DNA Storage in the Short Molecule Regime

Ran Tamir, *Member, IEEE*, Nir Weinberger, *Senior Member, IEEE*, and Albert Guillén i Fàbregas, *Fellow, IEEE*

***Abstract*—We study the amount of information that can be reliably stored in a DNA-based storage system composed of short DNA molecules. In this short molecule regime, Shomorony and Heckel (2022) put forward a conjecture on the scaling of the number of information bits that can be reliably stored. In this paper, we complete the proof of this conjecture. We analyze a random-coding scheme in which each codeword is obtained by quantizing a randomly chosen point from the probability simplex. By analyzing the optimal maximum-likelihood decoder, we derive an achievability bound that matches a recently established converse bound across the entire short-molecule regime. We also propose an alternative coding scheme, which operates at a significantly lower computational complexity, yet achieves the optimal scaling, except for a specific range of very short molecules.**

***Index Terms*—Data storage, DNA storage, molecular communication, permutation channel.**

I. Introduction

THE storage of information in DNA molecules offers extremely high information density[1] and longevity, and can address the ever-growing demand for digital storage. Several working prototypes and system proposals [1], [2], [3], [4], [5], [6], [7] have sparked a surge of information-theoretic and coding-theoretic research, including coding methods [8], channel capacity and error probability analysis [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], machine-learning based systems [29], [30], [31], [32], secrecy [33], [34], [35], and many more.

In this paper, we consider the commonly adopted DNA storage channel model, known as the *shuffling-sampling channel* [16]. In this channel, information is encoded as a codeword comprised of M molecules, where each molecule is of length L symbols drawn from an alphabet $\mathcal{A}$ (a natural choice is $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ representing the four DNA bases, however, other alphabets are also possible). The length of the molecules is assumed to grow with M and is parametrized as[2] $L = \beta \log M$ for some parameter $\beta > 0$. The M molecules are stored in a pool, without preservation of order. The retrieval of information is performed in two consecutive steps. First, one of the M molecules is chosen uniformly at random from the pool, with replacement. Second, the chosen molecule is sequenced, that is, the sequence of L nucleotides from which it is composed is reconstructed to obtain a read. These two steps are repeated for K times, where, typically, $K > M$. While sequencing is noisy in practice, in this work we consider the idealized case of noiseless sequencing. Even under this assumption, the list of K output reads is still random due to the sampling operation, since some molecules may be sampled multiple times, while others may not be sampled at all.

The length of the molecules is parametrized by $\beta > 0$. The parameter β affects channel capacity, as the effects of both the lack of order of molecules and non-ideal sampling are less severe as β increases. In the case that the molecule length parameter is large enough, specifically $\beta > \frac{1}{\log|\mathcal{A}|}$, a simple scheme achieves the channel capacity $C = \max\left\{\log|\mathcal{A}| - \frac{1}{\beta}, 0\right\}$ [15]. Start each molecule with a header of length $\log_{|\mathcal{A}|} M = \frac{L}{\beta \log|\mathcal{A}|}$ symbols from $\mathcal{A}$,[3] identifying the index of the molecule in $\{1, 2, \ldots, M\}$. The rest of the $L\left(1 - \frac{1}{\beta \log|\mathcal{A}|}\right)$ symbols are then used to encode the data. The resulting coding rate (or *information density*), i.e., total number of encoded bits divided by the total number of nucleotides used ML, is then given by $1 - \frac{1}{\beta \log|\mathcal{A}|}$, which can be proved to be the capacity. This scheme clearly breaks if the molecule length is shorter, i.e. $0 < \beta < \frac{1}{\log|\mathcal{A}|}$, because then each molecule is too short to encode even its index. Since the decoder can

Received 2 December 2025; revised 29 August 2026; accepted 27 September 2026. Date of publication 7 October 2026; date of current version 24 November 2026. The work of Ran Tamir and Albert Guillén i Fàbregas was supported in part by European Research Council under Grant 101142747 and Grant 101158232; and in part by Spanish Government under Grant PID2020-116683GB-C22, Grant PID2021-128373OB-I00, and Grant PID2025-167766OA-C22. The work of Nir Weinberger was supported in part by U.S.–Israel Binational Science Foundation (NSF-BSF) under Grant 2024763 and in part by Israel Science Foundation (ISF) under Grant 2032/26. An earlier version of this paper was presented in part at the 2026 International Zurich Seminar on Information and Communication and in part at the 2026 IEEE Information Theory Workshop (ITW) [DOI: 10.3929/ethz-c-000796719]. *(Corresponding author: Ran Tamir.)*

Ran Tamir is with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain (e-mail: ran.tamir@upc.edu).

Nir Weinberger is with the Department of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa 3200003, Israel (e-mail: nirwein@technion.ac.il).

Albert Guillén i Fàbregas is with the Department of Engineering, University of Cambridge, CB2 1PZ Cambridge, U.K., and also with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain (e-mail: guillen@ieee.org).

Communicated by M. Kovačević, Associate Editor for Coding and Decoding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2026.3741379.

Digital Object Identifier 10.1109/TIT.2026.3741379

[1] In information theory, the expression "information density" commonly refers to the random variable whose expectation is mutual information; In the current context, this expression should be understood as the amount of information bits per gram of DNA.

[2] Throughout the paper, logarithms are in the natural base, although we measure information in "bits" instead of "nats".

[3] For simplicity, here we ignore integer constraints.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

always ignore some nucleotides, it follows that the capacity is monotonically non-decreasing in β, thus the capacity equals *zero* for any $0 < \beta < \frac{1}{\log|\mathcal{A}|}$.

Nonetheless, the regime of $0 < \beta < \frac{1}{\log|\mathcal{A}|}$, which is called the *short molecule regime*, is still of interest. In this regime, the total number of different types of molecules $|\mathcal{A}|^L = |\mathcal{A}|^{\beta \log M} = M^{\beta \log|\mathcal{A}|}$ is less than the total number of molecules M. The pigeonhole principle then implies that each codeword must contain repeated molecules. Consequently, the information is encoded into a frequency vector, which contains the relative count of each of the $M^{\beta \log|\mathcal{A}|}$ types of molecules in the pool of M molecules. During reading, the sampling process yields a noisy version of this vector; for example, molecule types that appear once in the codeword may be sampled multiple times or not at all. The decoder then finds the codeword whose frequency vector is closest, in a manner to be made precise, to the frequency vector defined by the output reads.

As noted, the channel capacity of the shuffling-sampling channel is zero in the short molecule regime $\beta < \frac{1}{\log|\mathcal{A}|}$. This implies that the total number of reliably stored bits scales at most *sub-linearly* with the total number of nucleotides ML. However, just a few grams of DNA contain an enormous amount of nucleotides ML, and so for a given M and L, the potential total number of reliably stored bits may still be substantial. This observation motivated an analysis of this regime in [16, Sec. 7.3], leading to a conjecture on the maximal log-cardinality of a reliable codebook as a function of ML. Specifically, [16, Conjecture 4] states that for $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$ this log-cardinality is asymptotically

$$\frac{1-\beta\log|\mathcal{A}|}{2} \cdot M^{\beta\log|\mathcal{A}|}\log M. \tag{1}$$

Evidently, this total number of bits (assuming that the logarithm is in the binary base) is $o(ML) = o(M\log M)$, which increases with M. In [16, Sec. 7.3], a Poisson sampling model was considered. In this model, the total number of output reads is not fixed in advance, and the random number of times that each type of molecule appears in the output reads follows a Poisson distribution, whose parameter is given by the number of times that this type of molecule appears in the input codeword. This may be compared to the original sampling model, whose sampling operation can be described by a multinomial random variable. The conjecture in (1) is based on relating the frequency-based channel to a power-constrained Poisson channel, for which the asymptotic scaling of its capacity, as a function of the input power, is known [36]. However, (1) remained a conjecture since the Poisson channel obtained from the reduction is non-standard: Its power constraint increases with the blocklength (amounts to $M^{\beta\log|\mathcal{A}|}$), and its inputs are restricted to the *integers*.

In [37] a rigorous approach to this conjecture was taken, based on the original multinomial sampling model, rather than the Poisson sampling model. A converse result, based on a Poissonization of the multinomial (e.g., [38, Ch. 5]) showed that the log-cardinality cannot exceed (1), up to an additive $o\left(\frac{1}{\log M}\right)$ term. On the other hand, an achievability result showed that (1) can be achieved, however, under the additional condition that $\beta \in \left(\frac{1}{2\log|\mathcal{A}|}, \frac{1}{\log|\mathcal{A}|}\right)$, that is, for molecules that are not very short. The proof of achievability in [37] is based on Feinstein's maximal coding bound [39] [40, Thm. 20.7], and is relatively technical. It is based on several steps, aiming to rigorously address the reduction of the multinomial channel to a Poisson channel, the integer constraints on its input, and other constraints which stem from these reductions. This did not allow to prove the conjecture in the range $\beta \in \left(0, \frac{1}{2\log|\mathcal{A}|}\right)$.

Our first contribution in this paper is to rigorously establish [16, Conjecture 4] throughout the entire short-molecule regime $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$. We achieve this through a direct and conceptually distinct proof technique. We conduct a random coding analysis, in which codewords are drawn by randomly choosing points in the probability simplex, according to a Dirichlet distribution, and then rounding them to integer count vectors. Directly analyzing the average error probability of this ensemble leads to an achievable bound on the log-cardinality that matches (1).

