2 - z_0, 2z_2 - z_1 - z_3, \\
&0, 0, 0, z_0 + z_2 - z_1 - z_3, \\
&2z_0 - z_1 - z_2, z_0 - 2z_2).
\end{aligned}$$

In the derivative, the location of the tandem duplication contains a sequence of 0's.

We crucially observe that a tandem duplication of length k is equivalent to an injection of the factor 0^k into the derivative, at the same location (see Example 4 for an illustration of this point). This was already noted in [9].

Consider a sequence $\boldsymbol{y} \in \Xi_{\ell,q}^*$. We can always write it in the form

$$\boldsymbol{y} = (0^{m_0}, y_1, 0^{m_1}, y_2, 0^{m_2}, \ldots, y_s, 0^{m_s}), \tag{2}$$

where for all i, $y_i \in \Xi_{\ell,q} \setminus \{0\}$, and $m_i \geq 0$. We then define

$$\begin{aligned}
\mu_k(\boldsymbol{y}) \triangleq &(0^{m_0 \bmod k}, y_1, 0^{m_1 \bmod k}, y_2, 0^{m_2 \bmod k}, \ldots, \\
&y_s, 0^{m_s \bmod k}), \\
\sigma_k(\boldsymbol{y}) \triangleq &\left(\left\lfloor \frac{m_0}{k} \right\rfloor, \left\lfloor \frac{m_1}{k} \right\rfloor, \ldots, \left\lfloor \frac{m_s}{k} \right\rfloor \right). \tag{3}
\end{aligned}$$

Obviously, $\boldsymbol{y}$ is easily obtained from $\mu_k(\boldsymbol{y})$ and $\sigma_k(\boldsymbol{y})$.

Definition 3: Let $\boldsymbol{x} \in \mathbb{Z}_q^n$ be a sequence, and ℓ, k positive integers. We define the *nucleus* of $\boldsymbol{x}$ (w.r.t. the parameters k and ℓ) to be

$$N_{k,\ell}(\boldsymbol{x}) \triangleq \mu_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))).$$

Lemma 1: Let k, ℓ, and q be positive integers. Let $\boldsymbol{z} \in \Psi_{\ell,q}^*$, and assume $\boldsymbol{z} \Longrightarrow_k^t \boldsymbol{z}'$. Then

$$\mu_k(\Delta_k(\boldsymbol{z}')) = \mu_k(\Delta_k(\boldsymbol{z})).$$

Additionally, $\sigma_k(\Delta_k(\boldsymbol{z}')) - \sigma_k(\Delta_k(\boldsymbol{z}))$ contains only non-negative entries, and

$$\left\| \sigma_k(\Delta_k(\boldsymbol{z}')) - \sigma_k(\Delta_k(\boldsymbol{z})) \right\|_1 = t.$$

Proof: Consider first a single tandem duplication, namely, $t = 1$ and therefore $\boldsymbol{z} \Longrightarrow_k \boldsymbol{z}'$. By our previous observation, $\Delta_k(\boldsymbol{z}')$ differs from $\Delta_k(\boldsymbol{z})$ by an insertion of the factor 0^k. By definition, $\mu_k(\Delta_k(\boldsymbol{z}')) = \mu_k(\Delta_k(\boldsymbol{z}))$. Additionally, $\sigma_k(\Delta_k(\boldsymbol{z}'))$ differs from $\sigma_k(\Delta_k(\boldsymbol{z}))$ by a single entry that has increased by 1. The claim for general t follows by simple induction. ∎

With the last definition in place, we turn to show that for an unbounded number of tandem-duplication errors, the error balls around two sequences intersect if and only if they have the same nucleus.

Lemma 2: Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{Z}_q^*$ be two sequences, and let k and ℓ be two positive integers. Then

$$B_{k,\infty}(\mathcal{R}_\ell(\boldsymbol{x})) \cap B_{k,\infty}(\mathcal{R}_\ell(\boldsymbol{x}')) \neq \emptyset$$

if and only if $N_{k,\ell}(\boldsymbol{x}) = N_{k,\ell}(\boldsymbol{x}')$.

Proof: The proof is essentially the same as [9, Lemma 14]. For the first direction, assume that $z \in B_{k,\infty}(\mathcal{R}_\ell(\boldsymbol{x})) \cap B_{k,\infty}(\mathcal{R}_\ell(\boldsymbol{x}'))$. Thus,

$$\mathcal{R}_\ell(\boldsymbol{x}) \Longrightarrow_k^* z \qquad \text{and} \qquad \mathcal{R}_\ell(\boldsymbol{x}') \Longrightarrow_k^* z.$$

By Lemma 1,

$$\mu_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))) = \mu_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}'))).$$

Therefore, by definition

$$N_{k,\ell}(\boldsymbol{x}) = N_{k,\ell}(\boldsymbol{x}').$$

In the second direction, assume that as in (2) and (3),

$$N_{k,\ell}(\boldsymbol{x}) = N_{k,\ell}(\boldsymbol{x}') = (0^{m_0}, y_1, 0^{m_1}, y_2, 0^{m_2}, \ldots, y_s, 0^{m_s}),$$
$$\sigma_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))) = (a_0, a_1, \ldots, a_s),$$
$$\sigma_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}'))) = (b_0, b_1, \ldots, b_s).$$

Then,

$$\Delta_k(\mathcal{R}_\ell(\boldsymbol{x})) = (0^{m_0+a_0 k}, y_1, 0^{m_1+a_1 k}, y_2, 0^{m_2+a_2 k},$$
$$\ldots, y_s, 0^{m_s+a_s k}),$$
$$\Delta_k(\mathcal{R}_\ell(\boldsymbol{y})) = (0^{m_0+b_0 k}, y_1, 0^{m_1+b_1 k}, y_2, 0^{m_2+b_2 k},$$
$$\ldots, y_s, 0^{m_s+b_s k})$$

Since a tandem duplication of length k in $\mathcal{R}_\ell(\boldsymbol{x})$ corresponds to the injection of the factor 0^k in $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$ at the same position, the sequence

$$\Delta_k^{-1}((0^{m_0+(a_0+b_0)k}, y_1, 0^{m_1+(a_1+b_1)k},$$
$$y_2, \ldots, y_s, 0^{m_s+(a_s+b_s)k}))$$

is obtained by performing $\sum_{i=0}^{s} b_i$ tandem duplications in $\mathcal{R}_\ell(\boldsymbol{x})$ and it is also a sequence obtained by performing $\sum_{i=0}^{s} a_i$ tandem duplications in $\mathcal{R}_\ell(\boldsymbol{y})$. Therefore

$$\Delta_k^{-1}((0^{m_0+(a_0+b_0)k}, y_1, 0^{m_1+(a_1+b_1)k}, y_2, 0^{m_2+(a_2+b_2)k},$$
$$\ldots, y_s, 0^{m_s+(a_s+b_s)k})) \in B_{k,\infty}(\mathcal{R}_\ell(\boldsymbol{x})) \cap B_{k,\infty}(\mathcal{R}_\ell(\boldsymbol{x}')).$$

∎

The code construction follows immediately from this lemma.

Construction A: Let k, ℓ, n and $q \geqslant 2$ be positive integers. For any two sequences $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_q^n$, define the equivalence relation $\boldsymbol{x} \sim \boldsymbol{y}$ if and only if $N_{k,\ell}(\boldsymbol{x}) = N_{k,\ell}(\boldsymbol{y})$. Let $[\boldsymbol{x}]$ denote a fixed representative of the equivalence class containing $\boldsymbol{x}$. We construct the code:

$$\mathcal{C}_{k,\ell}(n) \triangleq \left\{ [\boldsymbol{x}] \ : \ \boldsymbol{x} \in \mathbb{Z}_q^n \right\},$$

namely, the code contains exactly one sequence from each of the equivalence classes of $\sim$.

Corollary 1: The code $\mathcal{C}_{k,\ell}(n)$ from Construction A is an optimal code capable of correcting any number of tandem duplications of length k in the ℓ-read vectors of its codewords.

Proof: The error-correction capability of the code is derived from Lemma 2 since its codewords have non-intersecting error balls. The code is optimal since it contains a codeword from each possible equivalence class. ∎

To better understand the efficiency of the code construction, our next goal is to find the asymptotic rate $\lim_{n\to\infty} R(\mathcal{C}_{k,\ell}(n))$.

The main obstacle here is the fact that the mapping from sequences to ℓ-read vectors, $\mathcal{R}_\ell : \mathbb{Z}_q^n \to \Psi_{\ell,q}^{n+\ell-1}$, is not always onto (see Example 4 and the short discussion following it). Further complications stem from the existence of two parameters k and ℓ that affect this asymptotic rate. We divide our discussion according to the values of these two parameters.

We begin with the case $\ell = 1$. In this case, we note that by identifying z_a with a, for all $a \in \mathbb{Z}_q$, we get $\mathcal{R}_1(\boldsymbol{x}) = \boldsymbol{x}$ and we sidestep the problem of $\mathcal{R}_\ell$ not being onto. Thus, Construction A reduces to [9, Construction A], whose asymptotic rate is known [9, eq. (4) and the Appendix], giving us

$$\lim_{n\to\infty} R(\mathcal{C}_{k,1}(n)) = \log_q \lambda_{k-1,q},$$

where $\lambda_{k,q}$ is the largest root of

$$f_{k,q}(x) \triangleq x^{k+2} - qx^{k+1} + q - 1. \tag{4}$$

We comment that $\log_q \lambda_{k-1,q}$ is in fact the capacity of the q-ary $(0, k-1)$-RLL constrained system (e.g., see [14]).

We divide some of the remaining non-trivial cases as follows. First, we study the case: $k = 1$ and $\ell \geqslant 2$ and obtain its asymptotic rate. We continue with the case of k being a multiple of ℓ. We carefully study the relationship between factors of the form 0^k in $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$ and the structure of $\boldsymbol{x}$. Finally, we use the probabilistic method to bound the asymptotic rate for the remaining cases.

As we proceed, we shall find the following definition useful.

