Torn-Paper Coding

Ilan Shomorony
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
ilans@illinois.edu

Alireza Vahid
Department of Electrical Engineering
University of Colorado Denver
alireza.vahid@ucdenver.edu

***Abstract*—We consider the problem of communicating over a channel that randomly "tears" the message block into small pieces of different sizes and shuffles them. For the binary torn-paper channel with block length n and pieces of length $\text{Geometric}(p_n)$, we characterize the capacity as $C = e^{-\alpha}$, where $\alpha = \lim_{n\to\infty} p_n \log n$. Our results show that the case of $\text{Geometric}(p_n)$-length fragments and the case of deterministic length-$(1/p_n)$ fragments are qualitatively different and, surprisingly, the capacity of the former is larger. Intuitively, this is due to the fact that, in the random fragments case, large fragments are sometimes observed, which boosts the capacity.**

I. Introduction

Consider the problem of transmitting a message by writing it on a piece of paper, which will be torn into small pieces of random sizes and randomly shuffled. This coding problem is illustrated in Figure 1. We refer to it as the *torn-paper coding*, in allusion to the classic dirty-paper coding problem [1].

This problem is mainly motivated by macromolecule-based (and in particular DNA-based) data storage, which has recently received significant attention due to several proof-of-concept DNA storage systems [2–7]. In these systems, data is written onto synthesized DNA molecules, which are then stored in solution. During synthesis and storage, molecules in solution are subject to random breaks and, due to the unordered nature of macromolecule-based storage, the resulting pieces are shuffled [8]. Furthermore, the data is read via high-throughput sequencing technologies, which is typically preceded by physical fragmentation of the DNA with techniques like *sonication* [9]. In addition, the torn-paper channel is related to the DNA shotgun sequencing channel, studied in [10–12], but in the context of variable-length reads, which are obtained in nanopore sequencing technologies [13, 14].

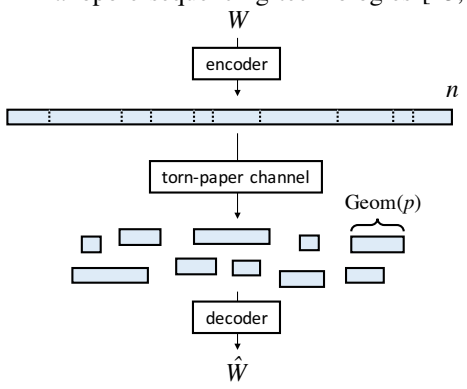

Fig. 1. The torn-paper channel.

We consider the scenario where the channel input is a length-n binary string that is then torn into pieces of lengths $N_1, N_2, ...$, each of which has a $\text{Geometric}(p_n)$ distribution. The channel output is the unordered set of these pieces. As we will see, even this noise-free version of the torn-paper coding problem is non-trivial.

To obtain some intuition, notice that $E[N_i] = 1/p_n$, and hence it is reasonable to compare our problem to the case where the tearing points are evenly separated, and $N_i = 1/p_n$ for $i = 1, 2, ..., np_n$ with probability 1. In this case, the channel becomes a *shuffling channel*, similar to the one considered in [15], but with no noise. Coding for the case of deterministic fragments of length $N_i = 1/p_n$ is easy: since the tearing points are known, we can prefix each fragment with a unique identifier, which allows the decoder to correctly order the np_n fragments. From the results in [15], such an index-based coding scheme is capacity-optimal for the shuffling channel, and any achievable rate must satisfy, for large n,

$$R < (1 - p_n \log n)^+. \tag{1}$$

If we let $\alpha = \lim_{n\to\infty} p_n \log n$, the capacity for the case of deterministic fragment lengths becomes $(1-\alpha)^+$.

It is not clear a priori whether the capacity of the torn-paper channel (with random fragment lengths) should be higher or lower than $(1-\alpha)^+$. The fact that the tearing points are not known to the encoder makes it challenging to place a unique identifier in each fragment, suggesting that the torn-paper channel is "harder" and should have a lower capacity. The main result of this paper contradicts this intuition and shows that the capacity of the torn-paper channel with $\text{Geometric}(p_n)$-length fragments is higher than $(1-\alpha)^+$. More precisely, we show that the capacity of the torn-paper channel is $C = e^{-\alpha}$. The comparison is shown in Figure 2. Intuitively, this boost in capacity comes from the tail of the geometric distribution, which guarantees that a fraction of the fragments will be significantly larger than the mean $E[N_i] = 1/p_n$. This allows the capacity to be positive even for $\alpha \geq 1$, in which case the capacity of the deterministic-tearing case in (1) becomes 0.

To prove the converse part of this result we partition the set of fragments into bins of fragments with roughly the same size and view the torn-paper channel as parallel shuffling channels. The achievability is based on a random code construction and optimal decoding. We also present an explicit code construction based on the idea of interleaving a synchronization pilot sequence with codewords from an erasure code. The synchronization sequence allows fragments that are long enough to have their location in the codeword determined [16]. As shown in Figure 2, the rates achieved

by this interleaved-pilot scheme have a similar shape to the capacity curve, but with a significant gap.

Related literature: The problem of reconstructing a string from a set of its subsequences has been studied in the context of the assembly problem [10, 11], the trace reconstruction problem [17–19], and the problem of reconstructing a string from its substring spectrum [12, 20]. In all of these settings, the set of observed strings have overlaps with each other, which is different from the case considered here.

Several recent works have designed codes tailored to specific aspects of DNA storage. These include DNA synthesis constraints such as sequence composition [6, 21, 22], the asymmetric nature of the DNA sequencing error channel [23], the need for codes that correct insertion errors [24], and the need for techniques to allow random access [22].

Motivated by DNA-based storage, a few recent works have considered the problem of coding across an unordered set of strings [25–27] and the problem of coding over sets [28, 29]. Channels that shuffle blocks of information were also recently studied in the context of the *bee-identification problem* [30] and noisy permutation channels [31].

Finally, the interleaved-pilot scheme presented in Section VII is related to the notion of phase detection sequences, which appear in the context of positioning systems [16]. Our proposed construction is based on *de Bruijn* sequences [32], which have been used in the problem of sequence reconstruction from substring profiles [21].

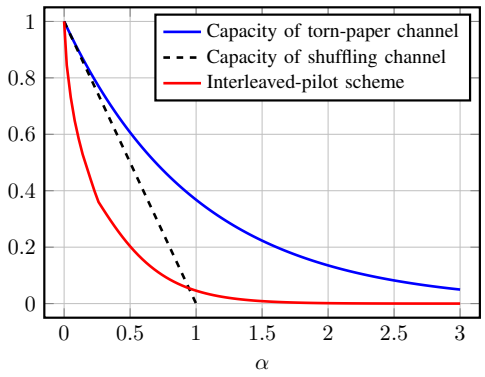

Fig. 2. Comparison between the capacity of the torn-paper channel $C = e^{-\alpha}$, the capacity of the shuffling channel with fragments of deterministic length $1/p_n$, and the rate achieved on the torn-paper channel by the explicit code construction based on the interleaved-pilot scheme.

II. Problem Setting

We consider the problem of coding for the torn-paper channel, illustrated in Figure 1. The transmitter encodes a message $W \in \{1, ..., 2^{nR}\}$ into a length-n binary codeword $X^n \in \mathbb{F}_2^n$. The channel output is a set of binary strings

$$\mathcal{Y} = \left\{ \vec{Y}_1, \vec{Y}_2, \ldots, \vec{Y}_K \right\}. \tag{2}$$

The process by which $\mathcal{Y}$ is obtained is described next.

1) The channel tears the input sequence into segments of Geometric(p_n)-length for a *tearing probability* p_n. More specifically, let $N_1, N_2, ...$ be i.i.d. Geometric(p_n) random variables, and K be the smallest index such that $\sum_{i=1}^{K} N_i \geq n$. Notice that K is also a random variable.

