A Practical Concatenated Coding Scheme for Noisy Shuffling Channels with Coset-based Indexing

Javad Haghighat[1] and Tolga M. Duman[2]
[1]*EEE Department, TED University,* Ankara, Turkey, javad.haghighat@tedu.edu.tr
[2]*EEE Department, Bilkent University,* Ankara, Turkey, duman@ee.bilkent.edu.tr

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437131

***Abstract*—Noisy shuffling channels capture the main characteristics of DNA storage systems where distinct segments of data are received out of order, after being corrupted by substitution errors. For realistic schemes with short-length segments, practical indexing and channel coding strategies are required to restore the order and combat the channel noise. In this paper, we develop a finite-length concatenated coding scheme that employs Reed-Solomon (RS) codes as outer codes and polar codes as inner codes, and utilizes an implicit indexing method based on cosets of the polar code. We propose a matched decoding method along with a metric for detecting the index that successfully restores the order, and correct channel errors at the receiver. Residual errors that are not corrected by the matched decoder are then corrected by the outer RS code. We derive analytical approximations for the frame error rate of the proposed scheme, and also evaluate its performance through simulations to demonstrate that the proposed implicit indexing method outperforms explicit indexing.**

I. Introduction

DNA storage systems are receiving significant attention from the research community, thanks to their longevity and their impressive storage density [1]- [6]. The basic idea in DNA storage is to employ a synthesizer that takes information bits as input and maps them to synthetic DNA strands. However, due to technical limitations in current synthesizing technologies, synthetic strands are limited to a few hundreds of nucleotides in length. Therefore, data has to be divided into short segments that are then written on short strands and are stored in a solution known as the DNA pool.

The DNA pool has a fundamental disadvantage compared to other storage environments such as disks and magnetic tapes; that is, DNA pool is not able to maintain the order of the stored strands (since the strands are floating in a solution and their physical position cannot be fixed). Consequently, when the information is being read from the pool, there is no guarantee that the strands are sequenced (read) in the same order as they are synthesized (written). In short, the output of the sequencer is a shuffled version of the strands generated by the synthesizer. Furthermore, errors are likely to occur during the synthesis, during the storage, and while sequencing the strands. For this reason, the end-to-end channel between the original data and the output of the sequencer may be modeled as a noisy shuffling channel [7]. The noisy shuffling channel model may be further modified by noting that the strands are amplified (copied several times) inside the DNA pool, via the Polymerase Chain Reaction (PCR) process. Hence, the sequencing process of a randomly selected subset of the stored strands is more accurately represented by a noisy shuffling-sampling channel model, in which each strand is sampled a random number of times. These channel models and their variations are studied in a number of recent papers including [7]- [17].

Since the ordering of the segments is not maintained by the noisy shuffling channel, a mechanism has to be implemented at the transmitter (i.e., during synthesis) to enable the receiver to restore their order. The most widely-suggested mechanism is to explicitly assign an index to each segment, through which the receiver may re-arrange the segments and restore their order. Although the explicit indexing approach is proved to be optimal for the asymptotic case [7], it has several disadvantages in practical finite-length regimes. For instance, explicit indexing may be prone to errors; i.e., when some indexes are corrupted by noise, the receiver may mis-detect those indexes and subsequently may arrange the segments in a (partially) incorrect order, giving rise to additional errors. This observation motivates several works including [13]- [17] to focus on error-resilient indexing methods. In [17] a concatenated coding scheme is proposed for transmission over noisy shuffling-sampling channels, where the inner code is partitioned into disjoint sub-codes and each data segment is encoded using a separate sub-code. It is assumed that a decoder implementation for such an inner code exists, which provides arbitrarily small decoding error probabilities. However, only an asymptotic case is considered, where the number of segments and the segment length grow arbitrarily large, and no practical implementation is given.

In this paper, we focus on a case where a sequence of information bits is sliced into a finite number of short-length segments, and the segments are transmitted over a noisy shuffling or a noisy shuffling-sampling channel. We implement a concatenated RS-polar coding scheme, where each data segment is encoded by a separate coset of a polar code. No explicit indexing is employed; instead, the decoder decides on the position of each segment by determining the coset by which that segment has been encoded. This task is accomplished by aid of a matched decoding method, where the segment is decoded by all cosets and a reliability metric is calculated based on

This work was funded by the European Union through the ERC Advanced Grant 101054904: TRANCIDS. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

978-1-6654-3540-6/22 © 2023 IEEE

which the correct (matched) coset is detected. Consequently, the matched decoding approach simultaneously finds the correct position of the segment and corrects the channel errors on that segment. However, the matched polar decoder is not able to correct all the channel errors on every segment, hence the residual errors are then corrected by the outer RS code.

To identify a proper reliability metric for detecting the matched coset, we take advantage of the concept of frozen bits in polar codes, which are set to zero in this paper. The reliabilities of the frozen bits at the output of a certain decoder may signal whether that decoder is matched to the input sequence or not. Motivated by this observation, we introduce a reliability metric in Section III and evaluate its performance in Section V. Also, through analytical approximations on the frame error rate (FER) of the proposed scheme, we demonstrate that the proposed implicit indexing method achieves lower FERs compared to the explicit indexing based solutions.

The paper is organized as follows. The system model is given in Section II. In Section III we present our proposed concatenated coding scheme. Section IV provides a performance analysis for the proposed scheme through an approximation derived on the FER. Section V includes numerical results and discussions. Finally, Section VI concludes the paper.

