Error Probability Bounds for Coded-Index DNA Storage Channels

Conference Paper

Author(s):
Weinberger, Nir

Publication date:
2022-03-02

Permanent link:
https://doi.org/10.3929/ethz-b-000535278

Error Probability Bounds for Coded-Index DNA Storage Channels

Nir Weinberger
The Viterbi Faculty of Electrical and Computer Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200004, Israel
nirwein@technion.ac.il

***Abstract*—The DNA storage channel is considered, where each codeword is comprised of M unordered DNA molecules. At reading time, the molecules are sampled N times with replacement, and then sequenced. A coded-index concatenated-coding scheme is proposed, in which the mth molecule of the codeword is restricted to an inner code, unique for each index. A low-complexity decoder is proposed that is based on separated decoding of each molecule (inner code), followed by decoding the sequence of molecules (outer code). Mild assumptions are made on the sequencing channel, in the form of the existence of an inner code and decoder with vanishing error probability. The error probability of a random code for the storage system is analyzed and shown to decay exponentially with N. This establishes the importance of high coverage depth N/M for achieving low error probability.**

I. Introduction

Various authors have recently proposed and analyzed coding methods for data storage systems based on a Deoxyribonucleic acid (DNA) medium (see a survey in [1]). In this channel model, information is stored in a pool of M DNA molecules, where each molecule is comprised of two complementary length-L strands of four nucleotides (Adenine, Cytosine, Guanine, and Thymine). The M molecules cannot be spatially ordered, and during reading, N molecules are independently sampled from the DNA pool, with replacement. Then, each of these sampled molecules is *sequenced* in order to obtain a length-L vector describing the synthesized nucleotides, and the N sequenced molecules is the channel output. Roughly speaking, the impairments of this channel include: (1) *Molecule errors* – e.g., the event in which some of the molecules are not sampled at all (erased). (2) *Symbol errors* – modeled by a channel $W^{(L)}$ which specifies the probability of sequencing some L-symbol vector conditioned that the information was (possibly other) L symbols. In this paper, we propose a random coding ensemble and a low-complexity decoder for this channel model, and analyze the average error probability.

In terms of fundamental limits, it was the capacity of such a channel which was first addressed [1], with the general conclusion that the capacity is positive only when $L = \beta \log M$, with $\beta > 1$. Under this scaling, [1]–[4], have derived bounds on the capacity, assuming a constant *coverage depth* N/M, and a discrete memoryless sequencing channel. In this paper, we focus on a somewhat different model for the following reasons: First, the tightest achievable bound for a discrete memoryless channel (DMC) [4] require a computationally intensive decoder, which is difficult to implement in practice. Second, in practice, the sequencing channel is not a DMC, and may include deletions and insertions [5], or constraints on the codeword symbols [6], [7]. Third, as was also established in [4], the error probability is dominated by molecule errors (erasures), and so the error probability decays as $e^{-\Theta(M)}$ rather than the $e^{-\Theta(ML)}$ decay rate anticipated from a blocklength of ML. This slow decay of the error probability is significant for practical systems of finite blocklength.

Accordingly, and following [2], in this paper, we theoretically analyze the error probability of a simple, yet general, coding method. The scheme follows a practical approach [8]–[11] in which the lack of order of the molecules is resolved by an *index*. The simplest version of indexing-based schemes uses the first $\log_2 M$ bits of each DNA molecule to specify its index $m \in [M]$, and is capacity achieving for noiseless sequencing channels, despite its rate loss of $1/\beta$, which seems to be inevitable [1], [3], [4], [12], [13]. Nonetheless, if the payload bits (the last $(\beta - 1)\log_2 M$ bits of the molecule) are arbitrary, then an erroneous ordering of the molecules can be caused by a single channel bit flip. This motivates us to consider in this paper *coded-indexing* based schemes for noisy sequencing channels. In such a scheme, the possible molecules of the codeword are chosen from a inner code – a sub-code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$ of all possible molecules. Moreover, this inner code is further partitioned into M equal size sub-codes $\mathcal{B}_m^{(L)}$ so that the mth molecule of a codeword is chosen only from $\mathcal{B}_m^{(L)}$. The inner code $\mathcal{B}^{(L)}$ thus also protects the index from sequencing errors. An outer code then specifies the valid sets of molecules.

