Communicating over the Torn-Paper Channel

Ilan Shomorony
ECE Department
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 *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].

We consider the scenario where the channel input is a length-n binary string, which 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.

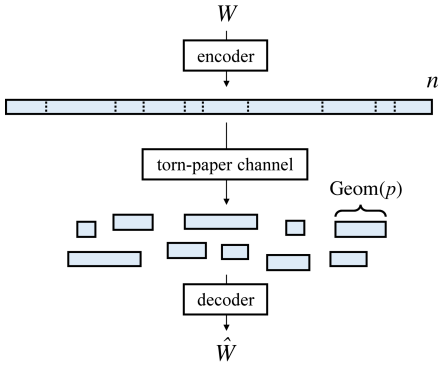

Fig. 1. The torn-paper channel.

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, and any achievable rate in this case 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 this case becomes $(1-\alpha)^+$.

It is not clear a priori whether the capacity of the torn-paper channel 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 $C = e^{-\alpha}$. This boost in capacity, illustrated in Figure 2, 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.

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 $X^n \in \{0,1\}^n$. 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 $\text{Geometric}(p_n)$-length for a *tearing probability* p_n. More

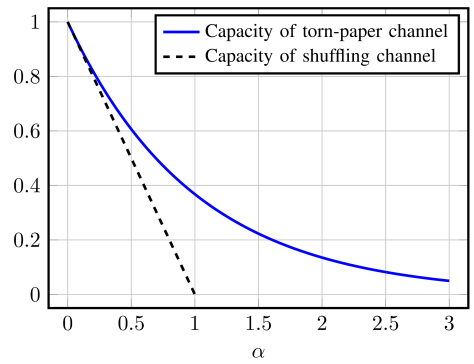

Fig. 2. Comparison between the capacity of the torn-paper channel $C = e^{-\alpha}$ and the capacity of the shuffling channel with fragments of fixed length $1/p_n$.

specifically, let $N_1, N_2, ...$ be i.i.d. Geometric(p_n) random variables. Let 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 $i = 1, ..., K$, $\vec{Y}_i = \vec{X}_{\pi_i}$.

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 $\{\mathcal{C}_n\}$, with blocklength $n \to \infty$, 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 broke evenly into np_n pieces of length $[N_1] = 1/p_n$, results from [15] imply that placing a unique index of length $\log(np_n)$ in each fragment is capacity optimal. In this case, the capacity is $(1-\alpha)^+$, where $\alpha = \lim_{n\to\infty} p_n \log n$ (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 long, 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 increases the channel capacity.

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

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

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

In Sections IV and V 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 inequalities, allows us to partition the set of fragments into multiple bins of fragments with roughly the same size and view the torn-paper coding, in essence, as parallel channels with fixed-size fragments. Our achievability is based on random coding arguments and does not provide much insight into efficient coding schemes. This opens up interesting avenues for future research.

IV. Converse

In order to prove the converse, 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 involves fragments of roughly the same size. More precisely, for an integer parameter L, we will let

$$\begin{aligned}\mathcal{X}_k &= \left\{\vec{X}_i : \tfrac{k-1}{L}\log n \leq N_i < \tfrac{k}{L}\log n\right\} \text{ and} \\ \mathcal{Y}_k &= \left\{\vec{Y}_i : \tfrac{k-1}{L}\log n \leq N_{\pi_i} < \tfrac{k}{L}\log n\right\},\end{aligned} \tag{3}$$

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 [16].

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

$$q_{k,n} = \Pr\left(\frac{k-1}{L} \leq \frac{N_1}{\log n} < \frac{k}{L}\right), \tag{4}$$

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

Lemma 1. *For any $\epsilon > 0$ and n large enough,*

$$\Pr\left(||\mathcal{Y}_k| - np_nq_{k,n}| > \epsilon np_n\right) \leq 4e^{-np_n^2\epsilon^2/4}. \quad (5)$$

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 $\text{Exponential}(\alpha)$ distribution. This known fact is stated as the following lemma, which we also prove in Section VII.