The channel tears X^n into segments $\vec{X}_1, ..., \vec{X}_K$, where

$$\vec{X}_i = \left[X_{1+\sum_{j=1}^{i-1} N_j}, ..., X_{\sum_{j=1}^{i} N_j} \right],$$

for $i = 1, ..., K-1$ and

$$\vec{X}_K = \left[X_{1+\sum_{j=1}^{K-1} N_j}, ..., X_n \right].$$

We note that this process is equivalent to independently tearing the message in between consecutive bits with probability p_n. More precisely, let $T_2, T_3, ..., T_n$ be binary indicators of whether there is a cut between X_{i-1} and X_i. Then, letting T_is be i.i.d. Bernoulli(p_n) random variables results in independent fragments of length Geometric(p_n). Also, $K = 1 + \sum_{i=2}^{n} T_i$, implying that $E[K] = 1 + (n-1)p_n = np_n + (1-p_n)$.

2) Given K, let $[\pi_1, ..., \pi_K]$ be a uniformly distributed random permutation on $[1, 2, \ldots, K]$. The output segments are then obtained by setting, for $i = 1, ..., K$,

$$\vec{Y}_i = \vec{X}_{\pi_i}. \tag{3}$$

We note that there are no bit-level errors, *e.g.*, bit flips, in this process. We also point out that we allow the tearing probability to be a function of the block length n, thus, including subscript n in p_n.

A code with rate R for the torn-paper channel is a set $\mathcal{C}$ of 2^{nR} binary codewords, each of length n, together with a decoding procedure that maps a set $\mathcal{Y}$ of variable-length binary strings to an index $\hat{W} \in \{1, ..., 2^{nR}\}$. The message W is assumed to be chosen uniformly at random from $\{1, ..., 2^{nR}\}$, and the error probability of a code is defined accordingly. A rate R is said to be achievable if there exists a sequence of rate-R codes $\{\mathcal{C}_n\}$, where $\mathcal{C}_n$ has blocklength n, whose error probability tends to 0 as $n \to \infty$. The capacity C is defined as the supremum over all achievable rates. Notice that C should be a function of the sequence of tearing probabilities $\{p_n\}_{n=1}^{\infty}$.

Notation: Throughout the paper, $\log(\cdot)$ represents the logarithm base 2, while $\ln(\cdot)$ represents the natural logarithm. For functions $f(n)$ and $g(n)$, we write $g(n) = o(f(n))$ if $g(n)/f(n) \to 0$ as $n \to \infty$. For an event A, we let $\mathbf{1}_A$ or $\mathbf{1}\{A\}$ be the binary indicator of A.

III. Main Results

If the encoder had access to the tearing locations ahead of time, a natural coding scheme would involve placing unique indices on every fragment, and using the remaining bits for encoding a message. In particular, if the message block was evenly broken into np_n pieces of length $N_i = 1/p_n$, results from [15] imply that placing a unique index of length $\log(np_n)$ in each fragment is capacity optimal. The rate achieved is

$$(N_i - \log(np_n))/N_i = 1 - p_n \log(np_n),$$

and the capacity is $(1-\alpha)^+$, where we define $\alpha = \lim_{n\to\infty} p_n \log(np_n) = \lim_{n\to\infty} p_n \log n$ (assuming the limit exists). If $\alpha \geq 1$, no positive rate is achievable.

However, in our setting, the fragment lengths are random and the same index-based approach cannot be used. Because we do not know the tearing points, we cannot place indices at the beginning of each fragment. Furthermore, while the expected fragment length may be large, some fragments may be shorter than $\log(np_n)$ and a unique index could not be placed in them even if we knew the tearing points. Our main result shows that, surprisingly, the random tearing locations and fragment lengths in fact increase the channel capacity.

Theorem 1. *The capacity of the torn-paper channel is*

$$C = e^{-\alpha},$$

where $\alpha = \lim_{n\to\infty} p_n \log n$.

At a high level, the reason for an exponential to appear in the capacity expression in Theorem 1 is that, if $N^{(n)}$ has a $\mathrm{Geometric}(p_n)$ distribution, as $n \to \infty$, $N^{(n)}/\log n$ converges in distribution to an $\mathrm{Exponential}(\alpha)$ random variable, where $\alpha = \lim_{n\to\infty} p_n \log n$ (provided the limit exists). In Section IV, we provide additional discussion on the intuition behind the capacity expression.

The rest of the paper is organized as follows. In Sections V and VI we prove Theorem 1. To prove the converse to this result we exploit the fact that, for large n, $N_i/\log n$ has an approximately exponential distribution. This, together with several concentration results, allows us to partition the set of fragments into multiple bins of fragments with roughly the same size and view the torn-paper channel, in essence, as parallel channels with fixed-size fragments. Our achievability is based on random coding arguments and does not provide much insight into practical coding schemes. Then, in Section VII we explore an explicit code construction based on "sprinkling" a synchronization sequence throughout all codewords, which allows fragments that are long enough to be ordered. A significant gap remains between the rate achieved by this explicit construction and the true capacity.

IV. Intuition for capacity expression

The capacity expression in Theorem 1 can be intuitively understood by considering a modified channel where the transmitter knows the locations of all tearing points. In that setting, a simple coding approach is the following: we ignore all fragments that are shorter than $\log n$ and we place a unique index at the beginning of every fragment longer than $\log n$. Since for large n, $N_i/\log n$ has approximately an $\mathrm{Exponential}(\alpha)$ distribution (which we formally state in Lemma 2),

$$\Pr\left(N_i \geq \log n\right) \approx e^{-\alpha}. \tag{4}$$

Since the total number of fragments is roughly $n/E[N_i] = np_n$, we need

$$\log(np_n e^{-\alpha}) < \log n$$

bits per fragment for the index, making it feasible to place a unique index in each fragment longer than $\log n$.

As we show later in Lemma 6, the number of bits from the original codeword X^n that end up in fragments of length at least $\log n$, for large n, is approximately

$$n(\alpha+1)e^{-\alpha}.$$

Out of those bits, since $\alpha \approx p_n \log n$, we use

$$(np_n e^{-\alpha})\log n \approx n\alpha e^{-\alpha}$$

for indices. Hence, we are left with

$$n(\alpha+1)e^{-\alpha} - n\alpha e^{-\alpha} = ne^{-\alpha}$$

message bits. Since there is no noise, message bits can be written directly onto the non-index parts of the fragments, yielding a data rate of $e^{-\alpha}$. The decoding procedure is straightforward: using the unique indices the fragments can be ordered and the message bits can then be read directly.

Notice that this scheme cannot be employed in the original torn-paper channel since the tearing points are not known at the transmitter. Furthermore, it is not obvious that throwing out fragments shorter than $\log n$ is capacity-optimal. Hence, this scheme is only included to provide intuition and place the capacity expression in context.

V. Converse

In order to prove the converse for Theorem 1, we first partition the input and output strings based on length. This allows us to view the torn-paper channel as a set of parallel channels, each of which with fragments of roughly the same size. More precisely, for an integer parameter L, we will let

$$\begin{aligned}\mathcal{X}_k &= \left\{\vec{X}_i : \tfrac{k-1}{L}\log n \leq N_i < \tfrac{k}{L}\log n\right\} \text{ and}\\ \mathcal{Y}_k &= \left\{\vec{Y}_i : \tfrac{k-1}{L}\log n \leq N_{\pi_i} < \tfrac{k}{L}\log n\right\},\end{aligned} \tag{5}$$

for $k = 1, 2, ...$, and we will think of the transformation from $\mathcal{X}_k$ to $\mathcal{Y}_k$ as a separate channel. Notice that the kth channel is intuitively similar to the shuffling channel with equal-length pieces considered in [25].

We will use the fact that the number of fragments in $\mathcal{Y}_k$ concentrates as $n \to \infty$. More precisely, we let

$$q_{k,n} = \Pr\left(\frac{k-1}{L} \leq \frac{N_1}{\log n} < \frac{k}{L}\right), \tag{6}$$

and we have the following lemma, proved in Section VIII.