As is standard in information theory, the random-coding proof is non-constructive, and the resulting decoder complexity is proportional to the total number of codewords in the codebook. Our second contribution is a low-complexity encoding-decoding method termed *partition coding*. In partition coding, the set of $M^{\beta\log|\mathcal{A}|}$ different types of molecule is partitioned into *ordered* subsets of equal sizes. In each of those subsets, all types of molecules have an identical number of copies stored in the molecular pool. The copy numbers governing the different subsets are determined by some fixed monotonic frequency vector. Thus, the codebook construction is made in a deterministic way. Moreover, the decoder operation is reduced to *sorting* the frequency vector of the output reads, which can be achieved in a computational complexity of $\Theta(M^{\beta\log|\mathcal{A}|}\log M)$ using standard sorting algorithms. This simple method is capable of asymptotically achieving (1) for any $\frac{1}{3\log|\mathcal{A}|} < \beta < \frac{1}{\log|\mathcal{A}|}$.

The works most directly related to this study are as follows. In [41], which was also motivated by the short-molecule regime with Poisson sampling, the capacity of Poisson channels with integer (lattice) inputs was considered. In [42], we studied the short-molecule regime without enforcing integer input constraints and derived random-coding bounds on the error probability. This setup is motivated by the fact that the actual synthesizing costs related to the DNA molecules storing the message are proportional to the number of *different* molecules that have to be synthesized. The reason for this is that once a molecule is synthesized, the costs of duplicating it are relatively low. It then follows that any arbitrary molecule frequency vector can be accurately approximated. In [42], we also briefly discussed the connection of the short molecule regime to composite DNA storage [31], [35], [43], [44], [45], [46], [47], and the *permutation channel* [48], [49], [50], [51], [52]. The storage system based on the short-molecule regime with a *noisy sequencing* was recently studied in [53] and [54].

The remainder of the paper is organized as follows. In Section II we establish notation conventions, formulate problem settings, and define the objective of this work. In Section III we introduce and discuss the main result of this work and compare it to the main results of [37]. In Section IV we introduce

partition coding along with its main theoretical guarantees. The proofs of the main results are provided in Sections V and VI, and in Section VII we conclude the article with a summary and future research directions.

II. Notation and Problem Formulation

A. Notation

For a positive integer n, the set $\{1,2,\ldots,n\}$ is denoted by $[n]$. The probability of an event $\mathcal{A}$ is denoted by $\mathbb{P}[\mathcal{A}]$. The expectation of a random variable X is denoted by $\mathbb{E}[X]$. The indicator function of an event $\mathcal{A}$ is denoted by $\mathbb{1}[\mathcal{A}]$. The cardinality of a finite set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. The *floor* function of a real number x, denoted by $\lfloor x \rfloor$, is defined as $\lfloor x \rfloor = \max\{y \in \mathbb{Z} : \ y \le x\}$. The $(n-1)$-dimensional probability simplex, denoted by $\mathcal{P}_n$, is defined as

$$\mathcal{P}_n = \left\{(x_1,\ldots,x_n) \in [0,1]^n : \ \sum_{i=1}^{n} x_i = 1\right\}. \tag{2}$$

The Kullback–Leibler (KL) divergence between two probability mass functions (PMFs) $\{P(x)\}_{x\in\mathcal{X}}$ and $\{Q(x)\}_{x\in\mathcal{X}}$ is defined by

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}, \tag{3}$$

with the common conventions that $0\log\frac{0}{q} = 0$ for any $q \ge 0$ and $p\log\frac{p}{0} = +\infty$ for any $p > 0$, which implies that $D(P\|Q) = +\infty$ when P is not absolutely continuous with respect to Q.

The gamma function is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t. \tag{4}$$

The Dirichlet distribution of order $n \ge 2$ with positive parameters $\alpha_1,\ldots,\alpha_n$ has a probability density function with respect to Lebesgue measure on the Euclidean space $\mathbb{R}^{n-1}$ given by

$$f(x_1,\ldots,x_n) = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n x_i^{\alpha_i - 1}, \tag{5}$$

for any $(x_1,\ldots,x_n) \in \mathcal{P}_n$ and zero otherwise.

B. Problem Formulation

Let $\mathcal{C}_M$ be a codebook for data storage in a system that relies on short molecules. Each codeword in $\mathcal{C}_M$ is composed of at most M molecules. Observe that different codewords may be of different sizes. However, we impose a uniform upper bound on codeword sizes because the cost of the input is related to the number of molecules synthesized. More specifically, for any $m \in \{1,2,\ldots,|\mathcal{C}_M|\}$, the codeword $\boldsymbol{x}(m)$ is given by a multiset[4] of sequences of the form

$$(\boldsymbol{x}_1^L(m), \boldsymbol{x}_2^L(m), \ldots, \boldsymbol{x}_{J(m)}^L(m)), \tag{6}$$

where $J(m) \le M$ and for every $i \in [J(m)]$, $\boldsymbol{x}_i^L \in \mathcal{A}^L$. In the short-molecule regime, we assume that for some[5] $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$

$$L = \beta \log M, \tag{7}$$

and then the cardinality of $\mathcal{A}^L$ is given by

$$|\mathcal{A}^L| = |\mathcal{A}|^{\beta\log M} = M^{\beta\log|\mathcal{A}|}. \tag{8}$$

It is assumed that the message m is drawn with a uniform distribution from the set $\{1,2,\ldots,|\mathcal{C}_M|\}$ and that all the molecules that make up the codeword $\boldsymbol{x}(m)$ are grouped with no preservation of order. When the message is restored, we assume that exactly K sequences $\boldsymbol{y} = (\boldsymbol{y}_1^L, \boldsymbol{y}_2^L, \ldots, \boldsymbol{y}_K^L)$ are independently sampled (with replacement) from the DNA pool. We assume that the *coverage depth* $\xi = \frac{K}{M}$ is fixed.

Based on the sampled sequences, the decoder estimates the message as $\hat{m}(\boldsymbol{y})$. The probability of error of any decoder is given by

$$\varepsilon_M = \mathbb{P}\left[\hat{m}(\boldsymbol{Y}) \ne m\right], \tag{9}$$

which is taken with respect to the randomness of the message selection, the (possibly) random codebook generation, and the sampling process.

Our main objective in this work is to resolve the direct part of [16, Conjecture 4], which states that for $\mathcal{A} = \{0,1\}$, there exists a sequence of codes $\{\mathcal{C}_M\}_{M\ge1}$ with $\xi = 1$ and a vanishing error probability, for which

$$\limsup_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^\beta L} = \frac{1-\beta}{2\beta}. \tag{10}$$

Note that by substituting $L = \beta\log M$, (10) is also equivalent to

$$\limsup_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^\beta \log M} = \frac{1-\beta}{2}. \tag{11}$$

Although [16, Conjecture 4] was stated for the special cases $\mathcal{A} = \{0,1\}$ and $\xi = 1$, in this work we address the general case of an arbitrary $\mathcal{A}$ (with $|\mathcal{A}| < \infty$) and $\xi > 0$.

III. Random Coding

Our first encoding-decoding scheme is close in spirit to the common methodology adopted when studying ordinary channel coding. In ordinary channel coding, a channel input vector is transmitted over a (discrete memoryless) channel, which produces a channel output vector. In order to study the channel capacity or various properties of the error probability, it is customary to generate a random codebook since the structure of the optimal code is unknown. For example, in i.i.d. random coding, codewords are drawn independently from the input space using a product input distribution.

In our current model, each codeword is composed of M short DNA molecules, and the system designer is allowed to choose how many copies to take from each possible string in $\mathcal{A}^L$. In other words, each codeword is equivalent to an empirical PMF over $M^{\beta\log|\mathcal{A}|}$ entries; hence, generating a codebook means choosing a relatively large number of empirical PMFs.

[4] We use the term "multiset" since we would like to allow for repetitions.

[5] In general, we ignore integer constraints on L. If one still insists on an integer L, then a possible rigorous approach is to consider, for a given M, only those values of $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$ for which $\beta\log M \in \mathbb{N}$.

While such a codebook may be chosen deterministically, it turns out that the random coding methodology may be adapted to the case at hand: Instead of drawing vectors from $\mathcal{X}^n$, one can draw PMFs from the probability simplex and quantize the given realizations to ensure that the resulting PMF corresponds to an empirical PMF.

We now describe the encoding–decoding scheme in detail. Let us define $\mathcal{A}^L = \{\boldsymbol{a}(1), \ldots, \boldsymbol{a}(n)\}$ with $n = M^{\beta \log |\mathcal{A}|}$. Each codeword in $\mathcal{C}_M$ is generated in the following procedure. For the message m, a random PMF $\boldsymbol{P}_m = (P_m(1), \ldots, P_m(n))$ is independently drawn from the $(n-1)$-dimensional simplex $\mathcal{P}_n$ according to the Dirichlet distribution with vector parameters $\boldsymbol{\alpha} = (1, \ldots, 1)$,[6] which is equivalent to the uniform measure over $\mathcal{P}_n$.