Definition 4: Let ℓ, k be positive integers. For a sequence $\boldsymbol{x} \in \mathbb{Z}_q^*$, we define its *depth* (w.r.t. the parameters k and ℓ) as

$$d_{k,\ell}(\boldsymbol{x}) \triangleq \|\sigma_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x})))\|_1.$$

The set of all sequences of depth i will be denoted by

$$\mathcal{D}_{k,\ell}^{(i)} \triangleq \left\{ \boldsymbol{x} \in \mathbb{Z}_q^* \ : \ d_{k,\ell}(\boldsymbol{x}) = i \right\}.$$

We partition the code from Construction A, $\mathcal{C}_{1,\ell}(n)$, according to the depth of the codewords, and define for all $i \geqslant 0$,

$$\mathcal{C}_{k,\ell}^{(i)}(n) \triangleq \mathcal{C}_{k,\ell}(n) \cap \mathcal{D}_{k,\ell}^{(i)}.$$

Moreover, $\mathcal{C}_{k,\ell}^{(0)}$ consists of all sequences $\boldsymbol{x} \in \mathbb{Z}_q^*$ such that $\mathcal{R}_\ell(\boldsymbol{x})$ does not contain the factor 0^k. Since the ℓ-read vectors are of length $n+\ell-1$, we cannot have any codeword of depth larger than $\lfloor \frac{n+\ell-1}{k} \rfloor$, so we can certainly write

$$\mathcal{C}_{k,\ell}(n) = \bigcup_{i=0}^{\lfloor \frac{n+\ell-1}{k} \rfloor} \mathcal{C}_{k,\ell}^{(i)}(n). \tag{5}$$

Theorem 1: Assume that $\ell \geqslant 2$, and let $\mathcal{C}_{1,\ell}(n)$ be the code from Construction A. Then

$$\lim_{n\to\infty} R(\mathcal{C}_{1,\ell}(n)) = \log_q(q-1).$$

Proof: Fix a value for n. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{C}_{1,\ell}^{(i)}(n)$ for some $i \geqslant 0$. We explicitly compute

$$\Delta_1(\mathcal{R}_\ell(\boldsymbol{x})) = (z_{x_1}, \ldots, z_{x_\ell} \mid z_{x_{\ell+1}} - z_{x_1}, \ldots, z_{x_n} - z_{x_{n-\ell}} \mid$$
$$- z_{x_{n-\ell+1}}, \ldots, -z_{x_{n-1}}), \tag{6}$$

where the purpose of the vertical lines is notational only, allowing us to refer to the first, second, and third parts of

$\Delta_1(\mathcal{R}_\ell(\boldsymbol{x}))$. We note that a 0 may only occur at the middle part of the derivative, exactly when $x_i = x_{i-\ell}$. Additionally, the first two parts of $\Delta_1(\mathcal{R}_\ell(\boldsymbol{x}))$ equal to the ℓ-step derivative of $\mathcal{R}_1(\boldsymbol{x})$,

$$\Delta_\ell(\mathcal{R}_1(\boldsymbol{x})) = (z_{x_1}, \ldots, z_{x_\ell} \mid z_{x_{\ell+1}} - z_{x_1}, \ldots, z_{x_n} - z_{x_{n-\ell}}). \quad (7)$$

Let us now define

$$\mathcal{N}^{(i)} \triangleq \left\{ \mu_1(\Delta_\ell(\mathcal{R}_1(\boldsymbol{x}))) \;:\; \boldsymbol{x} \in \mathcal{C}_{1,\ell}^{(i)}(n) \right\}.$$

By (6) and (7) we have

$$\mu_1(\Delta_1(\mathcal{R}_\ell(\boldsymbol{x}))) = (\mu_1(\Delta_\ell(\mathcal{R}_1(\boldsymbol{x}))), -z_{x_{n-\ell+1}}, \ldots, -z_{x_{n-1}}).$$

It therefore follows that

$$\left| \mathcal{N}^{(i)} \right| \leqslant \left| \mathcal{C}_{1,\ell}^{(i)}(n) \right| \leqslant q^{\ell-1} \left| \mathcal{N}^{(i)} \right|. \quad (8)$$

To find the size of $\mathcal{N}^{(i)}$, we use the same approach as discussed in the case of $\ell = 1$. For each $a \in \mathbb{Z}_q$, we identify z_a with a. Then, the mapping $\Delta_\ell(\mathcal{R}_1(\cdot)) : \mathbb{Z}_q^n \to \mathbb{Z}_q^n$ is bijective, and we have

$$\mathcal{N}^{(i)} = \big\{ \boldsymbol{y} = (y_1, \ldots, y_{n-i}) \in \mathbb{Z}_q^{n-i} \;:\; y_j \neq 0$$
$$\text{for } \ell + 1 \leqslant j \leqslant n - i. \big\}$$

Thus,

$$\left| \mathcal{N}^{(i)} \right| = q^\ell (q-1)^{n-\ell-i} \leqslant q^\ell (q-1)^{n-\ell} = \left| \mathcal{N}^{(0)} \right|,$$

and, by (8),

$$\left| \mathcal{N}^{(0)} \right| \leqslant \left| \mathcal{C}_{1,\ell}^{(0)}(n) \right| \leqslant \left| \mathcal{C}_{1,\ell}(n) \right|$$
$$= \sum_{i=0}^{n+\ell-1} \left| \mathcal{C}_{1,\ell}^{(i)}(n) \right| \leqslant (n+\ell) q^{\ell-1} \left| \mathcal{N}^{(0)} \right|.$$

It follows that

$$\lim_{n\to\infty} R(\mathcal{C}_{1,\ell}(n)) = \lim_{n\to\infty} \frac{\log_q \left| \mathcal{N}^{(0)} \right|}{n} = \log_q(q-1).$$

∎

We continue with the remaining cases, namely, when $k, \ell \geqslant 2$. In order to bound the size $\mathcal{C}_{k,\ell}(n)$, we look at its partition into $\mathcal{C}_{k,\ell}^{(i)}$. In turn, to bound the size of $\mathcal{C}_{k,\ell}^{(i)}$ we need to better understand the number of sequences $\boldsymbol{x}$ with a prescribed number of occurrences of the factor 0^k in $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$. Let us first consider the existence of a single 0 at position $i+\ell+k-1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$. This position contains $R_{i+\ell+k-1,\ell}(\boldsymbol{x}) - R_{i+\ell-1,\ell}(\boldsymbol{x})$. We note that the subtracted value, $R_{i+\ell-1,\ell}(\boldsymbol{x})$, is the composition of $\boldsymbol{x}[i, i + \ell - 1]$. If the result of the subtraction is to be 0, and assuming a long enough sequence, we must have $i \geqslant 1$, or else the subtracted window overlaps only partially with $\boldsymbol{x}$. Thus, in order to get a 0, we must have

$$R_{i+\ell-1,\ell}(\boldsymbol{x}) = R_{i+\ell+k-1,\ell}(\boldsymbol{x}). \quad (9)$$

Now, to get a factor 0^k in $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$, we need to require (9) for k consecutive indices, namely,

$$R_{i+\ell-1+j,\ell}(\boldsymbol{x}) = R_{i+\ell+k-1+j,\ell}(\boldsymbol{x}) \text{ for all } j \in [0, k-1]. \quad (10)$$

The implications of this are detailed in the following lemma.

Lemma 3: Let $k, \ell \geqslant 2$ and $i \geqslant 1$ be positive integers. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{Z}_q^n$. If 0^k occurs starting at position $i+\ell+k-1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$, then one of the following holds:
1) $\ell \leqslant k$, and there exists a subset $\mathcal{I} \subseteq [0, \ell - 1]$ such that for all $r \in \left[0, \lfloor \frac{k+\ell-s-2}{\ell} \rfloor \right]$,
    a) if $s \in [0, k-2] \cap \mathcal{I}$,

$$x_{i+s+r\ell} = x_{i+k+s+r\ell},$$

    b) if $s \in [0, k-2] \cap ([0, \ell - 1] \setminus \mathcal{I})$,

$$x_{i+s} = x_{i+s+r\ell} \text{ and } x_{i+k+s} = x_{i+k+s+r\ell}.$$

    If additionally $\ell | k$, then only $\mathcal{I} = [0, \ell - 1]$ is possible, and $\boldsymbol{x}[i, i + 2k + \ell - 2]$ is k-periodic.
2) $\ell > k$, and there exists a subset $\mathcal{I} \subseteq [0, k - 2]$ such that,
    a) if $s \in \mathcal{I}$,

$$x_{i+s} = x_{i+\ell+s} \text{ and } x_{i+k+s} = x_{i+k+\ell+s},$$

    b) if $s \in [0, k - 2] \setminus \mathcal{I}$,

$$x_{i+s} = x_{i+k+s} \text{ and } x_{i+\ell+s} = x_{i+k+\ell+s}.$$

Proof: Assume that 0^k occurs starting at position $i+\ell+k-1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$. We then have

$$R_{i+\ell-1+j,\ell}(\boldsymbol{x}) = R_{i+k+\ell-1+j,\ell}(\boldsymbol{x}) \text{ for all } j \in [0, k-1]. \quad (11)$$

By definition, we have $R_{m-k,\ell}(\boldsymbol{x}) \neq R_{m,\ell}(\boldsymbol{x})$ for all $m \geqslant n + 1$. Then $i+k+\ell-1+k-1 = i+2k+\ell-2 \leqslant n$. These equalities involve the symbols of $\boldsymbol{x}[i, i + 2k + \ell - 2]$. Since

$$R_{i+\ell-1+j,\ell}(\boldsymbol{x}) = c(x_{i+j}, \ldots, x_{i+\ell-1+j})$$
$$= c(x_{i+k+j}, \ldots, x_{i+k+\ell-1+j})$$
$$= R_{i+k+\ell-1+j,\ell}(\boldsymbol{x}),$$
$$R_{i+\ell+j,\ell}(\boldsymbol{x}) = c(x_{i+1+j}, \ldots, x_{i+\ell+j})$$
$$= c(x_{i+k+1+j}, \ldots, x_{i+k+\ell+j})$$
$$= R_{i+k+\ell+j,\ell}(\boldsymbol{x}),$$

for any $j \in [0, k-2]$, we have

$$\{\{x_{i+j}, x_{i+k+\ell+j}\}\} = \{\{x_{i+\ell+j}, x_{i+k+j}\}\}, \quad (12)$$

where $\{\ldots\}$ denotes a multiset. We divide our proof depending on ℓ and k.