II. System Model

Schematics of the employed channel models are shown in Fig. 1. We assume that M packets with length L (bits) are inputs to the channel. The input packets are generated by slicing a codeword of an outer channel code with length ML bits, into M equal-length segments. The M segments are passed through a noisy channel. Although, in practice, DNA storage schemes may be affected by several types of errors, including substitution, deletion and insertion errors, different works on current DNA storage technologies confirm that substitution errors are dominant [2]- [6]. For this reason and for the sake of simplicity, in this paper, we focus on noisy channels with substitution errors. Specifically, we consider a binary symmetric channel (BSC) with crossover probability δ. For the noisy shuffling channel, the segments at the output of the BSC are shuffled before being received by the decoder; while for the noisy shuffling-sampling channel, the segments are sampled $N \geq M$ times with replacement, and the N samples are shuffled before being received by the decoder (see Fig. 1). The ratio $\alpha = \frac{N}{M}$ is called the coverage depth.

In the specific example depicted in Fig. 1-b, $\alpha = 1.5$, and the first, the second, the third, and the fourth packets are sampled $3, 2, 0,$ and 1 times, respectively. Observe that in the noisy shuffling-sampling channel model, some segments may not be sampled, and consequently may not be received by the decoder at all (e.g., the third segment is not sampled in Fig. 1-b). Due to this *missing segments* effect, the expected error rate experienced over a noisy shuffling-sampling channel is higher compared to a noisy shuffling channel. However, by increasing the coverage depth, α, the error rate will reduce.

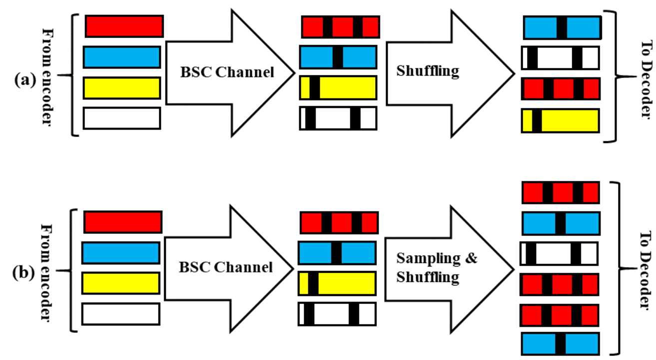

Fig. 1. Schematics of (a) noisy shuffling channel model, and (b) noisy shuffling-sampling channel model.

III. Proposed Coding Scheme

Due to the shuffling process, the order of segments is not preserved at the output of a noisy shuffling channel. The simplest solution to this problem is to add indexes to the segments at the encoder. This approach, which we refer to as *explicit indexing*, enables the decoder to retrieve the order of the transmitted segments. Although explicit indexing is a straightforward approach which is widely used in the literature, it is susceptible to channel noise. Motivated by this fact and inspired by the work of [17], in this section, we propose a concatenated encoding scheme along with an implicit indexing method that employs different cosets of a polar code to encode distinct segments; therefore, the position of each segment can be identified by detecting the coset by which that segment is encoded. We also propose a matched decoding method that detects the coset corresponding to a received noisy segment.

The block diagram of the proposed scheme is shown in Fig. 2. Let q, k_o be positive integers. The binary input sequence, $\mathbf{u}$, consisting of $q \times k_o$ bits, is partitioned into q-bit symbols, and then encoded using an (n_o, k_o) RS code, where $n_o = 2^q - 1$. The RS codeword length is n_o symbols, or equivalently $q \times n_o$ bits. This codeword is then zero-padded by $LM - qn_o$ bits to form a binary vector $\mathbf{s}$ with length LM, where $L = \left\lceil \frac{q \times n_o}{M} \right\rceil$ and $\lceil . \rceil$ denotes the ceiling function. The vector $\mathbf{s}$ is partitioned into M segments of length L bits, denoted by $\mathbf{s}^1$ through $\mathbf{s}^M$.

We consider a concatenated coding scheme, where a polar code is applied as the inner code. We integrate two different indexing implementations to the polar encoder and the polar decoder blocks of Fig. 2; namely, explicit indexing, and the proposed coset-based implicit indexing. In the explicit indexing scheme, an index with length $\lceil \log_2 M \rceil$ bits is appended to each segment; hence the length of the indexed segments is $k_i = L + \lceil \log_2 M \rceil$ bits. The indexed segments are encoded by an (n_i, k_i) polar code and are transmitted through the channel; i.e., the explicit index is encoded along with the information bits (the frozen bits are taken as zero). At the decoder, the received noisy segments are decoded and the $\lceil \log_2 M \rceil$ bits corresponding to the index are employed to determine the position of each segment in the entire decoded sequence, $\hat{\mathbf{s}}$. Note that if any of the $\lceil \log_2 M \rceil$ index bits is decoded with error, the position of the corresponding segment is determined incorrectly;

hence, $\hat{\mathbf{s}}$ is decoded with extra bit errors introduced by this incorrect ordering process (in addition to bit errors experienced at the output of a polar decoder).