Our proposed decoder is based on a decoder for the inner code $\mathcal{B}^{(L)}$, which is used to *independently* decode each of the N sequenced molecules to a sequence in $\mathcal{B}^{(L)}$. Since the decoder operates on a molecule-by-molecule basis, future design of codes based on this scheme is a feasible goal (L is typically on the order of $10^2 - 10^3$), and is much simpler than the decoder of [4] (the clustering-based decoder [12] also has a low complexity of $\Theta(N)$, but there are no guarantees on the decay rate of the error probability). A decoder for the outer

code is then used to resolve molecule erasures and undetected errors. Hence, the proposed coded-index based scheme is practically oriented, and its analysis is general, in the sense that very little is assumed on the sequencing channel. It is only required that a decoder for the inner code exists whose error probability decays to zero with increasing L. This addresses the first two issues raised above.

Regarding the third issue, as explained in [4], for fixed coverage depth ($N = \alpha M$ for some fixed $\alpha > 1$) the slow $e^{-\Theta(M)}$ decay rate of the error probability is the result of molecule errors (erasures), rather than sequencing errors. So, apparently, faster decay rate is only possible by increasing N. In accordance, we consider in this paper the scaling $N = \alpha_M M$, where α_M is (possibly) an increasing function of M (though rather slowly). Our main result is a single-letter upper bound on the error probability which decays as $e^{-\Theta(N)}$, achieved by a coded-index based scheme. An important consequence of this result is that operating at a large coverage depth N/M is of importance for low error probability. This is in opposed to capacity analysis, for which large N/M only provides marginal capacity gains [1, Sec. I]. We remark that our scheme is not capacity achieving under the DMC and fixed α model studied in [1]–[4], as it does not exploit multiple observations of the same molecule to increase the rate. However, the rate loss is small for sequencing channels which are fairly clean, as multiple observations only marginally increase the capacity in this case. Anyway, adapting our scheme to achieve capacity is an important open problem.

Previously, [12] has considered an (explicit) coded-indexing and concatenated coding scheme, whose decoder is based on (hard) output clustering, and so is mainly tailored to the binary symmetric channel. As described above, we consider here general sequencing channels, and focus on error probability analysis and simple decoding (see [14] for a detailed comparison with this, as well as with additional related work [15]). In our context, the conclusion is that the loss is more profound for small β. The rest of the paper is organized as follows. In Sec. II we establish notation conventions, formulate the DNA storage channel and coded-index based systems. In Sec. III we state our main result, and in Sec. IV we outline the proof. All proofs and further results and discussions are available in a full version of the paper [14]).

II. Problem Formulation

We begin with notation conventions. Random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be super-scripted by their dimension. The probability of the event $\mathcal{E}$ will be denoted by $\mathbb{P}(\mathcal{E})$, and its indicator function will be denoted by $\mathbb{1}(\mathcal{E})$. The expectation operator will be denoted by $\mathbb{E}[\cdot]$. Logarithms and exponents will be understood to be taken to the natural base. The binary Kullback–Leibler (KL) divergence $d_b\colon [0,1]\times(0,1) \to \mathbb{R}^+$ by $d_b(a||b) := a\log\frac{a}{b} + (1-a)\log\frac{(1-a)}{(1-b)}$. The number of *distinct* elements of a finite multiset $\mathcal{A}$ will be denoted by $|\mathcal{A}|$. The equivalence relation will be denoted by $\equiv$, and will mainly be used to simplify notation. Asymptotic Bachmann–Landau notation will be used. For a positive integer N we will denote $[N] := \{0, 1, \ldots, N-1\}$, where scalar multiplications of these sets will be used, e.g., as $\frac{1}{N}[N+1] = \{0, \frac{1}{N}, \ldots \frac{N-1}{N}, 1\}$.

Next, we formulate a sequence of channels, encoders and decoders for the DNA storage channel, indexed by M, the number of molecules in a codeword.

