ACHIEVING THE CAPACITY OF THE DNA STORAGE CHANNEL

Andreas Lenz [*] *, Paul H. Siegel* [†]*, Antonia Wachter-Zeh* [*]*, and Eitan Yaakobi* [‡]

[*] Institute for Communications Engineering, Technical University of Munich, Germany
[†] Department of Electrical and Computer Engineering, University of California, San Diego, California
[‡] Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel

ABSTRACT

Significant advances in biochemical technologies, such as synthesizing and sequencing devices, have made DNA a competitive medium for archival data storage. In this paper we analyze storage systems based on these macromolecules from an information theoretic perspective. Using an appropriate channel model for the synthesis and sequencing steps, we study the maximum achievable information density per nucleotide for reliable and error resilient data storage. The channel model features the main attributes that characterize DNA-based data storage. That is, information is synthesized onto many short DNA strands, and each strand is copied many times. Due to the storage and sequencing methods, the receiver draws strands from these synthesized strands in an uncontrollable manner, where it is possible that strands are drawn multiple times and also that some strands are not drawn at all. Additionally, due to imperfections, the obtained strands can be perturbed by errors. Here we settle the question of how to achieve a recently published upper bound on the Shannon capacity of this channel by proposing and analyzing a decoder that clusters received strands according to their similarity and then efficiently estimates the original strands based on these clusters.

1. INTRODUCTION

Recently, DNA-based data storage has emerged to a promising technology for long-term archival data storage. Several experiments [1–6] have demonstrated the viability of digital information storage in these macromolecules and addressed different aspects such as random access [2, 5], portability [4] and scalability [5]. While within these experiments, it has been possible to successfully recover the stored data, the question of fundamental limits on the storage and reading rate remain unknown. Recently, several works have addressed information and coding theoretic aspects of DNA-based data storage. Among these, the capacity of the storage channel has been found for the case, when there are no errors in the strands [7] and for the case, when each DNA strand is read exactly once [8] under presence of substitution errors. The channel capacity when sequences are drawn randomly under the presence of substitution errors has been bounded from above in [9]. Error-correcting codes for systems, where data is stored in unordered sets as in DNA-based storage systems has been discussed in [10–15]. An important aspect for decoding archives stored in DNA is to cluster output sequence based on their mutual Hamming, respectively edit distance [5,16]. This technique allows to obtain information about the original input sequence and will be an important aspect of the decoder presented later. In this paper we show how to achieve the capacity upper bound from [9] using a random coding argument and a decoder that is specifically designed to the DNA storage channel. The paper is organized as follows. We first introduce the notation and describe the channel, then state our result about its capacity and finish by proving the achievability.

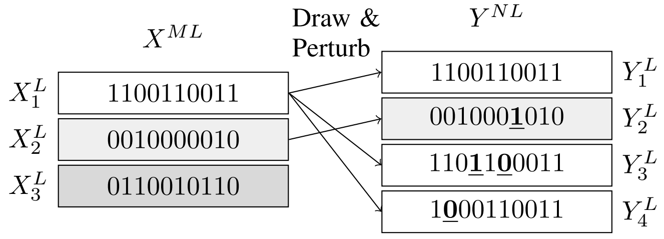

Fig. 1. Exemplary realization of the DNA storage channel

2. CHANNEL MODEL AND MAIN RESULT

Random variables are written in upper case letters, while their realizations are depicted in lower case. We denote by $\mathbb{P}\left(\bullet\right)$ the probability of an event and by $\mathbb{E}\left[\bullet\right]$ and $\mathbb{V}\left[\bullet\right]$ the expected value and variance of any random variable. Where it is clear from the context, we abbreviate the event $X = x$ by x. By $H(\bullet)$ we refer to the entropy of a random variable and by $H(p)$ for $0 \leq p \leq 1$ to the binary entropy function.

The input of the DNA storage channel are M sequences $X_1^L, \dots, X_M^L$ where each $X_i^L \in \Sigma^L$, $i \in [M]$ is a vector of length L over the alphabet Σ. From these input sequences, a total of N sequences are drawn with replacement, each uniformly at random, and received with errors. Denote by $I_j \in [M]$ i.i.d. uniform random draws with $\mathbb{P}\left(I_j = i\right) = \frac{1}{M}$ for all $j \in [N]$ and $i \in [M]$, where I_j depicts the input sequences which has been drawn in the j-th draw. The output sequences Y_j^L, $j \in [N]$ then satisfy $Y_j^L | X_{I_j}^L \sim \mathsf{BSC}(p)$, i.e., each received sequence Y_j^L is obtained by drawing a random input sequence $X_{I_j}^L$ and distorting it through a binary symmetric channel (BSC) with crossover probability p. For notational convenience, we comprise all input and output sequences to $X^{ML} = (X_1^L, \dots, X_M^L)$ and $Y^{NL} = (Y_1^L, \dots, Y_N^L)$.