Lemma 1. *The number of fragments in* $\mathcal{Y}_k$ *satisfies*

$$\Pr\left(||\mathcal{Y}_k| - np_n q_{k,n}| > \epsilon np_n\right) \leq 4e^{-np_n^2\epsilon^2/4}, \tag{7}$$

for any $\epsilon > 0$ *and* n *large enough,.*

Notice that, since $\lim_{n\to\infty} p_n \log n = \alpha$, $E\left[\frac{N_1}{\log n}\right] \to \alpha^{-1}$ as $n \to \infty$. Moreover, asymptotically, $\frac{N_1}{\log n}$ approaches an $\mathrm{Exponential}(\alpha)$ distribution. This known fact is stated as the following lemma, which we also prove in Section VIII.

Lemma 2. *If $N^{(n)}$ is a* Geometric(p_n) *random variable and* $\lim_{n\to\infty} E[N^{(n)}]/\log n = 1/\alpha$*, then*

$$\lim_{n\to\infty} \Pr\left(N^{(n)} \geq \beta \log n\right) = e^{-\alpha\beta}. \tag{8}$$

Lemma 1 implies that, with high probability, the number of fragments in the kth channel satisfies $||\mathcal{Y}_k| - np_nq_{k,n}| < \epsilon np_n$, which in particular implies that

$$\begin{aligned}\lim_{n\to\infty} \frac{E\left[|\mathcal{Y}_k|\right]}{np_n} &= \lim_{n\to\infty} \frac{np_nq_{k,n} + o(np_n)}{np_n} \\ &= \lim_{n\to\infty} \Pr\left(\tfrac{k-1}{L} \leq \tfrac{N_1}{\log n} < \tfrac{k}{L}\right) \\ &= e^{-\alpha(k-1)/L} - e^{-\alpha k/L}, \end{aligned} \tag{9}$$

where the last equality follows from Lemma 2. Next, we define the event $\mathcal{E}_{k,n} = \{||\mathcal{Y}_k| - np_nq_{k,n}| > \epsilon_n np_n\}$, where $\epsilon_n = 1/\log n$, which guarantees that, as $n \to \infty$, $\epsilon_n \to 0$ and $\Pr(\mathcal{E}_{k,n}) \to 0$ from Lemma 1. Then,

$$\begin{aligned} H(\mathcal{Y}_k) &\leq H(\mathcal{Y}_k, \mathbf{1}_{\mathcal{E}_{k,n}}) \leq 1 + H(\mathcal{Y}_k|\mathbf{1}_{\mathcal{E}_{k,n}}) \\ &\leq 1 + 2n\Pr(\mathcal{E}_{k,n}) + H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n}), \end{aligned} \tag{10}$$

where we loosely upper bound $H(\mathcal{Y}_k|\mathcal{E}_k)$ with $2n$, since $\mathcal{Y}$ can be fully described by the binary string X^n and the $n-1$ tearing points indicators $T_2, ..., T_n$.

In order to bound $H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n})$, *i.e.*, the entropy of $\mathcal{Y}_k$ given that its size is close to $np_nq_{k,n}$, we first note that the number of possible distinct sequences in $\mathcal{Y}_k$ is

$$\sum_{i=\frac{k-1}{L}\log n}^{\frac{k}{L}\log n} 2^i < 2 \cdot 2^{\frac{k}{L}\log n} = 2n^{k/L}.$$

Moreover, given $\bar{\mathcal{E}}_{k,n}$,

$$\begin{aligned} |\mathcal{Y}_k| &\leq np_nq_{k,n} + \epsilon_n np_n \\ &= np_n\left[\epsilon_n + \Pr\left(\frac{k-1}{L} \leq \frac{N_1}{\log n} < \frac{k}{L}\right)\right] \triangleq M, \end{aligned} \tag{11}$$

and the set $\mathcal{Y}_k$ can be seen as a histogram $(x_1, ..., x_{2n^{k/L}})$ over all possible $2n^{k/L}$ strings with $\sum x_i = M$. Notice that we can view the last element of the histogram as containing "excess counts" if $|\mathcal{Y}_k| < M$. Hence, using a simple counting argument to bound the number of different possible histograms (see Lemma 1 in [25]),

$$\begin{aligned} H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n}) &\leq \log\binom{2n^{k/L} + M - 1}{M} \\ &\leq M\log\left(\frac{e(2n^{k/L} + M - 1)}{M}\right) \\ &= M\left[\log\left(2n^{k/L} + M - 1\right) + \log(e) - \log M\right] \\ &= M\left[\max(\tfrac{k}{L}\log n, \log M) - \log M + o(\log n)\right] \\ &= M\left[(\tfrac{k}{L}\log n - \log M)^+ + o(\log n)\right] \\ &= M\log n\left[(\tfrac{k}{L} - \log M/\log n)^+ + o(1)\right]. \end{aligned} \tag{12}$$

From (11), we have $\log M/\log n \to 1$ as $n \to \infty$. Combining (10) and (12), dividing by n, and letting $n \to \infty$ yields

$$\begin{aligned} \lim_{n\to\infty} \frac{H(\mathcal{Y}_k)}{n} &\leq \lim_{n\to\infty} \frac{H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n}) + 1 + 2n\Pr(\mathcal{E}_{k,n})}{n} \\ &\leq \lim_{n\to\infty} \frac{M\log n}{n}\left(\frac{k}{L} - 1\right)^+ \\ &= \lim_{n\to\infty} p_n\log n\,(q_{k,n} + \epsilon_n)\left(\frac{k}{L} - 1\right)^+ \\ &= \alpha\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right)\left(\frac{k}{L} - 1\right)^+. \end{aligned} \tag{13}$$

In order to bound an achievable rate R, we use Fano's inequality to obtain

$$nR \leq I(X^n; \mathcal{Y}) + o(n) \leq H(\mathcal{Y}) + o(n), \tag{14}$$

and we conclude that any achievable rate must satisfy $R \leq \lim_{n\to\infty} \frac{H(\mathcal{Y})}{n}$. In order to connect (14) and (13), we state the following lemma, which allows us to move the limit inside the summation. The proof is in Section VIII.

Lemma 3. *If $\mathcal{Y}_k$ is defined as in (5) for $k = 1, 2, ...,$*

$$\lim_{n\to\infty} \frac{H(\mathcal{Y})}{n} \leq \sum_{k=1}^{\infty} \lim_{n\to\infty} \frac{H(\mathcal{Y}_k)}{n}.$$

Using this lemma and (13), we can upper bound any achievable rate as

$$\begin{aligned} R &\leq \lim_{n\to\infty} \frac{H(\mathcal{Y})}{n} \leq \sum_{k=1}^{\infty} \lim_{n\to\infty} \frac{H(\mathcal{Y}_k)}{n} \\ &= \sum_{k=L+1}^{\infty} \alpha\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right)(\tfrac{k}{L} - 1) \\ &= \frac{\alpha}{L}\sum_{k=L+1}^{\infty} k\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) \\ &\quad - \alpha\sum_{k=L+1}^{\infty}\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) \\ &= \frac{\alpha}{L}\sum_{k=L+1}^{\infty} k\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) - \alpha e^{-\alpha}, \end{aligned} \tag{15}$$

where the last equality is due to a telescoping sum. The remaining summation can be computed as

$$\begin{aligned} &\sum_{k=L+1}^{\infty} k\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) \\ &\quad = (L+1)e^{-\alpha} + \sum_{k=L+2}^{\infty} e^{-\alpha(k-1)/L} \\ &\quad = Le^{-\alpha} + e^{-\alpha}\sum_{k=0}^{\infty} e^{-\alpha k/L} = Le^{-\alpha} + \frac{e^{-\alpha}}{1 - e^{-\alpha/L}}. \end{aligned}$$

We conclude that any achievable rate must satisfy

$$R < \frac{\alpha}{L}\left(Le^{-\alpha} + \frac{e^{-\alpha}}{1 - e^{-\alpha/L}}\right) - \alpha e^{-\alpha} = \frac{\alpha e^{-\alpha}}{L(1 - e^{-\alpha/L})},$$

for any positive integer L. Since

$$\lim_{L\to\infty} L(1-e^{-\alpha/L}) = \alpha,$$

we obtain the outer bound $R < e^{-\alpha}$.