The channel model (reading mechanism): A DNA molecule is a sequence of $L \equiv L_M \in \mathbb{N}_+$ nucleotides (symbols) chosen from an arbitrary alphabet $\mathcal{X}$ (in physical systems $\mathcal{X} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, and in some previous works [1]–[3] a binary alphabet $\mathcal{X} = \{0, 1\}$ was assumed for simplicity). Thus, each molecule is uniquely represented by a sequence $x^L \in \mathcal{X}^L$. An input to the DNA channel is a sequence of M molecules, $x^{LM} = (x_0^L, \ldots x_{M-1}^L)$ where $x_m^L \in \mathcal{X}^L$ for all $m \in [M]$. A message is synthesized into a sequence of M molecules, x^{LM}. The DNA storage channel model is determined by the number of molecule samples $N \equiv N_M \in \mathbb{N}_+$, and by the sequencing channel $W^{(L)}\colon \mathcal{X}^L \to \mathcal{Y}^L$. The operation of the channel on the stored codeword is modeled as a two-stage process:

1) Sampling: N molecules are sampled uniformly from the M molecules of x^{LM}, independently, with replacement. Let $U^N \sim \text{Uniform}([M]^N)$ be such that U_n is the sampled molecule at sampling event $n \in [N]$. The result of the sampling stage is the vector $(x_{U_0}^L, x_{U_1}^L, \ldots, x_{U_{N-1}}^L) \in (\mathcal{X}^L)^N$. We also denote by S_m the number of times that molecule m was sampled, to wit $S_m = \sum_{n\in[N]} \mathbb{1}\{U_n = m\}$, the empirical count of U^N. It then holds that $S^M = (S_0, \ldots, S_{M-1}) \sim \text{Multinomial}(N; (\frac{1}{M}, \frac{1}{M}, \ldots \frac{1}{M}))$.

2) Sequencing: For each $n \in [N]$, $x_{U_n}^L$ is sequenced to $Y_n^L \in \mathcal{Y}^L$, where the sequencing of $x_{U_n}^L$ is independent for all $n \in [N]$. Denoting the channel output by $Y^{LN} = (Y_0^L, \ldots, Y_{N-1}^L) \in (\mathcal{Y}^L)^N$ it thus holds that

$$\mathbb{P}\left[Y^{LN} = y^{LN} \mid x^{LM},\, U^N\right] = \prod_{n\in[N]} W^{(L)}\left(y_n^L \mid x_{U_n}^L\right). \tag{1}$$

We make the following assumptions on the channel: (1) $L \equiv L_M = \beta \log M$ where $\beta > 1$ is the *molecule length parameter*. (2) N/M where $\alpha \equiv \alpha_M > 1$ is the *coverage depth scaling function*.

The encoder: A codebook is a set of different possible codewords $\mathcal{C} = \{x^{LM}(j)\}$. We propose the following restricted set of *coded-index based codebooks*:

Definition 1. Let $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$ be a collection of pairwise disjoint sets $\mathcal{B}_m^{(L)} \subset \mathcal{X}^L$ of equal cardinality, and let $\mathcal{B}^{(L)} := \cup_{m\in[M]}\mathcal{B}_m^{(L)}$ be their union. A DNA storage codebook is said to be a *coded-index* based codebook if $x_m^L(j) \in \mathcal{B}_m^{(L)}$ for all $m \in [M]$ and all $j \in [|\mathcal{C}|]$.

To wit, a codeword contains exactly a single molecule from each of the M sets $\{\mathcal{B}_m^{(L)}\}_{m\in[M]}$. The identity of the set from

which $x_m^L(j)$ was chosen from is considered an "index" of the molecule that is used by the decoder to order the molecules that has been decoded. A coded-index based codebook, can be thought of as a concatenated code. The set $\mathcal{B}^{(L)}$ is an inner-code, which is used to clean the output molecules from sequencing errors, and the dependency between molecules of different index m can be considered an outer-code which is used to cope with erasures (mainly due to the sampling stage).

The decoder: A general decoder is a mapping $\mathsf{D}\colon (\mathcal{Y}^L)^N \to [|\mathcal{C}|]$. We propose the following class of decoders, which are suitable for coded-index based codebooks. A decoder from this class is equipped with an inner-code decoder $\mathsf{D}_b\colon \mathcal{Y}^L \to \mathcal{B}^{(L)}$, and a threshold $T \in \mathbb{R}^+$, and decodes the channel output y^{LN} in three steps:

1) Correction of individual molecules: The decoder employs the inner-code decoder for each of the received molecules y_n^L, for each $n \in [N]$, and set $z_n^L = \mathsf{D}_b(y_n^L)$. Following this stage, it holds that $z^{LN} = (z_0^L, \ldots, z_{N-1}^L)$ is such that $z_n^L \in \mathcal{B}^{(L)}$ for all $n \in [N]$.