In order to turn $\boldsymbol{P}_m$ into an empirical PMF $\hat{\boldsymbol{P}}_m$, we choose the mth codeword to be composed of $\lfloor MP_m(\ell) \rfloor$ copies of the string $\boldsymbol{a}(\ell)$, for each $\ell \in [n]$. The mth codeword is also represented by the empirical probability vector $\hat{\boldsymbol{P}}_m = (\hat{P}_m(1), \ldots, \hat{P}_m(n))$, where for any $\ell \in [n]$,

$$\hat{P}_m(\ell) = \frac{\lfloor MP_m(\ell) \rfloor}{\sum_{k=1}^{n} \lfloor MP_m(k) \rfloor}. \tag{12}$$

After sampling, the decoder observes $\boldsymbol{y} = (\boldsymbol{y}_1^L, \boldsymbol{y}_2^L, \ldots, \boldsymbol{y}_K^L)$ and first calculates the frequency vector

$$\hat{\boldsymbol{Q}}_{\boldsymbol{y}} = (\hat{Q}_{\boldsymbol{y}}(1), \ldots, \hat{Q}_{\boldsymbol{y}}(n)), \tag{13}$$

where for any $\ell \in [n]$,

$$\hat{Q}_{\boldsymbol{y}}(\ell) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}[\boldsymbol{y}_i^L = \boldsymbol{a}(\ell)]. \tag{14}$$

It can be seen [42, Sec. II.B] that the maximum likelihood decoder is equivalent to a decoder that estimates the message as the one whose codeword minimizes the KL divergence with $\hat{\boldsymbol{Q}}_{\boldsymbol{y}}$. To this end, the decoder estimates the transmitted message according to

$$\hat{m}(\boldsymbol{y}) = \underset{m \in [|\mathcal{C}_M|]}{\arg\min}\, D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{P}}_m). \tag{15}$$

In Section V we prove the following result.

Theorem 1: Consider a noiseless shuffling-sampling channel with $\beta \in \left(0, \frac{1}{\log |\mathcal{A}|}\right)$ and a coverage depth $\xi > 0$. There exists a sequence of codes $\{\mathcal{C}_M\}_{M \geq 1}$ with vanishing error probabilities ($\varepsilon_M \to 0$), such that

$$\lim_{M \to \infty} \frac{\log |\mathcal{C}_M|}{M^{\beta \log |\mathcal{A}|} \log(M)} = \frac{1 - \beta \log |\mathcal{A}|}{2}. \tag{16}$$

Although we have considered a general coverage depth $\xi > 0$, the asymptotic log-cardinality of the largest storage codebook is independent of ξ. While the optimal information density is independent of ξ, the error probability converges faster for larger values of ξ, as was also observed in the recent studies [14], [18], [19].

The idea of generating PMF codewords using the Dirichlet distribution has been borrowed from [42], but with one important modification: While the channel model in [42] was assumed to be with infinite input-resolution, this can no longer be assumed in the current work, because each codeword assigns an integer number of copies to each molecule type. Hence, the channel in our model is restricted to a finite input resolution. To satisfy this restriction, we first randomly draw independent PMF codewords from the probability simplex as in [42], and then, in a second step, we generate a quantized version of each PMF codeword, as in (12). As a result of this quantization operation, the number of DNA molecules composing each codeword is not fixed and lies in the range $\{M - n, \ldots, M\}$; on the one hand,

$$J(m) = \sum_{\ell=1}^{n} \lfloor MP_m(\ell) \rfloor \leq \sum_{\ell=1}^{n} MP_m(\ell) = M, \tag{17}$$

while on the other hand,

$$J(m) = \sum_{\ell=1}^{n} \lfloor MP_m(\ell) \rfloor \geq \sum_{\ell=1}^{n} (MP_m(\ell) - 1) = M - n. \tag{18}$$

Since $n = M^{\beta \log |\mathcal{A}|} \ll M$, all the codewords have roughly the same size. While several other coding schemes may be implemented to yield a codebook with a fixed number of DNA molecules in each codeword (like the one that will be implemented later in Section IV or the random coding-based schemes that will be described in Section VII), they are either provably suboptimal or hard to analyze. Hence, we choose to analyze the above mentioned quantization method, since it is easier to analyze and yields optimal results.

We briefly explain why the Dirichlet distribution was chosen to draw the PMF codewords. At the beginning of the proof of Theorem 1, when handling the pairwise error probability $\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{P}}) \leq D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}}_0)\right]$, where $\hat{\boldsymbol{p}}_0$ is the true codeword, $\hat{\boldsymbol{P}}$ is a competing codeword, and $\hat{\boldsymbol{Q}}_{\boldsymbol{y}}$ is the empirical distribution of the vector of samples, we upper-bound this probability using Markov's inequality, and then, in the next step, we need to handle an expectation of the form

$$\mathbb{E}\left[\prod_{i=1}^{n} \hat{P}(i)^{\theta \hat{Q}_{\boldsymbol{y}}(i)}\right], \tag{19}$$

where $\theta > 0$ is a parameter. While the expectation in (19) is not easy to solve, one can tightly upper-bound it by a constant times the expectation

$$\mathbb{E}\left[\prod_{i=1}^{n} P(i)^{\theta \hat{Q}_{\boldsymbol{y}}(i)}\right], \tag{20}$$

where $\boldsymbol{P} = (P(1), \ldots, P(n))$ is the original, unquantized PMF drawn from the Dirichlet distribution. Unlike products of independent random variables, whose expected value can be calculated by pulling the multiplication operation outside, this is no longer the case with products of dependent random variables. Although the various components of the vector $\boldsymbol{P}$ are statistically dependent, it turns out that the product moment in (20) can be precisely evaluated when $\boldsymbol{P}$ follows a Dirichlet distribution. The following result concerning product moments of the Dirichlet distribution can be found, e.g., in [55, p. 274].

[6] A simple mechanism to generate such a random PMF $(P(1), \ldots, P(n))$ is as follows: Draw n independent random variables $X_1, \ldots, X_n$ from the exponential distribution with parameter 1 and then set $P(i) = \frac{X_i}{\sum_{j=1}^{n} X_j}$ for any $i \in [n]$.

Proposition 1: Let $(\alpha_1,\ldots,\alpha_n) \in (0,\infty)^n$, $(\beta_1,\ldots,\beta_n) \in [0,\infty)^n$, and let $(X_1,\ldots,X_n) \sim \mathrm{Dir}(\alpha_1,\ldots,\alpha_n)$. Then, it holds that

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i^{\beta_i}\right] = \frac{\Gamma\left(\sum_{i=1}^{n}\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{n}(\alpha_i+\beta_i)\right)} \cdot \prod_{i=1}^{n} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)}. \tag{21}$$

As can be seen, the expectation in (20) can thus be evaluated exactly using the result of Proposition 1. This allows us to overcome a major difficulty and is the key to the analysis of the PMF quantization technique as defined in (12).

Two remarks with respect to practical DNA storage settings are now in order. First, it appears that the scenario considered, where codewords are represented via the frequencies of different molecule types, could be vulnerable to interference from other codewords in the same DNA pool, as might occur when multiple files are stored together in the same chemical solution. The easiest way to overcome this potential difficulty is to use disjoint sets of molecule types for each stored file. This way, each file is encoded using a codeword composed by unique molecule types, thus distinct codewords cannot interfere with each other. We conjecture that more complicated coding schemes exist, where distinct files may be encoded using shared molecule types. Second, the multinomial distribution is assumed for molecule sampling, which is simplified since it does not account for the possibility that certain molecules could be biased to have more samples than others. Specifically, since we assume an asymptotic regime where $L = \beta \log M$ and $M \to \infty$, then the number of molecule types grows without bound, and it is even not trivial to consistently define a model with sample biases between distinct molecule types. Finding a chemically suitable model which incorporates sampling biases between molecule types may be quite challenging and is left as an important avenue for future research. Generally speaking, a model which takes into account such sampling biases is more challenging to analyze than an unbiased model. It seems that such sampling biases cannot be addressed using the methods in the current paper and would likely require substantially different analysis techniques.

A. Comparison With [37]

The work in [37] considered a more general frequency-based channel, derived tight lower and upper bounds on its capacity, and then specialized these results to the DNA storage channel in the short molecule regime. To understand the differences between [37] and the current work, we first describe the system model in [37] and briefly outline its main proof techniques. In [37], a set of n distinguishable types of objects is considered. An input message is encoded as a pool of unordered objects from the various types. Thus, the channel input is represented by the count vector $\boldsymbol{x}^n = (x_1,\ldots,x_n) \in \mathbb{N}^n$, where x_i is the number of objects of the ith type in the pool of objects. It is assumed that $\sum_{i=1}^{n} x_i = ng_n$ for all possible messages for some given sequence g_n. The channel model in [37] is the same shuffling-sampling channel considered in the current work; to read the message, nr_n samples are taken, where for each $i \in [nr_n]$, an object is chosen uniformly at random from the multiset of ng_n objects in the pool, with replacement. In other words, the channel output is distributed as a multinomial random variable, with a number of trials given by nr_n and probabilities proportional to the channel input count vector $\boldsymbol{x}^n$.

A code in [37] is defined as a set of $|\mathcal{C}_M|$ input count vectors $\mathcal{C}_M = \{\boldsymbol{x}^n(1),\ldots,\boldsymbol{x}^n(|\mathcal{C}_M|)\}$ for which $\sum_{i=1}^{n} x_i(m) = ng_n$ is constant for all $m \in [|\mathcal{C}_M|]$. The size of the largest code for n object types, normalized total count of input objects g_n, normalized number of sampled objects r_n, under a given error probability $\varepsilon_n \in (0,1)$ is denoted by $|\mathcal{C}_M|^\star(n|\varepsilon_n, g_n, r_n)$. The main results in [37] are tight upper and lower bounds on the rate defined by $\frac{1}{n}\log|\mathcal{C}_M|^\star(n|\varepsilon_n, g_n, r_n)$, where $\varepsilon_n \to 0$ as $n \to \infty$, possibly at an arbitrarily slow rate.