In the proposed implicit indexing method, we aim to introduce a scheme that is more robust to the previously-mentioned index decoding errors. For this, M cosets of an (n_i, L) polar code with coset leaders $\mathbf{e}^1$ through $\mathbf{e}^M$ are selected. For each m, $\mathbf{s^m}$ is encoded using the m-th coset. This encoding process can be performed by encoding $\mathbf{s^m}$ using the polar code, followed by modulo-2 addition of $\mathbf{e}^m$ to the generated codeword. The symbol π in Fig. 2 denotes a random permutation and aims to emphasize that the noisy codewords, $\mathbf{r}^1, \ldots, \mathbf{r}^M$, are received out of order. The decoding process for each received vector, $\mathbf{r}^{m'}$, $1 \leq m' \leq M$ is performed as follows. For all $1 \leq m \leq M$, the vector $\mathbf{r}^{m'} \oplus \mathbf{e}^m$ is fed to a decoder of the (n_i, L) polar code; i.e., $\mathbf{r}^{m'}$ is decoded using the m-th coset for all m ($\oplus$ denotes bit-wise modulo-2 addition). Although the M decoders are identical, for clarification purposes we denote them by M parallel decoders, denoted by decoders numbered 1 through M. After completing the decoding process, a metric $\xi_{m',m}$ is calculated to quantify the likelihood of the event that the mth decoder is matched to $\mathbf{r}^{m'}$ (i.e., the event that $\mathbf{r}^{m'}$ is in fact a noisy copy of a vector which is encoded using the m-th coset of the original code). After completing this process, $\hat{m} = \underset{m}{\mathbf{argmax}}\, \xi_{m',m}$ is determined, and the output of the $\hat{m}$-th decoder is given as $\mathbf{v}^{m'}$, the decoded version of $\mathbf{r}^{m'}$. This decoded sequence, $\mathbf{v}^{m'}$, is then written in the location corresponding to the $\hat{m}$th segment in $\hat{\mathbf{s}}$. After completing this matched decoding process for all m', the padded bits are removed from $\hat{\mathbf{s}}$ and the resulting sequence is decoded by the outer RS decoder to obtain the eventual decoded bit sequence, $\hat{\mathbf{u}}$.

In order to find a suitable metric, $\xi_{m',m}$, we propose to exploit the notion of frozen bits in polar codes. The frozen bits are the bits that are forced to specific values (e.g., 0's in our case) at the encoder. Each L-bit segment, $\mathbf{s}^m$ is encoded by taking an n_i-bit sequence and filling its L most reliable positions by the bits of $\mathbf{s}^m$. These positions are determined by a proper channel reliability sequence. The remaining $n_i - L$ positions are filled by frozen bits. The resulting n_i-bit sequence, denoted by $\tilde{\mathbf{s}}^m$, is then transformed into the n_i-bit codeword $\mathbf{x}^m$ by applying the polar transform. When a received vector $\mathbf{r}^{m'}$ is decoded using the m-th coset decoder, in addition to producing the hard-decision output, the decoder is capable of producing a vector $\underline{\boldsymbol{\mathcal{L}}}^{m',m} = \left(\mathcal{L}_1^{m',m}, \mathcal{L}_2^{m',m}, \ldots, \mathcal{L}_{n_i}^{m',m}\right)$ where $\mathcal{L}_j^{m',m}$ denotes the log-likelihood ratio (LLR) of the jth bit of $\tilde{\mathbf{s}}^{m'}$; i.e., $\mathcal{L}_j^{m',m} = \log \frac{Pr\left(\tilde{s}_j^{m'}=0|m\right)}{Pr\left(\tilde{s}_j^{m'}=1|m\right)}$, where $\tilde{s}_j^{m'}$ denotes the jth bit of $\tilde{\mathbf{s}}^{m'}$ and the conditioning on "m" aims to clarify the dependency of the LLRs on the employed decoder. These LLR values provide a natural way of measuring the reliability of the decoder output as follows: Let $\mathcal{F}$ denote the set of indexes of all frozen bits. Define:

$$\xi_{m',m} = \sum_{j \in \mathcal{F}} \mathcal{L}_j^{m',m}. \tag{1}$$

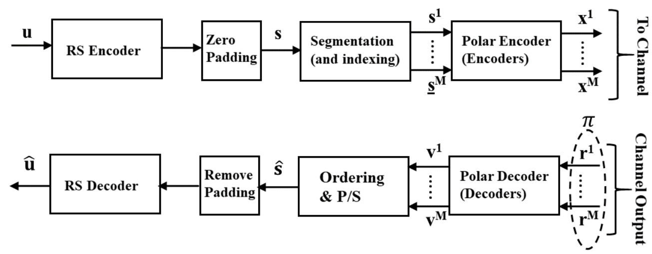

Fig. 2. Block diagram of the proposed encoding and decoding scheme.

When the matched decoder is applied, i.e., when $\mathbf{r}^{m'}$ is decoded in the same coset as the one in which it is encoded, $\xi_{m',m}$ is expected to be large. This is due to the fact that frozen bits are fixed to zero, hence, under matched decoding, their corresponding LLR values are expected to be high. On the other hand, for mismatched decoders, there is no such guarantee. Hence, we propose to detect the index by measuring the metric, $\xi_{m',m}$, for all m, and choosing the value of m that maximizes $\xi_{m',m}$.