2) Threshold for each index: For each index $m \in [M]$, if there exists a $b^L \in \mathcal{B}_m^{(L)}$ such that

$$\sum_{n\in[N]} \mathbb{1}\{z_n^L = b^L\} \geq T > \max_{\tilde{b}^L \in \mathcal{B}_m^{(L)} \setminus \{b_l^L\}} \sum_{n\in[N]} \mathbb{1}\{z_n^L = \tilde{b}^L\} \tag{2}$$

then the decoder sets $\hat{x}_m^L = b^L$. That is, $\hat{x}_m^L = b^L$ if b^L is a unique molecule in $\mathcal{B}_m^{(L)}$ whose number of appearances in z^{LN} is larger than T. Otherwise $\hat{x}_m^L = \mathsf{e}$, where e is a symbol representing an *erasure*.

3) Codeword decoding: Let

$$j^* = \arg\min_{j\in[|\mathcal{C}|]} \rho(\hat{x}^{LM}, x^{LM}(j)) \tag{3}$$

where (with a slight abuse of notation)

$$\rho(\hat{x}^L, x^L) := \begin{cases} \mathbb{1}\{\hat{x}^L \neq x^L\}, & \hat{x}^L \neq \mathsf{e} \\ 0, & \hat{x}^L = \mathsf{e} \end{cases} \tag{4}$$

and $\rho(\hat{x}^{LM}, x^{LM}) := \sum_{m\in[M]} \rho(\hat{x}_m^L, x_m^L)$, which is a Hamming distance with zero contribution in case of erasures.

The DNA storage channel is thus indexed by M and parameterized by $(\alpha_M, \beta, \{W^{(L)}\}_{L\in\mathbb{N}_+})$. The (storage) rate of the codebook $\mathcal{C}$ is given by $R = \frac{\log|\mathcal{C}|}{ML}$, and the error probability of D given that $x^{LM}(j) \in \mathcal{C}$ was stored is given by

$$\mathsf{pe}(\mathcal{C}, \mathsf{D} \mid x^{LM}(j)) := \mathbb{P}\left[\mathsf{D}(y) \neq j \mid x^{LM}(j)\right]. \tag{5}$$

Let $\psi_M\colon \mathbb{N}_+ \to \mathbb{N}_+$ be a monotonic strictly increasing sequence. An *error exponent* $E(R)$ *w.r.t. scaling* ψ_M is achievable for channel DNA at rate R, if there exists a sequence $\{\mathcal{C}_M, \mathsf{D}_M\}_{M\in\mathbb{N}_+}$ so that the average error probability is bounded as

$$-\log\left[\frac{1}{|\mathcal{C}_M|}\sum_{j\in[|\mathcal{C}_M|]} \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M \mid x^{LM}(j))\right] \geq \psi_M \cdot E(R) - o(\psi_M). \tag{6}$$

In this paper, we will obtain single-letter expressions for error exponents achieved under coded-index codebook and the class of decoders defined above. Throughout, we only make the following assumptions on the sequencing channel: 1) Inner code rate: $R_b := \frac{1}{L}\log|\mathcal{B}^{(L)}| > 1/\beta$. 2) Vanishing inner code (maximal) error probability:

$$\mathsf{pe}_b(\mathcal{B}^{(L)}) := \max_{b^L\in\mathcal{B}^{(L)}} W^{(L)}\left[\mathsf{D}_b(y^L) \neq b^L \mid b^L\right] = o(1). \tag{7}$$

As $|\mathcal{B}_m^{(L)}| = \frac{e^{R_b L}}{M} = \frac{\exp[R_b\beta\log M]}{\exp[\log M]}$, the assumption on R_b assures that $\mathcal{B}_m^{(L)}$ is not empty. The assumption on the error probability assures that the error probability at the first decoding step tends to zero as $L = \beta\log M \to \infty$. Thus, if the sequencing channel $W^{(L)}$ has capacity $C(W^{(L)})$ (with rate normalized to single symbol), then it must hold that $R_b \leq C(W^{(L)})$. For example, for sequencing DMC, the error probability decays as $e^{-E(R_b)\cdot L}$, where $E(R_b)$ is the error exponent. For general sequencing channels, the decay rate could be slower even for optimal codes. Thus, for concreteness, we set $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^\zeta)}$, where $\zeta > 0$, and as we shall see, ζ will not affect the achievable exponent of the DNA storage system. Therefore, even sub-optimal codes can be used, for example, *polar codes*, whose error scales as $e^{-\Theta(\sqrt{N})}$ for standard DMCs [16], and of $e^{-\Theta(N^{1/3})}$ for channel which include insertions, deletions, and substitutions [17].