VI. Achievability via Random Coding

A random coding argument can be used to show that any rate $R < e^{-\alpha}$ is achievable. Consider generating a codebook $\mathcal{C}$ with 2^{nR} codewords, by independently picking each symbol as Bernoulli$(1/2)$. Let $\mathcal{C} = \{\mathbf{x}_1, ..., \mathbf{x}_{2^{nR}}\}$, where $\mathbf{x}_i$ is the random codeword associated with message $W = i$. Notice that optimal decoding can be obtained by simply finding an index i such that $\mathbf{x}_i$ corresponds to a concatenation of the strings in $\mathcal{Y}$. If more than one such codewords exist, an error is declared.

Suppose message $W = 1$ is chosen and $\mathcal{Y} = \{\vec{Y}_1, ..., \vec{Y}_K\}$ is the random set of output strings. To bound the error probability, we consider a suboptimal decoder that throws out all fragments shorter than $\gamma \log n$, for some $\gamma > 0$ to be determined, and simply tries to find a codeword $\mathbf{x}_i$ that contains all output strings $\mathcal{Y}_\gamma = \{\vec{Y}_i : N_{\pi_i} \geq \gamma \log n\}$ as non-overlapping substrings. If we let $\mathcal{E}$ be the error event averaged over all codebook choices, we have

$$\begin{aligned}&\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W=1)\\ &= \Pr\left(\text{some } \mathbf{x}_j,\ j \neq 1, \text{ contains all strings in } \mathcal{Y}_\gamma | W = 1\right).\end{aligned}$$

Using a similar approach to the one used in Section V, it can be shown that $E[|\mathcal{Y}_\gamma|] = np_n \Pr(N_1 \geq \gamma \log n) + o(np_n)$. From Lemma 2, we thus have

$$\lim_{n\to\infty} \frac{E[|\mathcal{Y}_\gamma|]}{n \cdot p_n} = \lim_{n\to\infty} \Pr\left(N_1 \geq \gamma \log n\right) = e^{-\alpha\gamma}. \tag{16}$$

If we let Z_i be the binary indicator of the event $\{N_i \geq \gamma \log n\}$, then $|\mathcal{Y}_\gamma| = \sum_{i=1}^{K} Z_i$. In Section VIII, we prove the following concentration result.

Lemma 4. *The number of fragments in $\mathcal{Y}_\gamma$ satisfies*

$$\Pr\left(\left||\mathcal{Y}_\gamma| - e^{-\alpha\gamma} np_n\right| > \epsilon np_n\right) \to 0, \tag{17}$$

for any $\epsilon > 0$ and n large enough.

In addition to characterizing $|\mathcal{Y}_\gamma|$ asymptotically, we will also be interested in the total length of the sequences in $\mathcal{Y}_\gamma$. Intuitively, this determines how much of codeword $\mathbf{x}_1$ is "covered" by fragments in $\mathcal{Y}_\gamma$.

Definition 1. *The coverage of $\mathcal{Y}_\gamma$ is defined as*

$$c_\gamma = \frac{1}{n} \sum_{i=1}^{K} N_i \mathbf{1}_{\{N_i \geq \gamma \log n\}}. \tag{18}$$

Notice that $0 \leq c_\gamma \leq 1$ with probability 1.

In order to characterize c_γ asymptotically, we will again resort to the exponential approximation of a geometric distribution through the following lemma.

Lemma 5. *If $N^{(n)}$ is a* Geometric(p_n) *random variable and $\lim_{n\to\infty} E[N^{(n)}]/\log n = 1/\alpha$, then, for any $\beta \geq 0$,*

$$\begin{aligned}&\lim_{n\to\infty} E\left[N^{(n)} \mathbf{1}_{\{N^{(n)} \geq \gamma \log n\}}\right] / \log n\\ &\qquad = E\left[\tilde{N} \mathbf{1}_{\{\tilde{N} \geq \gamma\}}\right] = \left(\gamma + \frac{1}{\alpha}\right) e^{-\alpha\gamma},\end{aligned} \tag{19}$$

where $\tilde{N}$ is an Exponential(α) *random variable.*

Using Lemma 5, we can characterize the asymptotic value of $E[c_\gamma]$ and show that c_γ concentrates around this value. More precisely, we show the following lemma in Section VIII.

Lemma 6. *If c_γ is defined as in (18), then*

$$\Pr\left(\left|c_\gamma - (\alpha\gamma + 1)e^{-\alpha\gamma}\right| > \epsilon\right) \to 0, \tag{20}$$

as $n \to \infty$ for any $\epsilon > 0$.

In particular, Lemma 6 implies that

$$\lim_{n\to\infty} E[c_\gamma] = (\alpha\gamma + 1)e^{-\alpha\gamma}, \tag{21}$$

and that c_γ cannot deviate much from this value with high probability. If we let $B_1 = (1+\epsilon)e^{-\alpha\gamma} np_n$ and $B_2 = (1-\epsilon)(\alpha\gamma+1)e^{-\alpha\gamma}$, and we define the event

$$\mathcal{B} = \{|\mathcal{Y}_\gamma| > B_1\} \cup \{c_\gamma < B_2\}, \tag{22}$$

then (17) and (20) imply that $\Pr(\mathcal{B}) \to 0$ as $n \to \infty$. Since $\mathcal{B}$ is independent of $\{W = 1\}$, we can upper bound the probability of error as

$$\begin{aligned}\Pr(\mathcal{E}) &\leq \Pr\left(\text{some } \mathbf{x}_j \text{ contains all strings in } \mathcal{Y}_\gamma | W = 1\right)\\ &\leq \Pr\left(\text{some } \mathbf{x}_j \text{ contains all strings in } \mathcal{Y}_\gamma | \bar{\mathcal{B}}, W = 1\right)\\ &\qquad + \Pr(\mathcal{B})\\ &\overset{(i)}{\leq} |\mathcal{C}| \frac{n^{B_1}}{2^{nB_2}} + \Pr(\mathcal{B})\\ &\leq 2^{nR}\, 2^{B_1 \log n}\, 2^{-nB_2} + o(1)\\ &= 2^{nR}\, 2^{(1+\epsilon)e^{-\alpha\gamma} np_n \log n - n(1-\epsilon)(\alpha\gamma+1)e^{-\alpha\gamma}} + o(1)\\ &= 2^{-n((1-\epsilon)(\alpha\gamma+1)e^{-\alpha\gamma} - (1+\epsilon)e^{-\alpha\gamma} p_n \log n - R)} + o(1).\end{aligned}$$

Inequality (i) follows from the union bound and from the fact that thre are at most n^{B_1} ways to align the strings in $\mathcal{Y}_\gamma$ to a codeword $\mathbf{x}_j$ in a non-overlapping way and, given this alignment, 2^{nB_2} bits in $\mathbf{x}_j$ must be specified. Since $p_n \log n \to \alpha$ as $n \to \infty$, we see that we can achieve a rate R as long as

$$R < (1-\epsilon)(1+\alpha\gamma)e^{-\alpha\gamma} - (1+\epsilon)\alpha e^{-\alpha\gamma},$$

for some $\epsilon > 0$ and $\gamma > 0$. Letting $\epsilon \to 0$, yields

$$R < (1 + \alpha\gamma - \alpha)e^{-\alpha\gamma}$$

for some $\gamma > 0$. The right-hand side is maximized by setting $\gamma = 1$, which implies that we can achieve any rate $R < e^{-\alpha}$. We point out that this choice of γ justifies the optimality of discarding fragments of length less than $\log n$, first mentioned in Section IV.

