An Upper Bound on 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, CMRR, University of California, San Diego, California
[‡]Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel
Emails: andreas.lenz@mytum.de, psiegel@ucsd.edu, antonia.wachter-zeh@tum.de, yaakobi@cs.technion.ac.il

***Abstract*—Paved by recent advances in sequencing and synthesis technologies, DNA has evolved to a competitive medium for long-term data storage. In this paper we conduct an information theoretic study of the storage channel - the entity that formulates the relation between stored and sequenced strands. In particular, we derive an upper bound on the Shannon capacity of the channel. In our channel model, we incorporate the main attributes that characterize DNA-based data storage. That is, information is synthesized on many short DNA strands, and each strand is copied many times. Due to the storage and sequencing methods, the receiver draws strands from the original sequences in an uncontrollable manner, where it is possible that copies of the same sequence are drawn multiple times. Additionally, due to imperfections, the obtained strands can be perturbed by errors. We show that for a large range of parameters, the channel decomposes into sub-channels from each input sequence to multiple output sequences, so-called *clusters*. The cluster sizes hereby follow a Poisson distribution. Furthermore, the ordering of sub-channels is unknown to the receiver. Our results can be used to guide future experiments for DNA-based data storage by giving an upper bound on the achievable rate of any error-correcting code. We further give a detailed discussion and intuitive interpretation of the channel that provide insights about the nature of the channel and can inspire new ideas for error-correcting codes and decoding methods.**

I. Introduction

The design of error-correcting codes for common channels in communications, such as the additive white Gaussian noise (AWGN) channel or binary symmetric channel (BSC) has been guided by the channel *capacity*, that has been found by Shannon [1] in 1948. It allows researchers to choose the information rate of error-correcting codes according to this fundamental limit. While the capacity for the above mentioned channels has been known for a long time, this is different for recently relevant channels, such as the DNA storage channel.

Recently, DNA-based data storage has emerged as promising technology for long-term archival data storage. Several experiments [2]–[6] have demonstrated the viability of digital information storage in these macromolecules and addressed different aspects such as random access [3], [6], portability [5] and scalability [6]. 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 remains open. More 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. Error-correcting codes for systems where data is stored in unordered sets, as in DNA-based storage systems, has been discussed in [9]–[14]. An important aspect for decoding archives stored in DNA is to cluster output sequence based on their mutual Hamming, respectively edit distance [6], [15]. This technique allows to perform an accurate estimation of the original input sequence and is an important aspect for decoding DNA-based archives. Here we extend the work from [7] and study the channel capacity when sequences are drawn randomly under the presence of substitution errors. The paper is organized as follows. We first define the channel, then state and interpret our result about its capacity and finish by proving the statements.

This work was supported by the German-American Fulbright Commission which funded the visit of A. Lenz to the Center for Memory and Recording Research, University of California San Diego. 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.

II. Channel Model and Main Result

Random variables are written in upper case letters, while their realizations are depicted in lower case. We denote by $\mathbb{P}(\bullet)$ the probability of an event and by $\mathbb{E}[\bullet]$ and $\mathbb{V}[\bullet]$ 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. For a permutation $\pi : [n] \mapsto [n]$ and a vector $x^n = (x_1, \ldots, x_n)$, we write $\pi x^n = (x_{\pi(1)}, \ldots, x_{\pi(n)})$ as the permutation of x^n.

The DNA storage channel, which is depicted in Fig. 1, has M input sequences $X_1^L, \ldots, X_M^L$ where each input sequence $X_i^L \in \Sigma^L$, $i \in [M]$ 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, resulting in the output sequences Y_j^L, $j \in [N]$ with

$$Y_j^L = X_{I_j}^L \oplus E_j^L$$

where $I_j \in [M]$ are i.i.d. uniform random draws with $\mathbb{P}(I_j = i) = \frac{1}{M}$ for all $j \in [N]$ and $i \in [M]$ and $E_j^L \in \Sigma^L$ denote independent error vectors with i.i.d. Bernoulli entries $E_{j,k} \sim \mathsf{Ber}(p)$ with error probability p for all $j \in [N]$ and $k \in [L]$. Further $\oplus$ denotes the binary XOR operation. In other words, 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. The input and output of the channel is hence

$$X^{ML} = \{X_1^L, \ldots, X_M^L\},$$
$$Y^{NL} = \{Y_1^L, \ldots, Y_N^L\}.$$

978-1-5386-6900-6/19/$31.00 ©2019 IEEE

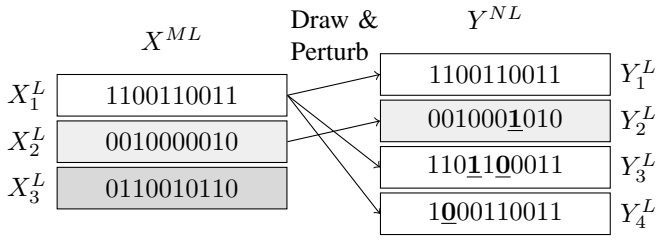

Fig. 1. Realization of the DNA storage channel with $M = 3$, $N = 4$ and $L = 10$. The shades and arrows indicate the origins. Errors are in bold.