In order to specialize this general frequency-based channel into the DNA-based storage channel with short molecules, the following choices were made: For the DNA storage channel, the number of unique objects is the number of unique molecules of length $L = \beta \log M$, given by $n \equiv |\mathcal{A}|^L = M^{\beta \log|\mathcal{A}|}$, and the total number of objects is $ng_n \equiv M$, which is also the number of sampled objects. The achievability bound in [37], yields[7]

$$\frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}L} \geq \frac{1-\beta\log|\mathcal{A}|}{2\beta} - \frac{2.773}{2\beta}\cdot\frac{1}{\log M} + o\left(\frac{1}{\log M}\right), \tag{22}$$

which agrees with Theorem 1 when $M \to \infty$ for an arbitrary $\xi > 0$. Since the bound for the general frequency-based channel holds under the condition $n = \Omega(g_n^{1+\zeta})$ for some $\zeta > 0$, the bound in (22) holds as long as $\beta \in \left(\frac{1}{2\log|\mathcal{A}|}, \frac{1}{\log|\mathcal{A}|}\right)$, while Theorem 1 holds for any $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$, including the very short molecule regime. Hence, in contrast to the achievability result in [37], the result in the current paper holds also for very short molecules.

Furthermore, the converse bound in [37] gives

$$\frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}L} \leq \frac{1-\beta\log|\mathcal{A}|}{2\beta} + o\left(\frac{1}{\log M}\right), \tag{23}$$

which holds for any $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$, and thus, when combined with Theorem 1, proves the conjecture for the scaling of the log-cardinality of the largest codebook in a DNA-based storage system with short molecules.

A few words regarding the proof techniques of the achievability bound in [37] are now in order. The proof of the achievability bound in [37] is based on Feinstein's maximal coding bound [39], which bounds the maximal error probability of the optimal codebook of a given cardinality via the cumulative distribution function of the information density random variable [56]. Specifically, the authors of [37] use the extended version stated in [40, Theorem 20.7], which also takes into account input constraints. Since the frequency-based channel from X^n to Y^n is a multinomial channel, which is not memoryless, a direct analysis of the information spectrum is challenging. To bypass this difficulty, known results (e.g., [38, Corollary 5.9]) are used to relate the multinomial distribution to a memoryless Poisson distribution. However, the analysis of the resulting Poisson channel, along with integer-input

[7]For two positive sequences f_n and g_n, we denote $f_n = o(g_n)$ whenever $\lim_{n\to\infty} f_n/g_n = 0$.

constraints requires additional technical steps, which makes the proof fairly technical and relatively long.

In contrast, Theorem 1 is proved via a direct route; the pairwise error probability of the optimal maximum likelihood decoder is upper-bounded using Markov's inequality and the resulting product moment admits a closed-form expression due to Proposition 1 and the fact that the codewords are independently drawn from the Dirichlet distribution. The proof of Theorem 1 is shorter and more direct than the proof in [37] and can be found in Section V.

IV. Partition Coding

In this section, we propose a different encoding-decoding scheme. While the system model in Section III uses random codebook generation and maximum likelihood decoding; the scheme below is much simpler for the following reasons:

1) The codebook is generated purely deterministically and therefore requires no randomness. Comparing to the random codebook in Section III, recall that the number of potential PMF codewords is given approximately by[8]
$$|\mathcal{C}_M| \approx \exp\left\{\frac{1-\beta\log|\mathcal{A}|}{2} M^{\beta\log|\mathcal{A}|}\log M\right\}, \tag{24}$$
which means that we need to generate this many Dirichlet random variables $\mathrm{Dir}(\boldsymbol{\alpha})$ with $\boldsymbol{\alpha}=(1,\ldots,1)$, and each one of those requires drawing $M^{\beta\log|\mathcal{A}|}$ exponentially distributed random variables.
2) The decoder complexity is significantly reduced, which is explained as follows. In both coding schemes, the random coding-based in Section III and the one that will be described in the sequel, the respective decoders count how many times each type of molecule appears within the K samples (In Section III, we normalized to get an empirical PMF, because the decoder was based on KL divergence scores), to yield $M^{\beta\log|\mathcal{A}|}$ count values, according to the number of distinct molecule types. In Section III, the maximum-likelihood decoder calculates $|\mathcal{C}_M|$ decoding measures (as in (15)), while the decoder which is described below only has to sort the count values from the largest to the smallest, which can be achieved in a computational complexity of $\Theta(M^{\beta\log|\mathcal{A}|}\log M)$ using standard sorting algorithms. It follows that the decoding complexity of the proposed scheme in the current section is only logarithmic with respect to the decoding complexity of the random coding-based scheme.

The lower computational complexity of the proposed coding scheme comes with a price; its theoretical guarantees match the theoretical guarantees of the encoding-decoding scheme in Section III only for $\beta\in\left(\frac{1}{3\log|\mathcal{A}|},\frac{1}{\log|\mathcal{A}|}\right)$, but when $\beta\in\left(0,\frac{1}{3\log|\mathcal{A}|}\right)$, i.e. for very short molecules, the proposed coding scheme attains a smaller information density than random coding.

The main idea in partition coding is that each message is mapped into a distinct *partition* of the set of $n=M^{\beta\log|\mathcal{A}|}$ molecule types. A partition of $[n]$ is defined as a set of *ordered* disjoint subsets of $[n]$, where the union of all of them is $[n]$. Without loss of optimality, we assume that all subsets composing a partition of $[n]$ have the same size. The encoding is performed as follows. A given message determines a specific partition of all molecule types. The copy numbers for all molecule types in a given subset are identical, but differ between different subsets, and are determined by some fixed monotonic frequency vector. All copies from all molecule types are stored in the molecular pool. At the decoder end, a partition of $[n]$ is estimated as follows. The decoder counts the number of samples of each molecule type. Then, it sorts these counts, which determines a specific sequence of all molecule types. Finally, this sequence is partitioned into a desired amount of subsets, i.e., into a partition of $[n]$.

We now describe more specifically the proposed coding scheme. We fix a design parameter $\rho\in[0,1]$, which controls the trade-off between the number of subsets in a partition and the size of each one of them. For a given $n\in\mathbb{N}$ and $\rho\in[0,1]$, the number of distinct molecule types that are going to be used is $n_{\text{eff}}=\lfloor n^{\rho}\rfloor\cdot\left\lfloor n^{1-\rho}\right\rfloor$, which is smaller or equal to the number n of available molecule types. The message set is given by all partitions of $\{1,\ldots,n_{\text{eff}}\}$ into $\lfloor n^{\rho}\rfloor$ ordered sets, each one of size $\left\lfloor n^{1-\rho}\right\rfloor$. The encoder operates as follows. For a given message m, its respective partition is denoted by
$$m=\{\mathcal{S}(1),\ldots,\mathcal{S}(\lfloor n^{\rho}\rfloor)\}, \tag{25}$$
where $\bigcup_{i=1}^{\lfloor n^{\rho}\rfloor}\mathcal{S}(i)=[n_{\text{eff}}]$ and $\mathcal{S}(i)\cap\mathcal{S}(j)=\emptyset$, for any $i\neq j$. In Appendix A, we show how to efficiently generate a partition for a given message index. We also write each set $\mathcal{S}(i)$, where $i\in[\lfloor n^{\rho}\rfloor]$, as a set of individual indices:
$$\mathcal{S}(i)=\left\{\ell_{(i-1)\lfloor n^{1-\rho}\rfloor+1},\ldots,\ell_{i\lfloor n^{1-\rho}\rfloor}\right\}. \tag{26}$$

For a given $\rho\in[0,1]$, let us define the PMF $\{R(i)\}_{i=1}^{\lfloor n^{\rho}\rfloor}$ as a decreasing arithmetic sequence with a common difference d between successive terms. This PMF will be used in the sequel to determine the copy numbers of the various molecule types. Since the total number of messages is given by the number of distinct partitions, this specific PMF choice is without loss of optimality in terms of the attained information density, as it can only control the error probability of the proposed coding scheme. Specifically, we choose the last term as
$$R(\lfloor n^{\rho}\rfloor)=\frac{1}{\lfloor n^{\rho}\rfloor^2}, \tag{27}$$
and the common difference by
$$d=\frac{2}{\lfloor n^{\rho}\rfloor^2}. \tag{28}$$
Then, for any $\ell\in\{1,\ldots,\lfloor n^{\rho}\rfloor-1\}$,
$$R(\ell)=R(\lfloor n^{\rho}\rfloor)+d(\lfloor n^{\rho}\rfloor-\ell), \tag{29}$$
and it can be verified that $\sum_{\ell=1}^{\lfloor n^{\rho}\rfloor}R(\ell)=1$.

For any $i\in[\lfloor n^{\rho}\rfloor]$, define the index sets
$$\mathcal{A}_i=\left\{(i-1)\left\lfloor n^{1-\rho}\right\rfloor+1,\ldots,i\left\lfloor n^{1-\rho}\right\rfloor\right\}. \tag{30}$$

[8]For two positive sequences f_n and g_n, we denote $f_n\approx g_n$ whenever $\lim_{n\to\infty}f_n/g_n=1$.