VII. Interleaved-pilot scheme

While the scheme presented in Section VI achieves the capacity of the torn-paper channel, it is far from being a practical scheme. In principle, it requires one to consider all possible $K!$ orderings of the K fragments and trying to align each one to each of the 2^{nR} codewords.

A natural way to design schemes for a channel that shuffles fragments of the message involves placing "indices" on the different pieces, which allows properly ordering them. However, as previously discussed, the randomness in the tearing locations and in the length of the fragments makes this approach not straightforward for the torn-paper channel. In particular, if we place indices at evenly separated points of the input string X^n, they will appear at random locations of the fragments, and a fraction of the indices will be fragmented, making the recovery more difficult. For that reason, in this section we explore the idea of *interleaving* a pilot, or a *phase detection sequence* [33] throughout the input codewords.

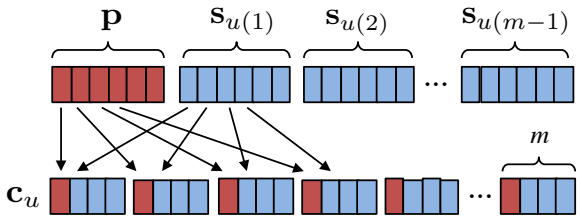

Fig. 3. Interleaving a pilot sequence $\mathbf{p}$ with codewords $\mathbf{s}_{u(1)}, ..., \mathbf{s}_{u(m-1)}$ to form the codeword $\mathbf{c}_u$.

The interleaving procedure to construct codewords is illustrated in Figure 3. As we describe in more detail below, the pilot block $\mathbf{p}$ and the message blocks $\mathbf{s}_i$ are designed so that no string of length $2\log n$ appears in both $\mathbf{p}$ and $\mathbf{s}_j$ for some j. The "sprinkled" nature of the pilot sequence in $\mathbf{c}_u$ prevents it from being fragmented by the tearing process. More precisely, we let n/m be the length of the pilot block $\mathbf{p}$ and of each message block $\mathbf{s}_j$, where $m \geq 2$ is a positive integer. Notice that a fragment of length N_i must contain at least N_i/m pilot symbols in it. As we will see, provided that N_i is long enough, this will allow its location on X^n to be uniquely determined.

A. Codebook construction

For a fixed value of m, we will construct a pilot sequence $\mathbf{p}$ of length n/m. Notice that m controls the fraction of the codeword that is dedicated to pilot symbols. The pilot sequence $\mathbf{p}$ is constructed as a *de Bruijn sequence* of order $\log(n/m)$ [32]. This sequence has length $2^{\log(n/m)} = n/m$ and it has the property that each length $\log(n/m)$ substring appears exactly once. For example, a de Bruijn sequence of order 4 is $S = 0000100110101111$. Notice that each binary string of length 4 appears exactly once (when we view S as a cyclic sequence). In order to simplify the exposition, we will assume that $\log(n/m)$ is an integer. The results can be extended to the general case, by considering a de Bruijn sequence of order $\lceil \log(n/m) \rceil$.

In order to build our codebook, we will interleave codewords from an erasure code with the pilot sequence $\mathbf{p}$. Suppose we have an erasure code $\mathcal{C}_{\text{er}}$ with rate R_{er} and blocklength n/m. We consider applying a random shift to $\mathcal{C}_{\text{er}}$. More precisely, we generate a length-n/m i.i.d. $\text{Ber}(1/2)$ sequence $Z^{n/m}$ and take the modulo-2 sum of every codeword in $\mathcal{C}_{\text{er}}$ with $Z^{n/m}$ to form a modified codebook $\tilde{\mathcal{C}}_{\text{er}}$. Notice that this effectively does not change the code, as the shift $Z^{n/m}$ is the same for all codewords and can be subtracted at the receiver side. The probability that a randomly shifted codeword $\mathbf{s} \in \tilde{\mathcal{C}}_{\text{er}}$ shares an identical length-k segment with the pilot sequence can be upper bounded as

$$\begin{aligned}&\Pr\left(\mathbf{p}[i:i+k-1] = \mathbf{s}[j:j+k-1],\right.\\ &\qquad\left.\text{for } 1 \leq i \leq n/m-k,\ 1 \leq j \leq n/m-k\right)\\ &\leq (n/m)^2 2^{-k}\end{aligned}$$

Therefore, if we let $k = (2+\delta)\log n$ for $\delta > 0$,

$$\begin{aligned}&\Pr\left(\mathbf{p}[i:i+k-1] = \mathbf{s}[j:j+k-1],\right.\\ &\qquad\left.\text{for } 1 \leq i \leq n/m-k,\ 1 \leq j \leq n/m-k\right) \to 0, \quad (23)\end{aligned}$$

as $n \to \infty$. This means that for any $\epsilon > 0$, for n large enough, it is possible to choose $Z^{n/m}$ so that at least a $(1-\epsilon)$ fraction of the shifted codewords in $\tilde{\mathcal{C}}_{\text{er}}$ contain no length-$(2+\delta)\log n$ segment that is also in the pilot sequence $\mathbf{p}$.

Let $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_{|\mathcal{S}|}\} \subset \tilde{\mathcal{C}}_{\text{er}}$ be a set with $(1-\epsilon)2^{\frac{n}{m}R_{\text{er}}}$ such sequences. We build each codeword $\mathbf{c}_u$ by taking $m-1$ sequences $\mathbf{s}_{u(1)}, ..., \mathbf{s}_{u(m-1)}$ from $\mathcal{S}$ and interleaving their symbols with the symbols from $\mathbf{p}$. More precisely, for each $u \in \{1, ..., |\mathcal{S}|\}^{m-1}$ we build the codeword $\mathbf{c}_u = (\mathbf{c}_u[0], ..., \mathbf{c}_u[n-1])$ as

$$\mathbf{c}_u[mt+j] = \begin{cases} \mathbf{p}[t], & \text{for } j = 0,\\ \mathbf{s}_{u(j)}[t], & \text{for } j = 1, ..., m-1,\end{cases}$$

for $t = 0, ..., n/m-1$, as illustrated in Figure 3. The resulting codebook $\mathcal{C}$ has $|\mathcal{S}|^{m-1} = (1-\epsilon)^{m-1}2^{(1-1/m)nR_{\text{er}}}$ codewords. Notice that, for any fixed m and any small $\epsilon > 0$, such a codebook can be constructed for n large enough, yielding a coding rate of approximately $(1-1/m)R_{\text{er}}$.

B. Decoding and analysis

As illustrated in Figure 3, a codeword $\mathbf{c}_u$ will contain one symbol of $\mathbf{p}$ every m bits. Hence, if a given output fragment has length N_i, it must contain at least N_i/m symbols from the pilot sequence (though at unknown locations).

Suppose a random fragment has length $N_i > (2+\delta)m\log n$. By the previous argument, it must contain at least $(2+\delta)\log n$ pilot symbols. We claim that the location of such a fragment in its original codeword $\mathbf{c}_u$ can be uniquely identified by aligning it to a "generic" codeword $\mathbf{c}_?$ that only contains the pilot symbols, as illustrated in Figure 4. Suppose by contradiction that the fragment can be properly aligned to $\mathbf{c}_?$ at an incorrect location. Since sequences of $\log n$ consecutive symbols of $\mathbf{p}$ are unique, it must be the case that $N_i/m > (2+\delta)\log n$ pilot symbols of $\mathbf{c}_?$ align with N_i/m non-pilot symbols of the fragment. However, these N_i/m symbols must correspond to

consecutive symbols in one of the codewords $\mathbf{s}_i$ from $\mathcal{S}$. Since no block of length $(2+\delta)\log n$ of $\mathbf{p}$ appears in any $\mathbf{s}_i \in \mathcal{S}$, this is a contradiction.