Case 1: $\ell \leqslant k$. We define the following set of integers:

$$\mathcal{I} \triangleq \big\{ s \in [0, \ell - 1] \;:\; x_{i+s} = x_{i+k+s} \big\}$$

By using (12), we obtain the following properties:
1) For all $s \in [0, k-2] \cap \mathcal{I}$, $x_{i+s} = x_{i+k+s}$, which implies $x_{i+\ell+s} = x_{i+k+\ell+s}$.
2) For all $s \in [0, k-2] \cap ([0, \ell-1] \setminus \mathcal{I})$, $x_{i+s} \neq x_{i+k+s}$, which implies $x_{i+s} = x_{i+\ell+s}$ and $x_{i+k+s} = x_{i+k+\ell+s}$. We also have $x_{i+\ell+s} \neq x_{i+k+\ell+s}$.

We can iterate this process by comparing the set $\{x_{i+s+r\ell} : s \in [0, \ell-1]\}$ with $\{x_{i+k+s+r\ell} : s \in [0, \ell-1]\}$ for $1 \leqslant r \leqslant \lfloor \frac{k+\ell-s-2}{\ell} \rfloor$, and get:
    a) For all $s \in [0, k-2] \cap \mathcal{I}$, $x_{i+s} = x_{i+k+s}$ and $x_{i+s+r\ell} = x_{i+k+s+r\ell}$ for all $1 \leqslant r \leqslant \lfloor \frac{k+\ell-s-2}{\ell} \rfloor$.
    b) For all $s \in [0, k-2] \cap ([0, \ell-1] \setminus \mathcal{I})$, $x_{i+s} = x_{i+s+r\ell}$ and $x_{i+k+s} = x_{i+k+s+r\ell}$ for all $1 \leqslant r \leqslant \lfloor \frac{k+\ell-s-2}{\ell} \rfloor$,

where (a) and (b) are iterated according to properties (1) and (2), respectively.

In addition, we consider the sub-case of $\ell | k$. Assume to the contrary that $\mathcal{I} \neq [0, \ell-1]$. Recall, from (11), that $R_{i+\ell-1,\ell}(\boldsymbol{x}) = R_{i+k+\ell-1,\ell}(\boldsymbol{x})$. This means that $x_i, \ldots, x_{i+\ell-1}$ is a permutation of $x_{i+k}, \ldots, x_{i+k+\ell-1}$. Thus, if $\mathcal{I} \neq [0, \ell-1]$, there exists one position, say j, such that $x_{j+k} \neq x_{j+k+\ell-1}$, then there exists another position, say m, such that $x_{m+k} \neq x_{m+k+\ell-1}$. Thus, we cannot have $|\mathcal{I}| = \ell - 1$, and we must have $|\mathcal{I}| \leqslant \ell - 2$. Thus, there exists $s \in [0, \ell-2]$ such that $x_{i+s} \neq x_{i+k+s}$. Note that, necessarily, $s \in [0, k-2] \cap ([0, \ell-1] \setminus \mathcal{I})$. This, however, contradicts b since we can choose $r = \frac{k}{\ell} \leqslant \lfloor \frac{k+\ell-s-2}{\ell} \rfloor$, and we get $x_{i+s} = x_{i+s+r\ell} = x_{i+k+s}$.

Thus, when $\ell | k$ we have $\mathcal{I} = [0, \ell-1]$. Then a implies that $x_{i+j+k} = x_{i+j}$ for all $j \in [0, k+\ell-2]$. Hence, $\boldsymbol{x}[i, i+2k+\ell-2]$ is k-periodic.

Case 2: $\ell > k$. We once again look at (11). Taking $j = 0$,

$$R_{i+\ell-1,\ell}(\boldsymbol{x}) = c(x_i, \ldots, x_{i+\ell-1})$$
$$= c(x_{i+k}, \ldots, x_{i+k+\ell-1})$$
$$= R_{i+k+\ell-1,\ell}(\boldsymbol{x}).$$

Since $\ell \geqslant k+1$, we can omit the common part $x_{i+k}, \ldots, x_{i+\ell-1}$ and obtain,

$$\{\{x_i, \ldots, x_{i+k-1}\}\} = \{\{x_{i+\ell}, \ldots, x_{i+\ell+k-1}\}\}.$$

Let us now define

$$\mathcal{I} \triangleq \{s \in [0, k-2] \ : \ x_{i+s} = x_{i+\ell+s}\},$$

where we emphasize the fact that we do not consider the case of $s = k - 1$. Therefore, by using (12):

1) For all $s \in \mathcal{I}$, $x_{i+s} = x_{i+\ell+s}$ and $x_{i+k+s} = x_{i+k+\ell+s}$.
2) For all $s \in [0, k-2] \setminus \mathcal{I}$, $x_{i+s} = x_{i+k+s}$ and $x_{i+\ell+s} = x_{i+k+\ell+s}$.

This completes all the cases of the theorem. ∎

To further understand Lemma 3, we provide the following example.

Example 5:
1) For $(\ell, k) = (2, 3), q = 3$, let

$$\boldsymbol{v} \in \mathbb{Z}_3^n \text{ with } \boldsymbol{v}[i, i+6] = (0, 1, 2, 0, 1, 2, 0)$$

and

$$\boldsymbol{w} \in \mathbb{Z}_3^n \text{ with } \boldsymbol{w}[j, j+6] = (0, 1, 0, 1, 0, 1, 0)$$

be two sequences. Then 0^3 occurs at position $i + 4$ of $\Delta_3(\mathcal{R}_2(\boldsymbol{v}))$ with the set $\mathcal{I} = \{0, 1\}$, and 0^3 occurs at position $j + 4$ of $\Delta_3(\mathcal{R}_2(\boldsymbol{w}))$ with the set $\mathcal{I} = \emptyset$.
2) For $(\ell, k) = (2, 2), q = 3$, let

$$\boldsymbol{x} \in \mathbb{Z}_3^n \text{ with } \boldsymbol{x}[i, i+4] = (0, 1, 0, 1, 0)$$

be a sequence. Then 0^2 occurs at position $i + 3$ of $\Delta_2(\mathcal{R}_2(\boldsymbol{x}))$ with the set $\mathcal{I} = \{0, 1\}$.
3) For $(\ell, k) = (3, 2), q = 3$, let

$$\boldsymbol{y} \in \mathbb{Z}_3^n \text{ with } \boldsymbol{y}[i, i+5] = (0, 1, 2, 0, 1, 2)$$

and

$$\boldsymbol{z} \in \mathbb{Z}_3^n \text{ with } \boldsymbol{z}[j, j+5] = (0, 1, 0, 1, 0, 1)$$

be two sequences. Then 0^2 occurs at position $i + 4$ of $\Delta_2(\mathcal{R}_3(\boldsymbol{y}))$ with the set $\mathcal{I} = \{0\}$, and 0^3 occurs at position $j + 4$ of $\Delta_2(\mathcal{R}_3(\boldsymbol{z}))$ with the set $\mathcal{I} = \emptyset$.

Lemma 3 shows the importance of sequences with no periodic factor, with proper parameters k and ℓ. We formalize such sequences in the following definition.

Definition 5: Let $k, \ell \geqslant 2$, and let $\boldsymbol{x} \in \mathbb{Z}_q^n$ be a sequence. We say that $\boldsymbol{x}$ is (k, ℓ)-*fine* if it does not contain any k-periodic factor of length $2k+\ell-1$. Namely, either the sequence is short $n \leqslant 2k+\ell-2$, or $n \geqslant 2k+\ell-1$ and for all $1 \leqslant i \leqslant n-2k-\ell+2$, $\boldsymbol{x}[i, i+2k+\ell-2]$ is not k-periodic. We denote the set of all (k, ℓ)-fine sequences of length n by $\mathcal{F}_{k,\ell}(n)$.

Some simple monotonicity is easy to show.

Lemma 4: Let $k, \ell \geqslant 2$, and $n \geqslant 2k+\ell-1$. Then

$$\left| \mathcal{F}_{k,\ell}(n) \right| \leqslant \left| \mathcal{F}_{k,\ell}(n+1) \right|.$$

Proof: We simply show that any sequence in $\mathcal{F}_{k,\ell}(n)$ can be extended by one symbol to a sequence in $\mathcal{F}_{k,\ell}(n+1)$. Let $\boldsymbol{x} \in \mathcal{F}_{k,\ell}(n)$, and let $a \in \mathbb{Z}_q$ be a symbol such that $a \neq \boldsymbol{x}[n-k+1]$ (which is always possible since $q \geqslant 2$, namely, the alphabet contains at least two letters). We contend that $\boldsymbol{x}a \in \mathcal{F}_{k,\ell}(n+1)$. Obviously any factor of length $2k+\ell-1$ of $\boldsymbol{x}a$ that does not involve the last symbol, is not k-periodic since $\boldsymbol{x} \in \mathcal{F}_{k,\ell}(n)$. The only factor in question is therefore $\boldsymbol{x}[n-2k-\ell+3, n]a$, but since $a \neq \boldsymbol{x}[n-k+1]$, it is certainly not k-periodic. ∎

Using these tools, we can now bound the rate of the code we constructed.