Throughout the paper we will use the random variables $D_i = |\{j \in [N] : I_j = i\}|$, $i \in [M]$, which count the number of times the i-th input sequence has been drawn and $Q_d = |\{i \in [M] : D_i = d\}|$, $d = 0, \dots, N$, that denote the number of input sequences that have been drawn a total of d times.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 801434). This work was also supported by NSF Grant CCF-BSF-1619053 and by the United States-Israel BSF grant 2015816.

Note that insertion, deletion errors and non-binary alphabets are not discussed here. The latter extension for symmetric channels however is directly obtained using similar methods as in this paper and is omitted for brevity. An error-correcting code for this channel is a set $\mathcal{C} \subseteq \Sigma^{ML}$ and we define its rate to be $R = \frac{\log|\mathcal{C}|}{ML}$. With these definitions, the Shannon capacity can be defined as usual as the supremum over all achievable rates. Here we show that the capacity upper bound from [9] is tight by proving its achievability. In particular, our main result is stated in Theorem 1.

Theorem 1. *Let $N = cM$, and $M = 2^{\beta L}$ for some fixed constants $0 < c$, $0 \le p < \frac{1}{8}$ and $0 < \beta < \frac{1-H(4p)}{2}$. Then, the Shannon capacity is given by*

$$C = \sum_{d=0}^{\infty} p_c(d) C_d - \beta(1 - \mathrm{e}^{-c}), \tag{1}$$

where $p_c(d) = \frac{e^{-c}c^d}{d!}$ is the probability mass function of the Poisson distribution and C_d denotes the capacity of the binomial channel with d draws and error probability p

$$C_d = 1 + \sum_{k=0}^{d} \binom{d}{k} p^k (1-p)^{d-k} \log \frac{1}{1 + p^{d-2k}(1-p)^{2k-d}}.$$

3. PROOF OF ACHIEVABILITY

Before proving the achievability rigorously, we first explain the idea of the proof. First note that the achievability of the upper bound is non-trivial as the DNA-storage channel does not fall under classical discrete (memoryless) channels. We will analyze the error probability of a decoder using a random codebook [17] and a special decoder that is suitable for the DNA-storage channel. The decoder first combines the received sequences Y^{NL} to *clusters* $\widehat{Z}^{\widehat{M}} = (\widehat{\mathcal{Z}}_1, \dots, \widehat{\mathcal{Z}}_{\widehat{M}})$ of sequences with a maximum Hamming distance of roughly $2pL$, i.e., $d(Y_i^L, Y_j^L) \lesssim 2pL$ for all $Y_i^L, Y_j^L \in \widehat{\mathcal{Z}}_k$. We then assign a measure of typicality between a codeword $X^{ML} \in \Sigma^{ML}$ and clusters $\widehat{Z}^{\widehat{M}}$ by counting the number of sequences X_i^L that could potentially be the unique origin of a cluster $\widehat{\mathcal{Z}}_k$. If, for a codeword, the number of matches is larger than roughly $M(1 - \mathrm{e}^{-c})$, which corresponds of the number of expected non-empty clusters, the decoder will decide for this codeword. The analysis shows that the decoder decides for the correct codeword with high probability and for any other codeword with very small probability.

To start with, we summarize some standard bounds on the conditional probability, which will be used several times later.

Lemma 1. *For any events $\mathscr{A}, \mathscr{B}$, $\mathbb{P}(\mathscr{A}|\mathscr{B})$ satisfies*

$$\mathbb{P}(\mathscr{A}) - \mathbb{P}(\mathscr{B}^{\mathrm{c}}) \le \mathbb{P}(\mathscr{A}|\mathscr{B}) \le \frac{\mathbb{P}(\mathscr{A})}{\mathbb{P}(\mathscr{B})}.$$

Let in the following $\delta = p + \epsilon$, $\alpha = 2\delta + \epsilon$ and $\gamma = 2\alpha + \epsilon$ be constants for some $\epsilon > 0$. Consider Algorithm 1, which greedily picks the first output sequence and then adds other output sequences, such that their maximum distance is less than αL. These sequences are combined to a cluster $\widehat{z}_1$ and all elements in $\widehat{z}_1$ are removed as candidates for succeeding clusters. The procedure successively continues to form clusters $\widehat{z}_2, \dots, \widehat{z}_{\widehat{M}}$ on the remaining sequences with the same procedure until no more sequences are present.