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

$$\lim_{n\to\infty} \Pr(N^{(n)} \geq \beta \log n) = e^{-\alpha\beta}. \quad (6)$$

Lemma 1 implies that $E[|\mathcal{Y}_k|] = np_nq_{k,n} + o(np_n)$, and

$$\begin{aligned}\lim_{n\to\infty} \frac{E\left[|\mathcal{Y}_k|\right]}{np_n} &= \lim_{n\to\infty} \frac{np_nq_{k,n} + o(np_n)}{np_n} \\ &= \lim_{n\to\infty} \Pr\left(\tfrac{k-1}{L} \leq \tfrac{N_1}{\log n} < \tfrac{k}{L}\right) \\ &= e^{-\alpha(k-1)/L} - e^{-\alpha k/L}, \end{aligned} \quad (7)$$

where the last equality follows from Lemma 2. Next, we define 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,

$$\begin{aligned} H(\mathcal{Y}_k) &\leq H(\mathcal{Y}_k, \mathbf{1}_{\mathcal{E}_{k,n}}) \leq 1 + H(\mathcal{Y}_k|\mathbf{1}_{\mathcal{E}_{k,n}}) \\ &\leq 1 + 2n\Pr(\mathcal{E}_{k,n}) + H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n}), \end{aligned} \quad (8)$$

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

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

Moreover, given $\bar{\mathcal{E}}_k$,

$$\begin{aligned} |\mathcal{Y}_k| &\leq np_nq_{k,n} + \epsilon np_n \\ &= np_n\left[\epsilon + \Pr\left(\tfrac{k-1}{L} \leq \tfrac{N_1}{\log n} < \tfrac{k}{L}\right)\right] \triangleq M, \end{aligned} \quad (9)$$

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, from Lemma 1 in [16],

$$\begin{aligned} H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n}) &\leq \log\binom{2n^{k/L} + M - 1}{M} \\ &\leq M\log\left(\frac{e(2n^{k/L} + M - 1)}{M}\right) \\ &= M\left[\log\left(2n^{k/L} + M - 1\right) + \log(e) - \log M\right] \\ &= M\left[\max(\tfrac{k}{L}\log n, \log M) - \log M + o(\log n)\right] \\ &= M\left[(\tfrac{k}{L}\log n - \log M)^+ + o(\log n)\right] \\ &= M\log n\left[(\tfrac{k}{L} - \log M/\log n)^+ + o(1)\right]. \end{aligned} \quad (10)$$

From (9), we have $\log M/\log n \to 1$ as $n \to \infty$. Combining (8) and (10), dividing by n, and letting $n \to \infty$ yields

$$\begin{aligned}\lim_{n\to\infty} \frac{H(\mathcal{Y}_k)}{n} &= \lim_{n\to\infty} \frac{H(\mathcal{Y}_k|\bar{\mathcal{E}}_{k,n}) + 1 + 2n\Pr(\mathcal{E}_{k,n})}{n} \\ &\leq \lim_{n\to\infty} \frac{M\log n}{n}\left(\frac{k}{L} - 1\right)^+ \\ &= \lim_{n\to\infty} p_n\log n\,(q_{k,n} + \epsilon_n)\left(\frac{k}{L} - 1\right)^+ \\ &= \alpha\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right)\left(\frac{k}{L} - 1\right)^+. \end{aligned} \quad (11)$$

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

$$nR \leq I(X^n; \mathcal{Y}) + o(n) \leq H(\mathcal{Y}) + o(n), \quad (12)$$

and we conclude that any achievable rate must satisfy $R \leq \lim_{n\to\infty} \frac{H(\mathcal{Y})}{n}$. In order to connect (12) and (11), we state the following lemma, which allows us to move the limit inside the summation. The proof is in Section VII.

Lemma 3. *If $\mathcal{Y}_k$ is defined as in (3) 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 (11), we can upper bound any achievable rate as

$$\begin{aligned} R &\leq \lim_{n\to\infty} \frac{H(\mathcal{Y})}{n} \leq \sum_{k=1}^{\infty} \lim_{n\to\infty} \frac{H(\mathcal{Y}_k)}{n} \\ &= \sum_{k=L+1}^{\infty} \alpha\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right)(\tfrac{k}{L} - 1) \\ &= \frac{\alpha}{L}\sum_{k=L+1}^{\infty} k\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) \\ &\quad - \alpha\sum_{k=L+1}^{\infty}\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) \\ &= \frac{\alpha}{L}\sum_{k=L+1}^{\infty} k\left(e^{-\alpha(k-1)/L} - e^{-\alpha k/L}\right) - \alpha e^{-\alpha}, \end{aligned} \quad (13)$$

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 $R < e^{-\alpha}$.