Here we choose to define the input and output sequences as multi-sets for notional convenience. However, it can directly be verified that by defining the input and output as matrices of sizes $M \times L$, respectively $N \times L$, one obtains an equivalent channel. 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 $Q_d = |\{i \in [M] : D_i = d\}|$, $d = 0, \ldots, N$, that denote the number of input sequences that have been drawn a total of d times. 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. In [7] the capacity is found for $p = 0$, and [8] considers the case where each sequence is drawn exactly once. Our work is therefore a non-trivial extension of [7]. For our derivations we use, among others, methods from [7], [8] and [16]. The 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 bounded from above by*

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

where $p_c(d) = \frac{\mathrm{e}^{-c}c^d}{d!}$ is the Poisson distribution and C_d denotes the capacity of the binomial channel with d draws and error probability p (see Lemma 3).

For $p = 0$ this implies the bound $(1-\beta)(1-\mathrm{e}^{-c})$ from [7]. To simplify the derivation and discussion we consider a genie-aided receiver that receives along with each output sequence Y_j^L a label $\Pi(I_j)$, where $\Pi : [M] \mapsto [M]$ is a uniform random and unknown permutation. These labels allow to group the output sequences into *clusters* $Z_i = \{Y_j^L : \Pi(I_j) = i\}$ of sequences that originate from the same input sequence, but do not give any information about the original sequence[1]. Notably, the additional information from the genie is considerably small, as especially for well-separated input sequences, a receiver without knowledge of the labels can cluster the output sequences [15]. The expressions in Theorem 1 allow for a vivid interpretation, which we present in the following.

Poissonization and the binomial channel: We start with discussing the first summand of (1) on an intuitive level. It is known [17] that when $c = \frac{N}{M}$ is fixed, the marginals $D_i \to \mathsf{Poi}(c)$ approach Poisson random variables as $M \to \infty$. This effect is known as *Poissonization*. In Lemma 2, we will show that, although the D_i are statistically dependent, the quantities $\frac{Q_d}{M}$ jointly converge to $\frac{Q_d}{M} \to p_c(d)$ as $M \to \infty$ and thus can be viewed as asymptotically deterministic. Consider now a receiver that, additionally to $\Pi(I_j)$, knows Π and thus the origin $\Pi^{-1}(i)$ of each cluster Z_i. This allows to view the overall channel as M parallel channels. Each such subchannel has one input sequence X_i^L and D_i output sequences, which result from transmission of X_i^L over independent BSCs with error probability p. For a fixed number of draws d, this channel is known as the *binomial channel* [16] and has capacity C_d, (see Lemma 3). Together with the fact that the variables $\frac{Q_d}{M} \to p_c(d)$ become asymptotically deterministic, the capacities of the individual binomial channels add up and the overall capacity of this channel becomes $p_c(0)C_0 + p_c(1)C_1 + \ldots$. In particular, for large c, this capacity approaches 1, as each input sequence can be accurately estimated using a bit-wise majority decision over the sequences in each cluster.

Loss of ordering information: Without the additional information about Π, the receiver cannot directly allocate the clusters with their input sequences anymore. As there are $M - Q_0$ non-empty clusters, the receiver has to decide between roughly M^{M-Q_0} possible allocations of clusters and input sequences. Therefore, the actual overall capacity of the channel is smaller by factor of $\frac{1}{ML}\log(M^{M-Q_0}) \to \beta(1-\mathrm{e}^{-c})$.

Parameter range: Theorem 1 only holds for the low-noise scenario, as in this case, different clusters have a smaller overlap, which simplifies the maximization of the mutual information. For more details, see the end of Section III.

III. Proof of Theorem 1

We now prove Theorem 1. Let $\mathcal{C} \subseteq \Sigma^{ML}$ be a code with rate $R = \frac{\log|\mathcal{C}|}{ML}$. From Fano's inequality, we have

$$MLR \le 1 + P_e MLR + I(X^{ML}; Y^{NL}),$$

where P_e is the decoding error probability. Since Y^{NL} can be obtained from $Z^M = (Z_1, \ldots, Z_M)$ by ignoring the labels, the mutual information $I(X^{ML}; Y^{NL})$ can be bounded from above by the data processing inequality and we obtain

$$I(X^{ML}; Y^{NL}) \le I(X^{ML}; Z^M).$$

The main difficulty in maximizing the mutual information is the non-regularity of the channel, i.e., the entropy $H(Z^M|X^{ML})$ depends on the channel input X^{ML}. This is because if X^{ML} contains input sequences that are close in Hamming distance, this will result in a small channel entropy as different channel realizations may result in the same output. On the contrary, if X^{ML} has large mutual Hamming distances, the channel entropy is maximized, as each channel realization results in a distinct channel outcome with high probability. Similarly, the output entropy $H(Z^M)$ is maximized when the distribution of X^{ML} favors a large Hamming distance between its sequences. As it turns out, the mutual information maximizing input distribution will be the one that maximizes $H(Z^M)$. For some deterministic constant $\alpha > 0$, denote by $\mathcal{T}(\alpha) \subseteq \mathcal{D} \triangleq \{i \in [M] : D_i > 0\}$ the largest subset of drawn input sequences such that $d(X_i, X_j) \ge \alpha L$ for all

[1]Note that for the derivation in [7] the same genie has been used which allowed the receiver to identify duplicates.