Algorithm 1 Clustering algorithm

```
Input: $N$ received sequences $y^{NL}$; cluster radius $\alpha L$
Output: $\widehat{M}$ Clusters $\widehat{z}_1, \dots, \widehat{z}_{\widehat{M}}$
$\widehat{M} \leftarrow 0$
$\mathcal{S} \leftarrow [N]$
while $\mathcal{S} \neq \emptyset$ do
    $\widehat{M} \leftarrow \widehat{M} + 1$
    for $j \in \mathcal{S}$ do
        if $d(y_j^L, y_k^L) < \alpha L \; \forall\, k \in \widehat{z}_{\widehat{M}}$ then
            $\widehat{z}_{\widehat{M}} \leftarrow \widehat{z}_{\widehat{M}} \cup \{j\}$
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \widehat{z}_{\widehat{M}}$
```

Based on Algorithm 1, we now define the following decoder $\mathsf{dec}(\mathcal{C}, y^{NL})$ that decodes y^{NL} into a code $\mathcal{C} \subseteq \Sigma^{ML}$. We first introduce the notion of typicality between a cluster of d received sequences $z^d = \{y_1^L, \dots, y_d^L\}$ and an input sequence $x^L \in \Sigma^L$. Let $X^L = (X_1, \dots, X_L)$ be a uniform random vector with $\mathbb{P}\left(X^L = x^L\right) = 2^{-L}$ for all $x^L \in \Sigma^L$. Further, for $d \in \mathbb{N}_0$ let $\mathcal{Z}^d = \{Y_1^L, \dots, Y_d^L\}$ be the output of the binomial channel with d draws, where $Y_i^L | X^L \sim \mathsf{BSC}(p)$ are independent realizations of the BSC with the same input X^L. Abbreviate by $\mathbb{P}\left(x^L, z^d\right) = \mathbb{P}\left(X^L = x^L, \mathcal{Z}^d = z^d\right)$ the joint distribution of the binomial channel (cf. [18]) and by $\mathbb{P}\left(x^L\right)$ and $\mathbb{P}\left(z^d\right)$ the marginal distributions. Based on this distribution, we define jointly typical sequences (cf. [17]) by

$$A_\epsilon^d = \left\{ (x^L, z^d) : \begin{vmatrix} -\log \mathbb{P}\left(x^L\right) - H(X^L) \big| < \epsilon L, \\ -\log \mathbb{P}\left(z^d\right) - H(\mathcal{Z}^d) \big| < \epsilon L, \\ -\log \mathbb{P}\left(x^L, z^d\right) - H(X^L, \mathcal{Z}^d) \big| < \epsilon L \end{vmatrix} \right\}$$

for $d > 0$. Further we set $A_\epsilon^0 = \emptyset$ such that empty clusters cannot be typical with any input sequence. Note that the first condition is always fulfilled, as $-\log \mathbb{P}\left(x^L\right) = H(X^L) = L$. Let $\widehat{z}^{\widehat{M}} = (\widehat{z}_1, \dots, \widehat{z}_{\widehat{M}})$ be $\widehat{M} \in \mathbb{N}_0$ clusters, where each cluster $\widehat{z}_i \subseteq \Sigma^L$ is a multi-set of sequences of length L. Using the joint typicality between an input sequence and a cluster, we can define the bipartite graph $G_{\mathrm{typ}}(x^{ML}, \widehat{z}^{\widehat{M}})$ with M vertices on the left and $\widehat{M}$ vertices on the right, where two vertices i and k are connected if $(x_i^L, \widehat{z}_k) \in A_\epsilon^{|\widehat{z}_k|}$. The graph naturally induces a measure of typicality between x^{ML} and $\widehat{z}^{\widehat{M}}$ by the size of its largest matching $\nu(G_{\mathrm{typ}}(x^{ML}, \widehat{z}^{\widehat{M}}))$. Abbreviate the joint typicality by $\mathsf{T}((x^{ML}, \widehat{z}^{\widehat{M}})) \triangleq \nu(G_{\mathrm{typ}}(x^{ML}, \widehat{z}^{\widehat{M}}))$ in the following. $\mathsf{T}((x^{ML}, \widehat{z}^{\widehat{M}}))$ can be viewed as the number of clusters that can be matched to a unique input sequence that is typical with respect to the multi-draw channel. With this measure, we define a measure of typicality between M input sequences and N output sequences by

$$A_\epsilon = \{(x^{ML}, y^{NL}) : \mathsf{T}(x^{ML}, \widehat{z}^{\widehat{M}}) \ge M(1 - \mathrm{e}^{-c} - \epsilon)\},$$

where $\widehat{z}^{\widehat{M}}$ is the result of Algorithm 1 with input y^{NL}.