Theorem 2: Let $k, \ell \geqslant 2$, and let $\mathcal{C}_{k,\ell}(n)$ be the code from Construction A. Then

$$\lim_{n \to \infty} R(\mathcal{C}_{k,\ell}^{(0)}(n)) \leqslant \lim_{n \to \infty} R(\mathcal{C}_{k,\ell}(n)) \leqslant \lim_{n \to \infty} R(\mathcal{F}_{k,\ell}(n)).$$

Moreover, if $\ell | k$ then,

$$\lim_{n \to \infty} R(\mathcal{C}_{k,\ell}(n)) = \lim_{n \to \infty} R(\mathcal{C}_{k,\ell}^{(0)}(n)) = \lim_{n \to \infty} R(\mathcal{F}_{k,\ell}(n)).$$

Proof: The lower bound is trivial since $\mathcal{C}_{k,\ell}^{(0)}(n) \subseteq \mathcal{C}_{k,\ell}(n)$. We therefore turn to prove the upper bound. We first contend that $\mathcal{C}_{k,\ell}^{(0)} \subseteq \mathcal{F}_{k,\ell}(n)$ when $n \geqslant 2k+\ell-1$. Assume to the contrary that is not the case, namely, that there exists $\boldsymbol{x} \in \mathcal{C}_{k,\ell}^{(0)}(n)$ such that $\boldsymbol{x} \notin \mathcal{F}_{k,\ell}(n)$. Thus, $\boldsymbol{x}$ contains a factor of length $2k+\ell-1$ that is k-periodic, say, $\boldsymbol{x}[i, i+2k+\ell-2]$. Then $x_{i+s} = x_{i+k+s}$ for all $0 \leqslant s \leqslant k+\ell-2$. Hence, $R_{i+\ell-1+t,\ell}(\boldsymbol{x}) = R_{i+k+\ell-1+t,\ell}(\boldsymbol{x})$ for $0 \leqslant t \leqslant k-1$. That would mean that $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$ contains 0^k starting at position $i+k+\ell-1$, which contradicts the definition of $\mathcal{C}_{k,\ell}^{(0)}$. Thus, $\mathcal{C}_{k,\ell}^{(0)}(n) \subseteq \mathcal{F}_{k,\ell}(n)$ and then

$$\left| \mathcal{C}_{k,\ell}^{(0)}(n) \right| \leqslant \left| \mathcal{F}_{k,\ell}(n) \right|.$$

Now, let us consider $\boldsymbol{x} \in \mathcal{C}_{k,\ell}^{(i)}(n)$ for $i \geqslant 1$. If $\boldsymbol{x}$ is (k, ℓ)-fine, namely, then $\boldsymbol{x} \in \mathcal{F}_{k,\ell}(n)$. Otherwise, $\boldsymbol{x}$ has a factor of length $2k+\ell-1$ that is k-periodic. Assume the first such factor is $(x_i, \ldots, x_{i+2k+\ell-2})$. We then define $\mathrm{Del}_{k,\ell}(\boldsymbol{x})$ as

$$\mathrm{Del}_{k,\ell}(\boldsymbol{x}) \triangleq M_{[i,i+k-1]}(\boldsymbol{x}) = \{x_1, \ldots, x_{i-1}, x_{i+k}, \ldots, x_n\},$$

namely, $\mathrm{Del}_{k,\ell}(\boldsymbol{x})$ deletes the first k symbols of the first factor of length $2k+\ell-1$ that is k-periodic. We call this operation a *k-periodic deletion at position i*.

Turning our focus to $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$, the k-periodic factor of length $2k+\ell-1$ at position i of $\boldsymbol{x}$, manifests as a 0^k

factor at position $i + k + \ell - 1$ at $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$. Let us define $\mathrm{Rem}_{k,\ell}(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x})))$ as the operator removing the said 0^k factor. Thus,

$$
\begin{aligned}
\mathrm{Rem}_{k,\ell}(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))) &= M_{[i+k+\ell-1,i+2k+\ell-2]}(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))) \\
&= (R_{1,\ell}(\boldsymbol{x}) - R_{1-k,\ell}(\boldsymbol{x}), \\
&\qquad \ldots, \\
&\qquad R_{i+k+\ell-2,\ell}(\boldsymbol{x}) - R_{i+\ell-2,\ell}(\boldsymbol{x}), \\
&\qquad R_{i+2k+\ell-1,\ell}(\boldsymbol{x}) - R_{i+k+\ell-1,\ell}(\boldsymbol{x}), \\
&\qquad \ldots, \\
&\qquad R_{n+\ell-1,\ell}(\boldsymbol{x}) - R_{n+\ell-k-1,\ell}(\boldsymbol{x})).
\end{aligned}
$$

We call this the 0^k *removal at position* $i + k + \ell - 1$.

We contend the following claim.

Claim 1:

$$
\mathrm{Rem}_{k,\ell}(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))) = \Delta_k(\mathcal{R}_\ell(\mathrm{Del}_{k,\ell}(\boldsymbol{x}))). \tag{13}
$$

Namely, that deleting the first factor of $\boldsymbol{x}$ that has length $2k + \ell - 1$ and is k-periodic (which is responsible for a 0^k factor in the derivative) and then taking the k-step derivative of the ℓ-read vector, is the same as first taking the k-step derivative of the ℓ-read vector and then removing the said 0^k factor.

Proof of claim: The proof of this claim is by brute-force inspection: For convenience, denote

$$
\begin{aligned}
\mathcal{R}_\ell(\boldsymbol{x}) &= (R_1, R_2, \ldots, R_{n+\ell-1}), \\
\mathcal{R}_\ell(\mathrm{Del}_{k,\ell}(\boldsymbol{x})) &= (R'_1, R'_2, \ldots, R'_{n+\ell-k-1}).
\end{aligned}
$$

Then:

- $R'_j \overset{(a)}{=} R_j$ for $1 \leqslant j \leqslant i - 1$,
- $R'_j = c(x_{j-\ell+1}, x_{j-\ell+2}, \cdots, x_{i-1}, x_{i+k}, \ldots, x_{j+k}) \overset{(b)}{=} c(x_{j-\ell+1}, x_{j-\ell+2}, \cdots, x_{i-1}, x_i, \ldots, x_j) = R_j$ for $i \leqslant j \leqslant i + \ell - 2$,
- $R'_j \overset{(c)}{=} R_j$ for $i + \ell - 1 \leqslant j \leqslant i + k + \ell - 2$,
- $R'_j \overset{(d)}{=} R_{j+k}$ for $i + \ell - 1 \leqslant j \leqslant n + \ell - k - 1$,

where $(a), (d)$ come from the definition and $(b), (c)$ come from the form of the k-periodic factor $(x_i, \ldots, x_{i+2k+\ell-2})$. Hence:

- $R'_j - R'_{j-k} = R_j - R_{j-k}$ for $1 \leqslant s \leqslant i + k + \ell - 2$,
- $R'_j - R'_{j-k} = R_{j+k} - R_j$ for $i + k + \ell - 1 \leqslant j \leqslant n + \ell - k - 1$,

thus proving the claim. ∎

An important consequence of (13) is that the nucleus of $\boldsymbol{x}$ (see Definition 3) remains unchanged under $\mathrm{Del}_{k,\ell}(\cdot)$, namely,

$$
N_{k,\ell}(\mathrm{Del}_{k,\ell}(\boldsymbol{x})) = N_{k,\ell}(\boldsymbol{x}).
$$

We can therefore repeatedly apply $\mathrm{Del}_{k,\ell}$ until we reach a (k, ℓ)-fine sequence. We denote that sequence by $\mathrm{Del}^*_{k,\ell}(\boldsymbol{x})$. Obviously,

$$
N_{k,\ell}(\mathrm{Del}^*_{k,\ell}(\boldsymbol{x})) = N_{k,\ell}(\boldsymbol{x}). \tag{14}
$$

Additionally,

$$
\mathrm{Del}^*_{k,\ell}(\boldsymbol{x}) \in \bigcup_{j=0}^{\lfloor n/k \rfloor} \mathcal{F}_{k,\ell}(n - jk). \tag{15}
$$

We observe that

$$
\begin{aligned}
\left| \mathcal{C}^{(i)}_{k,\ell}(n) \right| &\overset{(a)}{=} \left| \left\{ N_{k,\ell}(\boldsymbol{x}) \ : \ \boldsymbol{x} \in \mathcal{C}^{(i)}_{k,\ell}(n) \right\} \right| \\
&\overset{(b)}{=} \left| \left\{ N_{k,\ell}(\mathrm{Del}^*_{k,\ell}(\boldsymbol{x})) \ : \ \boldsymbol{x} \in \mathcal{C}^{(i)}_{k,\ell}(n) \right\} \right|
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(c)}{\leqslant} \left| \left\{ \mathrm{Del}^*_{k,\ell}(\boldsymbol{x}) \ : \ \boldsymbol{x} \in \mathcal{C}^{(i)}_{k,\ell}(n) \right\} \right| \\
&\overset{(d)}{\leqslant} \sum_{j=0}^{\lfloor n/k \rfloor} \left| \mathcal{F}_{k,\ell}(n - jk) \right| \\
&\overset{(e)}{\leqslant} n |\mathcal{F}_{k,\ell}(n)|,
\end{aligned}
$$

where (a) follows from the code definition in Construction A, (b) follows from (14), (c) follows from the fact that several sequences may have the same nucleus, (d) follows from (15), and (e) follows from Lemma 4 and $k \geqslant 2$. Thus, by also using (5),

$$
\left| \mathcal{C}_{k,\ell}(n) \right| = \sum_{i=0}^{\lfloor \frac{n+\ell-1}{k} \rfloor} \left| \mathcal{C}^{(i)}_{k,\ell}(n) \right| \leqslant n^2 |\mathcal{F}_{k,\ell}(n)|.
$$

Therefore,

$$
\lim_{n \to \infty} R(\mathcal{C}_{k,\ell}(n)) \leqslant \lim_{n \to \infty} R(\mathcal{F}_{k,\ell}(n)),
$$

as claimed.