IV. Performance Analysis of the Proposed Scheme

In this section, we bound the FER of the proposed coset-based scheme for a noisy shuffling channel under the minimum distance decoding. Since the RS code has a minimum distance of $n_o - k_o + 1$ symbols, a minimum distance decoder definitely corrects up to $\frac{n_o - k_o}{2}$ symbol errors in every RS codeword (selecting $n_o - k_o$ even). According to Fig. 2, if zero-padding is neglected, the sequence of symbols delivered to the RS decoder is $\hat{\mathbf{s}}$. Therefore, if a minimum distance decoder is applied, the probability of observing a frame error is less than or equal to the probability of observing more than $\frac{n_o - k_o}{2}$ symbol errors in $\hat{\mathbf{s}}$. Let us consider two events as follows:

(i) *A mis-detection event (defined as the event where at least one of the indexes (cosets) is detected incorrectly).* In this case, at least one segment of $\hat{\mathbf{s}}$ resembles a randomly generated sequence. This is due to the fact that when the m-th decoder is not matched to $\mathbf{s}^m$, the corresponding segment of $\hat{\mathbf{s}}$ either will be filled with another (mismatched) sequence, or is left without a candidate sequence, in which case without loss of generality we assume that it is filled with an all-zero sequence. In both cases, the expected number of symbol errors over the mth segment is large (with a mean value of $\frac{L}{2}$). Therefore, we consider the worst-case scenario; i.e., when a mis-detection event occurs we assume that the frame is always decoded erroneously.

(ii) *The event that all the indexes are detected correctly.* In such a case, the probability that each bit of $\hat{\mathbf{s}}$ is in error, is equal to the bit error rate (BER) of the polar code. If the bit error events are assumed to be independent (that is realistic given an interleaver is employed), since each symbol consists of q consecutive bits, the symbol error rate in such a case can be evaluated as:

$$p_s(\mathcal{H}) = 1 - \{1 - p_b(\mathcal{H})\}^q, \tag{2}$$

where $p_b(\mathcal{H})$ denotes the BER of the polar code and $\mathcal{H}$ is a vector specifying the system parameters (including n_i, L, q, M, δ).

Let $p_d(\mathcal{H})$ denote the probability of a mis-detection event. Then, based on the discussion on cases (i) and (ii), and assuming the symbol error events are independent, the FER achieved by a minimum distance decoder can be bounded as:

$$\begin{aligned} P_e(\mathcal{H}) \leq \; & p_d(\mathcal{H}) + (1 - p_d(\mathcal{H})) \\ & \times \left(1 - \sum_{j=0}^{\frac{n_o-k_o}{2}} \binom{n_o}{j} p_s^j(\mathcal{H}) (1 - p_s(\mathcal{H}))^{n_o-j}\right) \end{aligned} \quad (3)$$

To evaluate the right hand side of (3), one needs the values (or estimates) of $p_b(\mathcal{H})$ and $p_d(\mathcal{H})$. To estimate $p_b(\mathcal{H})$, one may either employ existing bounds on the error probability of polar codes (e.g., the bounds provided in [18]- [20]), or employ Monte-Carlo simulations.

While the evaluation of $p_d(\mathcal{H})$ for polar codes is left for future work, in the following, we derive a bound on $p_d(\mathcal{H})$ for a random coding scheme and a minimum distance decoding approach, explained below, and employ that bound for approximating the right hand side of (3).

Let $\mathcal{C} = \{\mathcal{C}^1, \ldots, \mathcal{C}^M\}$ be a set of M random codes with rate R. Each code, $\mathcal{C}^m$, contains 2^{nR} codewords and each codeword is drawn uniformly at random (with replacement) from $\mathcal{B}_n = \{\mathbf{b}_1, \ldots, \mathbf{b}_{2^n}\}$, the set of all possible realizations of an n-bit sequence. For each codeword $\mathbf{x} \in \mathcal{C}$, let $\mathcal{M}(\mathbf{x})$ denote the index of the code to which $\mathbf{x}$ belongs; i.e., $\mathcal{M}(\mathbf{x}) = m$ if and only if $\mathbf{x} \in \mathcal{C}^m$. Also, let

$$\hat{\mathbf{x}} = \operatorname*{\mathbf{argmin}}_{\mathbf{x} \in \mathcal{C}} d_H(\mathbf{r}, \mathbf{x}) \quad (4)$$

denote the output of a minimum distance decoder, where $d_H(\mathbf{r}, \mathbf{x})$ denotes the Hamming distance between a codeword $\mathbf{x}$ and the received vector, $\mathbf{r}$. If more than one codeword is at a minimum distance from $\mathbf{r}$, one of them is selected uniformly at random and is declared as $\hat{\mathbf{x}}$.

In order to derive a bound on $p_d(\mathcal{H})$, we first derive an upper bound on $Pr(\mathcal{M}(\hat{\mathbf{x}}) \neq \mathcal{M}(\mathbf{x}))$ for a randomly selected codeword $\mathbf{x}$, and then, we employ the union bound to derive an upper bound on $p_d(\mathcal{H})$. For this purpose, we follow the approach taken in [21] to find the exact average error probability of a random code ensemble over a BSC.