$i, j \in \mathcal{T}(\alpha)$, which provides a measure of the scattering of the input sequences[2]. For brevity, we simply write $\mathcal{T} \triangleq \mathcal{T}(\alpha)$ in the following. By this definition $\mathcal{T}$ is a random variable that solely depends on α and the distribution of X^{ML} and $D^M = (D_1, \dots, D_M)$. An upper bound on $I(X^{ML}; Z^M)$ based on the expected value of $|\mathcal{T}|$ is given in Lemma 1.

Lemma 1. *Let $T \triangleq \mathbb{E}[|\mathcal{T}|]$ be the expected size of $\mathcal{T}$. Then*

$$I(X^{ML}; Z^M) \leq ML \sum_{d=1}^{N} p_c(d) C_d - 2T \log T + TL(1 - H(\gamma)) + M(1 - \mathrm{e}^{-c}) \log T - ML(1 - \mathrm{e}^{-c})(1 - H(\gamma)) + o(ML).$$

for any γ, δ, α with $\frac{1}{2} > \gamma$, $\gamma = \alpha + 2\delta$, $\alpha > 2\delta$, and $\delta > p$.

Proof. We start by bounding the output entropy $H(Z^M)$ from above. We can assume w.l.o.g. that the input distribution is constant over all permutations $\pi : [M] \mapsto [M]$, i.e., $\mathbb{P}(X_i^L = x_{\pi(i)}^L \ \forall\, i \in [M]) = \mathbb{P}(X_i^L = x_i^L \ \forall\, i \in [M])$. This is because $\mathbb{P}(Z^M = z^M)$ only depends on the sums $\sum_\pi \mathbb{P}(X_i^L = x_{\pi(i)}^L \ \forall\, i \in [M])$ and not the individual input probabilities $\mathbb{P}(X_i^L = x_i^L \ \forall\, i \in [M])$. Denote by $\bar{Z}_i = \{Y_j^L : I_j = i\}$ the output clusters, ordered by their original sequence. Under the previous assumption, it can be verified that $\mathbb{P}\left(Z^M = z^M\right) = \mathbb{P}\left(\bar{Z}^M = z^M\right)$ for all z^M and thus

$$H(Z^M) = H(\bar{Z}^M).$$

Since marginalization reduces entropy we obtain

$$\begin{aligned} H(\bar{Z}^M) &\leq H(\bar{Z}^M | \mathcal{T}, D^M) + 2M + N \\ &= \sum_{d^M} \sum_{\tau \subseteq \mathcal{D}} H(\bar{Z}^M | \tau, d^M) \mathbb{P}\left(\tau, d^M\right) + 2M + N, \end{aligned} \quad (2)$$

where we additionally used that the number of possible values for D^M and $\mathcal{T}$ is given by $\binom{N+M-1}{M-1}$, respectively 2^M and thus $H(\mathcal{T}, D^M) \leq \log \binom{N+M-1}{M-1} + M \leq 2M + N$. Denote now by $\tau = \{i_1, \dots, i_{|\tau|}\} \subseteq \mathcal{D}$ with $i_1 < \dots < i_{|\tau|}$ the realization of $\mathcal{T}$. We decompose the entropy $H(\bar{Z}^M | \mathcal{T} = \tau, D^M = d^M)$ into free output clusters $\bar{Z}_\tau^M = (\bar{Z}_{i_1}, \dots, \bar{Z}_{i_{|\tau|}})$, and output clusters $\bar{Z}_{\tau^\mathsf{c}}^M$, with $\tau^\mathsf{c} = \mathcal{D} \setminus \tau$ and obtain

$$\begin{aligned} H(\bar{Z}^M | \mathcal{T} = \tau, D^M = d^M) &= H(\bar{Z}_\tau^M | \mathcal{T} = \tau, D^M = d^M) \\ &+ H(\bar{Z}_{\tau^\mathsf{c}}^M | \mathcal{T} = \tau, D^M = d^M, \bar{Z}_\tau^M). \end{aligned} \quad (3)$$

Conditioned on $D^M = d^M$, the marginal distributions of $\bar{Z}_i$ with $i \in [M]$ follow the output distribution of the binomial channel from Lemma 3 with $d \triangleq d_i$ draws. Note that the $\bar{Z}_i$ can be correlated through the distribution of X^{ML}. Denote by H_d the maximum output entropy of this channel. As the channel entropy of the binomial channel is given by $H(\mathsf{Bin}(d, p))$, it follows that $H_d = C_d + H(\mathsf{Bin}(d, p))$. Hence the entropy of the clusters $i \in \tau$ is at most $H(\bar{Z}_i | \mathcal{T} = \tau, D^M = d^M) \leq LH_{d_i}$. For the remaining $i \in \tau^\mathsf{c}$, we know that there always exists $J_i \in \tau$ such that $d(X_i^L, X_{J_i}^L) < \alpha L$, since otherwise X_i^L has distance at least αL to all other input sequences in τ and thus i would be contained in τ. Denote by