Our achievable error exponent will be based on the following *coded-index based random coding ensemble*:

Definition 2. Following Definition 1, let $\mathcal{C} = \{X^{LM}(j)\}$ be a random coded such that $X_m^L(j)$ is chosen uniformly at random from $\mathcal{B}_m^{(L)}$ independently for all $m \in [M]$ and all $j \in [|\mathcal{C}|]$.

III. Main Result

Our main result is as follows:

Theorem 3. *Let an inner code $\mathcal{B}^{(L)} \subset \mathcal{X}^L$, and let D_b be a decoder which satisfy the assumptions on the inner code ($R_b > 1/\beta$, $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^\zeta)}$). Then, there exists a sequence of codebooks $\{\mathcal{C}_M\}$ and corresponding threshold-based decoders $\{\mathsf{D}_M\}$ (as described in Sec. II) so that the following holds: If $N/M = \Theta(1)$ then for any $R < (R_b - 1/\beta)(1 - e^{-\frac{N}{M}})$,*

$$-\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq M \cdot d_b\left(1 - \frac{R}{R_b - 1/\beta}\,\middle\|\, e^{-\frac{N}{M}}\right) - O\left(\frac{M}{\log M}\right). \tag{8}$$

If $N/M = \omega(1)$ then for any $R < R_b - 1/\beta$,

$$-\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq \frac{N}{2}\left[1 - \frac{R}{R_b - 1/\beta}\right] - O(M) \tag{9}$$

if $\frac{N}{ML} < 2(R_b - 1/\beta)$, and

$$-\log \mathsf{pe}(\mathcal{C}_M, \mathsf{D}_M) \geq ML\left[R_b - 1/\beta - R\right] - O\left(\frac{N}{\log M}\right) \tag{10}$$

if $2(R_b - 1/\beta) \leq \frac{N}{ML}$.

Discussion:

1) The bound is not continuous in N (that is, there is a phase transition), and a the behavior is different between $N = \Theta(M)$ and $N = \omega(M)$. As stems from the proof, in both regimes, the threshold is chosen as $T \equiv T_M = o(M)$. This follows since the error probability of the inner code is $e^{-\Theta(L^\zeta)} = e^{-\Theta(\log^\zeta M)}$, and so the number of erroneously sequenced molecules is $o(M)$, with an average of less than a single erroneous molecule per index.

2) The result does not depend on ζ, the assumed scaling of the inner code error probability ($\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-\Theta(L^\zeta)}$), and manifests the fact that sampling events dominate the error probability, compared to sequencing error events.

3) For the standard channel coding problem over DMCs with blocklength N, the method of types leads to random coding and expurgated bounds which tend to their asymptotic values up to a $O((\log N)/N)$ term (this can be avoided for Gallager's method [18, Ch. 5], see also [19, Problem 10.33]). Here, it is evident that the decay is much slower, and could be as slow as $O(1/\log M)$. As discussed in [4, Sec. VII] this seems to be an inherent property of this channel.

4) Proving tightness of Theorem 3 is challenging, even for optimal decoders. The main difficulty is in the *Poissonization of the multinomial* effect which is used to upper bound the large-deviations behavior of the number of under-sampled number of molecules in Lemma 4 to follow (as proposed in [1], [12]). This upper bound is tight at the center of the multinomial distribution, but may be loose at its tails. Developing lower bounds on the error probability is thus an open problem.

5) An expurgated bound is also proved in [14], which improves the error probability at the regime $\frac{N}{ML} > 4(R_b - 1/\beta)$.