Let us begin by assuming that a fixed codeword $\mathbf{x}_0$ is transmitted over a BSC with channel crossover probability δ, and the vector $\mathbf{r} = \mathbf{x}_0 \oplus \mathbf{z}$ is received where $\mathbf{z}$ is the noise vector that is an independent and identically distributed (i.i.d.) binary sequence with $Pr(z_j = 1) = \delta$. Without loss of generality and for simplicity of notation, let $\mathbf{x}_0 \in \mathcal{C}^1$. Let $W_H(\mathbf{z})$ denote the Hamming weight of $\mathbf{z}$. Since $\mathbf{r} = \mathbf{x}_0 \oplus \mathbf{z}$, then $d_H(\mathbf{r}, \mathbf{x}_0) = W_H(\mathbf{z}) = w$. Note that $\mathbf{x}_0 \in \mathcal{C}^1$; therefore, a sufficient condition for obtaining $\mathcal{M}(\hat{\mathbf{x}}) = 1$ is that all the codewords in $\cup_{m=2}^M \mathcal{C}^m$ have a Hamming distance greater than w from $\mathbf{r}$. Since there are $2^{nR}(M-1)$ codewords in $\cup_{m=2}^M \mathcal{C}^m$ which are realizations of independent and uniformly distributed random vectors, we have:

$$\begin{aligned} & Pr(\mathcal{M}(\hat{\mathbf{x}}) = 1 | W_H(\mathbf{z}) = w) \\ & \qquad \geq \{Pr(d_H(\mathbf{r}, \mathbf{x}') > w | W_H(\mathbf{z}) = w)\}^{2^{nR}(M-1)} \end{aligned} \quad (5)$$

where $\mathbf{x}'$ is a random vector with a uniform distribution over $\mathcal{B}_n$.

By applying the law of total probability and by noting that $d_H(\mathbf{r}, \mathbf{x}') = W_H(\mathbf{z} \oplus \mathbf{x}_0 \oplus \mathbf{x}')$, we find:

$$\begin{aligned} & Pr(\mathcal{M}(\hat{\mathbf{x}}) = 1) \geq \sum_{w=0}^n Pr(W_H(\mathbf{z}) = w) \times \\ & \qquad \{Pr(W_H(\mathbf{z} \oplus \mathbf{x}_0 \oplus \mathbf{x}') > w | W_H(\mathbf{z}) = w)\}^{2^{nR}(M-1)} \end{aligned} \quad (6)$$

Let us define:

$$\mathcal{G}(\mathbf{z}) = \{\mathbf{b}_j \in \mathcal{B}_n \; s.t. \; W_H(\mathbf{z} \oplus \mathbf{b}_j) > W_H(\mathbf{z})\}. \quad (7)$$

Then:

$$\begin{aligned} & Pr(W_H(\mathbf{z} \oplus \mathbf{x}_0 \oplus \mathbf{x}') > w | W_H(\mathbf{z}) = w) \\ & \quad = Pr(\mathbf{x}_0 \oplus \mathbf{x}' \in \mathcal{G}(\mathbf{z}) | W_H(\mathbf{z}) = w) \end{aligned} \quad (8)$$

If we define $\mathcal{A}_w = \{\mathbf{b}_l \in \mathcal{B}_n \; s.t. \; W_H(\mathbf{b}_l) > w\}$, then for every $\mathbf{z}$ with $W_H(\mathbf{z}) = w$, there exists a one to one mapping between $\mathcal{A}_w$ and $\mathcal{G}(\mathbf{z})$ as follows. If $\mathbf{b}_j \in \mathcal{G}(\mathbf{z})$, then by definition $W_H(\mathbf{b}_j \oplus \mathbf{z}) > w$; i.e., $\mathbf{b}_j \oplus \mathbf{z} \in \mathcal{A}_w$. Also, if $\mathbf{b}_l \in \mathcal{A}_w$, then $W_H(\mathbf{z} \oplus (\mathbf{b}_l \oplus \mathbf{z})) = W_H(\mathbf{b}_l) > w$; i.e., $\mathbf{b}_l \oplus \mathbf{z} \in \mathcal{G}(\mathbf{z})$. Therefore, $|\mathcal{G}(\mathbf{z})| = |\mathcal{A}_w|$, which gives:

$$|\mathcal{G}(\mathbf{z})| = 2^n - \mathcal{N}(n, W_H(\mathbf{z})), \quad (9)$$

where $\mathcal{N}(n, w) = \sum_{h=0}^w \binom{n}{h}$. Notice that all vectors $\mathbf{z}$ with equal Hamming weights have an identical $|\mathcal{G}(\mathbf{z})|$. Also, since $\mathbf{x}'$ is uniformly distributed over $\mathcal{B}_n$, $\mathbf{x}_0 \oplus \mathbf{x}'$ is uniformly distributed over $\mathcal{B}_n$; hence:

$$Pr(\mathbf{x}_0 \oplus \mathbf{x}' \in \mathcal{G}(\mathbf{z}) | W_H(\mathbf{z}) = w) = \frac{2^n - \mathcal{N}(n, w)}{2^n}. \quad (10)$$

Using (8), (10), (6) and since $Pr(W_H(\mathbf{z}) = w) = \binom{n}{w} \delta^w (1-\delta)^{n-w}$, we obtain:

$$\begin{aligned} & Pr(\mathcal{M}(\hat{\mathbf{x}}) = 1) \geq \\ & \quad \sum_{w=0}^n \binom{n}{w} \delta^w (1-\delta)^{n-w} (1 - 2^{-n}\mathcal{N}(n, w))^{(M-1)2^{nR}} \end{aligned} \quad (11)$$

For BSCs with crossover probabilities $\delta \ll 1$ (e.g., for $\delta < 0.05$), $\binom{n}{w} \delta^w (1-\delta)^{n-w}$ attains its maximum for small w's (i.e., $w \ll n$).

