# Towards more precise and reliable estimations for the capacity of DNA storage channels

Giacomo Borin
Università di Trento
d.borin68.gb@gmail.com

Alessia Marelli and Rino Micheloni
DNAalgo
{alessia.marelli, rino.micheloni}@dnaalgo.com

***Abstract*—DNA-based data storage emerged as a new competitive technique for long-term storage and modeling it properly to study its capacity is a challenging problem for information theory. This paper investigates the capacity of DNA-storage channel for more general models. In the first part a capacity upper bound is obtained for a noise model insertions, deletions, and substitutions (IDS). The bound is found via reducing the problem to a known instance, even if with convoluted conditions on the channel parameters. In the second part we define a new adaptable channel model that generalize DNA-storage channel peculiarities. We then briefly show how known capacity bounding techniques can be leverage for this new model.**

***Post-scriptum*: after the write up of this work (May to June 2023) we were made aware that a similar channel model generalization is contained in [1] (published already in May 2023), that should be considered the first example of non-uniform standard sampling distribution.**

## I. Introduction

The key idea of DNA-based data storage is to encode information in nucleotides molecules, that present good resistance to deterioration and information density rates order of magnitude greater than classical SSD [2]. With the Illumina sequencing process the molecules are amplified using Polymerase Chain Reaction (PCR) and read in a random order [3]. Several noise sources are observed during this phase and their intensity increases whit the speed of this process. To be practical this noisy random sampling reading technique need to be combined with the use of dedicated error-correcting codes and to measure their quality we need to know the capacity of the channel.

The DNA-storage channel capacity is an open topics in information theory. Today its peculiar structure is abstracted as a noisy shuffling-sampling channel, e.g. see a complete survey in [4]. Noise model take an important role in the adequacy and the feasibility of this studies. In particular the shuffling-sampling channel (no noise) is studied in [5]. The case of a BSC noisy model without sampling is solved in [6], while a generalization with also the sampling is showed in [7], [8]. These last results are proven for a general discrete memoryless noise channel in [9].

The next step in this direction would be the use of a noise model allowing not only substitution errors, but also the insertion or deletion of nucleotides during the sequencing. The study of channel with these kind of errors, also called synchronization errors, dates back to the 60' with [10]. In recent decades they received intensive study [11]–[13], mainly thanks also to numerical methods e.g. [14], but still it is an open problem for several channel models.

There are also other concerns related to the noisy shuffling-sampling channel abstraction when the results are confronted with real observations. For example as a consequence of uniform sampling assumption the number of cluster of a fixed size should follow a Poisson distribution, but simulations in [15, Figure 10] prove a non negligible statistical distance. Also errors may happen also during synthesis phase, violating the noise independence assumption, for example an important constrain noted in [2] is that long homopolymers or particular distribution of nucleotides in a molecule render it inoperable.

*Contributions:* We leverage the previous results on the capacity of noisy shuffling-sampling channel to obtain a converse bound for a general noise model that consider also synchronization errors using an auxiliary channel technique.

Then, partially inspired by underlying assumptions used in previous literature and analyisis in [16], we propose a new channel abstract model that allows to chose the sampling distribution and different noise channel for each cluster size.

*Paper outline:* In Section II we recall notions for the DNA-storage channel and the synchronization errors. In Section III we generalize previous upper bounds for channels with synchronization errors, then in Section IV we introduce a new channel model. In Section V we take some conclusions and explore future directions.

## II. Problem and Channel Setting

*Notation Conventions:* Random variables will be denoted by capital letters $X$, specific values they may take will be denoted by the corresponding lower case letters $x$, and their alphabets will be denoted by calligraphic letters $\mathcal{X}$. Random sequences $X^n$ and their realizations $x^n$ will be super-scripted by their dimension $n$, where $*$ means that $x^*$ has unspecified length, and $\mathcal{X}^*$ is the set of sequences of arbitrary length from $\mathcal{X}$. The mutual information for a DMC $V$ with input distribution $P_A$ will be denoted also by $I(P_A, V)$. The Poisson probability mass functions with parameter $c$ is denoted by $\mathrm{Poi}_c(d) = \frac{e^c c^d}{d!}$. We label numerical sets as $[M] := \{1, ..., M\}$ and the integers as $\mathbb{N}$.

