Capacity of the Torn Paper Channel with Lost Pieces

Aditya Narayan Ravi
University of Illinois, Urbana-Champaign
anravi2@illinois.edu

Alireza Vahid
University of Colorado, Denver
alireza.vahid@ucdenver.edu

Ilan Shomorony
University of Illinois, Urbana-Champaign
ilans@illinois.edu

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518272

***Abstract*—We study the problem of transmitting a message over a channel that randomly breaks the message block into small fragments, deletes a subset of them, and shuffles the remaining fragments. We characterize the capacity of the binary torn-paper channel under arbitrary fragment length distribution and fragment deletion probabilities. We show that, for a message with block length n, discarding fragments shorter than $\log(n)$ does not affect the achievable rates, and that the capacity is given by a simple closed-form expression that can be understood as "coverage minus reordering-cost".**

I. Introduction

Consider the problem of transmitting a message by writing it on a piece of paper, which is then torn up into pieces of random sizes. Some of these pieces are lost (randomly, depending on their size in general) and the remaining pieces are shuffled. This problem is a generalisation of the Torn Paper Channel (TPC) introduced in [1], which is motivated by DNA-based data storage [2–7]. In DNA-based storage, data is written onto synthesized DNA molecules, which are then stored in solution. During synthesis and storage, molecules in the solution are subject to random breaks. Moreover, when retrieving the data via sequencing, molecules are read in a random order, and many fragments are lost [8].

Concretely, we study a TPC where the input is a length-n binary string and a "tearing" occurs between any bits with some probability. A sequence of i.i.d. random variables $N_1, N_2, \dots$ determine the length of the fragments obtained. Moreover each piece can be independently deleted with a probability d. In general, we allow d to be a function $d(\cdot)$ of the fragment length N_i. The channel output is an unordered multiset of all the fragments that remain. The torn-paper channel with lost pieces is illustrated in Figure 1.

In order to build up some intuition, we consider two previously established results. The first one is the capacity of the *shuffling channel* [9]. The input to the shuffling channel are strings of a *fixed length*, which are shuffled. Let us define the expected fragment length to be $E[N_i] \triangleq \ell_n$. We now consider a *shuffling channel* with a *fixed length* ℓ_n (which is actually just a special case of the TPC with $N_i = \ell_n$). The results in [9] imply that the capacity of this shuffling channel is

$$C_{\text{shuf}} = \left(1 - \lim_{n\to\infty} \frac{\log n}{\ell_n}\right)^+, \tag{1}$$

where $(x)^+ \triangleq \max(0, x)$. As explained in [9], the term $(\log n)/\ell_n$ can be understood as the fraction of bits in each length-ℓ_n fragment that must be used for a unique index, which allows the reordering of the shuffled fragments.

The second relevant result is the capacity of the TPC with no fragment deletions, in the special case where the fragment lengths are $N_i \sim \text{Geometric}(1/\ell_n)$ [1], which is given by

$$C_{\text{TPC}} = \exp\left(-\lim_{n\to\infty} \frac{\log n}{\ell_n}\right). \tag{2}$$

By letting $\alpha = \lim_{n\to\infty}(\log n)/\ell_n$, we see that $C_{\text{TPC}} = e^{-\alpha} > (1-\alpha)^+ = C_{\text{shuf}}$. This is surprising because, in the case of random fragment lengths, it is not possible to place a unique identifier at the beginning of each fragment, as the tearing locations are not known a priori. Moreover, the results in (1) and (2) feel qualitatively different, and it is not clear how to reconcile these two capacity expressions.

The main result in this paper generalizes the capacity of the TPC to (i) accommodate the case of lost fragments with a general deletion probability function $d(\cdot)$ and (ii) allow any distribution for the fragment length N_i, as long as some mild regularity conditions hold. In Section III we obtain closed form expressions for various choices of $d(\cdot)$ and N_i. Moreover, in doing so we provide a capacity expression that allows us to reconcile (1) and (2). More precisely, we prove that discarding fragments of length $\log n$ or shorter at the output does not affect the capacity, and that the capacity of the TPC with Lost

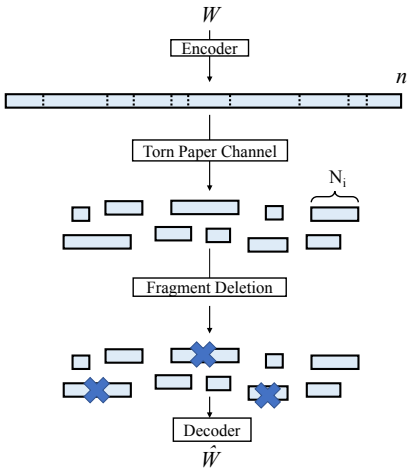

Fig. 1. The torn-paper channel with lost pieces.