IV. Main Steps of the Proof

The proof begins by analyzing the probability of channel-related events, and specifically, the event in which some of the molecules are not sampled enough times, or the event of excessive number of sequencing errors. Let the threshold $T \equiv T_\tau := \frac{N}{M}(1-\sqrt{2\tau})$ of the decoder D be parameterized by a parameter $\tau \in (0, 1/2)$. In coded-index based coding, each codeword $x^{LM}(j)$ contains exactly a single molecule from each of the sub-codes $\mathcal{B}_m^{(L)}$, and the molecule $x_m^L(j)$ is sampled S_m times. Let $K_m \in [S_m + 1]$ be the number of copies of the molecule $x_m^L(j)$ that have been erroneously sequenced, let $K := \sum_{m\in[M]} K_m \in [N+1]$ be the total number of molecules which have been erroneously sequenced, and let $V_m \in [K+1]$ be the number of molecules $x_{m'}^L(j)$ for $m' \in [M]\backslash\{m\}$ which have been erroneously sequenced to have index m. Note that $\sum_{m\in[M]} V_m = K$ holds too. The event in which the molecule x_m^L was not decoded correctly in the second stage of the operation of the decoder is included in a union of the following events:

1) $S_m < T_\tau$, that is, the molecule have not been sampled enough times in the sampling stage.

2) $S_m \geq T_\tau$ yet $S_m - K_m < T_\tau$, that is, the molecule have been sampled enough times in the sampling stage step, but K_m sequencing errors have caused the number of appearances of $x_m^L(j)$ to drop below the threshold T.

3) $V_m \geq T_\tau$, that is, there are more than T molecules with index m, which are not the correct molecule $x_m^L(j)$.

On the face of it, the event $V_m \geq T_\tau$ can lead to a crude upper bound, since the V_m molecules which are erroneously mapped to index m are not likely to be the *exact* same molecule in $\mathcal{B}_m^{(L)}$. However, a more precise analysis of this event would require making assumptions on the structure of the sub-codes $\{\mathcal{B}_m^{(L)}\}$, which we avoid here altogether.

Corresponding to these events, we define the following sets:

$$\mathcal{M}_\sigma := \{m \in [M] \colon S_m < T_\tau\} \tag{11}$$

$$\mathcal{M}_\kappa := \{m \in [M] \colon S_m \geq T_\tau,\ S_m - K_m < T_\tau\} \tag{12}$$

$$\mathcal{M}_\nu := \{m \in [M] \colon V_m \geq T_\tau\}, \tag{13}$$

The next lemma addresses the cardinality of $\mathcal{M}_\sigma$:

Lemma 4. *Let $x^{LM}(j)$ be a codeword from a coded-index codebook. Let $\tilde{S} \sim \mathrm{Pois}(N/M)$ and*

$$\varphi_\tau := -\frac{1}{N/M} \log \mathbb{P}\left[\tilde{S} \leq T_\tau\right]. \tag{14}$$

If $N/M = \Theta(1)$ then

$$\mathbb{P}\left[|\mathcal{M}_\sigma| \geq \sigma M \mid x^{LM}(j)\right] \leq 3 \cdot \exp\left[-M \cdot d_b\left(\sigma \,\middle\|\, e^{-\varphi_\tau \frac{N}{M}}\right)\right] \tag{15}$$

for $\sigma \in (e^{-\varphi_\tau \frac{N}{M}}, 1]$. If $N/M = \omega(1)$ then

$$\mathbb{P}\left[|\mathcal{M}_\sigma| \geq \sigma M \mid x^{LM}(j)\right] \leq 4e^{-\sigma\tau N} \tag{16}$$

for $\sigma \in (e^{-\tau \frac{N}{M}}, 1]$.

Proof outline: The empirical count vector S^M follows a multinomial distribution, whose components are dependent. The proof utilizes the *Poissonization* of the multinomial distribution effect [20, Thm. 5.6]: If $\tilde{N} \sim \mathrm{Pois}(\lambda)$ and $\tilde{S}^M \sim \mathrm{Multinomial}(\tilde{N}, (\frac{1}{M}, \ldots, \frac{1}{M}))$ conditioned on $\tilde{N}$, then $\{\tilde{S}_m\}_{m\in[M]}$ are independent and identically distributed (i.i.d.) $\tilde{S}_m \sim \mathrm{Pois}(\frac{\lambda}{M})$ (unconditioned on $\tilde{N}$). Let $A \equiv \sum_{m\in[M]} \mathbb{1}\{S_m < T_\tau\}$ and $\tilde{A} \equiv \sum_{m\in[M]} \mathbb{1}\{\tilde{S}_m < T_\tau\}$. The Poissonization effect is used to prove the upper bound (see also [20, Exercise 5.14])

$$\mathbb{P}\left[A \geq \sigma M\right] \leq 2 \cdot (1 + o(1)) \cdot \mathbb{P}\left[\tilde{A} \geq \sigma M\right], \tag{17}$$

and as $\tilde{A}$ is a sum of i.i.d. random variables $\{\tilde{S}_m\}$, the right probability is then evaluated by a standard Chernoff bound on the binomial distribution. ■