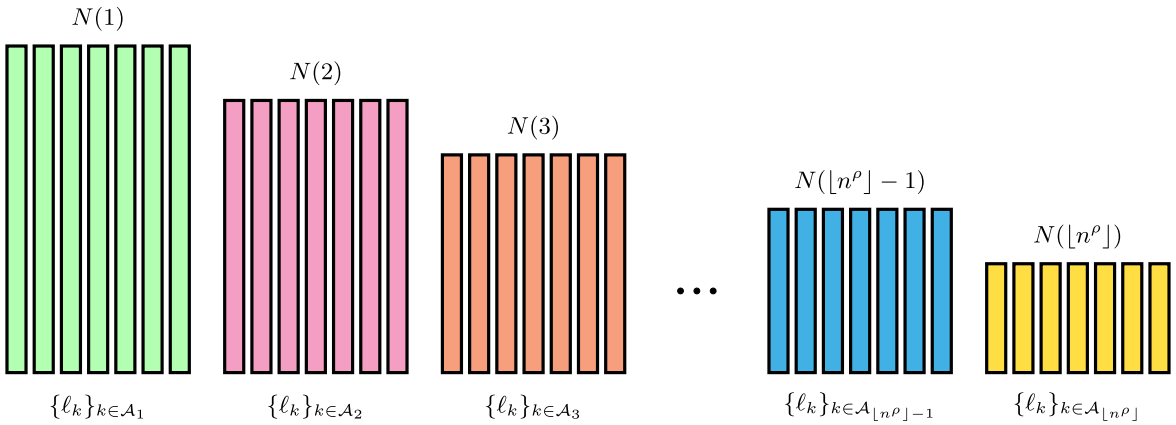

Fig. 1. A description of the partition code structure; the set $\{1, 2, \ldots, n_{\text{eff}}\}$ of molecule types is partitioned into $\lfloor n^\rho \rfloor$ equal-size subsets of molecule types. In each subset, all types within a subset have the same number of copies.

As in Section III, let $\boldsymbol{a}(1), \ldots, \boldsymbol{a}(n)$ be an ordered set of all strings in $\mathcal{A}^L$. The mth codeword is constructed as follows. For any molecule type subset $i \in \{1, 2, \ldots, \lfloor n^\rho \rfloor\}$,

$$N(i) = \left\lfloor \frac{MR(i)}{\lfloor n^{1-\rho} \rfloor} \right\rfloor \tag{31}$$

copies of each of the strings $\boldsymbol{a}(\ell_j)$, $j \in \mathcal{A}_i$, are added to the molecular pool.

At this point, we would like to find a sufficient condition on the parameters β and ρ under which the copy numbers $N(1), \ldots, N(\lfloor n^\rho \rfloor)$ form a *strictly* decreasing sequence as M grow without bound. To this end, we lower-bound the difference between consecutive values in $\{N(1), N(2), \ldots, N(\lfloor n^\rho \rfloor)\}$. For any $i \in \{1, 2, \ldots, \lfloor n^\rho \rfloor - 1\}$,

$$\begin{aligned}
N(i) - N(i+1) &= \left\lfloor \frac{MR(i)}{\lfloor n^{1-\rho} \rfloor} \right\rfloor - \left\lfloor \frac{MR(i+1)}{\lfloor n^{1-\rho} \rfloor} \right\rfloor && (32)\\
&\geq \frac{MR(i)}{\lfloor n^{1-\rho} \rfloor} - 1 - \frac{MR(i+1)}{\lfloor n^{1-\rho} \rfloor} && (33)\\
&= \frac{M}{\lfloor n^{1-\rho} \rfloor}(R(i) - R(i+1)) - 1 && (34)\\
&= \frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1 && (35)\\
&= \frac{2M}{\lfloor n^{1-\rho} \rfloor \lfloor n^\rho \rfloor^2} - 1 && (36)\\
&\geq \frac{2M}{n^{1+\rho}} - 1 && (37)\\
&= 2M^{1-(1+\rho)\beta \log |\mathcal{A}|} - 1, && (38)
\end{aligned}$$

where (33) holds since $t - 1 \leq \lfloor t \rfloor \leq t$ for any $t \geq 0$, and (38) follows from the fact that $n = M^{\beta \log |\mathcal{A}|}$. Thus, from this point onward, we assume that

$$(1 + \rho)\beta \log |\mathcal{A}| < 1 \tag{39}$$

holds and call it the *monotonicity condition*.

The total number of stored molecules is then given by

$$T = \lfloor n^{1-\rho} \rfloor \sum_{i=1}^{\lfloor n^\rho \rfloor} N(i), \tag{40}$$

and we note that all codewords are composed of a fixed number of DNA molecules. In addition, this amount of molecules is upper-bounded by

$$T = \lfloor n^{1-\rho} \rfloor \sum_{i=1}^{\lfloor n^\rho \rfloor} \left\lfloor \frac{MR(i)}{\lfloor n^{1-\rho} \rfloor} \right\rfloor \leq \sum_{i=1}^{\lfloor n^\rho \rfloor} MR(i) = M. \tag{41}$$

An illustration of the partition code is provided in Figure 1.

At the decoder end, upon observing $\boldsymbol{y} = (\boldsymbol{y}_1^L, \boldsymbol{y}_2^L, \ldots, \boldsymbol{y}_K^L)$, the decoder first calculates the count vector

$$\boldsymbol{U}_{\boldsymbol{y}} = (U_{\boldsymbol{y}}(1), \ldots, U_{\boldsymbol{y}}(n_{\text{eff}})), \tag{42}$$

where for any $\ell \in [n_{\text{eff}}]$,

$$U_{\boldsymbol{y}}(\ell) = \sum_{i=1}^{K} \mathbb{1}[\boldsymbol{y}_i^L = \boldsymbol{a}(\ell)]. \tag{43}$$

The expected values of those enumerators are equal for all molecule types in the same partition subset. For any $i \in [\lfloor n^\rho \rfloor]$ and any $k \in \mathcal{A}_i$,

$$\bar{U}(i) = \mathbb{E}[U_{\boldsymbol{y}}(\ell_k)] = \frac{KN(i)}{T}. \tag{44}$$

If at least one member of $\boldsymbol{U}_{\boldsymbol{y}}$ equals zero, a decoding error is declared. This condition may be relaxed to include only those events where more than $\lfloor n^{1-\rho} \rfloor$ members of $\boldsymbol{U}_{\boldsymbol{y}}$ equal zero. The difference between the probabilities of the two error events can be shown to be negligible. Otherwise, if all members of $\boldsymbol{U}_{\boldsymbol{y}}$ are positive, the decoder determines the order

$$U_{\boldsymbol{y}}(\hat{\ell}_1) \geq U_{\boldsymbol{y}}(\hat{\ell}_2) \geq \cdots \geq U_{\boldsymbol{y}}(\hat{\ell}_{n_{\text{eff}}-1}) \geq U_{\boldsymbol{y}}(\hat{\ell}_{n_{\text{eff}}}), \tag{45}$$

then estimates the transmitted message as the partition[9]

$$\hat{m}(\boldsymbol{y}) = \left\{\hat{\mathcal{S}}_{\boldsymbol{y}}(1), \ldots, \hat{\mathcal{S}}_{\boldsymbol{y}}(\lfloor n^\rho \rfloor)\right\}, \tag{46}$$

where for any $i \in [\lfloor n^\rho \rfloor]$

$$\hat{\mathcal{S}}_{\boldsymbol{y}}(i) = \left\{\hat{\ell}_{(i-1)\lfloor n^{1-\rho} \rfloor + 1}, \ldots, \hat{\ell}_{i\lfloor n^{1-\rho} \rfloor}\right\}. \tag{47}$$

Assuming that all members of $\boldsymbol{U}_{\boldsymbol{y}}$ are positive, such that a specific partition of $\{1, 2, \ldots, n_{\text{eff}}\}$ can be determined, we

[9] Analogous to Appendix A, one can *efficiently* transform a given partition into its respective message index.

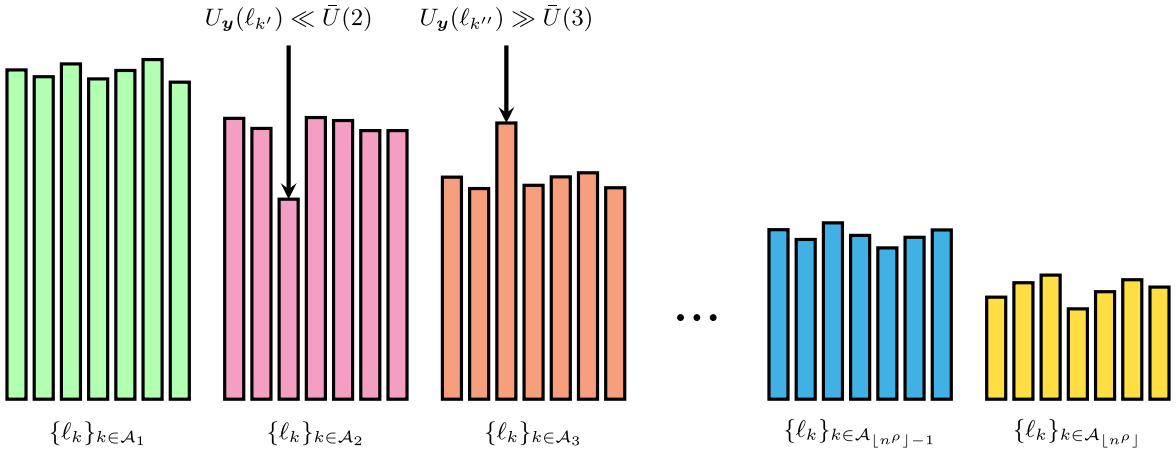

Fig. 2. The counts $\{U_{\boldsymbol{y}}(1),\ldots,U_{\boldsymbol{y}}(n_{\text{eff}})\}$ of all n_{eff} molecule types, as placed in the order of the encoded partition. Random fluctuations in the count numbers are usually sufficiently low, such that each molecule type is decoded in the correct subset. In the presented case, the count number of a molecule type from $\mathcal{S}_2$ is relatively low, and the count number of a molecule from $\mathcal{S}_3$ is relatively high; due to these two large deviations, the decoded partition is incorrect.

expect the effective noise due to sampling to be sufficiently low, such that all the (random) counts $\{U_{\boldsymbol{y}}(1),\ldots,U_{\boldsymbol{y}}(n_{\text{eff}})\}$ will be relatively close to their respective expected values in $\{\bar{U}(1),\ldots,\bar{U}(\lfloor n^\rho\rfloor)\}$. An error occurs in such cases where the number of counts of at least one type of molecule deviates significantly from expectation and becomes closer to the typical count values of a different subset of molecules. For example, consider the molecule type $\ell_k \in \mathcal{S}(i)$. The count number of this molecule type (and all other molecule types in $\mathcal{S}(i)$), $U_{\boldsymbol{y}}(\ell_k)$, is expected to be close to $\bar{U}(i)$, and this molecule type should be decoded to $\hat{\mathcal{S}}_{\boldsymbol{y}}(i)$. In rare cases (as shown in Section VI), where $U_{\boldsymbol{y}}(\ell_k)$ is much higher than its expected value, it will probably be decoded to $\hat{\mathcal{S}}_{\boldsymbol{y}}(i-1)$, and if $U_{\boldsymbol{y}}(\ell_k)$ is much lower than its expected value, it will probably be decoded to $\hat{\mathcal{S}}_{\boldsymbol{y}}(i+1)$. In both cases, the decoded partition will be in error. We illustrate this phenomenon visually in Figure 2.