Pieces (TPC-LP) is given by

$$C_{\text{TPC-LP}} = \text{coverage} - \text{reordering cost}, \tag{3}$$

where "coverage" represents the fraction of the original length-n string that is covered by the output fragments (after discarding those shorter than $\log n$), and "reordering cost" represents the fraction of the output fragments that would need to be dedicated for the placing of a unique index if we knew the tearing locations and were able to do so. Surprisingly, it turns out that the reordering cost does not change even when the tearing points are unknown. As a consequence we can state that $C_{\text{TPC-LP}} = 0$, whenever $N_i \leq \log n$ with probability 1.

Consider the capacity expression for the deterministic case with fragments of length $\ell_n > \log n$ and no lost fragments. Notice that the coverage for this case is 1, since all the fragments are retained. Now the reordering cost is the fraction of bits in each length-ℓ_n piece used for indexing purposes given by $\lim_{n\to\infty} \log n/\ell_n = \alpha$ as discussed before. Thus our result "coverage − reordering cost", yields (1) as a special case. Similarly, the result in [1] for a TPC with geometric piece lengths and no lost fragments is a special case of (3). More specifically, for the setting in [1] the coverage can be calculated as $(1+\alpha)e^{-\alpha}$ and the reordering cost can be shown to be $\alpha e^{-\alpha}$, which yields (2).

Related Work: Several recent papers have proposed new code constructions based on unique aspects of DNA data storage. Some of these works focus on DNA synthesis constraints such as sequence composition [6, 10, 11], the asymmetric nature of the DNA sequencing error channel [12], the need for codes that correct insertion errors [13], and the need for techniques to allow random access [11].

Also motivated by DNA-based storage, the problem of coding across unordered sets (of strings or points in a vector space) has recently received considerable attention [9, 14–20]. A channel that breaks the message down to its symbols and shuffles them was also studied in the context of the noisy permutation channel [21].

II. Problem Setting

We consider the TPC-LP as shown in Figure 1. The transmitter encodes a binary codeword $X^n \in \{0,1\}^n$, corresponding to the message $W \in [1 : 2^{nR}]$. The channel output is a set of binary strings $\mathcal{Y}$. The process by which $\mathcal{Y}$ is obtained from X^n is described as follows: The channel first breaks the input sequence into pieces of a random length. Specifically, define $N_1, N_2, \ldots$ to be i.i.d. random variables. We assume $E[N_i] = \ell_n \ \forall \ i$. Let K be the smallest index such that $\sum_{i=1}^{K} N_i \geq n$. Note that K is also a random variable. The channel tears the string X_n into $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_K$ where

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

We define the (unordered) multiset $\mathcal{Y}'$ as

$$\mathcal{Y}' = \{\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_K\}. \tag{4}$$

The multiset $\mathcal{Y}'_d$ is defined as follows. Each element $\vec{X}_i$ is removed from the set $\mathcal{Y}'$ with probability $d(N_i)$. The new set obtained is $\mathcal{Y}'_d$. The output of the channel is $\mathcal{Y} = \mathcal{Y}'_d$.

Note here that there are no bit-level errors (erasures or bit-flips). Moreover ℓ_n in general depends on the value of n.

For the purposes of this paper we assume (i) the limit $\alpha \triangleq \lim_{n\to\infty} \log n/\ell_n$ exists, $\alpha \in (0,\infty)$ and (ii) $E[N_1^2/(\log n)^2]$ is finite and bounded for all n. Our results hold irrespective of the first assumption (i.e. when $\alpha = 0, \infty$), but would require different steps to prove. Let $\mathcal{E}$ be the event that the decoder makes an error.

Definition 1. *The capacity C is the supremum of rates R such that there exists a sequence of encoder-decoder pairs with $P(\mathcal{E}) \to 0$ as $n \to \infty$.*

Notice that C is a function of the sequence $\{1/\ell_n\}_{n=1}^{\infty}$.

Notation: $\log(\cdot)$ represents the logarithm in base 2. For functions $a(n)$ and $b(n)$, we say $a(n) = o(b(n))$ if $a(n)/b(n) \to \infty$. For an event A, we let $\mathbf{1}_A$ be the binary indicator of A.