To simplify the calculations, we employ the inequality:

$$(1-a)^n \geq (1 - na) \times u_{-1}(1 - na), \; -1 < a < 1, \, n \geq 1 \quad (12)$$

where $u_{-1}(.)$ is the unit step function. Using (11), (12), we obtain $Pr(\mathcal{M}(\hat{\mathbf{x}}) \neq 1) \leq 1 - f(n, R, \delta)$ where:

$$\begin{aligned} & f(n, R, \delta) = \sum_{w=0}^n \binom{n}{w} \delta^w (1-\delta)^{n-w} \times \\ & \left(1 - 2^{-n(1-R)}\mathcal{N}(n, w)\right)^{(M-1)} u_{-1}\left(1 - 2^{-n(1-R)}\mathcal{N}(n, w)\right) \end{aligned} \quad (13)$$

We may take a similar approach to show that $Pr(\mathcal{M}(\hat{\mathbf{x}}) \neq m) \leq 1 - f(n, R, \delta)$ if a fixed codeword $\mathbf{x}_0 \in \mathcal{C}^m$ is transmitted. Since in such a case $\mathcal{M}(\mathbf{x}_0) = m$, we obtain $Pr(\mathcal{M}(\hat{\mathbf{x}}) \neq \mathcal{M}(\mathbf{x}_0)) \leq 1 - f(n, R, \delta)$ for all m; i.e., for any fixed codeword $\mathbf{x}_0 \in \mathcal{C}$. Eventually, by observing that $f(n, R, \delta)$ does not depend on the choice of $\mathbf{x}_0$, we may generalize the upper bound for a randomly selected codeword $\mathbf{x} \in \mathcal{C}$, as:

$$Pr\left(\mathcal{M}\left(\hat{\mathbf{x}}\right) \neq \mathcal{M}\left(\mathbf{x}\right)\right) \leq 1 - f\left(n, R, \delta\right). \tag{14}$$

Now, assume that M randomly selected codewords, $\mathbf{x}^1 \in \mathcal{C}^1, \ldots, \mathbf{x}^M \in \mathcal{C}^M$, are transmitted over a noisy shuffling channel and a permutation of their noisy versions, $\mathbf{r}^1, \ldots, \mathbf{r}^M$, is received at the decoder as shown in Fig. 2. By definition, an index detection error occurs if the index of at least one of these M codewords is detected incorrectly; i.e.,

$$\begin{aligned} p_d\left(\mathcal{H}\right) &= Pr\left(\bigcup_{m=1}^{M} \left\{\mathcal{M}\left(\hat{\mathbf{x}}^m\right) \neq \mathcal{M}\left(\mathbf{x}^m\right)\right\}\right) \\ &\overset{(a)}{\leq} \textstyle\sum_{m=1}^{M} Pr\left(\mathcal{M}\left(\hat{\mathbf{x}}^m\right) \neq \mathcal{M}\left(\mathbf{x}^m\right)\right) \\ &\overset{(b)}{=} M \times \left(1 - f\left(n, R, \delta\right)\right), \end{aligned} \tag{15}$$

where $\hat{\mathbf{x}}^m$ is the output of a minimum distance decoder with input $\mathbf{r}^m$; (a) follows from the union bound, and (b) follows from (14). The right hand side of (15) may be replaced in (3) as an approximation for the FER as follows:

$$\begin{aligned} P_e\left(\mathcal{H}\right) \approx M\left(1 - f\left(n, R, \delta\right)\right) + \left(1 - M + M \times f\left(n, R, \delta\right)\right) \\ \times \left(1 - \textstyle\sum_{j=0}^{\frac{n_o - k_o}{2}} \binom{n_o}{j} p_s^j\left(\mathcal{H}\right)\left(1 - p_s\left(\mathcal{H}\right)\right)^{n_o - j}\right) \end{aligned} \tag{16}$$

Note that we cannot claim that (16) gives an upper bound on the FER, since (15) is derived assuming random coding and minimum distance decoding; whereas, in the proposed scheme, polar codes and their corresponding decoder are implemented. Nonetheless, implementation of the minimum distance decoder, which is the optimal decoder for the BSC channel, is prohibitive due to its exponentially growing complexity with the block length. Also, in general polar codes have better distance properties compared to random codes; i.e., it is expected that if a minimum distance decoder is applied for polar codes, the proposed RS-polar coding scheme would achieve FER values lower than those suggested by (15). Therefore, (15) is useful in the sense that it provides an insight on achievable FER values of the proposed scheme, if optimal decoding is applied.

V. Numerical Results

In this section, we provide numerical examples to quantify the performance of the proposed scheme, and compare it with that of the explicit indexing method. Throughout the simulations, we consider a setup with $q = 8$, $n_o = 255$ symbols, $M = 32$, and $n_i = 128$, unless otherwise stated. The segment length (without the index bits) is $L = 64$ bits. As explained in Section III, in the explicit indexing scheme, each segment is appended by 5 index bits and the resulting 69 bit sequence is encoded by a $(128, 69)$ polar code. In the proposed coset-based scheme, $M = 32$ distinct cosets of a $(128, 64)$ polar code are selected, and the mth segment is encoded by the mth coset. We simulate a benchmark scheme, where for each input bit stream, the M cosets are selected uniformly at random (but of course, they are assumed to be known at the decoder). For polar encoding purposes, the most reliable positions are picked