*DNA Channel Model:* To capture the peculiarities of the DNA-storage channel a general model is used in literature, e.g. see [4], referred as noisy shuffling-sampling channel. For a $(c, \beta, W)$-noisy shuffling-sampling channel the input is composed by a sequence of $M$ strands of $L$ nucleotides, i.e. of values in $\mathcal{X} = \{A, C, T, G\}$, with $\beta = \log_2(M)/L$. Then $N = cM$ random sampled sequences are passed through a noise channel $W$ independently.

A rate $R$ is achievable for the $(c, \beta, W)$-noisy shuffling-sampling channel if we can define error correcting codes $\mathcal{C} \subseteq \mathcal{X}^{ML}$ with error probability arbitrarily close to zero and $\frac{\log_2 |\mathcal{C}|}{ML} \geq R$. The supremum of the achievable rates is referred as the capacity of the channel. To our knowledge the more general results for the capacity of noisy shuffle-sampling channel consider a DMC noise and are contained in [9].

**Theorem II.1** (Theorem 10 [9])**.** *Assume that the DNA channel* $\mathsf{DNA}(c, \beta, W)$ *satisfies* $W(y \mid x) > 0$ *for all* $x \in \mathcal{X}, y \in \mathcal{Y}$. *Then, its capacity* $C(\mathsf{DNA}(c, \beta, W))$ *is upper with*

$$\max_{P_X \in \mathcal{P}(\mathcal{X})} \sum_{d \in \mathbb{N}^+} \mathrm{Poi}(c, d) [I\left(P_X, W^{\oplus d}\right) + \\ + \Omega_d\left(\beta, P_X, W\right)] - \beta\left(1 - \mathrm{Poi}(c, 0)\right) \ ; \quad (1)$$

*where* $\Omega_d\left(\beta, P_X, W\right)$ *is defined as:*

$$\Omega_d\left(\beta, P_X, W\right) := \left[\beta \wedge (2\beta - 2 \cdot I(P_X, W^{\oplus d}) + I(P_X, W^{\oplus 2d}))\right] . \quad (2)$$

The term $\Omega_d$ is necessary to counter to the possibility that for high-noise short-molecule regime there is not enough redundancy for implement an index based coding strategy (usually used instead to prove rate achievability), see [4, Section 4.3] and [9, Section IV.8)]. To our knowledge all known capacity results for noisy shuffling-sampling channel fail to consider also errors due to insertion or deletion of nucleotides from the sample molecules, leaving it as an open question, e.g. see [4, Section 7.2]. To close this gap in this paper we consider the following channels as noise model:

**Definition II.2.** A *discrete memoryless synchronization channel* (DMSC) with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is a channel that can be expressed by the probability transition matrix $W(y^*|x)$, where $x \in \mathcal{X}$ and $y^* \in \mathcal{Y}^*$ (including the empty string).

We say that a DMSC has has *finite drift* if for each $n \in \mathbb{N}$ there exists a *drift* $D \in \mathbb{N}$ such that $W(y^* \mid x) = 0$ for all $y^*$ of length more than $D$.

For DMSC there is no general single letter formula, like for DMC, but in [10], [17], they have generalized Shannon's theorem as:

**Theorem II.3.** *For any DMSC* $W$ *with finite drift its capacity exists and is given by:*

$$C = \lim_{n \to \infty} \frac{1}{n} \max_{P_{X^n}} I\left(P_{X^n}; W\right) \ .$$

To generalize results from the DMC setting to the DMSC we use a technique proposed in [14] recalled here.