$$\mathsf{maj}(\bar{Z}_i) = \operatorname*{arg\,min}_{x^L \in \Sigma^L} \sum_{Y^L \in \bar{Z}_i} d(x^L, Y^L)$$

the bit-wise majority function over the d_i sequences in $\bar{Z}_i$, which represents the estimated center of the cluster $\bar{Z}_i$. Denote further for any sequence $Y^L \in \bar{Z}_i$ by $\mathcal{E}_i$ the event that $d(X_i, Y^L) \leq \delta L$, by $\mathcal{F}_j$ the event that $d(X_j, \mathsf{maj}(\bar{Z}_j)) \leq \delta L$ with indicators $\mathbb{1}_{\mathcal{E}_i}, \mathbb{1}_{\mathcal{F}_{J_i}}$ and abbreviate $\mathcal{M} \triangleq (\mathcal{T} = \tau, D^M = d^M, \bar{Z}_\tau^M)$. The entropy of Y^L is then bounded by

$$\begin{aligned} H(Y^L | \mathcal{M}) &\leq H(Y^L | \mathcal{M}, J_i, \mathbb{1}_{\mathcal{E}_i}, \mathbb{1}_{\mathcal{F}_{J_i}}) + H(J_i, \mathbb{1}_{\mathcal{E}_i}, \mathbb{1}_{\mathcal{F}_{J_i}}) \\ &\leq H(Y^L | \mathcal{M}, J_i, \mathcal{E}_i, \mathcal{F}_{J_i}) + L\mathbb{P}\left((\mathcal{E}_i, \mathcal{F}_{J_i})^\mathsf{c} | J_i\right) + \log|\tau| + 2, \end{aligned}$$

where $(\mathcal{E}_i, \mathcal{F}_{J_i})^\mathsf{c}$ is the complementary event of $(\mathcal{E}_i, \mathcal{F}_{J_i})$. Since $d(Y^L, \mathsf{maj}(\bar{Z}_{J_i})) \leq d(Y^L, X_i) + d(X_i, X_{J_i}) + d(X_{J_i}, \mathsf{maj}(\bar{Z}_{J_i})) \leq (2\delta + \alpha)L = \gamma L$, the conditional entropy of Y^L is bounded by $H(Y^L | \mathcal{M}, J_i, \mathcal{E}_i, \mathcal{F}_{J_i}) \leq LH(\gamma)$. Using [18, Lemma 4.7.2], we obtain $\mathbb{P}\left((\mathcal{E}_i, \mathcal{F}_{J_i})^\mathsf{c} | J_i\right) \leq 2^{-LD(\delta||p)+1}$, where $D(\delta||p)$ is the binary Kullback-Leibler divergence. Using the second part of Lemma 3, we therefore obtain for the entropy of all clusters $i \in \tau^\mathsf{c}$

$$\begin{aligned} H(\bar{Z}_i | \mathcal{M}) &\leq H(\bar{Z}_i | \mathcal{M}, Y^L) + H(Y^L | \mathcal{M}) \\ &\leq L(H_{d_i} - 1 + H(\gamma)) + \log|\tau| + \log d_i + o(L) \end{aligned}$$

Next, we bound the entropy in (3) by the sum of all marginal entropies of $\bar{Z}_i$, $i \in [M]$, combine all clusters with the same number of draws $d_i = d$ and plug this result into the upper bound for $H(\bar{Z}_M)$ in (2) such that

$$\begin{aligned} H(\bar{Z}^M) &\overset{(a)}{\leq} \sum_{d^M} \mathbb{P}\left(d^M\right) \Big(L \sum_{d=1}^{N} Q_d H_d - (M - Q_0 - T_{d^M}) \\ &\qquad \cdot (L - \log T_{d^M} - LH(\gamma)) \Big) + o(ML) \\ &\overset{(b)}{=} ML \sum_{d=1}^{N} p_c(d) H_d - (M - M\mathrm{e}^{-c} - T)(L - \log T - LH(\gamma)) \\ &\qquad + o(ML), \end{aligned}$$