Define now a decoder that decodes received sequences y^{NL} to codewords, that are typical with the transmitted word $\mathsf{dec}(\mathcal{C}, y^{NL}) = \{x^{ML} \in \mathcal{C} : (x^{ML}, y^{NL}) \in A_\epsilon\}$. The decoder outputs the unique codeword $\mathsf{dec}(\mathcal{C}, y^{NL})$, if $|\mathsf{dec}(\mathcal{C}, y^{NL})| = 1$ and fails otherwise. We are now ready to compute the average error probability over all codebooks under this decoder. We evaluate the average error probability using a random coding argument. Let $\mathcal{C} = \{X^{ML}(1), \dots, X^{ML}(2^{MLR})\}$ be a random codebook, where each codeword $X^{ML}(w)$ is chosen uniformly and independently with probability distribution

$$\mathbb{P}\left(X^{ML}(w) = x^{ML}\right) = 2^{-ML},$$

for all $x^{ML} \in \Sigma^{ML}$ and $w \in [2^{MLR}]$. Let $W \in [2^{MLR}]$ be the message, chosen uniform at random. The average error probability is then given by

$$\mathbb{P}\left(\mathcal{E}\right) = \mathbb{P}\left(\mathcal{E} | W = 1\right).$$

due to the symmetric nature of the random codebook. Let now Y^{NL} be the result of transmitting $X^{ML}(1)$ over the channel and define the event $\mathscr{J}_w = \{(X^{ML}(w), Y^{NL}) \in A_\epsilon\}$. Additionally, we omit the index for the first codeword for simplicity and thus write, e.g., X^{ML} instead of $X^{ML}(1)$ in the following. The average error probability is then given by

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}|W=1\right) &= \mathbb{P}\left(\mathscr{J}_1^{\mathsf{c}} \cup \mathscr{J}_2 \cup \dots \cup \mathscr{J}_{2^{MLR}} | W = 1\right) \\ &\le 1 - \mathbb{P}\left(\mathscr{J}_1 | W = 1\right) + \mathbb{P}\left(\mathscr{J}_2 \cup \dots \cup \mathscr{J}_{2^{MLR}} | W = 1\right).\end{aligned}$$

We start by bounding $\mathbb{P}\left(\mathscr{J}_1 | W = 1\right)$ from below. Let in the following $Z^M = (\mathcal{Z}_1, \dots, \mathcal{Z}_M)$ with $\mathcal{Z}_i = \{Y_j^L : I_j = i\}$ be the clusters of sequences with the same origin and $\widehat{Z}^{\widehat{M}} = (\widehat{\mathcal{Z}}_1, \dots, \widehat{\mathcal{Z}}_{\widehat{M}})$ with $\widehat{\mathcal{Z}}_1, \dots, \widehat{\mathcal{Z}}_{\widehat{M}} \subseteq \Sigma^L$ be $\widehat{M}$ disjoint clusters, i.e., $\widehat{\mathcal{Z}}_{k_1} \cap \widehat{\mathcal{Z}}_{k_2} = \emptyset$ for all $k_1 \neq k_2$. For now, think of these as arbitrary clusters, however later these will be the output of Algorithm 1. Consider now the tripartite layered graph $G_{\text{lay}}(X^{ML}, Z^M, \widehat{Z}^{\widehat{M}})$ with M vertices on the left, $M - Q_0$ vertices in the middle and $\widehat{M}$ vertices on the right. We connect a vertex i on the left with a vertex j in the middle, if $(X_i^L, \mathcal{Z}_j) \in A_\epsilon^{|\mathcal{Z}_j|}$. We further connect a vertex j in the middle with a vertex k on the right, if $\mathcal{Z}_j = \widehat{\mathcal{Z}}_k$. In the following, we call a cluster $\widehat{\mathcal{Z}}_k$ correct, if the degree of k in the above graph is at least 1, i.e., it contains exactly all sequences that originate from one input sequence. Let H be the size of the largest matching between vertices on the left and the middle and G be the number of correct clusters, i.e., the number of all vertices on the right that have degree at least 1. The following lemma establishes a connection between G, H and the joint typicality of input sequences and clusters $\mathsf{T}(X^{ML}, \widehat{Z}^{\widehat{M}})$.

Lemma 2. *The joint typicality of X^{ML} and $\widehat{Z}^{\widehat{M}}$ is at least*

$$\mathsf{T}(X^{ML}, \widehat{Z}^{\widehat{M}}) \ge H + G - (M - Q_0).$$

Conversely, the joint typicality is bounded from above by

$$\mathsf{T}(X^{ML}, \widehat{Z}^{\widehat{M}}) \le \widehat{M} - G + H.$$

Proof. We start with the observation that each correct cluster $\widehat{\mathcal{Z}}_k$, whose corresponding original cluster $\mathcal{Z}_j$ is jointly typical with an input sequence X_j^L, accounts for one element in $\mathsf{T}(X^{ML}, \widehat{Z}^{\widehat{M}})$. Let $\mathcal{G}$ be the vertices in the middle which belong to the largest matching between the middle and right vertices and let $\mathcal{H}$ be the vertices in the middle which belong to the largest matching between the middle and left vertices. With this definition,