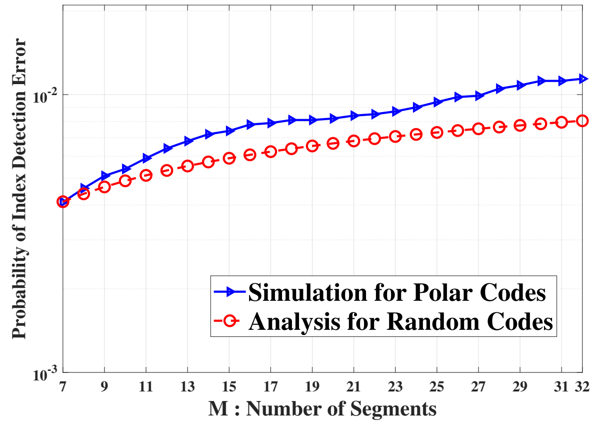

Fig. 3. Comparison between $Pr\left(\mathcal{M}\left(\hat{\mathbf{x}}\right) \neq \mathcal{M}\left(\mathbf{x}\right)\right)$ values found by simulations and by analysis, when $n = n_i = 128$, $R = \frac{1}{2}$, $\delta = 0.05$.

according to the 5G standardization unique channel-reliability sequence [22].

Figure 3 shows the value of $Pr\left(\mathcal{M}\left(\hat{\mathbf{x}}\right) \neq \mathcal{M}\left(\mathbf{x}\right)\right)$ for polar codes found by simulation, compared to the analytical upper bound derived in (14) for random coding and minimum distance decoding. It is observed that the analysis gives lower error probabilities; that may be justified by noting that minimum distance decoding which is the maximum likelihood (ML) decoding over the BSC is employed for the analysis; i.e., optimal decoding is applied. Also, there is no proof that the metric introduced in (1) for detecting the matched decoder is optimal (although our empirical results suggest that it is a good metric). However, we note that unlike the proposed polar encoding and matched decoding approach, the minimum distance decoding cannot be implemented in practice, except for very small block lengths. This is due to the fact that (4) has a computational complexity that grows exponentially with the block length, n (since the number of codewords is 2^{nR}).

Figure 4 shows the FER values found by simulations for $k_o = 225$ and $k_o = 235$. It is observed that the proposed scheme (labeled as "matched decoder") outperforms the explicit indexing scheme. The gain offered by the proposed scheme is more significant for case of $k_o = 225$; also, as expected, the FER in all cases reduces by decreasing k_o at the cost of a reduced code rate. The black curve shows the analytical result for $k_o = 225$ (Eq. (15)). Similar to Fig. 3, the analysis offers lower FERs compared to the simulation results based on polar codes (dashed red curve). This observation is justified by noting that the analytically derived values of $p_d\left(\mathcal{H}\right)$ are smaller than the actual values found for polar codes (see Fig. 3); hence, using them in (3) leads to optimistic results. Furthermore, the bound in (3) is found by applying minimum distance decoding which offers lower error probabilities compared to practical decoding methods such as the Berlekamp-Massey algorithm implemented for RS codes in simulations.

Fig. 5 shows the bit error rate results for $k_o = 215$ and noisy shuffling-sampling channels with $N = 120$ and $N = 150$ samples. Again, it is observed that the proposed

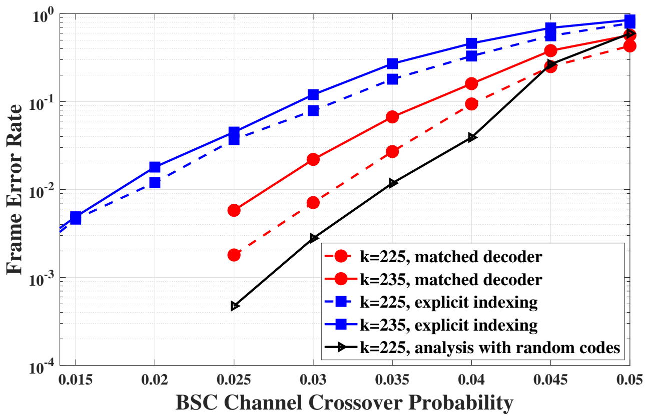

Fig. 4. FERs achieved by explicit indexing and matched decoding methods for noisy permutation channel, when $n_0 = 255$, $M = 32$, $n_i = 128$.

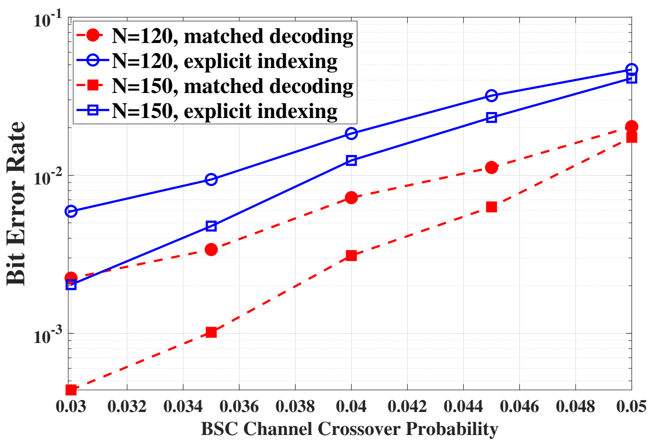

Fig. 5. BER results for $k_o = 215$ and noisy shuffling-sampling channels.