The following lemma is used to bound the total number of sequencing errors K, which, in turn, is used to bound the cardinalities of $\mathcal{M}_\kappa$ and $\mathcal{M}_\nu$:

Lemma 5. *Let K be the total number of erroneously sequenced molecules. Let $\mathcal{U} \subseteq [M]^N$ be a sampling event, and assume that $\mathsf{pe}_b(\mathcal{B}^{(L)}) = e^{-c\cdot L^\zeta}$. Then, $\mathbb{P}[K \geq \kappa N \mid \mathcal{U}] \leq e^{-c\cdot\kappa N L^\zeta}$ for any $\kappa \in (0, 1]$.*

Proof outline: The proof is based on a Chernoff bound over the N independent sequencing operations, for which the

probability of error is at most $e^{-c\cdot L^{\zeta}}$. It requires, however, a more refined argument, since the sequencing errors are not be independent for a given codebook $\mathcal{B}^{(L)}$. ■

The channel/decoder operation is more directly defined by the set of erased molecules and the set of molecules with undetected errors as

$$\mathcal{M}_{\mathsf{e}} := \left\{m \in [M]\colon \hat{x}_m^L = \mathsf{e}\right\}, \tag{18}$$

$$\mathcal{M}_{\mathsf{u}} := \left\{m \in [M]\colon \hat{x}_m^L \neq \mathsf{e},\ \hat{x}_m^L \neq x_m^L(j)\right\}. \tag{19}$$

Lemmas 4, and 5 are utilized to analyze the cardinality of $\mathcal{M}_{\mathsf{e}}$ and $\mathcal{M}_{\mathsf{u}}$. As it turns out, the dominating event is $\mathbb{P}[|\mathcal{M}_\sigma| \geq \sigma M]$, to wit, the probability that the molecules have not been amplified enough times, which is on the exponential order of N, compared to the probability evaluated in Lemma 5 which are on the exponential order of $LN = N\beta \log M$.

Lemma 6. *Consider a decoder* D *for a coded-index based codebook. For the erasure set* $\mathcal{M}_{\mathsf{e}}$*: If* $N/M = \Theta(1)$ *then*

$$-\log \mathbb{P}\left[|\mathcal{M}_{\mathsf{e}}| \geq \theta M\right] \geq M d_b\left(\theta || e^{-\varphi_\tau \frac{N}{M}}\right) + o(M) \tag{20}$$

for all $\theta \in (e^{-\varphi_\tau \frac{N}{M}}, 1]$*. If* $N/M = \omega(1)$ *then*

$$-\log \mathbb{P}\left[|\mathcal{M}_{\mathsf{e}}| \geq \theta M\right] \geq \theta \tau N \cdot [1 + o(1)] \tag{21}$$

for all $\theta \in (e^{-\tau \frac{N}{M}}, 1]$*. For the undetected error set* $\mathcal{M}_{\mathsf{u}}$*:*

$$-\log \mathbb{P}\left[|\mathcal{M}_{\mathsf{u}}| \geq \theta M\right] \geq c \cdot (1 - \sqrt{2\tau})\theta N L^{\zeta}. \tag{22}$$

Proof outline: By deriving relations between K and $|\mathcal{M}_\kappa|, |\mathcal{M}_\nu|$, and then between these sets and $|\mathcal{M}_\sigma|$, to $|\mathcal{M}_{\mathsf{e}}|$ and $|\mathcal{M}_{\mathsf{u}}|$, and utilizing Lemmas 4 and 5. ■

Thus, as apparent from Lemma 6, and as discussed in the introduction, for coded-index based codebooks, the type of decoders, and the analysis in this paper, the effect of sequencing errors is much less profound compared to erasures.

The random coding analysis is based on the following lemma, which bounds the probability that an erroneous codeword will be decoded, conditioned on a given number of channel erasures and undetected errors.