$$\begin{aligned}\nu(G_{\text{typ}}(X^{ML}, \widehat{Z}^{\widehat{M}})) &\ge |\mathcal{H} \cap \mathcal{G}| = |\mathcal{H}| + G - |\mathcal{H} \cup \mathcal{G}| \\ &\overset{(a)}{\ge} |\mathcal{H}| + G - (M - Q_0) \overset{(b)}{=} H + G - (M - Q_0),\end{aligned}$$

where in inequality (a) we used that $\mathcal{H}$ and $\mathcal{G}$ contain only clusters, which have at least 1 sequence and thus $|\mathcal{H} \cup \mathcal{G}| \le M - Q_0$. Equality (b) follows from the fact that $\mathcal{H} = H$, since all vertices in the middle have right-degree at most 1, since $\widehat{\mathcal{Z}}_{k_1} \neq \widehat{\mathcal{Z}}_{k_2}$ for any $k_1 \neq k_2 \in [\widehat{M}]$ by the definition of the clusters $\widehat{Z}^{\widehat{M}}$. On the other hand

$$\mathsf{T}(X^{ML}, \widehat{Z}^{\widehat{M}}) \le H + \widehat{M} - G,$$

since the number of correct clusters, which can be matched to an input sequence is at most the size of the largest matching on the left H. Finally, there are $\widehat{M} - G$ incorrect clusters, which potentially could also add to the joint typicality. □

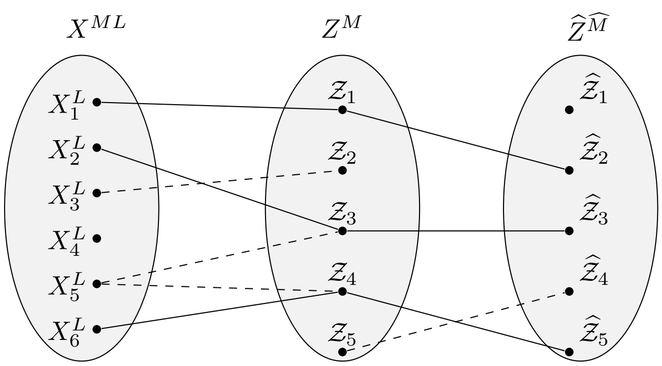
X^{ML}
Z^M
$\widehat{Z}^{\widehat{M}}$
X_1^L
X_2^L
X_3^L
X_4^L
X_5^L
X_6^L
$\mathcal{Z}_1$
$\mathcal{Z}_2$
$\mathcal{Z}_3$
$\mathcal{Z}_4$
$\mathcal{Z}_5$
$\widehat{\mathcal{Z}}_1$
$\widehat{\mathcal{Z}}_2$
$\widehat{\mathcal{Z}}_3$
$\widehat{\mathcal{Z}}_4$
$\widehat{\mathcal{Z}}_5$

Fig. 2. Illustration of the graph $G_{\text{lay}}(X^{ML}, Z^M, \widehat{Z}^{\widehat{M}})$. The solid lines highlight edges, which contribute to the joint typicality $\mathsf{T}(X^{ML}, \widehat{Z}^{\widehat{M}})$.

We continue by analyzing Algorithm 1 and bound the number of correct clusters it produces from below. To do so, we require the following entities. Denote by $\mathcal{U} \subseteq \Sigma^L$ the random variable which holds all output sequences, which were received with an untypical amount of errors, i.e., $\mathcal{U} = \{Y_j^L : j \in [N], |d(X_{I_j}^L, Y_j^L) - pL| > \epsilon L\}$. Further denote by $\mathscr{S}$ the event that $d(X_i^L, X_j^L) \ge \gamma L$ for all $i, j \in [M]$ with $i \neq j$.

Lemma 3. *Let Y^{NL} be the result of X^{ML}. Then, given the event $\mathscr{S}$, Algorithm 1 produces at least $G \ge M - Q_0 - 2|\mathcal{U}|$ correct clusters. Further, the total number of clusters is at most $\widehat{M} \le M - Q_0 + |\mathcal{U}|$.*

The proof follows from a careful analysis of the clustering algorithm and is omitted for brevity. Hence, by Lemma 2 and 3, given $\mathscr{S}$, Algorithm 1 produces at least $H - 2|\mathcal{U}|$ typical clusters. We obtain

$$\begin{aligned}\mathbb{P}\left(\mathscr{J}_1|W=1\right) &= \mathbb{P}\left((X^{ML},Y^{NL})\in A_\epsilon\right)\\ &\geq \mathbb{P}\left((X^{ML},Y^{NL})\in A_\epsilon|\mathscr{S}\right)\mathbb{P}\left(\mathscr{S}\right)\\ &\geq \mathbb{P}\left(H-2|\mathcal{U}|\geq M(1-\mathrm{e}^{-c}-\epsilon)|\mathscr{S}\right)\mathbb{P}\left(\mathscr{S}\right)\end{aligned}$$