where $T_{d^M} = \mathbb{E}\left[|\mathcal{T}| \ | D^M = d^M\right]$. Note that here both Q_d and T_{d^M} depend on d^M. In (a) we used $\sum_d \log(d) Q_d \leq \sum_d d Q_d = N$ and Jensen inequality on the expected value over $\mathcal{T}$. In (b) we split the expectation over d^M into the events $\mathcal{Q}$ and $\mathcal{Q}^\mathsf{c}$ (for the definition of $\mathcal{Q}$ see proof of Lemma 2) and used the asymptotic Poissonization of Q_d from Lemma 2. We now bound the channel entropy $H(Z^M | X^{ML})$ from below. In particular we will show that for a well-separated input, i.e., for a large $|\mathcal{T}|$ the channel entropy is also high. This is in direct correspondence with the findings in [11], where a similar property has been shown for a combinatorial channel. Let $\mathcal{U} = \{i \in \mathcal{D} : d(X_i^L, \bar{Z}_i) \leq \delta L\}$ where $d(X_i^L, \bar{Z}_i) \triangleq \max_{Y^L \in \bar{Z}_i} d(X_i^L, Y^L)$. Further let $\mathcal{S} = \mathcal{U} \cap \mathcal{T}$ with realization $\mathcal{S} = \sigma$. Then for all $i, j \in \sigma$ with $i \neq j$ on the one hand we have $d(X_i^L, X_j^L) \geq \alpha L$, and on the other hand $d(X_i^L, \bar{Z}_i) < \alpha L/2$. Therefore, $d(X_i^L, \bar{Z}_j) > \alpha L/2$. It follows that $\mathbb{P}\left(\bar{Z}^M = \pi \bar{z}^M | X^{ML}, \mathcal{S} = \sigma\right)$ is non-zero for at most one permutation π on the clusters in σ. This allows to employ Lemma 4, which gives a lower bound on the entropy $H(Z^M | X^{ML})$ of the permuted clusters based on $\mathbb{E}[|\mathcal{S}|]$. It remains to compute $\mathbb{E}[|\mathcal{S}|]$. Since $|\mathcal{U} \cup \mathcal{T}| \leq M - Q_0$, we have that $|\mathcal{S}| \geq |\mathcal{T}| + |\mathcal{U}| - (M - Q_0)$ and we obtain

$$\mathbb{E}[|\mathcal{S}|] \geq \mathbb{E}[|\mathcal{T}|] - (M - \mathbb{E}[|\mathcal{U}| + Q_0]).$$

As there are in total Q_d clusters with d drawn sequences, $|\mathcal{U}| = U_1 + \dots + U_N$ is the sum of binomial variables U_d.

[2] A similar quantity has been used in [11] to bound the redundancy of codes for a combinatorial channel and in [8], where $\mathcal{T}$ was defined over the output.

Each U_d has Q_d trials and success probability p_d, where $p_d = \mathbb{P}\left(d(X_i, \bar{Z}_i) \leq \delta L\right)$, which only depends on $d_i = d$. Using [18, Lemma 4.7.2], it holds that $p_d \geq (1 - \mathrm{e}^{-LD(\delta||p)L})^d \geq 1 - d\mathrm{e}^{-LD(\delta||p)}$ and hence the expected value of $\mathcal{U}$ is at least

$$\mathbb{E}\left[|\mathcal{U}|+Q_0\right] \geq \sum_{d=0}^{N} Q_d(1-d\mathrm{e}^{-2(\delta-p)^2 L}) = M(1-c\mathrm{e}^{-2(\delta-p)^2 L}).$$

Therefore, the expected value of $|\mathcal{S}|$ is bounded from below by $\mathbb{E}\left[|\mathcal{S}|\right] \geq \mathbb{E}\left[|\mathcal{T}|\right] - cM\mathrm{e}^{-L(\delta||p)}$. Using this bound on the expected value for Lemma 4 yields

$$H(Z^M|X^{ML}) \geq H(\bar{Z}^M|X^{ML}) + T\log T + O(M).$$

Here we used that $|\mathcal{S}| \leq |\mathcal{T}| \leq M$ to obtain the asymptotic estimation for the logarithmic expression. Finally, we have

$$\begin{aligned} H(\bar{Z}^M|X^{ML}) &\geq \sum_{d^M} \mathbb{P}\left(D^M = d^M\right) H(\bar{Z}^M|X^{ML}, D^M = d^M) \\ &= ML \sum_{d=0}^{N} p_c(d) H(\mathsf{Bin}(d,p)) + o(ML), \end{aligned}$$

where we used the conditional independence of $\bar{Z}_i$ given X^{ML} and D^M and Lemma 2 for the Poissonization of the Q_d. $\square$

In order to find the maximizing T note that $0 \leq T \leq M - \mathbb{E}\left[Q_0\right] \leq M(1 - \mathrm{e}^{-c}) + 1$. Denoting by $g(T)$ all terms of the upper bound, which contain T, we find that its derivative satisfies $g'(T) \geq -2\log(T/\mathrm{e}) + L(1 - H(\gamma))$. It can be verified that given $2\beta < 1 - H(\gamma)$ and for large enough M, we have $g'(T) > 0$ for all $0 \leq T \leq M(1 - \mathrm{e}^{-c}) + 1$ and thus $g(T)$ is monotonically increasing in T. It follows that $T^* = M(1 - \mathrm{e}^{-c}) + 1$ maximizes $g(T)$ under these conditions. Plugging T^* into the upper bound from Lemma 1 and using $P_e \to 0$ when $M \to \infty$ yields Theorem 1.