III. Main Results

Intuitively, the capacity of the TPC-LP should be affected by two distinct sources of uncertainty: (i) some of the pieces (which potentially carry information) are discarded by the channel and (ii) the remaining pieces are observed as an unordered set. As it turns out, (i) will be captured in the capacity expression by a quantity that represents the fraction of bits in X^n retained at the output, which can be written as

$$\frac{1}{n}\sum_{i=1}^{K} N_i \mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}\}}. \tag{5}$$

The limit as $n \to \infty$ of the expected value of (5) turns out to be fundamental in calculating the capacity of the TPC-LP.

Definition 2. *The coverage Φ_g is defined as*

$$\Phi_g \triangleq \lim_{n\to\infty} E\left[\frac{1}{n}\sum_{i=1}^{K} N_i \mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}'_g\}}\right]. \tag{6}$$

The capacity of the TPC-LP will also involve a quantity that captures (ii). To build intuition, let us imagine a channel where the tearing points and the set of pieces that are discarded are known a priori. For this channel, to preserve the ordering, a simple coding scheme is to include an index at the beginning of each fragment. There are $\approx n/\ell_n$ pieces. We therefore need

$$\log(n/\ell_n) = \log n - \log \ell_n \overset{(a)}{=} \log n - o(\log n)$$

bits per piece for indexing, where (a) holds since $\alpha \in (0,\infty)$ and $\ell_n = \alpha/\log n$ asymptotically. Therefore for a fragment $\vec{X}_i$ that exists in $\mathcal{Y}$, $\log n$ bits are needed for indexing. This can be succinctly written as $\mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}\}} \log n$. As it turns out this "reordering-cost" remains unchanged even when the tearing locations are not known. The empirical average of this quantity also plays a key role in the capacity of the TPC-LP.

Definition 3. *The reordering cost Ω_g is defined as*

$$\Omega_g \triangleq \lim_{n\to\infty} E\left[\frac{\log n}{n}\sum_{i=1}^{K}\mathbf{1}_{\{\vec{X}_i\in\mathcal{Y}'_g\}}\right]. \tag{7}$$

We now state our main result.

Theorem 1. *The capacity of the* TPC-LP *is*

$$C = \Phi_{\tilde{d}} - \Omega_{\tilde{d}}, \tag{8}$$

where $\tilde{d}(x) = \begin{cases} 1 & \text{for } x \le \log n \\ d(x) & \text{for } x > \log n \end{cases}$.

The capacity expression is "coverage − reordering cost," which intuitively is the fraction of bits that carry information about the message. Notice that the use of $\tilde{d} \in [0,1]$ instead of d in the coverage and reordering cost computations implies the surprising fact that pieces of size $\le \log n$ do not contribute to improving the capacity of this channel. The following corollary allows us to compute the capacity numerically.

Corollary 2. *The capacity of the* TPC-LP *is equivalently*

$$C = \alpha\int_1^\infty (\beta-1)\left(1-\hat{d}(\beta)\right)h(\beta)d\beta, \tag{9}$$

where we define $\hat{d}(\beta) \triangleq \lim_{n\to\infty} d(\beta\log n)$ *and* $h(\beta) \triangleq \lim_{n\to\infty}\Pr(N_1 = \beta\log n)\log n$*, provided the limits exist.*

We prove Corollary 2 in section VI. Note that the result in [1] can be obtained by taking $\hat{d}(\beta) = 0$, which implies that no pieces are lost. Table I shows the capacity expression evaluated for several choices of N_i and $\hat{d}(\cdot)$. In the next two sections, we prove Theorem 1.

TABLE I
CAPACITY EXPRESSIONS

$\hat{d}(\beta)$	N_i	$h(\beta)$	$C_{\text{TPC-LP}}$
0	Geometric($1/\ell_n$)	$\alpha e^{-\alpha\beta}$	$e^{-\alpha}$
ϵ	Geometric($1/\ell_n$)	$\alpha e^{-\alpha\beta}$	$(1-\epsilon)e^{-\alpha}$
$e^{-\gamma\beta}$	Geometric($1/\ell_n$)	$\alpha e^{-\alpha\beta}$	$e^{-\alpha}\left(1-\frac{\alpha^2 e^{-\gamma}}{(\alpha+\gamma)^2}\right)$
0	U[$0:\gamma\log n$], $\gamma\ge 1$	$1/\gamma$	$((\gamma-1)/\gamma)^2$
0	Fixed(ℓ_n), $\ell_n \ge \log n$	NA[1]	$1-\alpha$

[1] $h(\cdot)$ does not exist, hence we directly employ Theorem 1.