Lemma 7. *Let* $\mathcal{C}$ *be drawn from the coded-index based random coding ensemble. Let* $X^{LM}(0) = x^{LM}(0)$ *be arbitrary, and let* $\hat{X}^{LM}$ *be the output of the decoder conditioned on the input* $x^{LM}(0)$*. Then, for* $\theta_{\mathsf{e}}, \theta_{\mathsf{u}} \in \frac{1}{M}[M+1]$ *such that* $\theta_{\mathsf{e}} + \theta_{\mathsf{u}} \leq 1$ *and any* $j \in [|\mathcal{C}|]\backslash\{0\}$ *it holds that*

$$-\frac{1}{M}\log \mathbb{P}\Bigg[\rho(\hat{X}^{LM}, X^{LM}(j)) \leq \rho(\hat{X}^{LM}, x^{LM}(0)) \Bigg| |\mathcal{M}_{\mathsf{e}}| = \theta_{\mathsf{e}} M,\ |\mathcal{M}_{\mathsf{u}}| = \theta_{\mathsf{u}} M\Bigg] \geq (R_b\beta - 1)(1 - \theta_{\mathsf{e}} - \theta_{\mathsf{u}})\log M - \Theta(1). \tag{23}$$

Proof outline: The proof is based on an argument which counts the relative number of competing codewords $\tilde{x}^{LM}$ in the coded-index based ensemble which have distance $\rho(\hat{X}^{LM}, \tilde{x}^{LM})$ smaller than $\rho(\hat{X}^{LM}, x^{LM}(0))$, followed by an analysis of its asymptotic behavior with M. ■

The proof of Theorem 3 then follows from Lemma 7, by conditioning over $\theta_{\mathsf{e}}, \theta_{\mathsf{u}}$, taking a clipped union bound over the probability that one of the $\lceil e^{MLR}\rceil - 1$ competing codewords causes an error, averaging over $\theta_{\mathsf{e}}, \theta_{\mathsf{u}}$ via Lemma 6, and analyzing the asymptotic behavior of the resulting expressions for the different regimes of $\alpha_M = N/M$.

References

[1] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3675–3689, 2021.

[2] A. Lenz, L. Welter, and S. Puchinger, "Achievable rates of concatenated codes in DNA storage under substitution errors," in *International Symposium on Information Theory and Its Applications*, pp. 269–273, IEEE, 2020.

[3] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *IEEE Information Theory Workshop*, pp. 1–5, IEEE, 2019.

[4] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability bounds," 2021. Available at https://arxiv.org/pdf/2109.12549.pdf.

[5] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[6] K. A. S. Immink and K. Cai, "Design of capacity-approaching constrained codes for DNA-based storage systems," *IEEE Communications Letters*, vol. 22, no. 2, pp. 224–227, 2017.

[7] Y. Wang, M. Noor-A-Rahim, E. Gunawan, Y. L. Guan, and C. L. Poh, "Construction of bio-constrained code for DNA data storage," *IEEE Communications Letters*, vol. 23, no. 6, pp. 963–966, 2019.

[8] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. S., "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[9] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific reports*, vol. 5, no. 1, pp. 1–10, 2015.

[10] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.

[11] L. Organick, S. D. Ang, Y. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, and B. Nguyen, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[12] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the DNA storage channel," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8846–8850, IEEE, 2020.

[13] L. C. Meiser, P. L. Antkowiak, W. D. Koch, J.and Chen, A. X. Kohll, W. J. Stark, R. Heckel, and R. Grass, "Reading and writing digital data in DNA," *Nature protocols*, vol. 15, no. 1, pp. 86–101, 2020.

[14] N. Weinberger, "Error probability bounds for coded-index DNA storage channels," 2021. Available at https://drive.google.com/file/d/1tuEGj4852slCPvNq6xteTgZz8TgCx1ME/view?usp=sharing.

[15] M. Kovačević and V. Y. F. Tan, "Codes in the space of multisets – coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, 2018.

[16] S. H. Hassani, R. Mori, T. Tanaka, and R. L. Urbanke, "Rate-dependent analysis of the asymptotic behavior of channel polarization," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2267–2276, 2012.

[17] I. Tal, H. D. Pfister, A. Fazeli, and A. Vardy, "Polar codes for the deletion channel: Weak and strong polarization," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 1362–1366, IEEE, 2019.

[18] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley and Sons, 1968.

[19] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge University Press, 2011.

[20] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.