Note that the random variables H and $|\mathcal{U}|$, are in general statistically dependent as sequences with an untypical amount of errors relate to untypical clusters. However, we circumvent this obstacle as follows.

$$\begin{aligned}&\mathbb{P}\left(H-2|\mathcal{U}|\geq M(1-\mathrm{e}^{-c}-\epsilon)|\mathscr{S}\right)\\ &\geq \mathbb{P}\left(H\geq M(1-\mathrm{e}^{-c}-\epsilon/2)\wedge|\mathcal{U}|\leq M\epsilon/4|\mathscr{S}\right)\\ &= 1-\mathbb{P}\left(H< M(1-\mathrm{e}^{-c}-\epsilon/2)\vee|\mathcal{U}|> M\epsilon/4|\mathscr{S}\right)\\ &\overset{(a)}{\geq} 1-\mathbb{P}\left(H< M(1-\mathrm{e}^{-c}-\epsilon/2)|\mathscr{S}\right)-\mathbb{P}\left(|\mathcal{U}|> M\epsilon/4\right)\\ &\overset{(b)}{\geq} \mathbb{P}\left(H\geq M(1-\mathrm{e}^{-c}-\epsilon/2)\right)-\mathbb{P}\left(\mathscr{S}^{\mathsf{c}}\right)-\mathbb{P}\left(|\mathcal{U}|>M\epsilon/4\right),\end{aligned}$$

where we used in inequality (a) that $|\mathcal{U}|$ only depends on the realizations of the errors from the BSC and thus is independent of $\mathscr{S}$ and in (b) we used Lemma 1. We now compute $\mathbb{P}\left(\mathscr{S}\right)$.

Lemma 4. *Let $2\beta < 1-H(\gamma)$. For any $\epsilon>0$ and large L,*

$$\mathbb{P}\left(\mathscr{S}\right)\geq 1-\epsilon.$$

The proof uses standard counting arguments and is omitted for brevity. Next, we bound the probability of having many pairs of jointly typical input sequences and output clusters.

Lemma 5. *For any $\epsilon>0$ and for sufficiently large M*

$$\mathbb{P}\left(H\geq M(1-\mathrm{e}^{-c}-\epsilon/2)\right)\geq 1-\epsilon.$$

Proof. Let $H' = |\{i\in[M]:(X_i^L,\mathcal{Z}_i\in A_\epsilon^{D_i})\}|$ be the number of clusters that are jointly typical with their original input sequence. Using $H'\leq H$ and $\mathbb{P}\left((X_i^L,\mathcal{Z}_i)\in A_\epsilon^{d_i}\right)\to 1$ due to the joint asymptotic equipartition property [17, Thm. 7.6.1] of the binomial channel yields the lemma. □

Finally $|\mathcal{U}|$ is a binomial distribution with N trials and success probability $\mathbb{P}\left(|d(X_{I_j}^L,Y_j^L)-pL|\geq\epsilon L\right)\leq \mathrm{e}^{-L\epsilon^2/2}\to 0$ for large L and thus $\mathbb{P}\left(|\mathcal{U}|>M\epsilon/4\right)\leq\epsilon$ for large enough M. Putting everything together, for large enough M,

$$\mathbb{P}\left(\mathscr{J}_1|W=1\right)\geq(1-3\epsilon)(1-\epsilon)$$

and thus $\mathbb{P}\left(\mathscr{J}_1|W=1\right)\to 1$ for $M\to\infty$.

Now we turn to bound $\mathbb{P}\left(\mathscr{J}_2\cup\dots\cup\mathscr{J}_{2^{MLR}}|W=1\right)$ from above. Denote by $\mathscr{U}=\{\mathcal{U}:|\mathcal{U}|\leq\epsilon M\}$ the event that the number of sequences in $\mathcal{U}$ is less than ϵM. Then,

$$\begin{aligned}&\mathbb{P}\left(\mathscr{J}_2\cup\dots\cup\mathscr{J}_{2^{MLR}}|W=1\right)\\ &\overset{(a)}{\leq}\mathbb{P}\left(\mathscr{J}_2\cup\dots\cup\mathscr{J}_{2^{MLR}}|W=1,\mathscr{S},\mathscr{U}\right)+\mathbb{P}\left(\mathscr{S}^{\mathsf{c}}\cup\mathscr{U}^{\mathsf{c}}\right).\end{aligned}$$

where (a) follows from Lemma 1. Denoting by $\mathscr{Q}$ the event from [9, Lemma 2], we obtain

$$\begin{aligned}&\mathbb{P}\left(\mathscr{J}_2\cup\dots\cup\mathscr{J}_{2^{MLR}}|W=1,\mathscr{S},\mathscr{U}\right)\\ &\leq\sum_{w=2}^{2^{MLR}}\sum_{d^M\in\mathscr{Q}}\mathbb{P}\left(\mathscr{J}_w|W{=}1,\mathscr{S},D^M{=}d^M,\mathscr{U}\right)\mathbb{P}\left(d^M\right)+\mathbb{P}\left(\mathscr{Q}^{\mathsf{c}}\right).\end{aligned}$$