IV. CONVERSE

In order to prove the converse, we partition the set $\mathcal{Y}$ into sets that contain pieces of roughly the same length. This allows us to view the TPC-LP as a set of parallel channels that process pieces of roughly the same length. More precisely we define

$$\mathcal{Y}_k \triangleq \left\{\vec{X}_i \in \mathcal{Y} : \frac{k-1}{L}\log n \le N_i < \frac{k}{L}\log n\right\}, \tag{10}$$

where L is a fixed integer. We then split the set of "channels" into two sets, one with pieces of smaller sizes and the other with larger sizes. Specifically, we fix another integer $J > L$, and define $\mathcal{Y}_{\ge J} = \{\vec{X}_i : N_i \ge (J/L)\log n\}$. Then, by Fano's inequality, we have

$$R \le \lim_{n\to\infty}\frac{I(X^n;\mathcal{Y})}{n} \le \lim_{n\to\infty}\frac{H(\mathcal{Y})}{n}$$

$$\overset{(a)}{\le} \lim_{n\to\infty}\sum_{k=1}^{J}\frac{H(\mathcal{Y}_k)}{n} + \lim_{n\to\infty}\frac{H(\mathcal{Y}_{\ge J})}{n}, \tag{11}$$

where (a) holds from the independence bound on partition $\mathcal{Y} = (\cup_{k=1}^{J}\mathcal{Y}_k)\cup\mathcal{Y}_{\ge J}$.

The idea is that for fixed large values of J the second term in equation (11) is arbitrarily small. We will now use the fact that $|\mathcal{Y}_k|$ concentrates around its mean to tackle the first term in (11). To that end, we define the event $B = \{X_1 \in \mathcal{Y}'_d\}$ and

$$q_{k,n} = \Pr\left(\frac{k-1}{L}\log n \le N_1 < \frac{k}{L}\log n\right)$$
$$e_{k,n} = \Pr\left(B \,\middle|\, \frac{k-1}{L}\log n \le N_1 < \frac{k}{L}\log n\right). \tag{12}$$

Additionally, we define the event

$$\mathcal{E}_{k,n} = \{||\mathcal{Y}_k| - nq_{k,n}e_{k,n}/\ell_n| > \epsilon_n n/\ell_n\}.$$

We establish that $|\mathcal{Y}_k|$ concentrates in the following lemma.

Lemma 3. *For $\epsilon_n > 0$ and n large enough,*

$$\Pr(\mathcal{E}_{k,n}) \le 2e^{-n\epsilon_n^2/(2\ell_n)} + 2e^{-\frac{8\epsilon_n^2\ell_n}{(1+2\epsilon_n)}n}. \tag{13}$$

We provide the proof of this lemma in the longer version of this manuscript [22]. The lemma indicates that with high probability, $|\mathcal{Y}_k|$ is close to $nq_{k,n}e_{k,n}/\ell_n$. We set $\epsilon_n = 1/\log n$, ensuring that $\epsilon_n \to 0$ and $\Pr(\mathcal{E}_{k,n}) \to 0$ from Lemma 3. Then

$$H(\mathcal{Y}_k) \le H(\mathcal{Y}_k, \mathbf{1}_{\mathcal{E}_{k,n}}) \le 1 + H(\mathcal{Y}_k|\mathbf{1}_{\mathcal{E}_{k,n}})$$
$$\le 1 + 2n\Pr(\mathcal{E}_{k,n}) + H(\mathcal{Y}_k|\overline{\mathcal{E}}_{k,n}). \tag{14}$$

Here we loosely upper bound $H(\mathcal{Y}_k|\mathcal{E}_{k,n})$ with $2n$ since $\mathcal{Y}_k$ is fully described by X^n and $n-1$ binary variables that indicate whether there is a tear between the $(i-1)$th and ith bits. We now need to upper bound $H(\mathcal{Y}_k|\overline{\mathcal{E}}_{k,n})$. We first note that the total number of possible distinct sequences in $\mathcal{Y}_k$ are

$$\sum_{i=\frac{k-1}{L}\log n}^{\frac{k}{L}\log n} 2^i < 2.2^{\frac{k}{L}}\log n = 2n^{k/L}. \tag{15}$$