V. 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 IV, 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

$$\lim_{n\to\infty} \frac{E[|\mathcal{Y}_\gamma|]}{n \cdot p_n} = \lim_{n\to\infty} \Pr\left(N_1 \geq \gamma \log n\right) = e^{-\alpha\gamma}. \quad (14)$$

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 VII, we prove the following concentration result.

Lemma 4. *For any $\epsilon > 0$, as $n \to \infty$,*

$$\Pr\left(||\mathcal{Y}_\gamma| - e^{-\alpha\gamma} np_n| > \epsilon np_n\right) \to 0. \quad (15)$$

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 well the fragments in $\mathcal{Y}_\gamma$ *cover* their codeword of origin $\mathbf{x}_1$.

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

$$c_\gamma = \frac{1}{n} \sum_{i=1}^{K} N_i \mathbf{1}_{\{N_i \geq \gamma \log n\}}. \quad (16)$$

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 to 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 $\beta \geq 0$,*

$$\begin{aligned}\lim_{n\to\infty} E\left[N^{(n)} \mathbf{1}_{\{N^{(n)} \geq \gamma \log n\}}\right] / \log n \\ = E\left[\tilde{N} \mathbf{1}_{\{\tilde{N} \geq \gamma\}}\right] = \left(\gamma + \frac{1}{\alpha}\right) e^{-\alpha\gamma}, \quad (17)\end{aligned}$$

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 VII.

Lemma 6. *For any $\epsilon > 0$, as $n \to \infty$,*

$$\Pr\left(\left|c_\gamma - (\alpha\gamma + 1)e^{-\alpha\gamma}\right| > \epsilon\right) \to 0. \quad (18)$$

In particular, Lemma 6 implies that

$$\lim_{n\to\infty} E[c_\gamma] = (\alpha\gamma + 1)e^{-\alpha\gamma}, \quad (19)$$

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

$$\mathcal{B} = \{|\mathcal{Y}_\gamma| > B_1\} \cup \{c_\gamma < B_2\}, \quad (20)$$

then (15) and (18) 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

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

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 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 $R < e^{-\alpha}$.

VI. Concluding Remarks

We introduced the torn-paper channel as a model to study macromolecule-based data storage, where the molecule (typically DNA) is subject to breaks at random locations, and the pieces are recovered out of order. We considered the setting in which breaks occur between any pair of consecutive symbols independently with probability p_n, which leads to a model in which each of the fragments have length Geometric(p_n). A natural direction for future work is to consider the impact of different fragment length distributions. If the fragments are

not geometrically distributed, we will no longer be able to take advantage of the connection with the exponential distribution to establish the capacity. Another future direction is to include symbol errors (*e.g.*, via a BSC) and to consider a model where not all fragments are recovered at the output, which are both important aspects of macromolecule-based storage systems.

Finally, we point out that the achievability argument from Section V is based on a random code construction. In an extended version of this paper [18], we started exploring explicit code constructions for the torn-paper channel based on *de Bruijn* sequences [19], which appear in the context of phase detection sequences [20].

VII. Proofs of Lemmas

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,

$$\begin{aligned}
&\Pr(|K - np_n| > \delta np_n) \\
&\quad = \Pr\left(|K - E[K] + (1 - p_n)| > \delta np_n\right) \\
&\quad \le \Pr\left(|K - E[K]| > \delta np_n - (1 - p_n)\right) \\
&\quad = \Pr\left(\left|\sum_{i=2}^{n}(T_i - p_n)\right| > (n-1)\frac{\delta np_n - (1-p_n)}{n-1}\right) \\
&\quad \le 2e^{-2(n-1)\left(\frac{\delta np_n - (1-p_n)}{n-1}\right)^2} \le 2e^{-2n\left(\frac{\delta np_n - (1-p_n)}{n}\right)^2} \\
&\quad \le 2e^{-np_n^2\delta^2}, \qquad (21)
\end{aligned}$$

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 (21). □

Proof of Lemma 2. By definition,

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

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

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

$$\begin{aligned}
\lim_{n\to\infty}\frac{H(\mathcal{Y})}{n} &\le \lim_{n\to\infty}\sum_{k=1}^{A}\frac{H(\mathcal{Y}_k)}{n} + \lim_{n\to\infty}\frac{H(\mathcal{Y}_{\ge A})}{n} \\
&= \sum_{k=1}^{A}\lim_{n\to\infty}\frac{H(\mathcal{Y}_k)}{n} + \lim_{n\to\infty}\frac{H(\mathcal{Y}_{\ge A})}{n}. \qquad (22)
\end{aligned}$$