(matched decoder based) scheme outperforms the explicit indexing scheme. Also, the BER reduces by increasing the number of samples (i.e., by increasing the sampling depth α). However, this reduction in BER comes at the cost of larger complexity at the decoder, since more samples are required to be decoded in order to generate the ordered sequence $\hat{\mathbf{s}}$.

VI. Conclusions

We propose an implicit indexing approach for data transmission over a noisy shuffling channel, where data is encoded by an outer RS code, then the RS codeword is sliced into short-length segments, which are encoded by separate cosets of a polar code. We devise a matched decoding method that detects the correct coset for each received noisy segment. We also derive an upper bound for the probability of index detection error for the case of random codes being employed to encode the M segments. Through this bound, we find an approximation for the FER of the proposed scheme, which provides insights on the potential performance of the proposed scheme if optimal decoding is implemented for the inner code. Performance analysis of the proposed scheme for noisy shuffling channels with insertion, deletion and substitution errors, and design of suitable inner codes, are among interesting directions for future research.

References

[1] G. M. Church, Y. Gao, and S. Kosuri, “Next-generation digital information storage in DNA,” Science, vol. 337, no. 6102, pp. 1628–1628, Sep. 2012.

[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, “Towards practical, high-capacity, low- maintenance information storage in synthesized DNA,” Nature, vol. 494, no. 7435, pp. 77–80, 2013.

[3] S. Kosuri and G. Church, “Large-scale de novo DNA synthesis: Technologies and applications”, Nature Methods, vol. 11, no. 5, pp. 499-507, May 2014.

[4] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, “Robust chemical preservation of digital information on DNA in silica with error correcting codes,” Angew. Chem. Int. Ed., vol. 54, no. 8, pp. 2552–2555, 2015.

[5] L. Organick *et al.*, “Random access in large-scale DNA data storage,” Nature Biotechnology, vol. 36, pp. 242–248, Feb. 2018.

[6] Heckel, R., Mikutis, G., and Grass, R.N. A characterization of the DNA data storage channel. Sci Rep 9, 9663 (2019).

[7] I. Shomorony and R. Heckel, “Capacity results for the noisy shuffling channel,” 2019 IEEE International Symposium on Information Theory (ISIT), 2019, pp. 762-766.

[8] A. Lenz, P. H. Siegel, A. Wachter-Zeh and E. Yaakobi, “Coding over sets for DNA storage,” in IEEE Transactions on Information Theory, vol. 66, no. 4, pp. 2331-2351, April 2020.

[9] I. Shomorony and R. Heckel, “DNA-based storage: Models and fundamental limits,” in IEEE Transactions on Information Theory, vol. 67, no. 6, pp. 3675-3689, June 2021.

[10] M. Kovačević and V. Y. F. Tan, “Codes in the space of multisets – Coding for permutation channels with impairments,” in IEEE Transactions on Information Theory, vol. 64, no. 7, pp. 5156-5169, July 2018.

[11] J. Sima, N. Raviv and J. Bruck, “On coding over sliced information,” in IEEE Transactions on Information Theory, vol. 67, no. 5, pp. 2793-2807, May 2021.

[12] K. Levick, R. Heckel and I. Shomorony, “Achieving the capacity of a DNA storage channel with linear coding schemes,” 2022 56th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 2022, pp. 218-223.

[13] A. Lenz, P. H. Siegel, A. Wachter-Zeh and E. Yaakobi, “Anchor-based correction of substitutions in indexed sets,” 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 2019, pp. 757-761.

[14] J. Sima, N. Raviv and J. Bruck, “Robust indexing - Optimal codes for DNA storage,” 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 2020, pp. 717-722.

[15] L. Welter, I. Maarouf, A. Lenz, A. Wachter-Zeh, E. Rosnes and A. G. I. Amat, “Index-based concatenated codes for the multi-draw DNA storage channel,” 2023 IEEE Information Theory Workshop (ITW), Saint-Malo, France, 2023, pp. 383-388.

[16] T. Shinkar, E. Yaakobi, A. Lenz and A. Wachter-Zeh, “Clustering-correcting codes,” in IEEE Transactions on Information Theory, vol. 68, no. 3, pp. 1560-1580, March 2022.

[17] N. Weinberger, “Error probability bounds for coded-index DNA storage systems,” in IEEE Transactions on Information Theory, vol. 68, no. 11, pp. 7005-7022, Nov. 2022.

[18] B. Shuval and I. Tal, “A lower bound on the probability of error of polar codes over BMS channels,” 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 2017, pp. 854-858.

[19] G. Ricciutelli, M. Baldi, F. Chiaraluce and G. Liva, “On the error probability of short concatenated polar and cyclic codes with interleaving,” 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 2017, pp. 1858-1862.

[20] S. Seyedmasoumian and T. M. Duman, “Approximate weight distribution of polarization-adjusted convolutional (PAC) codes,” 2022 IEEE International Symposium on Information Theory (ISIT), Espoo, Finland, 2022, pp. 2577-2582.

[21] S. J. MacMullan and O. M. Collins, “A comparison of known codes, random codes, and the best codes,” in IEEE Transactions on Information Theory, vol. 44, no. 7, pp. 3009-3022, Nov. 1998.

[22] V. Bioglio, C. Condo and I. Land, “Design of polar codes in 5G new radio,” in IEEE Communications Surveys & Tutorials, vol. 23, no. 1, pp. 29-40, Firstquarter 2021.