Now given $\overline{\mathcal{E}}_{k,n}$,

$$|\mathcal{Y}_k| \le M \triangleq (\epsilon_n + q_{k,n}e_{k,n})\, n/\ell_n. \tag{16}$$

Following the counting argument in [14], we note that the set $\mathcal{Y}_k$ can be viewed as a histogram over $2n^{k/L}$ sequences. Moreover, we can view the last element of the histogram as containing "excess counts" if $|\mathcal{Y}_k| < M$, so that the sum of the histogram entries is exactly M. This allows us to bound the term $H(\mathcal{Y}_k|\overline{\mathcal{E}}_{k,n})$ as

$$H(\mathcal{Y}_k|\overline{\mathcal{E}}_{k,n}) \le \log\binom{2n^{k/L}+M-1}{M}$$
$$\le M\log\left(\frac{e(2n^{k/L}+M-1)}{M}\right)$$
$$= M(\log(2n^{k/L}+M-1) + \log e - \log M)$$
$$\overset{(a)}{=} M\left[\max\left(\frac{k}{L}\log n, \log M\right) + \log e - \log M + P\right]$$

$$= M\left[\left(\frac{k}{L}\log n - \log M\right)^+ + \log e + P\right], \quad (17)$$

where $P \triangleq \min\left(\log\left(2+\frac{M-1}{n^{k/L}}\right), \log\left(1+\frac{2n^{k/L}-1}{M}\right)\right)$. In step (a) we employ the fact that if $a = b+c = d+e$, then $a = \max(b,d)+\min(c,e)$. From (14), this implies that

$$\begin{aligned}
\frac{H(\mathcal{Y}_k)}{n} &\le \frac{1+2n\Pr(\mathcal{E}_{k,n}) + H(\mathcal{Y}_k|\overline{\mathcal{E}}_{k,n})}{n} \\
&\le \frac{M}{n}\left(\frac{k}{L}\log n - \log M\right)^+ + A(k,n) \\
&\le \frac{M\log n}{n}\left(\frac{k}{L} - \frac{\log M}{\log n}\right)^+ + A(k,n) \\
&\overset{(a)}{\le} \frac{M\log n}{n}\left(\frac{k}{L}-1\right)^+ + A(k,n) \\
&\quad + \frac{M\log n \log(\ell_n \log n)}{n} \\
&\le \frac{\log n}{\ell_n}(\epsilon_n + q_{k,n}e_{k,n})\left(\frac{k}{L}-1\right)^+ + A(k,n) \\
&\quad + \frac{(\epsilon_n+1)\log(\ell_n\log n)}{\ell_n}\log n \quad (18)
\end{aligned}$$

where $A(k,n) \triangleq \frac{1}{n} + 2\Pr(\mathcal{E}_{k,n}) + \frac{M}{n}(\log e + P)$. (a) follows from the definition of M and $\epsilon_n = 1/\log n$. This allows us to bound $\sum_{k=1}^{J}\frac{H(\mathcal{Y}_k)}{n}$ as follows:

$$\begin{aligned}
&\sum_{k=1}^{J}\frac{H(\mathcal{Y}_k)}{n} \overset{(a)}{\le} \frac{\log n}{\ell_n}\sum_{k=1}^{J} q_{k,n}e_{k,n}\left(\frac{k}{L}-1\right)^+ + A(n) \\
&= \frac{\log n}{\ell_n}\sum_{k=L+1}^{J} q_{k,n}e_{k,n}\left(\frac{k}{L}-1\right) + A(n) \\
&\overset{(b)}{\le} \frac{\log n}{\ell_n}\sum_{k=L+1}^{\infty}\frac{k}{L}q_{k,n}e_{k,n} + A(n) - \frac{\log n}{\ell_n}E[\mathbf{1}_{\tilde{B}}], \quad (19)
\end{aligned}$$

where in (a) we define

$$\begin{aligned}
A(n) \triangleq &\sum_{k=1}^{J}\frac{(\epsilon_n+1)\log(\ell_n\log n)\log(n)}{\ell_n} \\
&+ \sum_{k=1}^{J}\epsilon_n\left(\frac{k}{L}-1\right)^+ + \sum_{i=1}^{J}A(k,n).
\end{aligned}$$