Finally, assume the case of $\ell | k$. By Lemma 3, every factor 0^k of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$ implies the existence of a k-periodic factor of $\boldsymbol{x}$ of length $2k + \ell - 1$. Thus, not only $\mathcal{C}^{(0)}_{k,\ell} \subseteq \mathcal{F}_{k,\ell}(n)$ as in the general case, but necessarily $\mathcal{C}^{(0)}_{k,\ell} = \mathcal{F}_{k,\ell}(n)$. By sandwiching we get

$$
\lim_{n \to \infty} R(\mathcal{C}_{k,\ell}(n)) = \lim_{n \to \infty} R(\mathcal{C}^{(0)}_{k,\ell}(n)) = \lim_{n \to \infty} R(\mathcal{F}_{k,\ell}(n)),
$$

which completes the proof. ∎

We proceed to obtain closed-form bounds on the asymptotic rate of $\mathcal{C}_{k,\ell}(n)$ from Construction A. For the upper bound, we note that $\mathcal{F}_{k,\ell}(n)$ is in fact related to run-length-limited (RLL) sequences, whose asymptotic rate is already known. As for the lower bound, we use the probabilistic method, and in particular, the Lovász Local Lemma in one of its many forms.

Lemma 5 ([1, Lemma 5.1.1]): Let $A_1, \ldots, A_m$ be events in an arbitrary probability space. Let $G = (V, E)$ be a directed graph with $V = [m]$ such that for every $i \in [m]$, the event A_i is mutually independent of all the events $\{A_j : (i, j) \notin E\}$. Suppose that there are real numbers $c_1, \ldots, c_m$ such that $0 \leqslant c_i < 1$ and $\Pr(A_i) \leqslant c_i \prod_{(i,j) \in E} (1 - c_j)$ for all $i \in [m]$. Then

$$
\Pr\left(\bigwedge_{i \in [m]} \overline{A_i} \right) \geqslant \prod_{i \in [m]} (1 - c_i).
$$

For any sequence $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{Z}_q^n$, deciding whether 0^k occurs starting at position $i + \ell + k - 1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$ depends only on $\boldsymbol{x}[i, i + 2k + \ell - 2]$. To analyze this, we introduce the following technical lemma to estimate the number of sequences $\boldsymbol{x}[i, i + 2k + \ell - 2] \in \Sigma_q^{2k+\ell-1}$ such that 0^k occurs starting at position $i + \ell + k - 1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$.

Lemma 6: Let $k, \ell \geqslant 2$ and $i \geqslant 1$ be positive integers. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{Z}_q^n$. If 0^k occurs starting at position $i + \ell + k - 1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$, then the number of possible values for $\boldsymbol{x}[i, i + 2k + \ell - 2]$ is upper bounded by

$$
\leqslant \begin{cases} \ell! q^k & \text{if } \ell \leqslant k, \\ k! q^\ell & \text{if } \ell > k. \end{cases}
$$

Proof: Assume that 0^k occurs starting at position $i+\ell+k-1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$. We then have (10) as follows,

$$R_{i+\ell-1+j,\ell}(\boldsymbol{x}) = R_{i+k+\ell-1+j,\ell}(\boldsymbol{x}) \text{for all } j \in [0, k-1],$$

which involves exactly the symbols of $\boldsymbol{x}[i, i+2k+\ell-2]$. We first contend that if we know the value of $\boldsymbol{x}[i, i+k+\ell-1]$, then there is exactly one way of completing $\boldsymbol{x}[i+k+\ell, i+2k+\ell-2]$ such that (10) holds. Indeed, since $R_{i+\ell,\ell}(\boldsymbol{x}) = R_{i+\ell+k,\ell}(\boldsymbol{x})$ by (10), we have

$$\begin{aligned}\{\{x_{i+k+\ell}\}\} &= \left(\{\{x_{i+k}\}\} \cup \{\{x_{i+k+1}\ldots, x_{i+k+\ell}\}\}\right) \setminus \\ &\quad \{\{x_{i+k}, \ldots, x_{i+k+\ell-1}\}\} \\ &= \left(\{\{x_{i+k}\}\} \cup \{\{x_{i+1}\ldots, x_{i+\ell}\}\}\right) \setminus \\ &\quad \{\{x_{i+k}, \ldots, x_{i+k+\ell-1}\}\},\end{aligned}$$

which determines the value of $x_{i+k+\ell}$ using the known values of $\boldsymbol{x}[i, i+k+\ell-1]$. Iterating this by increasing the indices, we determine the values of $\boldsymbol{x}[i+k+\ell, i+2k+\ell-2]$. We therefore focus on bounding the number of values $\boldsymbol{x}[i, i+k+\ell-1]$ can take.

For the first case, assume $\ell \leqslant k$. There are q^k ways of choosing $\boldsymbol{x}[i, i+k-1]$. By (10), $R_{i+\ell-1,\ell}(\boldsymbol{x}) = R_{i+k+\ell-1,\ell}(\boldsymbol{x})$, so

$$\{\{x_i, \ldots, x_{i+\ell-1}\}\} = \{\{x_{i+k}, \ldots, x_{i+k+\ell-1}\}\}.$$

Since $\ell \leqslant k$, the two multisets do not overlap. Thus, $\boldsymbol{x}[i, i+\ell-1]$ is the same as $\boldsymbol{x}[i+k, i+k+\ell-1]$ up to a permutation of the ℓ symbols. There are $\ell!$ ways of choosing a permutation, and we obtain the claimed upper bound, $\ell! q^k$.

For the second case, assume $\ell > k$. There are q^ℓ ways of choosing $\boldsymbol{x}[i, i+\ell-1]$. By (10), $R_{i+\ell-1,\ell}(\boldsymbol{x}) = R_{i+k+\ell-1,\ell}(\boldsymbol{x})$, so

$$\{\{x_i, \ldots, x_{i+\ell-1}\}\} = \{\{x_{i+k}, \ldots, x_{i+k+\ell-1}\}\}.$$

Because $\ell > k$, the two multisets overlap in $\{x_{i+k}, \ldots, x_{i+\ell-1}\}$. Removing this overlap we get

$$\{\{x_i, \ldots, x_{i+k-1}\}\} = \{\{x_{i+\ell}, \ldots, x_{i+k+\ell-1}\}\}.$$

Thus, $\boldsymbol{x}[i, i+k-1]$ is the same as $\boldsymbol{x}[i+\ell, i+k+\ell-1]$ up to a permutation of the k symbols. There are $k!$ ways of choosing a permutation, and we obtain the claimed upper bound, $k! q^\ell$. ∎

Theorem 3: Let $k, \ell \geqslant 2$, and let $\mathcal{C}_{k,\ell}(n)$ be the code from Construction A for alphabet size $q \geqslant 2$. Then

$$\lim_{n\to\infty} R(\mathcal{C}_{\ell,k}(n)) \leqslant \log_q \lambda_{k+\ell-2,q}.$$

If additionally $\ell | k$ then

$$\lim_{n\to\infty} R(\mathcal{C}_{\ell,k}(n)) = \log_q \lambda_{k+\ell-2,q},$$

where $\lambda_{k+\ell-2,q}$ is defined in (4).

Proof: Our starting point is the upper bound presented in Theorem 2. It is shown using the asymptotic rate of $\mathcal{F}_{k,\ell}(n)$. Here, we simply find the asymptotic rate of $\mathcal{F}_{k,\ell}(n)$. Using the same approach as before, we take any $\boldsymbol{x} \in \mathbb{Z}_q^n$ and look at $\Delta_k(\mathcal{R}_1(\boldsymbol{x}))$. Since we use the 1-read vector of $\boldsymbol{x}$, we identify the polynomial z_a with the element a, for any $a \in \mathbb{Z}_q$. Thus, $\Delta_k(\mathcal{R}_1(\boldsymbol{x}))$ is just a sequence over $\mathbb{Z}_q$ as well. We also note

that $\boldsymbol{x}$ contains a factor of length $2k+\ell-1$ that is k-periodic, if and only if $\Delta_k(\mathcal{R}_1(\boldsymbol{x}))$ contains a factor $0^{k+\ell-1}$. Thus, $\mathcal{F}_{k,\ell}(n)$ contains exactly those sequences whose k-step derivative of their 1-read vector does not contain the factor $0^{k+\ell-1}$.

The sequences in $\mathbb{Z}_q^n$ that do not contain the factor $0^{k+\ell-1}$ are called q-ary $(0, k+\ell-2)$-run-length-limited (RLL) sequences, and are denoted by $\mathrm{RLL}_q(0, k+\ell-2)(n)$. Since there is a clear bijection between $\mathbb{Z}_q^n$ and its image under $\Delta_k(\mathcal{R}_1(\cdot))$ we get

$$\left|\mathcal{F}_{k,\ell}(n)\right| = \left|\mathrm{RLL}_q(0, k+\ell-2)(n)\right|.$$

Thus,

$$\begin{aligned}\lim_{n\to\infty} R(\mathcal{F}_{k,\ell}(n)) &= \lim_{n\to\infty} R(\mathrm{RLL}_q(0, k+\ell-2)(n)) \\ &= \log_q \lambda_{k+\ell-2,q},\end{aligned}$$

where the last equality is from [9, (4) and the Appendix].

If $\ell | k$ then by Theorem 2 the asymptotic rate of $\mathcal{C}_{k,\ell}(n)$ equals the asymptotic rate of $\mathcal{F}_{k,\ell}(n)$, which equals the asymptotic rate of $\mathrm{RLL}_q(0, k+\ell-2)(n)$. ∎

Remark 1: We can give more pleasant upper and lower bounds on $\log_q \lambda_{k+\ell-2,q}$ that appears in Theorem 3. By [9, (4) and the Appendix] we can write

$$1 - \frac{\alpha(q-1)\log_q e}{q^{k+\ell}} \leqslant \log_q \lambda_{k+\ell-2,q} \leqslant 1 - \frac{(q-1)\log_q e}{q^{k+\ell}},$$

where

$$\alpha = e^{-W\left(-\frac{q-1}{q^{k+\ell}}(k+\ell)\right)},$$

and $W(\cdot)$ is the Lambert W-function. As $k + \ell \to \infty$ we have $\alpha \to 1$.