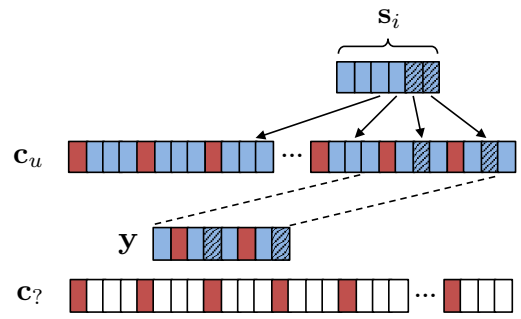

Fig. 4. Aligning a fragment $\mathbf{y}$ incorrectly to the generic codeword $\mathbf{c}_?$ requires $|\mathbf{y}|/m$ pilot symbols in $\mathbf{c}_?$ to align with $|\mathbf{y}|/m$ consecutive symbols of $\mathbf{s}_i$.

This suggests a straightforward decoding procedure for the code outlined above. Each of the received fragments with length at least $(2+\delta)m\log n$ is aligned to $\mathbf{c}_?$. Shorter fragments are discarded, and their locations on $\mathbf{c}_?$ are treated as erasures. This effectively converts the channel into an erasure channel (though not memoryless) with a total number of erasures given by

$$\sum_{i=1}^{K} N_i \mathbf{1}_{\{N_i<(2+\delta)\log n\}} = n(1-c_{(2+\delta)m}), \tag{24}$$

where $c_{(2+\delta)m}$ is the coverage by fragments of length at least $(2+\delta)m\log n$, as defined in Definition 1. Hence, as long as the rate of the original erasure code $\mathcal{C}_{\rm er}$ satisfies $R_{\rm er} < c_{(2+\delta)m}$, the code for the torn-paper channel can be decoded with small error probability as $n\to\infty$. From Lemma 6, we know that $c_{(2+\delta)m}$ concentrates around its mean as $n\to\infty$. Hence we can choose $R_{\rm er}$ arbitrarily close to $E[c_{(2+\delta)m}]$ and achieve arbitrarily close to rate

$$\left(1-\tfrac{1}{m}\right)\lim_{n\to\infty} E[c_{(2+\delta)m}]. \tag{25}$$

Since $\delta>0$ can be chosen arbitrarily small, and using (21), as $n\to\infty$, we can achieve any rate below

$$\left(1-\tfrac{1}{m}\right)\lim_{n\to\infty} E[c_{2m}] = \left(1-\tfrac{1}{m}\right)(2m\alpha+1)\,e^{-2m\alpha}. \tag{26}$$

This expression can be optimized over positive integers m, yielding the achievable curve shown in Figure 2. We notice that the gap between this efficient interleaved-pilot approach and the actual capacity is still very significant.

VIII. Proofs of Lemmas

Lemma 1. *The number of fragments in $\mathcal{Y}_k$ satisfies*

$$\Pr\left(||\mathcal{Y}_k| - np_nq_{k,n}| > \epsilon np_n\right) \le 4e^{-np_n^2\epsilon^2/4},$$

for any $\epsilon>0$ and n large enough.

Proof of Lemma 1. First notice that, since $K = 1+\sum_{i=2}^{n} T_i$, where $T_2,...,T_n$ are i.i.d. Bernoulli(p_n) random variables, $E[K] = np_n + (1-p_n)$, and using Hoeffding's inequality,

$$\begin{aligned}
&\Pr(|K-np_n| > \delta np_n)\\
&\quad= \Pr\left(|K-E[K]+(1-p_n)| > \delta np_n\right)\\
&\quad\le \Pr\left(|K-E[K]| > \delta np_n-(1-p_n)\right)\\
&\quad= \Pr\left(\left|\sum_{i=2}^{n}(T_i-p_n)\right| > (n-1)\frac{\delta np_n-(1-p_n)}{n-1}\right)\\
&\quad\le 2e^{-2(n-1)\left(\frac{\delta np_n-(1-p_n)}{n-1}\right)^2} \le 2e^{-2n\left(\frac{\delta np_n-(1-p_n)}{n}\right)^2}\\
&\quad\le 2e^{-np_n^2\delta^2},
\end{aligned} \tag{27}$$

where the last inequality holds for n large enough.

Now suppose the sequence $N_1, N_2, ...$ of independent Geometric(p_n) random variables is an infinite sequence (and does not stop at K). Let Z_i be the binary indicator of the event $\{(k-1)/L \le N_i/\log n < k/L\}$, and $\tilde{Z} = \sum_{i=1}^{np_n} Z_i$. Intuitively, $|\mathcal{Y}_k|$ and $\tilde{Z}$ should be close. In particular, $||\mathcal{Y}_k| - \tilde{Z}| \le |K-np_n|$. Moreover, $E[\tilde{Z}] = np_nq_{k,n}$. If $|\tilde{Z}-np_nq_{k,n}| < \frac{1}{2}\epsilon np_n$ and $||\mathcal{Y}_k|-\tilde{Z}| < |K-np_n| < \frac{1}{2}\epsilon np_n$, by the triangle inequality, $||\mathcal{Y}_k|-np_nq_{k,n}| < \epsilon np_n$. Therefore,

$$\begin{aligned}
&\Pr\left(||\mathcal{Y}_k| - np_nq_{k,n}| > \epsilon np_n\right)\\
&\quad\le \Pr\left(|\tilde{Z}-np_nq_{k,n}| > \tfrac{1}{2}\epsilon np_n\right)\\
&\qquad + \Pr\left(|K-np_n| > \tfrac{1}{2}\epsilon np_n\right)\\
&\quad\le 2e^{-np_n\epsilon^2/2} + 2e^{-np_n^2\epsilon^2/4} \le 4e^{-np_n^2\epsilon^2/4}
\end{aligned}$$

where we used Hoeffding's inequality and (27). □

Lemma 2. *If $N^{(n)}$ is a* Geometric(p_n) *random variable and* $\lim_{n\to\infty} E[N^{(n)}]/\log n = 1/\alpha$, *then*

$$\lim_{n\to\infty}\Pr\left(N^{(n)} \ge \beta\log n\right) = e^{-\alpha\beta}.$$

Proof of Lemma 2. By definition,

$$\begin{aligned}
\Pr\left(N^{(n)} \ge \beta\log n\right) &= (1-p_n)^{\lceil\beta\log n\rceil}\\
&= \left(1-\frac{1}{E[N^{(n)}]}\right)^{E[N^{(n)}](\lceil\beta\log n\rceil/E[N^{(n)}])}.
\end{aligned}$$

As $n\to\infty$, $\lceil\beta\log n\rceil/E[N^{(n)}] \to \alpha\beta$, $E[N^{(n)}]\to\infty$, and $(1-1/E[N^{(n)}])^{E[N^{(n)}]} \to e^{-1}$, implying the lemma. □

Lemma 3. *If $\mathcal{Y}_k$ is defined as in (5) for $k=1,...,\infty$,*

$$\lim_{n\to\infty}\frac{H(\mathcal{Y})}{n} \le \sum_{k=1}^{\infty}\lim_{n\to\infty}\frac{H(\mathcal{Y}_k)}{n}.$$

Proof of Lemma 3. For a fixed integer A, we define $\mathcal{Y}_{\ge A} = \{\vec{Y}_i : N_{\pi_i} \ge (A/L)\log n\}$ and we have

$$\begin{aligned}
\lim_{n\to\infty}\frac{H(\mathcal{Y})}{n} &\le \lim_{n\to\infty}\sum_{k=1}^{A}\frac{H(\mathcal{Y}_k)}{n} + \lim_{n\to\infty}\frac{H(\mathcal{Y}_{\ge A})}{n}\\
&= \sum_{k=1}^{A}\lim_{n\to\infty}\frac{H(\mathcal{Y}_k)}{n} + \lim_{n\to\infty}\frac{H(\mathcal{Y}_{\ge A})}{n}.
\end{aligned} \tag{28}$$