(b) holds if we define the event $\tilde{B} = \{X_1 \in \mathcal{Y}'_{\tilde{d}}\}$ and note that $\sum_{k=L+1}^{\infty} q_{k,n}e_{k,n} = E[\mathbf{1}_{\tilde{B}}]$. The first term in (19) is

$$\begin{aligned}
&\frac{\log n}{\ell_n}\sum_{k=L+1}^{\infty}\frac{k}{L}q_{k,n}e_{k,n} \\
&= \sum_{k=L+1}^{\infty}\frac{q_{k,n}}{\ell_n}E\left[\frac{k}{L}\log n\mathbf{1}_B\middle|\frac{N_1L}{\log n}\in[k-1,k)\right] \\
&\overset{(a)}{=} \sum_{k=L+1}^{\infty}\frac{q_{k,n}}{\ell_n}E\left[N_1\mathbf{1}_B\middle|\frac{N_1L}{\log n}\in[k-1,k)\right] \\
&+ \sum_{k=L+1}^{\infty}\frac{q_{k,n}}{\ell_n}E\left[\delta(N_1)\mathbf{1}_B\middle|\frac{N_1L}{\log n}\in[k-1,k)\right] \\
&= \frac{E[N_1\mathbf{1}_{\tilde{B}}]}{\ell_n} + \sum_{k=L+1}^{\infty}\frac{q_{k,n}}{\ell_n}E\left[\delta(N_1)\mathbf{1}_B\middle|\frac{N_1L}{\log n}\in[k-1,k)\right] \quad (20)
\end{aligned}$$

where, in (a), we define $\delta(N_1) \triangleq \frac{k}{L}\log n - N_1$.

Note that given $\frac{k-1}{L}\log n \le N_1 < \frac{k}{L}\log n$,

$$\delta(N_1) \le (\log n)/L. \quad (21)$$

The second summation in (20) can be upper bounded as

$$\begin{aligned}
&\frac{1}{\ell_n}\sum_{k=L+1}^{\infty} q_k E[\delta(N_1)\mathbf{1}_B|\tfrac{N_1L}{\log n}\in[k-1,k)] \\
&\overset{(a)}{\le} \frac{\log n}{\ell_n}\sum_{k=L+1}^{\infty}\frac{q_k}{L}E[\mathbf{1}_B|\tfrac{N_1L}{\log n}\in[k-1,k)] \\
&\le \frac{\log n}{\ell_n L}\sum_{k=L+1}^{\infty} q_k \overset{(b)}{\le} \frac{\log n}{\ell_n L}, \quad (22)
\end{aligned}$$

where (a) follows from (21) and (b) follows because q_k is a probability mass function over $k \in \{0,1,2,\dots\}$.

In summary equations (19)-(22) show that

$$\sum_{k=1}^{J}\frac{H(\mathcal{Y}_k)}{n} = \frac{E[N_1\mathbf{1}_{\tilde{B}}]}{\ell_n} - \frac{\log n}{n}E[\mathbf{1}_{\tilde{B}}] + \frac{\log n}{\ell_n L} + A(n). \quad (23)$$

Lemma 4 formalizes $H(\mathcal{Y}_{\ge J})/n$ being bounded as $n \to \infty$.

Lemma 4. *The entropy of the set $\mathcal{Y}_{\ge J}$ is upper bounded as*

$$\lim_{n\to\infty}\frac{H(\mathcal{Y}_{\ge J})}{n} \le 2\left(S\sqrt{L/J}+\delta\right),$$

for some finite S, and every J, L and $\delta > 0$.

Lemma 5 shows that $A(n)$ vanishes as $n \to \infty$.

Lemma 5. *The value of $A(n)$ as $n \to \infty$ is given by*

$$\lim_{n\to\infty} A(n) = 0.$$

We defer the proof of Lemmas 4 and 5 to a longer version of the paper [22]. From (11) and (23), we get

$$\begin{aligned}
R &\le \lim_{n\to\infty}\left(\frac{E[N_1\mathbf{1}_{\tilde{B}}]}{\ell_n} - \frac{\log n}{n}E[\mathbf{1}_{\tilde{B}}] + A(n) + \frac{H(\mathcal{Y}_{\ge J})}{n}\right) \\
&\quad + \lim_{n\to\infty}\frac{\log n}{\ell_n L} \\
&\overset{(a)}{\le} \Phi_{\tilde{d}} - \Omega_{\tilde{d}} + 2\left(S\sqrt{L/J}+\delta\right) + \alpha/L, \quad (24)
\end{aligned}$$

where (a) is due to Lemmas 4 and 5, followed by Definitions 2 and 3. Note here that the equivalence between the terms in (24) and Definitions 2 and 3 are implied by Lemmas 6 and 7 (which we state in the next section). Further, note that (24) holds for all integers $J > L$ and any $\delta > 0$. We can thus pick $L = \log J$ and let $\delta \to 0$ and $J \to \infty$. This proves the converse part of Theorem 1.