Theorem 4: Let $k, \ell \geqslant 2$, and let $\mathcal{C}_{k,\ell}(n)$ be the code from Construction A. Then

$$\lim_{n\to\infty} R(\mathcal{C}_{k,k}(n)) \geqslant 1 - \log_q\left(\frac{1}{1-c}\right).$$

If $\ell \leqslant k$, then

$$c = \frac{1 - \sqrt{1 - \frac{\ell!(16k+8\ell-16)}{q^{k+\ell-1}}}}{8k+4\ell-8},$$

and we require

$$q \geqslant \left\lceil (\ell!(16k + 8\ell - 16)^{\frac{1}{k+\ell-1}} \right\rceil.$$

If $\ell > k$,

$$c = \frac{1 - \sqrt{1 - \frac{k!(16k+8\ell-16)}{q^{2k-1}}}}{8k+4\ell-8},$$

and we require

$$q \geqslant \left\lceil (k!(16k + 8\ell - 16)^{\frac{1}{2k-1}} \right\rceil.$$

Proof: Our starting point, once again, is the lower bound presented in Theorem 2, which is shown using the asymptotic rate of $\mathcal{C}_{k,\ell}^{(0)}(n)$. We lower bound this asymptotic rate by using the Lovász Local Lemma.

Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{Z}_q^n$ be a random sequence, where for each $i \in [1, n]$, X_i is i.i.d. uniformly at random chosen from $\mathbb{Z}_q$. Thus, $\boldsymbol{X}$ is uniformly at random chosen from $\mathbb{Z}_q^n$. Let A_i be the event that $\Delta_k(\mathcal{R}_\ell(\boldsymbol{X}))$ has 0^k at position $i + \ell + k - 1$, where $1 \leqslant i \leqslant n - 2k - \ell + 2$. Let $G = (V, E)$ be a directed

TABLE I

THE ASYMPTOTIC RATE OF THE OPTIMAL CODE CORRECTING ANY NUMBER OF TANDEM
DUPLICATION OF LENGTH k IN THE NANOPORE ℓ-READ VECTOR

k	ℓ	Asymptotic Rate	Comment
$k \geqslant 2$	1	$= \log_q \lambda_{k-1,q}$	[9, (4)] (see also (4))
1	$\ell \geqslant 2$	$= \log_q(q-1)$	Theorem 12
$k \geqslant 2$	$\ell \geqslant 2$	$= \log_q \lambda_{k+\ell-2,q}$	Theorem 21, $\ell \mid k$ (see also (4) and Remark 22)
		$\leqslant \log_q \lambda_{k+\ell-2,q}$	Theorem 21 (see also (4) and Remark 22)
		$\geqslant 1 - \log_q\left(\frac{1}{1-c}\right)$ where $c = \frac{1 - \sqrt{1 - \frac{\ell!(16k+8\ell-16)}{q^{k+\ell-1}}}}{8k+4\ell-8}$	Theorem 23, $\ell \leqslant k$, $q \geqslant \left\lceil (\ell!(16k+8\ell-16))^{\frac{1}{k+\ell-1}} \right\rceil$
		$\geqslant 1 - \log_q\left(\frac{1}{1-c}\right)$ where $c = \frac{1 - \sqrt{1 - \frac{k!(16k+8\ell-16)}{q^{2k-1}}}}{8k+4\ell-8}$	Theorem 23, $\ell > k$, $q \geqslant \left\lceil (k!(16k+8\ell-16))^{\frac{1}{2k-1}} \right\rceil$
		$\geqslant 1 - \frac{1}{k} \log_q\left(\frac{q}{q-1}\right)$	Remark 24

graph with $V = [n - 2k - \ell + 2]$ and $(i, j) \in E$ if and only if $|i - j| \leqslant 2k + \ell - 2$. By the observation in the proof of Lemma 6, A_i depends on $X[i, i+2k+\ell-2]$ only, and so the event A_i is mutually independent of all the events $\{A_j \ : \ (i, j) \notin E\}$.

By Lemma 6,

$$\Pr(A_i) \leqslant \begin{cases} \frac{\ell! q^k}{q^{2k+\ell-1}} = \frac{\ell!}{q^{k+\ell-1}} & \text{if } \ell \leqslant k, \\ \frac{k! q^\ell}{q^{2k+\ell-1}} = \frac{k!}{q^{2k-1}} & \text{if } \ell > k. \end{cases} \quad (16)$$

Additionally,

$$\Pr\left(X \in \bigwedge_{i \in [n-2k-\ell+2]} \overline{A_i}\right) = \Pr\left(X \in \mathcal{C}_{k,\ell}^{(0)}\right) = \frac{|\mathcal{C}_{k,\ell}^{(0)}|}{q^n}. \quad (17)$$

For the first case, assume $\ell \leqslant k$. Define

$$c \triangleq \frac{1 - \sqrt{1 - \frac{\ell!(16k+8\ell-16)}{q^{k+\ell-1}}}}{8k+4\ell-8},$$

and for all $i \in [n - 2k - \ell + 2]$ set $c_i = c$. We then have for all i,

$$c_i \prod_{(i,j) \in E} (1 - c_j) \overset{(a)}{\geqslant} c(1-c)^{4k+2\ell-4}$$

$$\overset{(b)}{\geqslant} c(1 - (4k + 2\ell - 4)c)$$

$$= \frac{\ell!}{q^{k+\ell-1}} \overset{(c)}{\geqslant} \Pr(A_i),$$

where (a) follows from the fact that each vertex is connected to at most $4k + 2\ell - 4$ other vertices, (b) follows from Bernoulli's inequality $(1 - x)^r \geqslant 1 - rx$ for $r \geqslant 1$ and $x \in (0, 1)$, and (c) follows from (16). Note that $q^{k+\ell-1} \geqslant \ell!(16k+8\ell-16)$ for any $q \geqslant \lceil (\ell!(16k+8\ell-16))^{\frac{1}{k+\ell-1}} \rceil$, which makes sure that $c_i \in (0, 1)$.

By Lemma 5 and (17), we have

$$\left| \mathcal{C}_{k,\ell}^{(0)} \right| \geqslant q^n (1 - c)^{n-2k-\ell+2},$$

and therefore,

$$\lim_{n \to \infty} R(\mathcal{C}_{k,\ell}^{(0)}) \geqslant \lim_{n \to \infty} \frac{n + (n - 2k - \ell + 2) \log_q(1 - c)}{n}$$

$$= 1 - \log_q\left(\frac{1}{1-c}\right).$$

For the second case, when $k < \ell$, we choose

$$c \triangleq \frac{1 - \sqrt{1 - \frac{k!(16k+8\ell-16)}{q^{2k-1}}}}{8k+4\ell-8},$$

and $q \geqslant \lceil (k!(16k + 8\ell - 16))^{\frac{1}{2k-1}} \rceil$. The remainder of the proof is the same. ∎

Remark 2: While requirement on the alphabet size q in Theorem 4 is modest, we may wish to remove it to obtain a lower bound on the asymptotic rate for small alphabet sizes. We can do so, at the price of a lower rate guarantee, in the following fashion. We can construct sequences $\mathbf{x} \in \mathcal{C}_{k,\ell}^{(0)}$ by picking arbitrary symbols in all locations except those that are 0 modulo k. In those positions (the positions i that is congruent to 0 (mod k)), we choose x_i in a way that $R_{i,\ell}(\mathbf{x}) \neq R_{i-k,\ell}(\mathbf{x})$. Then $\mathbf{x} \in \mathcal{C}_{k,\ell}^{(0)}$, otherwise there exists a 0^k factor in $\Delta_k(\mathcal{R}_\ell(\mathbf{x}))$. Specifically, there exists some $j \in [n]$ such that $R_{j+s,\ell}(\mathbf{x}) = R_{j+k+s,\ell}(\mathbf{x})$ for all $s \in [0, k-1]$. Consequently, the set $\{j + k + s : s \in [0, k-1]\}$ must include at least one index that is 0 (mod k), which contradicts our construction. This removes at most one possible letter from the alphabet, per such position. We then get for all $q \geqslant 2$,

$$\lim_{n \to \infty} R(\mathcal{C}_{k,\ell}^{(0)}) \geqslant 1 - \frac{1}{k} \log_q\left(\frac{q}{q-1}\right).$$

The results of this section are summarized in Table I. Table II shows the bounds for common values associated with nanopore sequencing.

IV. CORRECTING A CONSTANT NUMBER OF ERRORS

In this section we turn to look at the case of a constant number of tandem-duplication error of length k. As discussed in the introduction, we shall be paralleling a construction of such a code, bringing it to the domain of ℓ-read vectors. Unlike the previous section, we shall parallel the construction

k	ℓ	Lower Bound	Upper Bound
1	5	= 0.792481	
2	5	0.896241	0.999868
3	5	0.995182	0.999967
4	5	0.998906	0.999992
5	5	= 0.999998	
6	5	0.999917	0.999999
7	5	0.999979	1.000000
8	5	0.999995	1.000000
9	5	0.999999	1.000000
1	9	= 0.792481	
2	9	0.896241	0.999999
3	9	0.994779	1.000000
4	9	0.998891	1.000000
5	9	0.999664	1.000000
6	9	0.999875	1.000000
7	9	0.999946	1.000000
8	9	0.999973	1.000000
9	9	= 1.000000	

given in [11]. As in the previous section, the challenge is not the construction itself, but rather finding its redundancy, and whether it is optimal.

The construction of [11], which we parallel here, uses Sidon sets. Let G be a finite Abelian group. A set $B = \{b_1, \ldots, b_r\} \subseteq G$ is said to be a *Sidon set of order t* if the sums $\sum_{i=1}^{r} a_i b_i$ have different values for all $a_1, \ldots, a_r \in \mathbb{Z}$ with $a_i \geqslant 0$ and $\sum_{i=1}^{r} a_i \leqslant t$. By the result of [4], we can find such an Abelian group of size $|G| = O(r^t)$, where we assume $r \to \infty$ while t is fixed. More precisely, by [4, Theorem 2], for r a prime power, we can take $G = \mathbb{Z}_q$ with $q = (r^{t+1} - 1)/(r - 1)$.

Let us introduce a useful notation. given two vectors of integers $\boldsymbol{a} \in \mathbb{Z}^m$ and $\boldsymbol{a}' \in \mathbb{Z}^{m'}$, $m, m' \geqslant 1$, not necessarily of the same length, we define their dot product as

$$\boldsymbol{a} \odot \boldsymbol{a}' \triangleq \sum_{i=1}^{\min(m,m')} a_i a_i'.$$

With this notation we can proceed with the construction.