### A. Auxiliary channels

Consider an additional output $V$ obtained from an auxiliary channel on $W$. By using basic results from information theory they obtain the following bounds:

$$I(X; Y, V) - H(V) \leq I(X; Y) \leq I(X; Y, V) \quad (3)$$

Now, by considering the capacity of the channel with auxiliary information $C(W, V) = \lim_{n \to \infty} \frac{1}{n} \max_{P_{X^n}} I(P_{X^n}; W, V)$ for the case in which $V$ does not depend on the distribution of $X$, they get the following lower and upper bounds:

$$C(W, V) - \lim_{n \to \infty} \frac{1}{n} H(V^n) \leq C(W) \leq C(W, V) \quad (4)$$

By proceeding we can use a particular auxiliary channel $V$ to render a DMSC to a DMC and leverage the bounds obtained before. as As in [14] consider an integer $v > 0$ and divide the input channel of length $n$ into blocks of length $v$: $X = X_1^{(v)}, ..., X_q^{(v)}$. We then define the *auxiliary drift channel* $V$ as follows: $V_i$ represents the total drift occurred in the block $X^{(i)}$, i.e. the difference between the number of output and input symbols associated to this block. Using this additional information, it is possible to divide the output channel $Y = Y_1^{(v)}, ..., Y_q^{(v)}$ in the same way, such that $Y_i^{(v)}$ depends *exclusively* on $X_i^{(v)}$. Thus, we have obtained a DMC $W^{(v)}$ where the input alphabet is $\mathcal{X}^v$, and the output alphabet is $\mathcal{X}^*$. For a finite drift DMSC we know also that $|y^{(v)}| \leq vD$.

Once taken into consideration $V$ we get that, given an input length of $n = qv$, we get the following equality:

$$\frac{1}{n} I(X; Y, V) = \frac{1}{n} I(X_1^{(v)}, ..., X_q^{(v)}; Y_1^{(v)}, ..., Y_q^{(v)}) = \\ = \frac{1}{v} I(X^{(v)}, Y^{(v)}) \quad (5)$$

thus we have that $C(W^{(v)}) = \frac{1}{v} \max_{P_{X^{(v)}}} I(P_{X^{(v)}}, W^{(v)})$. Since we are dealing with a DMC the term can be evaluated with the classical Blahut-Arimoto Algorithm (BAA) [18], [19].

Moreover from (4) we have that for all $v > 0$ hold

$$|C(W) - C(W^{(v)})| \leq \frac{1}{n} H(V) = \frac{H(V_1)}{v} \quad (6)$$

and we only need to bound the right torm to have convergence of the capacities. For a finite drift channel we have $V_1 \in \{0, ..., v + dv\}$, so $H(V_1) \leq \log((1 + d)v)$ so that $H(V_1)/v \to 0$ when $v \to \infty$. When looking at a particular channel we can improve the bound computing the entropy via numerical methods, for example for an IDS we may use a Markov walk on the integers.

## III. Upper bound by reduction to a DMC

For this section consider a DMSC $W$ with finite drift $D$, as the IDS channel, and output lenght independent on the input by reducing it to a DMC. Now we generalize Theorem II.1 by reducing $W$ to a DMC.

**Theorem III.1.** *Assume that the DNA channel* $\mathsf{DNA}(c, \beta, W)$ *satisfies* $W(y^* \mid x) > 0$ *for all* $x, y^* \in \mathcal{X} \times \mathcal{Y}^*$ *up to the* *drift length $D$. Thus, its capacity $C(\mathsf{DNA}(c,\beta,W)$ is upper bounded for all $v \in \mathbb{N}$ by:*

$$\frac{1}{v}\sup_{P_{X^v}}\sum_{d=0}^{\infty}\mathrm{Poi}(c,d)[I(P_{X^v},W^{\oplus d})+ \\ + \Omega_d\left(v\beta,P_{X^v},W^{(v)}\right)] \; - \; \beta\left(1-\mathrm{e}^{-c}\right) \; . \tag{7}$$