IV. Auxiliary Lemmas

Lemma 2. *For any fixed $c > 0$, there exist non-negative functions $f_d(M)$, $d = 0, \ldots, N$ with $f_d(M) \geq 0$ and $f_0(M) + f_1(M) + \cdots + f_N(M) = o(M)$ such that for $M \to \infty$*

$$\mathbb{P}\left(|Q_d - Mp_c(d)| \leq f_d(M) \ \forall \ d = 0, \ldots, N\right) \to 1.$$

Proof. Let $\mathcal{Q} = \{(Q_0, \ldots, Q_N) : |Q_d - Mp_c(d)| \leq f_d(M) \ \forall \ d = 0, \ldots, N\}$ be the sought-after event. Using $|Q_d - Mp_c(d)| \leq |Q_d - \mathbb{E}\left[Q_d\right]| + |\mathbb{E}\left[Q_d\right] - Mp_c(d)|$ together with the union bound and Chebyschev's inequality, we obtain

$$\mathbb{P}\left(\mathcal{Q}\right) \geq 1 - \sum_{d=0}^{N} \frac{\mathbb{V}\left[Q_d\right]}{\left(-|\mathbb{E}\left[Q_d\right] - Mp_c(d)| + f_d(M)\right)^2}.$$

It is known [17, ch. 2] that the first and second moment of Q_d can be computed explicitly to be

$$\begin{aligned} \mathbb{E}\left[Q_d\right] &= M\binom{N}{d}\frac{1}{M^d}\left(1 - \frac{1}{M}\right)^{N-d}, \\ \mathbb{E}\left[Q_d^2\right] &= \mathbb{E}\left[Q_d\right] + M(M-1)\frac{N^{[2d]}}{(d!)^2 M^{2d}}\left(1 - \frac{2}{M}\right)^{N-2d}, \end{aligned}$$

where $N^{[2d]} = N(N-1)\cdots(N-2d+1)$ is the falling factorial. Using the inequalities $1 - x \leq \mathrm{e}^{-x}$ for any $x \in \mathbb{R}$ and $\binom{N}{d} \leq \frac{N^d}{d!}$ for any $N, d \in \mathbb{N}_0$ with $N \geq d$ we directly find that $\mathbb{E}\left[Q_d\right] \leq Mp_c(d)\mathrm{e}^{-d/M}$. This allows to bound the variance of Q_d from above by

$$\begin{aligned} \mathbb{V}\left[Q_d\right] &\leq Mp_c(d)\mathrm{e}^{d/M} + M^2 p_d^2(c)\mathrm{e}^{4d/M}\left(1 - \left(1 - \frac{2}{M}\right)^d\right) \\ &\overset{(a)}{\leq} Mp_c(d)\mathrm{e}^{c} + M^2 p_d^2(c)\mathrm{e}^{4c}\frac{2d}{M} \overset{(b)}{\leq} 2Mp_c(d)\mathrm{e}^{4c}(1+c), \end{aligned}$$

where in (a) we used that for any $0 < x < 1$ and $d \in \mathbb{N}_0$ it holds that $(1-x)^d \geq 1 - dx$. In inequality (b) it has been used that $dp_c(d) \leq c$. Let us now define $f_d = |\mathbb{E}\left[Q_d\right] - Mp_c(d)| + (d+1)M^{\frac{1}{2}+\epsilon}\sqrt{p_c(d)}$ as the bounding function. It follows that the probability of $\mathcal{Q}$ is at least

$$\mathbb{P}\left(\mathcal{Q}\right) \geq 1 - M^{-2\epsilon}\sum_{d=0}^{N}\frac{2\mathrm{e}^{4c}(1+c)}{(d+1)^2} = 1 - o(1),$$

and consequently $\mathbb{P}\left(\mathcal{Q}\right) \to 1$, when $M \to \infty$, for any $\epsilon > 0$. It remains to prove the asymptotic behavior of $f_d(M)$. We start with the first summand of $f_d(M)$ and obtain on the one hand

$$\mathbb{E}\left[Q_d\right] - Mp_c(d) \leq Mp_c(d)(\mathrm{e}^{\frac{d}{M}} - 1) \triangleq \phi_d^+(M).$$

On the other hand, we bound the difference from above by

$$\begin{aligned} Mp_c(d) - \mathbb{E}\left[Q_d\right] &\overset{(a)}{\leq} M\left(p_c(d) - \frac{(c - d/M)^d}{d!}\mathrm{e}^{-\frac{c}{1-1/M}}\right) \\ &\leq Mp_c(d)\left(\frac{d^2}{N} + \frac{c}{M}\right) \triangleq \phi_d^-(M), \end{aligned}$$

where for inequality (a) we used that $\binom{N}{d} \geq \frac{(N-d)^d}{d!}$ for any $N, d \in \mathbb{N}_0$ with $N \geq d$ and $\ln(1-x) \geq -\frac{x}{1-x}$ for any $x < 1$. The sum over the bounds $\phi_d^+(M)$ is at most

$$\begin{aligned} \sum_{d=0}^{N}\phi_d^+(M) &\overset{(a)}{=} M\sum_{d=0}^{N}\sum_{k=1}^{\infty}\frac{c^d\mathrm{e}^{-c}d^k}{d!M^k k!} \overset{(b)}{\leq} M\sum_{k=1}^{\infty}\frac{1}{M^k k!}\mathrm{E}[D^k] \\ &\overset{(c)}{=} M\sum_{k=1}^{\infty}\frac{1}{M^k}[t^k]\mathrm{e}^{c(\mathrm{e}^t-1)} = M\left[\mathrm{e}^{c(\mathrm{e}^{\frac{t}{M}}-1)} - 1\right]_{t=1} \\ &= M\left(\mathrm{e}^{c(\mathrm{e}^{\frac{1}{M}}-1)} - 1\right) = o(M), \end{aligned}$$