Construction B: Let k, ℓ, t, n, and $q \geqslant 2$ be positive integers. Let G be an Abelian group containing a Sidon set $B = \{b_1, \ldots, b_{n+\ell}\}$ of order t. Define $\boldsymbol{b} \triangleq (b_1, \ldots, b_{n+\ell})$. For any $g \in G$ we construct the code

$$\mathcal{C}_{k,\ell}^g(n) \triangleq \left\{ \boldsymbol{x} \in \mathbb{Z}_q^n \ : \ \boldsymbol{b} \odot \sigma_k(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))) = g \right\},$$

where we recall the definition of $\sigma_k(\cdot)$ from (3).

Theorem 5: The code $\mathcal{C}_{k,\ell}^g(n)$ from Construction B can correct any t or fewer tandem duplications of length k in the ℓ-read vectors of its codewords.

Proof: We prove the claim by giving a decoding procedure. Assume $\boldsymbol{x} \in \mathcal{C}_{k,\ell}^g(n)$ was transmitted, and denote its ℓ-read vector by $\boldsymbol{z} = \mathcal{R}_\ell(\boldsymbol{x})$. By construction,

$$\boldsymbol{b} \odot \sigma_k(\Delta_k(\boldsymbol{z})) = g.$$

Assume, however, that $\boldsymbol{z}'$ is the ℓ-read vector obtained at the channel output, $\boldsymbol{z} \Longrightarrow_k^i \boldsymbol{z}'$, where $0 \leqslant i \leqslant t$. We can compute

$$\boldsymbol{b} \odot \sigma_k(\Delta_k(\boldsymbol{z}')) = g'.$$

By Lemma 1, $\sigma_k(\Delta_k(\boldsymbol{z}')) - \sigma_k(\Delta_k(\boldsymbol{z}))$ contains only non-negative entries, and

$$\|\sigma_k(\Delta_k(\boldsymbol{z}')) - \sigma_k(\Delta_k(\boldsymbol{z}))\|_1 = i.$$

Thus, by the properties of the Sidon set, knowing the value of the element

$$g' - g = \boldsymbol{b} \odot (\sigma_k(\Delta_k(\boldsymbol{z}')) - \sigma_k(\Delta_k(\boldsymbol{z})))$$

gives us exactly the entries of $\sigma_k(\Delta_k(\boldsymbol{z}')) - \sigma_k(\Delta_k(\boldsymbol{z}))$. We can therefore recover $\sigma_k(\Delta_k(\boldsymbol{z}))$, which together with $\mu_k(\Delta_k(\boldsymbol{z}')) = \mu_k(\Delta_k(\boldsymbol{z}))$ immediately allows us to recover $\boldsymbol{z}$. ∎

Corollary 2: Let k, ℓ, t, n, and $q \geqslant 2$ be positive integers. Then there exists a code $\mathcal{C}$ of length n over $\mathbb{Z}_q$ that can correct any t or fewer tandem duplications of length k in the ℓ-read vectors of its codewords, and has redundancy $t \log_q n + O(1)$.

Proof: By [4, Theorem 1], there exists a Sidon set of order t with $n + \ell$ elements over an Abelian group G of size $|G| = O(n^t)$. Consider the codes $\mathcal{C}_{k,\ell}^g$ from Construction B. By Theorem 5 they all possess the desired error-correction capability.

We note that

$$\bigcup_{g \in G} \mathcal{C}_{k,\ell}^g(n) = \mathbb{Z}_q^n.$$

By averaging, there exists $g \in G$ such that

$$\left| \mathcal{C}_{k,\ell}^g(n) \right| \geqslant \frac{|\mathbb{Z}_q^n|}{|G|} = \frac{q^n}{O(n^t)}.$$

By definition this code has redundancy $t \log_q n + O(1)$. ∎

We now turn to lower bound the redundancy required for a code that is capable of correcting t tandem duplications of length k in the ℓ-read vector. Unlike the upper bound obtained through Construction B, it appears to be difficult to get a lower bound. We do so for the case of $\ell | k$, and show that it matches the upper bound of Corollary 2 up to an additive number of $O(1)$ symbols.

Theorem 6: Let k, ℓ, t, n, and $q \geqslant 2$ be positive integers. Assume that $\ell | k$. Then any code $\mathcal{C}$ of length n over $\mathbb{Z}_q$ that can correct any t or fewer tandem duplications of length k in the ℓ-read vectors of its codewords, has redundancy at least $t \log_q n + O(1)$.

Proof: Assume $\mathcal{C}$ is a code as stated in the theorem. Then $\mathcal{R}_\ell(\mathcal{C}) \triangleq \{\mathcal{R}_\ell(\boldsymbol{x}) \ : \ \boldsymbol{x} \in \mathcal{C}\}$ is a set capable of correcting up to t duplications of length k. By the previous discussion, this means that $\Delta_k(\mathcal{R}_\ell(\mathcal{C})) \triangleq \{\Delta_k(\mathcal{R}_\ell(\boldsymbol{x})) \ : \ \boldsymbol{x} \in \mathcal{C}\}$ is a code capable of correcting up to t insertions of 0^k. By [11, Lemma 1], $\Delta_k(\mathcal{R}_\ell(\mathcal{C}))$ can correct up to t deletions of 0^k.

For us to use the deletion-correction capability stated above, we need to bound the number of sequences $\boldsymbol{x} \in \mathbb{Z}_q^n$ such that $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$ contains sufficiently many maximal runs of zeros of length at least k (which allow a deletion of 0^k to occur). At this point we start using the assumption that $\ell | k$, and we make the following observations. Assume $\boldsymbol{x} \in \mathbb{Z}_q^n$ is a sequence such that 0^k occurs at position $i + \ell + k - 1$ of $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$. By

Lemma 3, $\boldsymbol{x}[i, i + 2k + \ell - 2]$ is k-periodic. This means that $\Delta_k(\boldsymbol{x})$ (notice that we take the derivative of $\boldsymbol{x}$ here, and not of the ℓ-read vector of $\boldsymbol{x}$) contains $0^{k+\ell-1}$ at position $i + k$. Since the other direction is trivial, we get

$$\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))[i + k + \ell - 1, i + 2k + \ell - 2] = 0^k$$
$$\Updownarrow$$
$$\boldsymbol{x}[i, i + 2k + \ell - 2] \text{ is } k\text{-periodic}$$
$$\Updownarrow$$
$$\Delta_k(\boldsymbol{x})[i + k, i + 2k + \ell - 2] = 0^{k+\ell-1}. \quad (18)$$

We also note that since all relevant transformations are invertible,

$$\left|\mathbb{Z}_q^n\right| = \left|\Delta_k(\mathbb{Z}_q^n)\right| = \left|\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^n))\right| = q^n.$$

Another crucial fact is that, when $\ell | k$, if we delete the 0^k factor from $\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))$, what remains is indeed a k-step derivative of an ℓ-read vector of some sequence, specifically, by the definition and the k-periodicity of $\boldsymbol{x}[i, i + 2k + \ell - 2]$, it follows that

$$\Delta_k(\mathcal{R}_\ell(M_{[i,i+k-1]}(\boldsymbol{x}))) = M_{[i+k+\ell-1, i+2k+\ell-2]}(\Delta_k(\mathcal{R}_\ell(\boldsymbol{x}))). \quad (19)$$

We shall need the following notation. Assume Σ is some finite alphabet that contains the symbol 0. If $\boldsymbol{x} \in \Sigma^n$ is some sequence, we can uniquely factor in the following way:

$$\boldsymbol{x} = 0^{m_0} a_1 0^{m_1} a_2 0^{m_2} \ldots a_s 0^{m_s},$$

for some non-negative integers m_i, and where $a_i \in \Sigma \setminus \{0\}$. We say $\boldsymbol{x}$ contains m maximal runs of zeros of length $\geqslant r$ if there are exactly m indices i for which $m_i \geqslant r$. The set of all such sequences is denoted by $S_m^{\geqslant r}(\Sigma^n)$. Back to our problem, we shall be interested in the size of $S_m^{\geqslant k+\ell-1}(\mathbb{Z}_q^n)$.

By [11, Lemma 3][2], there exists a sub-linear function $f(n) = o(n)$ such that

$$\sum_{|m-\mu n| > f(n)} \left|S_m^{\geqslant k+\ell-1}(\mathbb{Z}_q^n)\right| \leqslant \frac{q^n}{n^{\log_2 n}},$$

where $\mu \triangleq \frac{q+1}{q^{k+\ell}}$. Loosely speaking, most sequences contain around μn maximal runs of zeros of length $\geqslant k + \ell - 1$. By (18), we have

$$\sum_{|m-\mu n| > f(n)} \left|S_m^{\geqslant k}(\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^n)))\right| \leqslant \frac{q^n}{n^{\log_2 n}}, \quad (20)$$

Let $\mathcal{C}' \subseteq \bigcup_{|m-\mu n| \leqslant f(n)} S_m^{\geqslant k}(\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^n)))$ be a code capable of correcting up to t deletions of 0^k. Since each $\boldsymbol{x} \in \mathcal{C}'$ contains at least $\mu n - f(n)$ maximal runs of zeros of length $\geqslant k$, after exactly t deletions of 0^k, the number of resulting sequences is at least $\binom{\mu n - f(n)}{t}$. Additionally, by (19), after these t deletions the sequence resides in $\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^{n-tk}))$. Since after the deletions, all sequences must be distinct,

$$|\mathcal{C}'| \binom{\mu n - f(n)}{t} \leqslant \left|\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^{n-tk}))\right|.$$

[2]We mention that [11, Lemma 3] contains an extra weight restriction which is of no consequence here.