*Proof:* We fix an integer $v > 0$, partition the strand length $L$ to sub-strands of length $v$ each, and assume that a genie reveals the decoder the output symbols for each sub-block in each of the output strands via an auxiliary drift channel, as in Section II-A. This reduces the noise channel into a DMC sequencing channel $W^{(v)}$, the strand-length parameter increases to $v\beta$, and the number of channel uses is reduced by a factor of $v$, thus the ratio between strands in output and input and it is unchanged. Thus we have a new DNA channel $\mathsf{DNA}(c,\beta_v,W^{(v)})$. If a rate $R$ is achieved by $\mathsf{DNA}(c,\beta,W)$ then it is also achievable in the DNA channel with auxiliary information, thus a capacity upper bound $\bar{C}_{A,v}$ for the second channel is also a bound for the first one $C(\mathsf{DNA}(c,\beta,W))$, modulo dividing it by $v$ to take into consideration that the alphabet is $\mathcal{X}^v$ instead of $\mathcal{X}$. Thus we have that:

$$C(\mathsf{DNA}(c,\beta,W)) \leq \frac{1}{v}C_{A,v} \; . \tag{8}$$

Since now the noise channel $W^{(v)}$ is a DMC we can use Theorem II.1. The hypothesis $W(y|x) > 0$ implies immediately that also $W^{(v)}(y^* \mid x^v) > 0$ for all output up to drift length $Dv$, thus we can plug the parameters $(c,\beta_v,W^{(v)})$ in Theorem II.1, after dividing by $v$ the obtained bound we can transport it to the initial channel with (8). By taking the infimum we obtain the desired inequality. ■

The formula is, to our knowledge, the only non trivial upper for the capacity of the DNA channel with synchronization errors, but it is unpractical to be evaluate. Moreover the term $\Omega_d(\cdot)$ take in consideration essentially only degenerate cases, avoided in a real situation, thus we interested in the case in which we can ignore it. In what follows, we loosen the bound to obtain an upper bound that depends on the capacity of the channels $\{W^{\oplus d}\}_{d\in\mathbb{N}}$, though under a condition on a minimal value of $\beta$.

**Theorem III.2.** *Consider DNA channel $\mathsf{DNA}(c,\beta,W)$ satisfying the assumptions of Theorem III.1. If it further holds for all $d > 0$ that*

$$2\beta \leq \liminf_{v\to\infty}\frac{1}{v}\Big[\max_{P_{X^v}} I(P_{X^v},(Y^*)^d)- \\ - \max_{P_{X^v}} I(P_{X^v},(Y^*)^d \mid (\bar{Y}^*)^d)\Big] \; ; \tag{9}$$

*where the variables $(Y^*)^d, (\bar{Y}^*)^d$ are two independent outputs of $X^{(v)}$ through $(W^{(v)})^{\oplus d}$. Then*

$$C(\mathsf{DNA}(c,\beta,W)) \leq \sum_{d=0}^{\infty}\mathrm{Poi}(c,d)C(W^{\oplus d}) - \beta\left(1-\mathrm{e}^{-c}\right) \; , \tag{10}$$

*where $C(W^{\oplus d})$ is the channel capacity for the multistrand channel $W^{\oplus d}$.*

As it will be clear by the proof the requirement in (9) is necessary to discard the term $\Omega_d(\cdot)$ in the upper bound. Theorem III.2 thus implies that there is an interval for $\beta$ in which the upper bound conjectured in [4, Section 7.2] holds. However, it appears to be computationally difficult to explicitly compute this critical value. This can be compared with the symmetric DMC setting in [9], where a simple interval can be found since it suffices to verify this condition for $d = 1$.

*Proof:* Consider the upper bound (7) of Theorem III.1.

To handle the term $\Omega_d$ now we use we use equations [9, (C.1) to (C.4) and (C.8)][1] the inequalities (C.1) to (C.4) and the equality (C.8) to rewrite (C.4) from Appendix C [9], obtaining that:

$$I(X^v,W^{\oplus d}) + \Omega_d\left(v\beta,P_{X^v},W^{(v)}\right) \leq \\ \leq \left(I(X^v,Y^d \mid \bar{Y}^d) + 2v\beta\right) \vee I(X^v,W^{\oplus d}) \; . \tag{11}$$

We thus clearly have that by fixing all the parameters, but $v$, (9) implies that we can ignore the $\Omega_d(\cdot)$ term for the maximization.