where we used the series expansion of the exponential in (a). The expected value after inequality (b) is taken with respect to the Poisson distributed variable D with mean c. In equality (c), we used that the moment generating function of the Poisson distribution is equal to $M(t) = \mathrm{e}^{c(\mathrm{e}^t-1)}$. Similarly, the sum over the bounds $\phi_d^-(M)$ satisfies

$$\sum_{d=0}^{N}\phi_d^-(M) \leq \frac{1}{c}\mathbb{E}\left[D^2\right] + c = 2c + 1,$$

where we used that the second moment of a Poisson distribution is $\mathbb{E}\left[D^2\right] = c^2 + c$. Now, we can use that both $\phi_d^+(M) \geq 0$ and $\phi_d^-(M) \geq 0$, and thus $|\mathbb{E}\left[Q_d\right] - Mp_c(d)| \leq \max\{\phi_d^+(M), \phi_d^-(M)\} \leq \phi_d^+(M) + \phi_d^-(M)$ to obtain

$$\sum_{d=0}^{N}|\mathbb{E}\left[Q_d\right] - Mp_c(d)| \leq \sum_{d=0}^{N}\phi_d^+(M) + \phi_d^-(M) = o(M).$$

The second summand of $f_d(M)$ can be bounded by

$$\sum_{d=0}^{N}(d+1)M^{\frac{1}{2}+\epsilon}\sqrt{p_c(d)} \overset{(a)}{\leq} M^{\frac{1}{2}+\epsilon}\sum_{d=0}^{N}(d+1)\frac{c^{\frac{d}{2}}\mathrm{e}^{-\frac{c}{2}}}{\lfloor\frac{d}{2}\rfloor!}$$

$$\overset{(b)}{\leq} M^{\frac{1}{2}+\epsilon}\mathrm{e}^{\frac{c}{2}} \sum_{d=0}^{\lceil N/2 \rceil} (d+1)\left(p_c(d)+\sqrt{c}p_c(d)\right)$$

$$\overset{(c)}{\leq} M^{\frac{1}{2}+\epsilon}\mathrm{e}^{\frac{c}{2}}(1+c)(1+\sqrt{c}) = o(M)$$

for $\epsilon < \frac{1}{2}$. Here we used in (a) that $\sqrt{d!} \geq \lfloor \frac{d}{2} \rfloor!$ for any $d \in \mathbb{N}_0$ and for inequality (b) we split the sum into terms with even and odd d. Inequality (c) follows by identifying the sums over $p_c(d)$ and $dp_c(d)$ as the cumulative distribution function, respectively mean of a Poisson distribution. Hence, using the proposed $f_d(M)$ with any $0 < \epsilon < \frac{1}{2}$ yields the lemma. □

Lemma 3 (c.f. [16])**.** *Let $X^L = (X_1, \ldots, X_L) \in \Sigma^L$ with i.i.d. entries $X_i \sim \mathsf{Ber}(p_X)$. Further, let $d \in \mathbb{N}_0$ and denote by $Y_i^L = X^L \oplus E_i^L$, $i \in [d]$ with $E_i^L = (E_{i,1}, \ldots, E_{i,L})$ and i.i.d. $E_{i,k} \sim \mathsf{Ber}(p)$ outcomes from d-fold repeated transmission of X^L over independent binary symmetric channels with crossover probability p and $Z = \{Y_1^L, \ldots, Y_d^L\}$. The capacity $C_d = \max_{p_X} I(X^L; Z)$ of the channel is given by*

$$C_d = 1 + \sum_{k=0}^{d} B_{d,p}(k) \log \frac{B_{d,p}(k)}{B_{d,p}(k) + B_{d,p}(d-k)},$$

where $B_{d,p}(k) = \binom{d}{k} p^k (1-p)^{d-k}$ is the binomial distribution. Further, $H(Z|Y_i^L) \leq L(C_d - 1 + H(\mathsf{Bin}(d,p))) + \log d$ for any $i \in [d]$, where $\mathsf{Bin}(d,p)$ is a binomial random variable with d trials and success probability p.