The probability of error of the partition code as a function of M, β, ξ, and ρ is given by

$$\varepsilon_{\text{PC}}(M,\beta,\xi,\rho) = \mathbb{P}\left[\bigcup_{i=1}^{n_{\text{eff}}}\{U_{\boldsymbol{Y}}(i)=0\}\cup\bigcup_{i=1}^{\lfloor n^\rho\rfloor}\left\{\hat{\mathcal{S}}_{\boldsymbol{Y}}(i)\neq\mathcal{S}(i)\right\}\right]. \tag{48}$$

Before introducing our upper bound for $\varepsilon_{\text{PC}}(M,\beta,\xi,\rho)$, we define

$$\Phi(M,\beta,\rho) = \frac{M^{1-(1+\rho)\beta\log|\mathcal{A}|}-1}{2M^{1-(1+\rho)\beta\log|\mathcal{A}|}}. \tag{49}$$

In Section VI we prove the following result.

Theorem 2: Consider the partition code with M molecules per codeword, molecule length parameter $\beta > 0$, coverage depth $\xi > 0$, and $\rho \in [0,1]$. Assume that the monotonicity condition in (39) holds. The following bound holds:

$$\varepsilon_{\text{PC}}(M,\beta,\xi,\rho) \le M^{\beta\log|\mathcal{A}|}\cdot\exp\left\{-\xi\cdot\left\lfloor M^{1-(1+\rho)\beta\log|\mathcal{A}|}\right\rfloor\right\} + M^{(2-\rho)\beta\log|\mathcal{A}|}\cdot\exp\left\{-\xi\Phi(M,\beta,\rho)M^{1-(1+2\rho)\beta\log|\mathcal{A}|}\right\}. \tag{50}$$

As can be seen in (50), the error probability is upper-bounded by a sum of two terms, which refers to the two possible types of errors. The first expression bounds the probability that at least one term in $\boldsymbol{U}_{\boldsymbol{y}}$ equals zero, while the second expression bounds the probability that the order of the n terms in $\boldsymbol{U}_{\boldsymbol{y}}$ is incorrect. For any given $\rho \in [0,1]$ and $\beta \in \left(0, \frac{1}{(1+\rho)\log|\mathcal{A}|}\right)$, it follows that $\Phi(M,\beta,\rho) \to \frac{1}{2}$ as $M \to \infty$, hence, this expression does not affect the convergence of the second term in (50). While the first term converges to 0 as $M \to \infty$ for any $\beta \in \left(0, \frac{1}{(1+\rho)\log|\mathcal{A}|}\right)$, the second term converges slower to 0, and only as long as $\beta \in \left(0, \frac{1}{(1+2\rho)\log|\mathcal{A}|}\right)$. Thus, for any $\beta \in \left(0, \frac{1}{(1+2\rho)\log|\mathcal{A}|}\right)$, the bound in (50) converges to 0 as $M \to \infty$. Note that the two expressions in (50) depend on the coverage depth; for increased coverage depth, the error probability converges faster to 0 as $M \to \infty$, as expected. It is important to mention that while the error probability related to the maximum likelihood decoder in the random coding-based scheme converges to 0 for any $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$, we were only able to prove that it is upper-bounded by an expression which converges to zero at a much slower rate (as can be seen in the error probability bound in (133)) than the convergence rate of (50).

Relying on the fact that $M^{1-(1+\rho)\beta\log|\mathcal{A}|}$ grows without bound as $M \to \infty$ due to the monotonicity condition, we may assume that $M^{1-(1+\rho)\beta\log|\mathcal{A}|} \ge 3$ for all M sufficiently large, which implies the following simplified bound.

Corollary 1: Consider the partition code with M molecules per codeword, molecule length parameter $\beta > 0$, coverage depth $\xi > 0$, and $\rho \in [0,1]$. Then, for all M sufficiently large,

$$\varepsilon_{\text{PC}}(M,\beta,\xi,\rho) \le 2M^{(2-\rho)\beta\log|\mathcal{A}|}\cdot\exp\left\{-\frac{\xi}{3}\cdot M^{1-(1+2\rho)\beta\log|\mathcal{A}|}\right\}. \tag{51}$$

With respect to the information density attained by the partition code, note that the codebook is given by all partitions of the set $\{1,2,\ldots,n_{\text{eff}}\}$ into $\lfloor n^\rho\rfloor$ ordered subsets of equal sizes, hence

$$|\mathcal{C}_M| = \frac{n_{\text{eff}}!}{\left(\left\lfloor n^{1-\rho}\right\rfloor!\right)^{\lfloor n^\rho\rfloor}}, \tag{52}$$

because we divide by the number of permutations inside each one of the subsets. Then, the following result, which is proved in Appendix B, is a direct consequence of Stirling bounds.

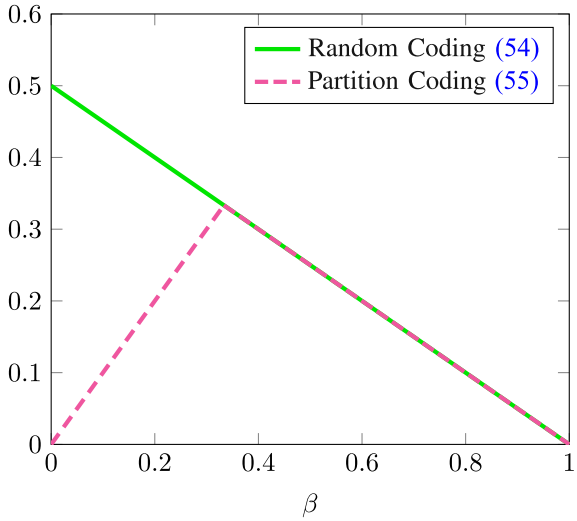

Fig. 3. Comparison between the leading factors in (54) and (55) as functions of β for $|\mathcal{A}| = 2$ and $\rho = \rho^*(\beta)$.

Proposition 2: Consider the partition code with M molecules per codeword, molecule length parameter $\beta > 0$, and $\rho \in [0, 1]$. The following holds:

$$\lim_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}\log(M)} = \rho\beta\log|\mathcal{A}|. \tag{53}$$

Next, we compare between (16) and (53): For random coding with maximum likelihood decoding, we have

$$\log|\mathcal{C}_M^{\text{RC}}| \approx \frac{1-\beta\log|\mathcal{A}|}{2}\cdot M^{\beta\log|\mathcal{A}|}\log(M), \tag{54}$$

while for partition coding,

$$\log|\mathcal{C}_M^{\text{PC}}| \approx \rho\beta\log|\mathcal{A}|\cdot M^{\beta\log|\mathcal{A}|}\log(M). \tag{55}$$

Thus, the two information densities have a similar scaling law with respect to M. Choosing ρ as[10]

$$\rho^*(\beta) = \min\left\{1, \frac{1-\beta\log|\mathcal{A}|}{2\beta\log|\mathcal{A}|}\right\} \tag{56}$$

yields an equality between the leading factors for any $\beta \in \left(\frac{1}{3\log|\mathcal{A}|}, \frac{1}{\log|\mathcal{A}|}\right)$, and hence we find that the low-complexity partition code achieves the optimal performance in this range; otherwise, the result of Theorem 1 strictly outperforms that of Prop. 2. To see how the leading factors of Theorem 1 and Prop. 2 compare (after substituting $\rho = \rho^*(\beta)$), see Figure 3.

We close this section with a remark on the choice of the PMF $\{R(i)\}_{i=1}^{\lfloor n^\rho\rfloor}$ in (27)-(29). Basically, one should only require this sequence to be a monotonic sequence; it does not necessarily have to be an arithmetic sequence. The main reason for which we specifically made this choice is technical. The probability of two molecule types that belong to two consecutive subsets of the partition to be switched when decoding (due to an atypical sampling) is upper-bounded by an expression which is proportional to the difference between their respective number of copies. Hence, it turns out that having the sequence of the number of copies across the various subsets with a common difference between each consecutive pair simplifies the analysis significantly. It follows from Proposition 2 that the information density attained by partition coding depends only on ρ and not on $\{R(i)\}_{i=1}^{\lfloor n^\rho\rfloor}$. Hence, we conclude that partition coding equipped with arithmetically decreasing weights is capable of achieving the optimal information density, except for a specific range of low β values. It may still be possible that by making other choices for $\{R(i)\}_{i=1}^{\lfloor n^\rho\rfloor}$, one can achieve a faster convergence rate for the probability of error. Since we are mainly concerned with achieving the optimal information density, this will not be explored further.