Now we would like to bring both the supremum and a limit inside the sum in (7). To this we use the Weistrass M-test we prove that the sum:

$$\sum_{d=0}^{\infty}\mathrm{Poi}(c,d)\frac{1}{v}\left[I(X^v,W^{\oplus d}) + \Omega_d\left(v\beta,P_{X^v},W^{(v)}\right)\right]$$

converges uniformly with respect to $v$. The term inside can be bound as:

$$\frac{\mathrm{Poi}(c,d)}{v}\left|I(X^v,W^{\oplus d}) + \Omega_d\left(v\beta,P_{X^v},W^{(v)}\right)\right| \leq \\ \leq \mathrm{Poi}(c,d)\frac{1}{v}\left(v\log(|\mathcal{X}|) + v\beta\right) \leq \mathrm{Poi}(c,d)(\log(|\mathcal{X}|) + \beta),$$

that clearly converges to a finite value when sum over $d$ since $\mathrm{Poi}(c,\cdot)$ is a probability distribution.

Thus by bringing the supremum inside (7) we get:

$$C(\mathsf{DNA}(c,\beta,W)) \leq \sum_{d=0}^{\infty}\mathrm{Poi}(c,d)\frac{1}{v}\sup_{P_{X^v}}\left[I(X^v,W^{\oplus d})+ \\ +\Omega_d\left(v\beta,P_{X^v},W^{(v)}\right)\right] \; . \tag{12}$$

At this point we can exploit again the uniform convergence to study the limit of the sum in (12) and obtain:

$$\sum_{d=0}^{\infty}\mathrm{Poi}(c,d)\lim_{v\to\infty}\left(\frac{1}{v}\max_{P_{X^v}} I(P_{X^v},W^{\oplus d})\right) - \\ \beta\left(1-\mathrm{e}^{-c}\right) \tag{13}$$

By Theorem II.3 we have the following convergence to the capacity

$$\lim_{v\to\infty}\frac{1}{v}\max_{P_{X^v}} I(X^v,W^{\oplus d}) = C_{W^{\oplus d}} \tag{14}$$

[1] The symbol $\vee$ therein is a typo, and should be $\wedge$, i.e., a minimum.

that we can substitute in (13) and (12) to obtain the required bound. ■

**Remark III.3.** Nonetheless, if the (9) condition holds, then the upper bound in (10) can be further upper bounded using upper bounds on the capacity $C_{W^{\oplus d}}$ To evaluate these we can use again the technique of the drift channel in combination to the BAA, by finding $C((W^{(v)})^{\oplus d})$. The transition matrix for $(W^{(v)})^{\oplus d}$ can be obtained from the one of the single strand channel via the Kronecker product on the rows. A main drawback is that the dimensions of this matrix increase exponentially in $d$, rendering it unfeasible even for not so large values. However when the error probabilities are close to 0 the capacity rapidly approaches the upperbound $\log(|\mathcal{X}|)$ reducing the effect of this computational limit. Moreover since the drift channels are independent the bound for the convergence of the has now the form of:

$$|C_{W^{\oplus d}} - C((W^{(v)})^{\oplus d})| \le d\frac{H(V_1)}{v} \ ; \tag{15}$$

again reducing the effectiveness of the bound when $d$ is large.

## IV. New Channel Model

When looking into the details of the proof strategy of Theorem II.1 in [9] you can see that another auxiliary channel with the knowledge of the cluster sets is used. This is used also in other converse bound, like for [5], [8]. Also part of the proof of the direct bound in [7] used a genie-aided decoder with this information, to then prove clustering being transparent for capacity calculations.

Thus we propose now a new abstract channel model where we assume that the clustering procedure is genie-aided (thus the *Clustered* label). We consider this a valid assumption since in reasonable situations ($\beta \ll 1$) the edit distance is big enough so that the classical greedy clustering procedure succeed with overwhelming probability, without necessity of additional redundancy clustering oriented. This is verified by the actual computations and some results can be found in Appendix A of [20]. However we point out that it is still an open question to precisely identify the maximum $\beta$ for a general channel and input distribution so that the probability of failing the clustering decreases exponentially in $M$.