This also proves that pieces of size $\le \log n$ are futile in increasing the achievable rates. This fact is used in the next section while designing a decoder.

V. Achievability

We use a random coding argument to prove the achievability of Theorem 1. We generate a codebook $\mathcal{C}$ with 2^{nR} codewords, by independently picking each letter as Bern$(1/2)$. Let the resulting random codebook be $\mathcal{C} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{2^{nR}}\}$.

Assume that $W = 1$ is the message that is transmitted. The output $\mathcal{Y}$ is available at the decoder. We follow steps similar to [1], but with considerable generalisation. We choose a sub-optimal decoder, which throws out all substrings of size $\leq \log n$. The set obtained is precisely $\mathcal{Y}'_{\tilde{d}}$. If elements of $\mathcal{Y}'_{\tilde{d}}$ exist as non-overlapping substrings in some $\mathbf{x}_i$, then the decoder declares the index of that codeword as the message. We bound the probability of error averaged over all codebook choices as

$$\begin{aligned}\Pr(\mathcal{E}) &= \Pr(\mathcal{E}|W=1)\\ &= \Pr\left(\exists\ x_j : j \neq 1 \text{ contains all } \in \mathcal{Y}'_{\tilde{d}} \middle| W = 1\right) \qquad (25)\end{aligned}$$

We now state two lemmas (the proofs of which are available in [22]) that are crucial to prove the achievable part of Theorem 1. They provide us with a concentration on the coverage and reordering cost. This intuitively can be used to bound the probability of error by restricting the possibility of the value of coverage being too high or the reordering-cost too low.

Lemma 6. *For any $\epsilon > 0$,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{K} N_i \mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}'_{\tilde{d}}\}} - \frac{E[N_1 \mathbf{1}_{\tilde{B}}]}{\ell_n}\right| > \epsilon\right) \to 0, \qquad (26)$$

as $n \to \infty$.

Lemma 7. *For any $\epsilon > 0$,*

$$\Pr\left(\left|\sum_{i=1}^{K} \mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}'_{\tilde{d}}\}} - \frac{nE[\mathbf{1}_{\tilde{B}}]}{\ell_n}\right| \geq \frac{n}{\ell_n}\epsilon\right) \to 0, \qquad (27)$$

as $n \to \infty$.

Now we let $B_1 = (1+\epsilon)\frac{nE[\mathbf{1}_{B'}]}{\ell_n}$ and $B_2 = (1-\epsilon)\frac{E[N_1 \mathbf{1}_{B'}]}{\ell_n}$ and define the event

$$\mathcal{B} = \left\{\sum_{i=1}^{K} \mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}'_{\tilde{d}}\}} > B_1\right\} \cup \left\{\frac{1}{n}\sum_{i=1}^{K} N_i \mathbf{1}_{\{\vec{X}_i \in \mathcal{Y}'_{\tilde{d}}\}} < B_2\right\}. \qquad (28)$$

Lemmas 6 and 7 imply that $\Pr(\mathcal{B}) \to 0$ as $n \to \infty$. Therefore

$$\begin{aligned}\Pr(\mathcal{E}) &= \Pr(\exists\ x_j \text{ contains all elements in } \mathcal{Y}'|W=1)\\ &\leq \Pr(\exists\ x_j \text{ contains all elements in } \mathcal{Y}'|W=1,\overline{\mathcal{B}})\\ &\quad + \Pr(\mathcal{B})\\ &\overset{(a)}{\leq} |\mathcal{C}|\frac{n^{B_1}}{2^{nB_2}} + \Pr(\mathcal{B})\\ &\leq 2^{nR}2^{B_1 \log n}2^{-nB_2} + o(1)\end{aligned}$$

Inequality (a) follows from the Union Bound and the fact that given $\overline{\mathcal{B}}$, there are at most n^{B_1} ways to align $\mathcal{Y}'$ to a codeword x_j. To see this note that, given $|\mathcal{Y}'_{\tilde{d}}| < B_1$, there are at most n places the fragments can start from to align each piece and at most B_1 such pieces. Since a non-overlapping alignment of the strings in $\mathcal{Y}'_{\tilde{d}}$ to a codeword x_j covers at least nB_2 positions of x_j, the probability that it matches x_j on all covered positions is at most 2^{-nB_2}. Now $\Pr(\mathcal{E}) \to 0$ if

$$\begin{aligned}R &\leq \lim_{n\to\infty}\left((1-\epsilon)\frac{E\left[N_1 \mathbf{1}_{\tilde{B}}\right]}{\ell_n} - (1+\epsilon)\frac{E\left[\mathbf{1}_{\tilde{B}}\right]\log n}{\ell_n}\right)\\ &= (1-\epsilon)\Phi_{\tilde{d}} - (1+\epsilon)\Omega_{\tilde{d}}. \qquad (29)\end{aligned}$$