Rearranging, and using $\binom{a}{b} \geqslant a^b/b^b$,

$$|\mathcal{C}'| \leqslant \frac{q^{n-tk} t^t}{(\mu n - f(n))^t}. \quad (21)$$

Finally, assume $\mathcal{C} \subseteq \mathbb{Z}_q^n$ is a code satisfying the conditions of the theorem. We have $|\mathcal{C}| = |\Delta_k(\mathcal{R}_\ell(\mathcal{C}))|$. The latter we partition into two parts:

$$\mathcal{C}_1' \triangleq \Delta_k(\mathcal{R}_\ell(\mathcal{C})) \cap \bigcup_{|m-\mu n| \leqslant f(n)} S_m^{\geqslant k}(\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^n))),$$
$$\mathcal{C}_2' \triangleq \Delta_k(\mathcal{R}_\ell(\mathcal{C})) \cap \bigcup_{|m-\mu n| > f(n)} S_m^{\geqslant k}(\Delta_k(\mathcal{R}_\ell(\mathbb{Z}_q^n))).$$

By (20) and (21) we have

$$|\mathcal{C}_1'| \leqslant \frac{q^{n-tk} t^t}{(\mu n - f(n))^t},$$
$$|\mathcal{C}_2'| \leqslant \frac{q^n}{n^{\log_2 n}}.$$

Thus,

$$|\mathcal{C}| = |\mathcal{C}_1'| + |\mathcal{C}_2'| \leqslant \frac{q^{n-tk} t^t}{(\mu n - f(n))^t} + \frac{q^n}{n^{\log_2 n}} = O\left(\frac{q^n}{n^t}\right),$$

and the redundancy of $\mathcal{C}$ is at least $t \log_q n + O(1)$. ∎

Remark 3: We note that if we remove the restriction of $\ell | k$, the proof of Theorem 6 may break. For example, we show a counter-example to (19). Assume $q = 2$, $\ell = 2$, $k = 3$, and $n = 7$. Consider the sequence $\boldsymbol{x} = (0, 1, 0, 1, 0, 1, 0)$. Then,

$$\mathcal{R}_2(\boldsymbol{x}) = (\mathsf{z}_0, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1,$$
$$\mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0),$$
$$\Delta_3(\mathcal{R}_2(\boldsymbol{x})) = (\mathsf{z}_0, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_1, 0, 0, 0, -\mathsf{z}_1).$$

Let us now delete the 0^3 factor in $\Delta_3(\mathcal{R}_2(\boldsymbol{x}))$ to obtain

$$\boldsymbol{z}' = (\mathsf{z}_0, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_1, -\mathsf{z}_1).$$

If we compute $\Delta_3^{-1}(\boldsymbol{z}')$ we get

$$\Delta_3^{-1}(\boldsymbol{z}') = (\mathsf{z}_0, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0 + \mathsf{z}_1, \mathsf{z}_0).$$

However, $\Delta_3^{-1}(\boldsymbol{z}')$ is not the 2-read vector of any binary sequence. Thus, $\boldsymbol{z}' \notin \Delta_3(\mathcal{R}_2(\mathbb{Z}_2^4))$.

V. Conclusion

In this paper we studied the asymptotic rate of the optimal code capable of correcting tandem duplications of length k in the ℓ-read vector of the nanopore channel. The number of correctable tandem duplications was either unbounded (in Section III), or bounded by a constant (in Section IV).

We observe that in the case of unbounded tandem duplication errors, the redundancy is linear in the length of the code n. This channel generalizes the channel described in [9]. Both the lower and upper bounds tend to 1 exponentially fast as $\ell, k \to \infty$. As Table II shows, even for practical values, the rate is nearly 1.

When the number of tandem duplications is bounded by t errors and $\ell | k$, our bounds match asymptotically, giving a redundancy of $t \log_q n + O(1)$. It is interesting to note that this expression does not depend on ℓ and k, which agrees with the

special case ($\ell = 1$) studied in [11]. However, when $\ell \nmid k$, we only have $t \log_q n + O(1)$ as an upper bound on the redundancy.

Some open questions remain. In particular, we are interested in finding explicit constructions for asymptotically optimal codes. The special case of $\ell = 1$ studied for unbounded number of errors in [9] does provide an explicit construction. In our case, there seems to be an interesting connection between the parameters k and ℓ, making it difficult to give an explicit construction similar to [9]. Thus, while the construction we describe is optimal, more work is required to flesh out an explicit one. When we restrict ourselves to a bounded number of errors, already [11] has an optimal construction, but that one is not explicit.

Another open question concerns the length of the duplication errors, i.e., the amount of backtracking in the nanopore sequencing. The codes we propose rely on codes that correct duplication errors of a fixed length. This is the case in almost all constructions for duplication-correcting codes. Constructing codes capable of correcting multiple duplications of varying lengths seems to require considerable more effort, which we leave for future work.

ACKNOWLEDGMENT

REFERENCES

[1] N. Alon and J. H. Spencer, *The Probabilistic Method*. Hoboken, NJ, USA: Wiley, 2016.

[2] A. Banerjee, Y. Yehezkeally, A. Wachter-Zeh, and E. Yaakobi, "Correcting a single deletion in reads from a nanopore sequencer," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Athens, Greece, Jul. 2024, pp. 103–108.

[3] A. Banerjee, Y. Yehezkeally, A. Wachter-Zeh, and E. Yaakobi, "Error-correcting codes for nanopore sequencing," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 4956–4967, Jul. 2024.

[4] R. C. Bose, S. Chowla, and B. Singer, "Theorems in the additive theory of numbers," Contract, 1960.

[5] Y. M. Chee, K. A. S. Immink, and V. K. Vu, "Coding scheme for noisy nanopore sequencing with backtracking and skipping errors," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Athens, Greece, Jul. 2024, pp. 458–463.

[6] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, Sep. 2012.

[7] F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of string-duplication systems," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 811–824, Feb. 2016.

[8] D. Goshkoder, N. Polyanskii, and I. Vorobyev, "Codes correcting a single long duplication error," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Taipei, Taiwan, Jun. 2023, pp. 2708–2713.

[9] S. Jain, F. Farnoud Hassanzadeh, M. Schwartz, and J. Bruck, "Duplication-correcting codes for data storage in the DNA of living organisms," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4996–5010, Aug. 2017.

[10] M. Kovacevic, "Zero-error capacity of duplication channels," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6735–6742, Oct. 2019.

[11] M. Kovacevic and V. Y. F. Tan, "Asymptotically optimal codes correcting fixed-length duplication errors in DNA storage systems," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2194–2197, Nov. 2018.

[12] A. H. Laszlo et al., "Decoding long nanopore sequencing reads of natural DNA," *Nature Biotechnol.*, vol. 32, no. 8, pp. 829–833, Aug. 2014.

[13] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Duplication-correcting codes," *Designs, Codes Cryptography*, vol. 87, nos. 2–3, pp. 277–298, Mar. 2019.

[14] D. Lind and B. H. Marcus, *An Introduction To Symbolic Dynamics and Coding*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[15] W. Mao, S. N. Diggavi, and S. Kannan, "Models and information-theoretic bounds for nanopore sequencing," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 3216–3236, Apr. 2018.

[16] Y. Sun and G. Ge, "Bounds and constructions of ℓ-read codes under the Hamming metric," 2024, *arXiv:2403.11754*.

[17] Y. Tang and F. Farnoud, "Error-correcting codes for noisy duplication channels," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3452–3463, Jun. 2021.

[18] Y. Tang, S. Wang, H. Lou, R. Gabrys, and F. Farnoud, "Low-redundancy codes for correcting multiple short-duplication and edit errors," *IEEE Trans. Inf. Theory*, vol. 69, no. 5, pp. 2940–2954, May 2023.

[19] Y. Tang, Y. Yehezkeally, M. Schwartz, and F. Farnoud, "Single-error detection and correction for duplication and substitution channels," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6908–6919, Nov. 2020.

[20] Y. Yehezkeally and M. Schwartz, "Reconstruction codes for DNA sequences with uniform tandem-duplication errors," *IEEE Trans. Inf. Theory*, vol. 66, no. 5, pp. 2658–2668, May 2020.

[21] W. Yu and M. Schwartz, "On duplication-free codes for disjoint or equal-length errors," *Designs, Codes Cryptography*, vol. 92, no. 10, pp. 2845–2861, Oct. 2024.

[22] M. Zeraatpisheh, M. Esmaeili, and T. A. Gulliver, "Construction of tandem duplication correcting codes," *IET Commun.*, vol. 13, no. 15, pp. 2217–2225, Sep. 2019.

Wenjun Yu received the Ph.D. degree in mathematics from the University of Science and Technology of China, Hefei, Anhui, China, in 2021. He is currently holding a postdoctoral position with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev. His research interests include combinatorics, graph theory, coding theory, and their interactions.

Zuo Ye (Member, IEEE) received the Ph.D. degree in mathematics from the University of Science and Technology of China, Hefei, Anhui, China, in 2021. He is currently holding a postdoctoral position with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev. His research interests include combinatorics, coding theory, and their interactions.

Moshe Schwartz (Fellow, IEEE) received the B.A. (summa cum laude), M.Sc., and Ph.D. degrees from the Computer Science Department, Technion-Israel Institute of Technology, Haifa, Israel, in 1997, 1998, and 2004, respectively.

He was a Fulbright Post-Doctoral Researcher with the Department of Electrical and Computer Engineering, University of California San Diego, and a Post-Doctoral Researcher with the Department of Electrical Engineering, California Institute of Technology. While on sabbatical 2012–2014, he was a Visiting Scientist with Massachusetts Institute of Technology (MIT). He is currently a Professor with the Department of Electrical and Computer Engineering, McMaster University, Canada, and a Professor with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel (on a leave of absence). His research interests include algebraic coding, combinatorial structures, and digital sequences.

Prof. Schwartz received the 2009 IEEE Communications Society Best Paper Award in Signal Processing and Coding for Data Storage and the 2020 NVMW Persistent Impact Prize. He served as an Associate Editor for Coding Techniques and Coding Theory for IEEE TRANSACTIONS ON INFORMATION THEORY from 2014 to 2021 and since 2021, he has been serving as an Area Editor for Coding and Decoding for IEEE TRANSACTIONS ON INFORMATION THEORY. He has been an Editorial Board Member of the *Journal of Combinatorial Theory, Series A* since 2021.