*Noisy Clustered Shuffling-Sampling Channel (NCSSC):* To define the NCSSC we consider a sequence of distributions $\pi^\infty := \{\pi^{(M)}\}_{M\in\mathbb{N}_+}$ so that the $M$ marginal distributions are identical (but not independent) and a sequence of noisy channels $V_\infty := \{V_d\}_{d\in\mathbb{N}_+}$ (even DMSC). The input to the channel is a list $(x_1^L, ..., x_M^L)$ of $M$ sequences of length $L$ over a finite input alphabet $\mathcal{X}$. The channel performs the following operations:

1) **Sampling**: Let $S^{(M)} = (S_1^{(M)}, S_2^{(M)}, \dots, S_M^{(M)}) \sim \pi^{(M)}$, for each $m \in [M]$ the strand $x_m^L$ is assigned to the non negative integer $S_m^{(M)}$.
2) **Noise**: for each $m \in [M]$ the strand $x_m^L$ is passed through a noise channel $V_{d_m} : \mathcal{X}^L \to \mathcal{Y}^*$, with $d_m$ given by $S_m^{(M)}$. For the strands associated to 0 the output is empty (i.e. $V_0$ is an erasure channel that erase the input with probability 1).
3) **Shuffling**: The remaining strands are shuffled maintaining the information over the assigned integer.

The classical $(c, \beta, W)$ DNA-storage channel can be obtained by fixing $S^{(M)} \sim \text{Multinomial}\left(cM; \left(\frac{1}{M}, \frac{1}{M}, \dots \frac{1}{M}\right)\right)$ and $V_d$ as the multistrand channel $W^{\oplus d}$. This way $S_m^{(M)}$ corresponds to the size of cluster in the classical random sampling model. To study its asymptotic behaviour we need additional property on the distributions.

**Definition IV.1.** A sequence of sampling distributions $\pi^\infty$ is *proper* if:

1) there exists a distribution $\pi = \{\pi_d\}_d$ such that $\pi^\infty$ *strongly converges* to $\pi$, i.e. with $\sum_{d=0}^{\infty} |\pi^{(M)}(d) - \pi(d)| \to 0$ for $M \to \infty$;
2) there exists a constant $\kappa > 0$ so that for all $d > 0$ and $m, m' \in [M]$, $m \ne m'$, it holds
$$\Pr(S_m = S_{m'} = d) - (\pi^{(M)}(d))^2 \le \kappa \frac{\pi^{(M)}(d)}{M} \ ; \tag{16}$$
3) there exists a $\delta \in (0, 1/2)$ so that
$$\sum_{d>0} \sqrt{\pi_d^{(M)}} = o(M^{2\delta}) \ , \ \sum_{d>0} \sqrt[4]{\pi_d^{(M)}} = o(M^{\frac{1}{2}-\delta}). \tag{17}$$

In analogy to [8], [9] we define $Q_d^{(M)} = |\{m \in \{1, ..., M\} \mid S_m^{(M)} = d\}| = \sum_{m=1}^{M} \mathbb{1}(S_m^{(M)} = d)$, i.e. the number of clusters associated to $d$. Since $\Pr(S_m^{(M)} = d) = \pi_d^{(M)}$ then $\mathbb{1}(S_m^{(M)} = d) \sim Be(\pi_d^{(M)})$. As done for the multinomial distribution with [8, Lemma 2] we need to control the behavior of the sum $\sum_{d>0} |\frac{Q_d^{(M)}}{M} - \pi_d|$ also for general distributions.