Proof. The proof of the capacity result follows standard methods and is omitted for brevity. We prove only the second part of the lemma. The entropy $H(Z|Y_i^L)$ satisfies

$$H(Z|Y_i^L) = H(\{X^L \oplus E_1^L, \ldots, X^L \oplus E_d^L\}|X^L \oplus E_i^L)$$
$$\overset{(a)}{=} H(\{E_1^L \oplus E_i^L, \ldots, E_d^L \oplus E_i^L\}|X^L \oplus E_i^L),$$

where (a) is due to [19, Problem 2.14]. Denote by $E^{dL} = \{E_1^L \oplus E_i^L, \ldots, E_d^L \oplus E_i^L\}$. We will show that E^{dL} is independent of $X^L \oplus E_i^L$ for $p_X = \frac{1}{2}$. Consider the distribution

$$\mathbb{P}\left(E^{dL}{=}e^{dL}|X^L{\oplus}E_i^L{=}a^L\right){=}\frac{\mathbb{P}\left(E^{dL}{=}e^{dL}, X^L{\oplus}E_i^L{=}a^L\right)}{\mathbb{P}\left(X^L \oplus E_i^L{=}a^L\right)}$$

$$\overset{(a)}{=} 2^L \sum_{e_i^L \in \Sigma^L} \mathbb{P}\left(E^{dL} = e^{dL}, X^L = e_i^L \oplus a^L, E_i^L = e_i^L\right)$$

$$\overset{(b)}{=} \sum_{e_i^L \in \Sigma^L} \mathbb{P}\left(E^{dL} = e^{dL}, E_i^L = e_i^L\right),$$

where in (a), we used that $\mathbb{P}\left(X^L \oplus E_i^L = a^L\right) = 2^{-L}$ for all $a^L \in \Sigma^L$. In equality (b) we used the independence of X^L and E_j^L for all $j \in [d]$ together with $\mathbb{P}\left(X^L = x^L\right) = 2^{-L}$ for any $x^L \in \Sigma^L$ under the condition $p_X = \frac{1}{2}$. Hence, E^{dL} is independent of $X^L \oplus E_i^L$ for $p_X = \frac{1}{2}$ and consequently the conditional entropy $H(Z|Y_i^L)$ is maximized for $p_X = \frac{1}{2}$. Since $H(Z)$ is also maximized by $p_X = \frac{1}{2}$ and $H(Y_i^L) = L$ in this case, we have $H(Z|Y_i) \leq H(Z) - H(Y_i^L) + \log d \leq LC_d + H(Z|X^L) - L + \log d = L(C_d + H(\mathsf{Bin}(d,k)) - 1) + \log d$. □

Lemma 4. *Let $F^M = (F_1, \ldots, F_M)$ be M random variables over the same space and denote for any $\sigma \subseteq [M]$ by $\mathcal{P}(\sigma) = \{\pi : [M] \mapsto [M] : \pi(i) = i \ \forall \ i \in [M] \setminus \sigma\}$ the set of all permutations that only permute positions in σ. Further let $\mathcal{S} \subseteq [M]$ be a random variable such that for any $\sigma \subseteq [M]$ and f^M the conditional probability $\mathbb{P}\left(F^M = \pi f^M | \mathcal{S} = \sigma\right) > 0$ for at most one permutation $\pi \in \mathcal{P}(\sigma)$. Then, the entropy of F^M after a uniform random permutation $\Pi : [M] \mapsto [M]$ is bounded from below by*

$$H(\Pi F^M) \geq H(F^M|\mathcal{S}) + \mathbb{E}\left[\log(|\mathcal{S}|!)\right].$$

The proof of this Lemma is omitted for brevity here and will be provided in the full version of this paper.

V. Conclusion and Outlook

In this paper we have derived an upper bound on the channel capacity of DNA-based storage systems. Due to the fact that for moderate error probabilities and appropriate input distributions an accurate clustering operation is indeed possible for the receiver, we believe that the upper bound on the channel capacity is tight. However, this problem remains open. It would also be interesting to examine, as in [8], the capacity for very large noise, as well as the case when the logarithm of the number of sequences, $\log M$, is very large compared to L. Another important aspect for future research is to discuss the presence of insertions and deletions, as these types of errors are common in DNA synthesis and sequencing.

References

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

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

[3] S. M. H. T. Yazdi, Y. Yuan, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific Reports*, vol. 5, 2015.

[4] M. Blawat *et al.*, "Forward error correction for DNA data storage," *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016.

[5] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific Reports*, vol. 7, 2017.

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

[7] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2017, pp. 3130–3134.

[8] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," *arXiv:1902.10832*, 2019. [Online]. Available: http://arxiv.org/abs/1902.10832

[9] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," vol. 62, no. 6, pp. 3125–3146, 2016.

[10] M. Kovačević and V. 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.

[11] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2018, pp. 2411–2415.

[12] W. Song and K. Cai, "Sequence-subset distance and coding for error control in DNA-based data storage," 2018. [Online]. Available: http://arxiv.org/abs/1809.05821

[13] J. Sima, N. Raviv, and J. Bruck, "On coding over sliced information," 2018. [Online]. Available: http://arxiv.org/abs/1809.02716

[14] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Anchor-based correction of substitutions in indexed sets," in *Proc. of the IEEE Int. Symp. on Inf. Theory*, 2019.

[15] C. Rashtchian *et al.*, "Clustering billions of reads for DNA data storage," in *Conf. Neural Information Processing Systems*, 2017, pp. 3360–3371.

[16] M. 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.

[17] V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov, *Random allocations*. Washington, D.C.: V. H. Winston & Sons, 1978.

[18] R. Ash, *Information Theory*. Dover Publications Inc., 1965.

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