If we define c_γ as in Definition 1, from Lemma 6, we have

$$\lim_{n\to\infty} E\left[c_{A/L}\right] = (\alpha A/L+1)e^{-\alpha A/L}.$$

Moreover, for any $\delta>0$, from Lemma 6, the event

$$\mathcal{A} = \{c_{A/L} > (\alpha A/L+1)e^{-\alpha A/L}+\delta\}$$

has vanishing probability as $n\to\infty$. This allows us to write

$$\begin{aligned} H(\mathcal{Y}_{\geq A}) &\leq H(\mathcal{Y}_{\geq A}|\bar{\mathcal{A}}) + H(\mathcal{Y}_{\geq A}|\mathcal{A})\Pr(\mathcal{A}) + 1 \\ &\leq H(\mathcal{Y}_{\geq A}|\bar{\mathcal{A}}) + 2n\Pr(\mathcal{A}) + 1 \\ &\leq 2n\left[(\alpha A/L+1)e^{-\alpha A/L}+\delta\right] + o(n). \end{aligned}$$

Hence, from (28), we have that for every A and $\delta>0$,

$$\lim_{n\to\infty}\frac{H(\mathcal{Y})}{n} \leq \sum_{k=1}^{A}\lim_{n\to\infty}\frac{H(\mathcal{Y}_k)}{n} + 2(\alpha A/L+1)e^{-\alpha A/L} + 2\delta.$$

Notice that $(\alpha A/L+1)e^{-\alpha A/L}\to 0$ as $A\to\infty$. Therefore, we can let $\delta\to 0$ and $A\to\infty$, and we conclude that

$$\lim_{n\to\infty}\frac{H(\mathcal{Y})}{n} \leq \sum_{k=1}^{\infty}\lim_{n\to\infty}\frac{H(\mathcal{Y}_k)}{n}.$$

□

Lemma 4. *The number of fragments in $\mathcal{Y}_\gamma$ satisfies*

$$\Pr\left(\left||\mathcal{Y}_\gamma| - e^{-\alpha\gamma}np_n\right| > \epsilon np_n\right) \leq 4e^{-np_n^2\epsilon^2/9}$$

for any $\epsilon>0$ and n large enough.

Proof of Lemma 4. Let $Z_i = \mathbf{1}_{\{N_i\geq\gamma\log n\}}$, for $i=1,2,....$ Then $|\mathcal{Y}_\gamma| = \sum_{i=1}^{K} Z_i$. Since K is random (and not independent of the N_is), we need to follow similar steps to those in the proof of Lemma 1.

Let us assume that the sequence $N_1, N_2, ...$ of independent $\text{Geometric}(p_n)$ random variables is an infinite sequence and let $\tilde{Z} = \sum_{i=1}^{np_n} Z_i$. Notice that $\tilde{Z}$ is a sum of i.i.d. Bernoulli random variables with

$$E[\tilde{Z}] = np_n\Pr(N_1\geq\gamma\log n), \tag{29}$$

and the standard Hoeffding's inequality can be applied. Moreover, from (29) and Lemma 2,

$$\lim_{n\to\infty} E[\tilde{Z}]/(np_n) = e^{-\alpha\gamma}$$

and, for any $\delta>0$, $|E[\tilde{Z}] - e^{-\alpha\gamma}np_n| < \delta np_n$, for n large enough. If we set $\delta=\epsilon/3$, for n large enough, we have $|E[\tilde{Z}]-e^{-\alpha\gamma}np_n| < \frac{1}{3}\epsilon np_n$. Moreover, if $|\tilde{Z}-E[\tilde{Z}]| < \frac{1}{3}\epsilon np_n$ and $||\mathcal{Y}_\gamma| - \tilde{Z}| < |K-np_n| < \frac{1}{3}\epsilon np_n$, by the triangle inequality (applied twice), $||\mathcal{Y}_\gamma| - e^{-\alpha\gamma}np_n| < \epsilon np_n$. Hence,

$$\begin{aligned} &\Pr\left(\left||\mathcal{Y}_\gamma| - e^{-\alpha\gamma}np_n\right| > \epsilon np_n\right) \\ &\leq \Pr\left(|\tilde{Z}-E[\tilde{Z}]| > \tfrac{1}{3}\epsilon np_n\right) + \Pr\left(\left||\mathcal{Y}_\gamma| - \tilde{Z}\right| > \tfrac{1}{3}\epsilon np_n\right) \\ &\leq \Pr\left(\left|\tilde{Z}-E|Z|\right| > \tfrac{1}{3}\epsilon np_n\right) + \Pr\left(|K-np_n| > \tfrac{1}{3}\epsilon np_n\right) \\ &\leq 2e^{-2np_n\epsilon^2/9} + 2e^{-np_n^2\epsilon^2/9} \leq 4e^{-np_n^2\epsilon^2/9} \end{aligned}$$

where we used Hoeffding's inequality and (27). □

Lemma 5. *If $N^{(n)}$ is a $\text{Geometric}(p_n)$ random variable and $\lim_{n\to\infty} E[N^{(n)}]/\log n = 1/\alpha$, then, for any $\beta\geq 0$,*

$$\begin{aligned} &\lim_{n\to\infty} E\left[N^{(n)}\mathbf{1}_{\{N^{(n)}\geq\gamma\log n\}}\right]/\log n \\ &\qquad = E\left[\tilde{N}\mathbf{1}_{\{\tilde{N}\geq\gamma\}}\right] = \left(\gamma+\frac{1}{\alpha}\right)e^{-\alpha\gamma}, \end{aligned}$$

where $\tilde{N}$ is an $\text{Exponential}(\alpha)$ random variable.

Proof of Lemma 5. We first notice that

$$\begin{aligned} &\frac{1}{\log n}E\left[N^{(n)}\mathbf{1}_{\{N^{(n)}\geq\gamma\log n\}}\right] \\ &\quad = \frac{1}{\log n}E\left[N^{(n)}\,\middle|\,N^{(n)}\geq\gamma\log n\right]\Pr\left(N^{(n)}\geq\gamma\log n\right) \\ &\quad = \frac{1}{\log n}\left(\lceil\gamma\log n\rceil + E[N^{(n)}]\right)\Pr\left(N^{(n)}\geq\gamma\log n\right), \end{aligned}$$

where we used the memoryless property of the Geometric distribution. As $n\to\infty$, we have $\lceil\gamma\log n\rceil/\log n\to\gamma$, $E[N^{(n)}]/\log n\to 1/\alpha$. Moreover, from Lemma 2, $\Pr\left(N^{(n)}\geq\gamma\log n\right)\to e^{-\alpha\gamma}$, and the lemma follows. □

Lemma 6. *If c_γ is defined as in (18), then, for any $\epsilon>0$,*

$$\Pr\left(\left|c_\gamma - (\alpha\gamma+1)e^{-\alpha\gamma}\right| > \epsilon\right) \leq \frac{19}{\epsilon^2 np_n^2}$$

for n large enough.

Proof of Lemma 6. Since $c_\gamma = \frac{1}{n}\sum_{i=1}^{K} N_i\mathbf{1}_{\{N_i\geq\gamma\log n\}}$, where K is a random variable, we once again follow an approach similar to the one in the proof of Lemma 1.