**Lemma IV.2.** *Consider a proper sequence of distributions $\pi^\infty$, then we can define an event $\mathcal{Q}_M$ such that for $M \to \infty$ $\Pr(\mathcal{Q}_M) \to 1$ and on it $\sum_{d=0}^{\infty} |\frac{Q_d}{M} - \pi_d| \to 0$.*

*Proof:* To improve readability we avoid using the apex $M$ for the random variables. We start by bounding the term $|\frac{Q_d}{M} - \pi_d^{(M)}|$ for each $d > 0$ using the Chebyshev's inequality:

$$\Pr\left(\left|\frac{Q_d}{M} - \pi_d^{(M)}\right| \ge a_d^{(M)}\right) \le \frac{Var(Q_d)}{(Ma_d^{(M)})^2} =: p_d^{(M)} \ ; \tag{18}$$

Suppose to have $a_d^{(M)}$ so that $\sum_{d>0} a_d^{(M)}$, $\sum_{d>0} p_d^{(M)} \to 0$ for $M \to \infty$. At this point we would define the set $\mathcal{Q}_M$ as:

$$\mathcal{Q}_M := \left\{\left|\frac{Q_d}{M} - \pi_d^{(M)}\right| \le a_d^{(M)} \mid d > 0\right\} \ . \tag{19}$$

It satisfies the convergence requirement since on it

$$\sum_d \left|\frac{Q_d}{M} - \pi_d\right| \le \sum_d \left|\frac{Q_d}{M} - \pi_d^{(M)}\right| + \left|\pi_d - \pi_d^{(M)}\right| \le \\ \le \sum_d a_d^{(M)} + \sum_d \left|\pi_d - \pi_d^{(M)}\right| \to 0 \ . \tag{20}$$

Let's now fix $a_d^{(M)} := \frac{(\pi_d^{(M)})^{1/4}}{M^{1/2-\delta}}$. Observe that:

$$\sum_{d>0} a_d^{(M)} = \frac{1}{M^{1/2-\delta}} \sum_{d>0} (\pi_d^{(M)})^{1/4} \overset{(17)}{=} \frac{o(M^{1/2-\delta})}{M^{1/2-\delta}} \to 0 \tag{21}$$

Since $Q_d$ is a sum of Bernoulli variables $\mathbb{1}(S_m^{(M)} = d)$, we can bound the variance term from (18) by using $\pi_d^{(M)}(1 - \pi_d^{(M)}) \leq \pi_d^{(M)}$ for the $M$ square terms, while for the $O(M^2)$ covariance terms

$$Cov(\mathbb{1}(S_m = d), \mathbb{1}(S_{m'} = d)) = \Pr(S_m = S_{m'} = d) - (\pi_d^{(M)})^2$$

we use (16). Thus we can bound

$$p_d^{(M)} \leq \frac{\pi_d^{(M)}(1+c)}{M} \cdot \frac{M^{1-2\delta}}{(\pi_d^{(M)})^{1/2}} = \frac{(\pi_d^{(M)})^{1/2}}{M^{2\delta}}(1+c) \ . \tag{22}$$

Then we $\Pr(\mathcal{Q}_M) \to 1$, as requires, since:

$$\Pr(\mathcal{Q}_M^c) \leq \sum_{d>0} p_d^{(M)} = \frac{1}{M^{2\delta}} \sum (\pi_d^{(M)})^{1/2}(1+\kappa) \leq$$
$$\leq \frac{1+\kappa}{M^{2\delta}} \sum (\pi_d^{(M)})^{1/2} \overset{(17)}{=} \frac{o(M^{2\delta})}{M^{2\delta}} \to 0 \ . \tag{23}$$

■

For most cases, like the multinomial distribution used in the previous literature, the previous hypothesis are verified.

*Example of Use:* Thanks to Lemma IV.2 the NCSSC with proper distributions can be studied using the same techniques of the classical DNA channel. For example, consider a symmetric DMC noise model $\{W^{\oplus d}\}_{d\in\mathbb{N}_+}$. For this channel it is straightforward to exploit same the decoding strategy (minus the greedy clustering algorithm) used in [7] to obtain the direct bound of the classical DNA channel. To prove the bound also for the $\mathsf{NCSSC}(\pi^\infty, \beta, \{W^{\oplus d}\}_{d\in\mathbb{N}_+})$ we simply have to use Lemma IV.2 instead of [8, Lemma 2] and ignore all terms used to control clustering failure.