Combining Lemma 2 and Lemma 3 we obtain for any $w\geq 2$

$$\begin{aligned}&\mathbb{P}\left(\mathscr{J}_w|W=1,\mathscr{S},D^M=d^M,\mathscr{U}\right)\\ &=\mathbb{P}\left(\mathsf{T}(X^{ML}(w),\widehat{Z}^{\widehat{M}})\geq M(1-\mathrm{e}^{-c}-\epsilon)|\mathscr{S},D^M{=}d^M,\mathscr{U}\right)\\ &\leq\mathbb{P}\left(H+3|\mathcal{U}|\geq M(1-\mathrm{e}^{-c}-\epsilon)|\mathscr{S},D^M=d^M,\mathscr{U}\right)\\ &\leq\mathbb{P}\left(H\geq M(1-\mathrm{e}^{-c}-4\epsilon)|\mathscr{S},D^M=d^M,\mathscr{U}\right)\\ &\overset{(a)}{\leq}\frac{\mathbb{P}\left(H\geq M(1-\mathrm{e}^{-c}-4\epsilon)|D^M=d^M\right)}{1-\mathbb{P}\left(\mathscr{S}^{\mathsf{c}}\right)-\mathbb{P}\left(\mathscr{U}^{\mathsf{c}}\right)},\end{aligned}$$

where we used Lemma 1 in inequality (a) to resolve the dependency on the events $\mathscr{S}$ and $\mathscr{U}$. We continue by discussing the distribution of H with the help of the following experiment. Let the clusters Z^M be given and choose a random codeword $X^{ML}(w)$. Each $X_i^L(w)$ of $X^{ML}(w)$ is an independent random sequence, that can be jointly typical with one of the clusters. Abbreviate $M'\triangleq M(1-\mathrm{e}^{-c}-4\epsilon)$ and let $\mathcal{M}\subseteq[M]$ be the set of indices such that $|\mathcal{M}|=M'$ and $0<d_i\leq d_j$ for all $i\in\mathcal{M}$ and all $j\in[M]\setminus\mathcal{M}$. Further let $\mathscr{H}$ be the event that there are at least M' positions $i\in\mathcal{M}$, where $X_i^L(w)$ is jointly typical with $\mathcal{Z}_i$. The probability of $\mathscr{H}$ is at most

$$\mathbb{P}\left(\mathscr{H}|D^M{=}d^M\right)\leq\sum_{j=M'}^{M-q_0}\binom{M-q_0}{j}\prod_{i\in\mathcal{M}}\mathbb{P}\left((X_i^L,\mathcal{Z}_i)\in A_\epsilon^{d_i}\right).$$

Since both X_i^L and $\mathcal{Z}_i$ are independent and have the marginal distributions of the input, respectively output of the binomial channel with d_i draws, the probability of joint typicality is given by $\mathbb{P}\left((X_i^L,\mathcal{Z}_i)\in A_\epsilon^{d_i}\right)\leq 2^{-L(C_{d_i}-3\epsilon)}$ [17, Thm. 7.6.1]. Consequently, for large enough M

$$\mathbb{P}\left(\mathscr{H}|D^M{=}d^M\right)\leq 2^M 2^{-L\sum_{i\in\mathcal{M}}(C_{d_i}-3\epsilon)},$$

where we used that the binomial sum is trivially at most 2^M. For the exponent, for $d^M\in\mathscr{Q}$ we obtain

$$\begin{aligned}\sum_{i\in\mathcal{M}}(C_{d_i}-3\epsilon)&\overset{(a)}{\geq}\sum_{i\in\mathcal{M}}(C_{d_i}-3\epsilon)+\sum_{i\notin\mathcal{M},D_i>0}(C_{d_i}-1)\\ &\geq\sum_{i:D_i>0}C_{d_i}-8M\epsilon\overset{(b)}{\geq}M\sum_{d=1}^{N}p_c(d)C_d-9M\epsilon\end{aligned}$$