[10]Since the exponential term in (51) does not vanish at $\rho = \rho^*(\beta)$, then more rigorously, one needs to choose $\rho = \rho^*(\beta) - \varepsilon$ for some arbitrarily small $\varepsilon > 0$ in order to assure that $\varepsilon_{\text{PC}}(M, \beta, \xi, \rho) \to 0$ as $M \to \infty$.

V. Proof of Theorem 1

We assume without loss of generality that the encoded message is $m = 1$. We first condition on the transmitted codeword $\hat{\boldsymbol{P}}_1 = \hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_n)$ and on the channel output sequence $\boldsymbol{Y} = \boldsymbol{y}$ (which induces the PMF $\hat{\boldsymbol{Q}}_{\boldsymbol{y}}$) and denote the competing codewords $\hat{\boldsymbol{P}}_m = (\hat{P}_m(1), \ldots, \hat{P}_m(n))$. The probability of error is given by[11]

$$\begin{aligned}
&\varepsilon_M(\hat{\boldsymbol{p}}, \boldsymbol{y}) \\
&\quad= \mathbb{P}\left[\bigcup_{m=2}^{|\mathcal{C}_M|}\{D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{P}}_m) \le D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{p}})\}\right] &(57)\\
&\quad\le \min\left\{1, \sum_{m=2}^{|\mathcal{C}_M|}\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{P}}_m) \le D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{p}})\right]\right\}, &(58)
\end{aligned}$$

using the clipped union bound, where the pairwise error probability $\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{P}}_m) \le D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{p}})\right]$ is the probability of deciding in favor of message m when message 1 was sent for a fixed channel output sequence $\boldsymbol{y}$.

A. Bounding the Pairwise Error Probability

Let $\theta \ge 0$ be an arbitrary parameter. The probability in (58) is given by

$$\begin{aligned}
&\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{P}}_m) \le D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}}\|\hat{\boldsymbol{p}})\right] \\
&\quad= \mathbb{P}\left[\sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\frac{\hat{Q}_{\boldsymbol{y}}(i)}{\hat{P}_m(i)} \le \sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\frac{\hat{Q}_{\boldsymbol{y}}(i)}{\hat{p}_i}\right] &(59)\\
&\quad= \mathbb{P}\left[\sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\hat{P}_m(i) \ge \sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\hat{p}_i\right] &(60)\\
&\quad= \mathbb{P}\left[\sum_{i=1}^n \log\hat{P}_m(i)^{\theta\hat{Q}_{\boldsymbol{y}}(i)} \ge \theta\sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\hat{p}_i\right] &(61)\\
&\quad= \mathbb{P}\left[\prod_{i=1}^n \hat{P}_m(i)^{\theta\hat{Q}_{\boldsymbol{y}}(i)} \ge \exp\left\{\theta\sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\hat{p}_i\right\}\right] &(62)\\
&\quad\le \frac{\mathbb{E}\left[\prod_{i=1}^n \hat{P}_m(i)^{\theta\hat{Q}_{\boldsymbol{y}}(i)}\right]}{\exp\left\{\theta\sum_{i=1}^n \hat{Q}_{\boldsymbol{y}}(i)\log\hat{p}_i\right\}}, &(63)
\end{aligned}$$

where (63) follows from Markov's inequality.

[11]An error is also declared if the minimal KL-divergence score is attained by more than a single message.

We upper-bound the empirical probabilities $\{\hat{P}_m(i)\}$ defined in (12) as

$$\hat{P}_m(i) = \frac{\lfloor MP_m(i) \rfloor}{\sum_{k=1}^{n} \lfloor MP_m(k) \rfloor} \tag{64}$$

$$\leq \frac{MP_m(i)}{\sum_{k=1}^{n} (MP_m(k) - 1)} \tag{65}$$

$$= \frac{MP_m(i)}{M-n}. \tag{66}$$

We then bound the expectation in (63) as

$$\mathbb{E}\left[\prod_{i=1}^{n} \hat{P}_m(i)^{\theta \hat{Q}_y(i)}\right]$$

$$\leq \mathbb{E}\left[\prod_{i=1}^{n} \left(\frac{MP_m(i)}{M-n}\right)^{\theta \hat{Q}_y(i)}\right] \tag{67}$$

$$= \left(\frac{M}{M-n}\right)^{\theta} \cdot \mathbb{E}\left[\prod_{i=1}^{n} P_m(i)^{\theta \hat{Q}_y(i)}\right]. \tag{68}$$

In order to evaluate the expectation in (68), we use Proposition 1 from Section III with $\alpha_1 = \ldots = \alpha_n = 1$ and $\beta_i = \theta \hat{Q}_y(i)$, giving

$$\mathbb{E}\left[\prod_{i=1}^{n} P_m(i)^{\theta \hat{Q}_y(i)}\right]$$

$$= \frac{\Gamma(n)}{\Gamma\left(\sum_{i=1}^{n} (1 + \theta \hat{Q}_y(i))\right)} \cdot \prod_{i=1}^{n} \frac{\Gamma\left(1 + \theta \hat{Q}_y(i)\right)}{\Gamma(1)} \tag{69}$$

$$= \frac{\Gamma(n)}{\Gamma(n+\theta)} \cdot \prod_{i=1}^{n} \Gamma\left(1 + \theta \hat{Q}_y(i)\right), \tag{70}$$

thanks to the fact that $\Gamma(1) = 1$.

Substituting back into (68) and then into (63), we arrive at

$$\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{P}}_m) \leq D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right]$$

$$\leq \left(\frac{M}{M-n}\right)^{\theta} \cdot \frac{\Gamma(n)}{\Gamma(n+\theta)} \cdot \left(\prod_{i=1}^{n} \Gamma\left(1 + \theta \hat{Q}_y(i)\right)\right)$$

$$\times \exp\left\{-\theta \sum_{i=1}^{n} \hat{Q}_y(i) \log \hat{p}_i\right\}. \tag{71}$$

Since the bound is valid for any $\theta > 0$, we choose $\theta = K$, which results in

$$\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{P}}_m) \leq D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right]$$

$$\leq \left(\frac{M}{M-n}\right)^{K} \cdot \frac{\Gamma(n)}{\Gamma(n+K)} \cdot \left(\prod_{i=1}^{n} \Gamma\left(1 + K \hat{Q}_y(i)\right)\right)$$

$$\times \exp\left\{-K \sum_{i=1}^{n} \hat{Q}_y(i) \log \hat{p}_i\right\}. \tag{72}$$

It follows from the definition of $\hat{Q}_y(i)$ in (14) that $K\hat{Q}_y(i) \in \{0, 1, \ldots, K\}$ for any $i \in [n]$. In order to bound the Gamma function factors in (72), we invoke the inequality [57]

$$\Gamma(1+x) \leq \omega \sqrt{2\pi \left(x + \frac{1}{6}\right)} \left(\frac{x}{e}\right)^{x} \tag{73}$$

which holds for every $x \geq 1$, where $\omega = e\sqrt{\frac{3}{7\pi}}$. It can be checked that (73) also holds at $x = 0$.[12]

The inequality in (73) yields

$$\Gamma\left(1 + K\hat{Q}_y(i)\right)$$

$$\leq \omega \sqrt{2\pi} (K\hat{Q}_y(i))^{K\hat{Q}_y(i)} e^{-K\hat{Q}_y(i)} \sqrt{K\hat{Q}_y(i) + \frac{1}{6}}, \tag{74}$$

and in turn,

$$\prod_{i=1}^{n} \Gamma\left(1 + K\hat{Q}_y(i)\right)$$

$$\leq \prod_{i=1}^{n} \omega \sqrt{2\pi} (K\hat{Q}_y(i))^{K\hat{Q}_y(i)} e^{-K\hat{Q}_y(i)} \sqrt{K\hat{Q}_y(i) + \frac{1}{6}} \tag{75}$$

$$= (\omega\sqrt{2\pi})^n K^K e^{-K} \prod_{i=1}^{n} \hat{Q}_y(i)^{K\hat{Q}_y(i)} \prod_{i=1}^{n} \sqrt{K\hat{Q}_y(i) + \frac{1}{6}}. \tag{76}$$

Now,

$$\prod_{i=1}^{n} \sqrt{K\hat{Q}_y(i) + \frac{1}{6}}$$

$$= \exp\left\{\frac{1}{2} \sum_{i=1}^{n} \log\left(K\hat{Q}_y(i) + \frac{1}{6}\right)\right\} \tag{77}$$

$$\leq \exp\left\{\frac{n}{2} \log\left(\frac{K}{n} + \frac{1}{6}\right)\right\} \tag{78}$$

$$\leq \exp\left\{\frac{n}{2} \log\left(\frac{2K}{n}\right)\right\} \tag{79}$$

$$= 2^{\frac{n}{2}} \cdot \left(\frac{K}{n}\right)^{\frac{n}{2}}, \tag{80}$$

where (78) follows from Jensen's inequality and the concavity of the logarithmic function and in (79) we used the fact that $K \geq n$.