This way we get the following direct bound for $C(\mathsf{NCSSC}(\pi^\infty, \beta, \{W^{\oplus d}\}_{d\in\mathbb{N}_+})$:

$$\sum_{d\in\mathbb{N}^+} \pi_d C(W^{\oplus d}) - \beta(1 - \pi_0) \ . \tag{24}$$

Moreover as discussed at the start of the section all previous results in Section III are actually proved for the NCSSC instantiated with the classsical DNA channel.

Bound like (24) could be leveraged to adapt the sampling distribution to the observed one when it is not coherent with the Poisson one. Moreover more general result that does not fix the channel $V_\infty$ could be used to model errors happening during the writing phase, not comprehended in the classica model since they are not independent between strands originating from the same molecule.

## V. Conclusion and Future Directions

Via the use of the auxiliary drift channel in some original ways we where able to reduce the DNA-channel with synchronization errors to the already solved DMC case, obtaining a novel generalization of known bounds. The following steps in this direction would require a, possibly tailored to a noise model, numerical analysis to actually evaluate a validity interval for $\beta$ and the bound in (7).

Then using genie-aided clustering we defined a more general channel model with additional freedom for the distribution and the noise channel. We also give reasonable hypothesis to control the asymptotic behaviour of the distributions and use them to prove a first straightforward direct bound exploiting a known technique in literature. As pointed out before this generality makes the model adaptable, a valuable characteristic for a cutting edge field like DNA-storage.

It remains open the use of the NCSSC model for more interesting noise model, like DMSC. A possible strategy would be to generalize the concept of typical sets (see [21, Section 7.6]) also to more general channels, as done for example in [17, Section 4] for the proof of Theorem 1.

## References

[1] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "The noisy drawing channel: Reliable data storage in dna sequences," *IEEE Transactions on Information Theory*, 2023.

[2] Y. Erlich and D. Zielinski, "Dna fountain enables a robust and efficient storage architecture," *bioRxiv*, 2016.

[3] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on dna in silica with error-correcting codes," *Angewandte Chemie International Edition*, 2015.

[4] I. Shomorony, R. Heckel, *et al.*, "Information-theoretic foundations of dna data storage," 2022.

[5] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of dna storage systems," 2017.

[6] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," 2019.

[7] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the dna storage channel," IEEE, 2020.

[8] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the dna storage channel," p. 1–5, IEEE Press, 2019.

[9] N. Weinberger and N. Merhav, "The dna storage channel: Capacity and error probability bounds," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5657–5700, 2022.

[10] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory. Translated by E. Hewitt," *Transl., Ser. 2, Am. Math. Soc.*, vol. 33, pp. 323–438, 1963.

[11] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Transactions on Information Theory*, 2020.

[12] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," 2009.

[13] S. Diggavi and M. Grossglauser, "On information transmission over a finite buffer channel," *IEEE Transactions on Information Theory*, 2006.

[14] D. Fertonani, T. M. Duman, and M. F. Erden, "Bounds on the capacity of channels with insertions, deletions and substitutions," *IEEE Transactions on Communications*, vol. 59, pp. 2–6, 2011.

[15] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Improved read/write cost tradeoff in dna-based data storage using ldpc codes," IEEE, 2019.

[16] N. Weinberger, "Error probability bounds for coded-index dna storage systems," *IEEE Transactions on Information Theory*, 2022.

[17] R. L. Dobrushin, "The Shannon's theorems for channels with synchronization errors," *Probl. Peredachi Inf.*, vol. 3, no. 4, pp. 18–36, 1967.

[18] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, 1972.

[19] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, 1972.

[20] C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss, "Clustering billions of reads for dna data storage," 2017.

[21] T. M. Cover and J. A. Thomas, *Elements of information theory*. Hoboken, NJ: John Wiley & Sons, 2nd ed. ed., 2006.