Letting $\epsilon \to 0$ we obtain $R \leq \Phi_{\tilde{d}} - \Omega_{\tilde{d}}$. This proves the achievable part of Theorem 1.

VI. Proof of Corollary 2

Corollary 2 provides us with an expression to numerically compute the TPC-LP capacity for certain classes of fragment length distributions (N_i). Intuitively it holds for distributions of N_i which have a continuous analog (for example the geometric or uniform distributions). We proceed as follows:

$$\begin{aligned}C = \Phi_{\tilde{d}} - \Omega_{\tilde{d}} &\overset{(a)}{=} \lim_{n\to\infty}\left(\frac{E[N_1 \mathbf{1}_{\tilde{B}}]}{\ell_n} - \frac{\log n E[\mathbf{1}_{\tilde{B}}]}{\ell_n}\right)\\ &= \lim_{n\to\infty}(\log n/\ell_n)\lim_{n\to\infty} E\left[\left(\frac{N_1}{\log n} - 1\right)\mathbf{1}_{\tilde{B}}\right]\\ &= \alpha \lim_{n\to\infty} E\left[\left(\frac{N_1}{\log n} - 1\right)E\left[\mathbf{1}_{\tilde{B}}|N_1\right]\right]\\ &= \alpha \lim_{n\to\infty}\sum_{x=\log n}^{\infty}\left(\frac{x}{\log n} - 1\right)(1-\tilde{d}(x))\Pr(N_1 = x).\end{aligned}$$

Note in (a) we use the equivalent definition of coverage that we employed in (24). Now we substitute $x = \beta \log n$ in the above equation, where $\beta \in \{1, 1+\frac{1}{\log n}, 1+\frac{2}{\log n}, \ldots\} \triangleq \mathcal{Q}$. Then we have

$$\begin{aligned}C &= \alpha \lim_{n\to\infty}\sum_{\beta\in\mathcal{Q}}(\beta-1)(1-\tilde{d}(\beta\log n))\Pr(N_1 = \beta\log n)\\ &= \alpha \lim_{n\to\infty}\sum_{\beta\in\mathcal{Q}}(\beta-1)(1-\tilde{d}(\beta\log n))\frac{\Pr(N_1=\beta\log n)}{\Delta_n}\Delta_n\\ &\overset{(a)}{=} \alpha\int_1^{\infty}(\beta-1)(1-\hat{d}(\beta))h(\beta)d\beta, \qquad (30)\end{aligned}$$

where $\lim_{n\to\infty}\Pr(N_1 = \beta\log n)\log n \triangleq h(\beta)$ and $\Delta_n = \beta + 1/\log n - \beta$. (a) follows from the definition of Riemann integration. This result holds when $h(\beta)$ exists and is finite.

VII. Conclusion

We have shown that the capacity of a TPC-LP can be expressed as "coverage $-$ reordering cost" (after throwing out fragments of size $\leq \log n$), where the channel deleted entire fragments randomly. It would be worthwhile to compare the effect of these fragment-level erasures to that of bit level erasures, where each bit is erased independently with a probability ϵ. This is essentially concatenating a binary erasure channel to the TPC-LP. Finding capacity expressions for this case is not straightforward and is the subject of future work.

Acknowledgements

The research of A. N. Ravi and 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] I. Shomorony and A. Vahid, "Communicating over the torn-paper channel," *GLOBECOM*, 2020.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] S. Shin, R. Heckel, and I. Shomorony, "Capacity of the erasure shuffling channel," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8841–8845, IEEE, 2020.

[16] 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.

[17] 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.

[18] 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.

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

[20] S. Nassirpour and A. Vahid, "Embedded codes for reassembling non-overlapping random DNA fragments," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2020.

[21] 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.

[22] A. Ravi, A. Vahid, and I. Shomorony, "Capacity of the torn paper channel with lost pieces," *https://adityanarayan.web.illinois.edu/Data/TPCLPlong.pdf*, 2021.