Let us assume that the sequence $N_1, N_2, ...$ of independent $\text{Geometric}(p_n)$ random variables is an infinite sequence. Let $Z_i = N_i\mathbf{1}_{\{N_i\geq\gamma\log n\}}$, and $\tilde{Z} = \sum_{i=1}^{np_n} Z_i$. Since $E[\tilde{Z}] = np_nE[N_1\mathbf{1}_{\{N_1\geq\gamma\log n\}}]$, by Lemma 5,

$$\lim_{n\to\infty}\frac{E[\tilde{Z}]}{n} \to \alpha\left(\gamma+\frac{1}{\alpha}\right)e^{-\alpha\gamma}. \tag{30}$$

Intuitively, $Z := nc_\gamma$ and $\tilde{Z}$ should be close. If $\tilde{Z} > Z$, then $np_n > K$, and

$$|Z-\tilde{Z}| = \sum_{i=K+1}^{np_n} Z_i \leq \sum_{i=K+1}^{np_n} N_i \leq \left|\sum_{i=1}^{np_n} N_i - n\right|. \tag{31}$$

If $Z > \tilde{Z}$, then $K > np_n$, and

$$|Z-\tilde{Z}| = \sum_{i=np_n+1}^{K} Z_i \leq \sum_{i=np_n+1}^{K} N_i \leq \left|\sum_{i=1}^{np_n} N_i - n\right|. \tag{32}$$

Hence, for any $\delta>0$, we have that

$$\begin{aligned} \Pr\left(|Z-\tilde{Z}| > \delta np_n\right) &\leq \Pr\left(\left|\sum_{i=1}^{np_n} N_i - n\right| > \delta np_n\right) \\ &\leq e^{-np_n(\delta-\ln(1+\delta))} + e^{-np_n(-\delta-\ln(1-\delta))} \\ &\leq 2e^{-np_n(\delta-\ln(1+\delta))}. \end{aligned} \tag{33}$$

where we used the Chernoff bound for geometrically distributed random variables [34], and the fact that $x-\ln(1+x) < -x-\ln(1-x)$ for $x>0$.

To bound the probability that $|\tilde{Z} - E[\tilde{Z}]| > \delta n$, we can use a Chernoff bound, which requires the computation of the rate function for $N_1 \mathbf{1}_{\{N_1 \geq \gamma \log n\}}$. A simpler approach is to use Chebyshev's inequality, which yields

$$\begin{aligned}\Pr\left(|\tilde{Z} - E[\tilde{Z}]| > \delta n\right) &\leq \frac{\mathrm{Var}(Z_1)}{\delta^2 n} \leq \frac{E[Z_1^2]}{\delta^2 n} \\ &= \frac{E[N_1^2 \mathbf{1}_{\{N_1 \geq \gamma \log n\}}]}{\delta^2 n} \leq \frac{E[N_1^2]}{\delta^2 n} = \frac{2 - p_n}{\delta^2 n p_n^2}. \end{aligned} \tag{34}$$

From (30), we know that for any $\delta > 0$ and n large enough,

$$|E[\tilde{Z}] - n(\alpha\gamma + 1)e^{-\alpha\gamma}| < \delta n.$$

Moreover, if $|\tilde{Z} - E[\tilde{Z}]| < \frac{1}{3}\epsilon n$, $|nc_\gamma - \tilde{Z}| < \frac{1}{3}\epsilon n$, and $|E[\tilde{Z}] - n(\alpha\gamma+1)e^{-\alpha\gamma}| < \frac{1}{3}\epsilon n$, then, by the triangle inequality, $|c_\gamma - (\alpha\gamma + 1)e^{-\alpha\gamma}| < \epsilon$. Therefore, for n large enough so that $|E[\tilde{Z}] - n(\alpha\gamma + 1)e^{-\alpha\gamma}| < \frac{1}{3}\epsilon n$,

$$\begin{aligned}&\Pr\left(\left|c_\gamma - (\alpha\gamma + 1)e^{-\alpha\gamma}\right| > \epsilon\right) \\ &\quad\leq \Pr\left(|\tilde{Z} - E[\tilde{Z}]| > \tfrac{1}{3}\epsilon n\right) + \Pr\left(|\tilde{Z} - Z| > \tfrac{1}{3}\epsilon n\right) \\ &\quad\leq \Pr\left(|\tilde{Z} - E[\tilde{Z}]| > \tfrac{1}{3}\epsilon n\right) + \Pr\left(|\tilde{Z} - Z| > \tfrac{1}{3}\epsilon n p_n\right) \\ &\quad\leq 18/(\epsilon^2 n p_n^2) + 2e^{-np_n(\epsilon/3 - \ln(1+\epsilon/3))} \leq 19/(\epsilon^2 n p_n^2),\end{aligned}$$

where we used (33) and (34), and the last inequality follows for n large enough. □

REFERENCES

[1] M. H. Costa, "Writing on Dirty Paper," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, May 1983.

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

[3] 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.

[4] R. 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," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[5] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-Based Archival Storage System," in *Proc. of ASPLOS*, (New York, NY, USA), pp. 637–649, ACM, 2016.

[6] Y. Erlich and D. Zielinski, "Dna fountain enables a robust and efficient storage architecture," *Science*, 2017.

[7] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, and et al., "Random access in large-scale DNA data storage," *Nature Biotechnology*, 2018.

[8] R. Heckel, G. Mikutis, and R. N. Grass, "A Characterization of the DNA Data Storage Channel," *arXiv:1803.03322*, 2018.

[9] K. R. Pomraning, K. M. Smith, E. L. Bredeweg, L. R. Connolly, P. A. Phatale, and M. Freitag, "Library preparation and data analysis packages for rapid genome sequencing," in *Fungal Secondary Metabolism*, pp. 1–22, Springer, 2012.

[10] A. Motahari, G. Bresler, and D. Tse, "Information Theory of DNA Shotgun Sequencing," *IEEE Transactions on Information Theory*, vol. 59, pp. 6273–6289, Oct. 2013.

[11] G. Bresler, M. Bresler, and D. Tse, "Optimal Assembly for High Throughput Shotgun Sequencing," *BMC Bioinformatics*, 2013.

[12] R. Gabrys and O. Milenkovic, "Unique reconstruction of coded sequences from multiset substring spectra," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2540–2544, IEEE, 2018.

[13] T. Laver, J. Harrison, P. O'neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme, "Assessing the performance of the oxford nanopore technologies minion," *Biomolecular detection and quantification*, vol. 3, pp. 1–8, 2015.

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

[15] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," in *IEEE International Symposium on Information Theory (ISIT)*, 2019.

[16] L. Wang, S. Hu, and O. Shayevitz, "Quickest sequence phase detection," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5834–5849, 2017.

[17] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results.," in *SODA*, vol. 8, pp. 389–398, 2008.

[18] S. R. Srinivasavaradhan, M. Du, S. Diggavi, and C. Fragouli, "On maximum likelihood reconstruction over multiple deletion channels," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 436–440, IEEE, 2018.

[19] M. Cheraghchi, J. Ribeiro, R. Gabrys, and O. Milenkovic, "Coded trace reconstruction," in *2019 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2019.

[20] S. Marcovich and E. Yaakobi, "Reconstruction of strings from their substrings spectrum," *arXiv preprint arXiv:1912.11108*, 2019.

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

[22] H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, p. 14138, 2015.

[23] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes: New bounds and constructions," in *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

[24] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Transactions on Information Theory*, 2017.

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

[26] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Anchor-based correction of substitutions in indexed sets," *arXiv preprint arXiv:1901.06840*, 2019.

[27] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for dna storage," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2411–2415, IEEE, 2018.

[28] M. Kovačević and V. Y. Tan, "Codes in the space of multisets—coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, 2018.

[29] W. Song and K. Cai, "Sequence-subset distance and coding for error control in dna data storage," *arXiv preprint arXiv:1809.05821*, 2018.

[30] A. Tandon, V. Y. Tan, and L. R. Varshney, "The bee-identification problem: Bounds on the error exponent," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7405–7416, 2019.

[31] A. Makur, "Information capacity of bsc and bec permutation channels," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1112–1119, IEEE, 2018.

[32] N. G. De Bruijn, "A combinatorial problem," in *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, vol. 49, pp. 758–764, 1946.

[33] L. Wang, S. Hu, and O. Shayevitz, "Quickest sequence phase detection," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5834–5849, 2017.

[34] S. Janson, "Tail bounds for sums of geometric and exponential variables," *Statistics & Probability Letters*, vol. 135, pp. 1–6, 2018.