where (a) follows from $C_{d_i}\leq 1$ for any $d_i\geq 0$ and (b) follows from the fact that $q_d/M\to p_c(d)$ (and in particular $q_0/M\to\mathrm{e}^{-c}$) for all $d^M\in\mathscr{Q}$ as proven in [9, Lemma 2]. By the definition of $\mathscr{H}$ only the joint typicality of an input sequence X_i^L with its corresponding cluster $\mathcal{Z}_i$ has been considered. Since any permutation of the input sequences is also possible, we obtain for large enough M

$$\begin{aligned}&\mathbb{P}\left(H\geq M'|D^M=d^M\right)\leq M^{M-q_0}\mathbb{P}\left(\mathscr{H}|D^M=d^M\right)\\ &\overset{(a)}{\leq}2^{-LM(\sum_{d=1}^{N}p_c(d)C_d+10\epsilon)+M(1-\mathrm{e}^{-c})\log M}=2^{-LM(C+10\epsilon)},\end{aligned}$$

where (a) follows again from the Poissonization of q_0. Putting everything together, we obtain

$$\mathbb{P}\left(\mathscr{J}_2\cup\dots\cup\mathscr{J}_{2^{MLR}}|W=1\right)\leq 2^{-LM(C-R+10\epsilon)}+3\epsilon$$

for large enough M. Since we can choose ϵ as small as we want, there exists a code family with $R<C$, such that the error probability tends to 0, which proves the achievability.

4. REFERENCES

[1] George M. Church, Yuan Gao, and Sriram Kosuri, “Next-generation digital information storage in DNA,” *Science*, vol. 337, no. 6102, pp. 1628, 2012.

[2] S. M. Hossein Tabatabaei Yazdi, Yongbo Yuan, Huimin Zhao, and Olgica Milenkovic, “A rewritable, random-access DNA-based storage system,” *Scientific Reports*, vol. 5, 2015.

[3] Meinolf Blawat, Klaus Gaedke, Ingo Huetter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin Pruitt, and George Church, “Forward error correction for DNA data storage,” *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016.

[4] S. M. Hossein Tabatabaei Yazdi, Ryan Gabrys, and Olgica Milenkovic, “Portable and error-free DNA-based data storage,” *Scientific Reports*, vol. 7, 2017.

[5] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss, “Random access in large-scale DNA data storage,” *Nature Biotechnology*, vol. 36, pp. 242–248, 2018.

[6] Shubham Chandak, Kedar Tatwawadi, Billy Lau, Jay Mardia, Matthew Kubit, Joachim Neu, Peter Griffin, Mary Wootters, Tsachy Weissman, and Hanlee Ji, “Improved read/write cost tradeoff in DNA-based data storage using LDPC codes,” *bioRxiv*, 2019.

[7] Reinhard Heckel, Ilan Shomorony, Kannan Ramchandran, and David N. C. Tse, “Fundamental limits of DNA storage systems,” in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2017, pp. 3130–3134.

[8] Ilan Shomorony and Reinhard Heckel, “Capacity results for the noisy shuffling channel,” in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2019.

[9] Andreas Lenz, Paul H. Siegel, Antonia Wachter-Zeh, and Eitan Yaakobi, “An upper bound on the capacity of the DNA storage channel,” in *IEEE Information Theory Workshop*, 2019.

[10] Han M. Kiah, Gregory J Puleo, and Olgica Milenkovic, “Codes for DNA sequence profiles,” vol. 62, no. 6, pp. 3125–3146, 2016.

[11] Mladen Kovačević and Vincent Y. F. Tan, “Codes in the space of multisets – coding for permutation channels with impairments,” *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5156–5169, 2016.

[12] Andreas Lenz, Paul H. Siegel, Antonia Wachter-Zeh, and Eitan Yaakobi, “Coding over sets for DNA storage,” in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2018, pp. 2411–2415.

[13] Wentu Song and Kui Cai, “Sequence-subset distance and coding for error control in DNA-based data storage,” 2018.

[14] Jin Sima, Netanel Raviv, and Jehoshua Bruck, “On coding over sliced information,” 2018.

[15] Andreas Lenz, Paul H. Siegel, Antonia Wachter-Zeh, and Eitan Yaakobi, “Anchor-based correction of substitutions in indexed sets,” in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2019.

[16] Cyrus Rashtchian, Konstantin Makarychev, Miklos Z. Racz, Djordje Jevdjic, Sergey Yekhanin, Siena Dumas Ang, Karin Strauss, and Luis Ceze, “Clustering billions of reads for DNA data storage,” in *Conf. Neural Information Processing Systems*, 2017, pp. 3360–3371.

[17] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, New York, NY, USA, 2006.

[18] Michael Mitzenmacher, “On the theory and practice of data recovery with multiple versions,” *Proc. of the IEEE Int. Symp. on Inf. Theory*, pp. 982–986, 2006.