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, 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

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

Hence, from (22), we have that for every A and $\delta > 0$,

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

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

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

Proof of Lemma 4. Let $Z_i = \mathbf{1}_{\{N_i \ge \gamma \log n\}}$, 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 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

$$E[\tilde{Z}] = np_n\Pr(N_1 \ge \gamma \log n), \qquad (23)$$

and the standard Hoeffding's inequality can be applied. Moreover, from 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 $\delta = \epsilon/3$ and, 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,

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

$$\leq 2e^{-2np_n\epsilon^2/9} + 2e^{-np_n^2\epsilon^2/9} \leq 4e^{-np_n^2\epsilon^2/9}$$

where we used Hoeffding's inequality and (21). □

Proof of Lemma 5. We first notice that

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

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. □

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 $E[\tilde{Z}] = np_nE[N\mathbf{1}_{\{N_1\geq\gamma\log n\}}]$, by Lemma 5,

$$\lim_{n\to\infty}\frac{E[\tilde{Z}]}{n}\to\alpha\left(\gamma+\frac{1}{\alpha}\right)e^{-\alpha\gamma}. \tag{24}$$

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

$$|Z-\tilde{Z}| = \sum_{i=K+1}^{np_n} Z_i \leq \sum_{i=K+1}^{np_n} N_i \leq \left|\sum_{i=1}^{np_n} N_i - n\right|. \tag{25}$$

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

$$|Z-\tilde{Z}| = \sum_{i=np_n+1}^{K} Z_i \leq \sum_{i=np_n+1}^{K} N_i \leq \left|\sum_{i=1}^{np_n} N_i - n\right|. \tag{26}$$

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

$$\begin{aligned}\Pr\left(|Z-\tilde{Z}|>\delta np_n\right) &\leq \Pr\left(\left|\sum_{i=1}^{np_n} N_i - n\right| > \delta np_n\right)\\ &\leq e^{-np_n(\delta-\ln(1+\delta))} + e^{-n(-\delta-\ln(1-\delta))}\\ &\leq 2e^{-np_n(\delta-\ln(1+\delta))}.\end{aligned} \tag{27}$$

where we used the Chernoff bound for exponentially distributed random variables [17], 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

$$\begin{aligned}\Pr\left(|\tilde{Z}-E[\tilde{Z}]|>\delta n\right) &\leq \frac{\text{Var}(Z_1)}{\delta^2 n} \leq \frac{E[Z_1^2]}{\delta^2 n}\\ &= \frac{E[N_1^2\mathbf{1}_{\{N_1\geq\gamma\log n\}}]}{\delta^2 n} \leq \frac{E[N_1^2]}{\delta^2 n} = \frac{2-p_n}{\delta^2 np_n^2}.\end{aligned} \tag{28}$$

From (24), 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 $|E[\tilde{Z}] - n(\alpha\gamma+1)e^{-\alpha\gamma}| < \frac{1}{3}\epsilon n$, and using (27) and (28),

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

where the last inequality follows for n large enough. □

Acknowledgements

The research of I. Shomorony was supported in part by NSF grant CCF-2007597. The research of A. Vahid was supported in part by NSF grant ECCS-2030285.

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, 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*, 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*, 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, 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*, 2012.
[10] A. Motahari, G. Bresler, and D. Tse, "Information Theory of DNA Shotgun Sequencing," *IEEE Trans. on Inf. Theory*, vol. 59, 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 *Proceedings of ISIT*, 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*, 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 Inf. Theory (ISIT)*, 2019.
[16] 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.
[17] S. Janson, "Tail bounds for sums of geometric and exponential variables," *Statistics & Probability Letters*, vol. 135, pp. 1–6, 2018.
[18] I. Shomorony and A. Vahid, "Torn-Paper Coding," *Submitted. Preprint:* ilanshom.github.io/papers/torn_paper_coding_two_columns.pdf
[19] N. G. De Bruijn, "A combinatorial problem," in *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, vol. 49, pp. 758–764, 1946.
[20] L. Wang, S. Hu, and O. Shayevitz, "Quickest sequence phase detection," *IEEE Transactions on Information Theory*, vol. 63, no. 9, 2017.