Before we proceed, we recall that the χ^2-divergence between two PMFs $\{P(x)\}_{x\in\mathcal{X}}$ and $\{Q(x)\}_{x\in\mathcal{X}}$, where P is assumed to be absolutely continuous with respect to Q, is defined by

$$\chi^2(P\|Q) = \sum_{\{x\in\mathcal{X}: Q(x)>0\}} \frac{(P(x) - Q(x))^2}{Q(x)} \tag{81}$$

$$= \sum_{\{x\in\mathcal{X}: Q(x)>0\}} \frac{P(x)^2}{Q(x)} - 1. \tag{82}$$

Substituting (80) back into (76) and then into (72), we arrive at

$$\mathbb{P}\left[D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{P}}_m) \leq D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right]$$

$$\leq \left(\frac{M}{M-n}\right)^{K} \cdot \frac{\Gamma(n)}{\Gamma(n+K)} \cdot (\omega\sqrt{2\pi})^n K^K e^{-K}$$

$$\times \prod_{i=1}^{n} \hat{Q}_y(i)^{K\hat{Q}_y(i)} 2^{\frac{n}{2}} \cdot \left(\frac{K}{n}\right)^{\frac{n}{2}}$$

$$\times \exp\left\{K \sum_{i=1}^{n} \hat{Q}_y(i) \log \frac{1}{\hat{p}_i}\right\} \tag{83}$$

[12]Where the expression 0^0 is interpreted as 1, based on the limit $\lim_{x\to 0^+} x^x = \lim_{x\to 0^+} e^{x\log(x)} = 1$.

$$= (2\omega\sqrt{\pi})^n \cdot \left(\frac{M}{M-n}\right)^K \cdot \frac{\Gamma(n)}{\Gamma(n+K)} \cdot K^K \cdot e^{-K} \times \left(\frac{K}{n}\right)^{\frac{n}{2}} \cdot \exp\left\{K \sum_{i=1}^{n} \hat{Q}_y(i) \log \frac{\hat{Q}_y(i)}{\hat{p}_i}\right\} \tag{84}$$

$$= (2\omega\sqrt{\pi})^n \cdot \left(\frac{M}{M-n}\right)^K \cdot \frac{\Gamma(n)}{\Gamma(n+K)} \cdot K^K \cdot e^{-K} \times \left(\frac{K}{n}\right)^{\frac{n}{2}} \cdot \exp\left\{K \cdot D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right\} \tag{85}$$

$$\leq (2\omega\sqrt{\pi})^n \cdot \left(\frac{M}{M-n}\right)^K \cdot \frac{\Gamma(n)}{\Gamma(n+K)} \cdot K^K \cdot e^{-K} \times \left(\frac{K}{n}\right)^{\frac{n}{2}} \cdot \exp\left\{K \cdot \chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right\} \tag{86}$$

$$\triangleq B(n, M, K) \cdot \exp\left\{K \cdot \chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right\}, \tag{87}$$

where (86) follows from the fact that $D(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}}) \leq \chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})$ [58, Theorem 5].

B. Averaging Over the Random Samples

Upper-bounding (58) with (87) gives

$$\varepsilon_M(\hat{\boldsymbol{p}}, \boldsymbol{y}) \leq \min\left\{1, \sum_{m=2}^{|\mathcal{C}_M|} B(n, M, K) \cdot \exp\left\{K \cdot \chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right\}\right\} \tag{88}$$

$$\leq \min\left\{1, |\mathcal{C}_M| \cdot B(n, M, K) \cdot \exp\left\{K \cdot \chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})\right\}\right\}. \tag{89}$$

Next, we take the expectation with respect to the random samples. If $\hat{\boldsymbol{p}}$ has support size one, then necessarily $\hat{\boldsymbol{Q}}_{\boldsymbol{y}} = \hat{\boldsymbol{p}}$ and therefore $\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}}) = 0$. In this case the bound in (89) is the lowest possible, and thus, we assume in the sequel that the support size of $\hat{\boldsymbol{p}}$ is strictly greater than one. To this end, we split the space of $\boldsymbol{y}$ vectors into two complement parts; $\boldsymbol{y}$ vectors for which $\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})$ is relatively small and $\boldsymbol{y}$ vectors for which $\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})$ is relatively large. Let $\{\Delta_K\}_{K=1}^{\infty}$ be a monotonically increasing sequence with $\lim_{K\to\infty} \Delta_K = \infty$ that will be chosen at a later point. The main purpose of this sequence is to determine the boundary between small and large values of $\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})$. For a given n, K, Δ_K, and $\hat{\boldsymbol{p}}$, let $\mathcal{F}_n = \mathcal{F}(n, K, \Delta_K, \hat{\boldsymbol{p}})$ be the set of all vectors $\boldsymbol{y} = (\boldsymbol{y}_1^L, \boldsymbol{y}_2^L, \ldots, \boldsymbol{y}_K^L) \in (\mathcal{A}^L)^{\otimes K}$ that induce a PMF $\hat{\boldsymbol{Q}}_{\boldsymbol{y}}$ which is relatively close to $\hat{\boldsymbol{p}}$ in the χ^2-divergence sense. More precisely,

$$\mathcal{F}(n, K, \Delta_K, \hat{\boldsymbol{p}}) = \left\{\boldsymbol{y} \in (\mathcal{A}^L)^{\otimes K} \;\middle|\; 0 \leq \chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}}) \leq \Delta_K \cdot \mathbb{E}\left[\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{Y}} \| \hat{\boldsymbol{p}})\right]\right\}. \tag{90}$$

The expectation in (90) is calculated as follows. Let $(Z_1, \ldots, Z_n) \sim \text{Multinomial}(K, \hat{\boldsymbol{p}})$. Then, by definition,

$$\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}}) = \sum_{\{i\in[n]:\hat{p}_i>0\}} \frac{\hat{Q}_y^2(i)}{\hat{p}_i} - 1 \tag{91}$$

$$= \sum_{\{i\in[n]:\hat{p}_i>0\}} \frac{Z_i^2}{K^2\hat{p}_i} - 1, \tag{92}$$

and the expectation of this χ^2-divergence is given by

$$\mathbb{E}\left[\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{Y}} \| \hat{\boldsymbol{p}})\right] = \sum_{\{i\in[n]:\hat{p}_i>0\}} \frac{\mathbb{E}\left[Z_i^2\right]}{K^2\hat{p}_i} - 1 \tag{93}$$

$$= \sum_{\{i\in[n]:\hat{p}_i>0\}} \frac{K\hat{p}_i(1-\hat{p}_i) + K^2\hat{p}_i^2}{K^2\hat{p}_i} - 1 \tag{94}$$

$$= \sum_{\{i\in[n]:\hat{p}_i>0\}} \frac{(1-\hat{p}_i)}{K} + \sum_{\{i\in[n]:\hat{p}_i>0\}} \hat{p}_i - 1 \tag{95}$$

$$= \frac{|\{i \in [n] : \hat{p}_i > 0\}| - 1}{K} \tag{96}$$

$$\leq \frac{n}{K}, \tag{97}$$

where (94) follows by expanding the second moment of a binomial random variable.

Averaging with respect to the random samples, we get

$$\varepsilon_M(\hat{\boldsymbol{p}}) = \sum_{\boldsymbol{y}\in(\mathcal{A}^L)^{\otimes K}} P_{\boldsymbol{Y}}(\boldsymbol{y}) \varepsilon_M(\hat{\boldsymbol{p}}, \boldsymbol{y}) \tag{98}$$

$$\leq \sum_{\boldsymbol{y}\in(\mathcal{A}^L)^{\otimes K}} P_{\boldsymbol{Y}}(\boldsymbol{y}) \min\left\{1, |\mathcal{C}_M| B(n, M, K) e^{K\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})}\right\} \tag{99}$$

$$= \sum_{\boldsymbol{y}\in\mathcal{F}_n} P_{\boldsymbol{Y}}(\boldsymbol{y}) \min\left\{1, |\mathcal{C}_M| B(n, M, K) e^{K\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})}\right\} + \sum_{\boldsymbol{y}\in\mathcal{F}_n^{\mathrm{c}}} P_{\boldsymbol{Y}}(\boldsymbol{y}) \min\left\{1, |\mathcal{C}_M| B(n, M, K) e^{K\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})}\right\} \tag{100}$$

$$\leq \sum_{\boldsymbol{y}\in\mathcal{F}_n} P_{\boldsymbol{Y}}(\boldsymbol{y}) \min\left\{1, |\mathcal{C}_M| B(n, M, K) \exp\{n\Delta_K\}\right\} + \sum_{\boldsymbol{y}\in\mathcal{F}_n^{\mathrm{c}}} P_{\boldsymbol{Y}}(\boldsymbol{y}) \tag{101}$$

$$= \mathbb{P}[\boldsymbol{Y} \in \mathcal{F}_n] \cdot \min\left\{1, |\mathcal{C}_M| B(n, M, K) \exp\{n\Delta_K\}\right\} + \mathbb{P}[\boldsymbol{Y} \in \mathcal{F}_n^{\mathrm{c}}] \tag{102}$$

$$\leq |\mathcal{C}_M| B(n, M, K) \exp\{n\Delta_K\} + \mathbb{P}[\boldsymbol{Y} \in \mathcal{F}_n^{\mathrm{c}}], \tag{103}$$

where (101) is due to the fact that for any $\boldsymbol{y} \in \mathcal{F}_n$, $\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{y}} \| \hat{\boldsymbol{p}})$ is upper-bounded by $\frac{n\Delta_K}{K}$, and the right-hand-side summation is bounded using $\min\{1, t\} \leq 1$.

It follows by Markov's inequality that

$$\mathbb{P}[\boldsymbol{Y} \in \mathcal{F}_n^{\mathrm{c}}] \leq \mathbb{P}\left[\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{Y}} \| \hat{\boldsymbol{p}}) \geq \Delta_K \cdot \mathbb{E}\left[\chi^2(\hat{\boldsymbol{Q}}_{\boldsymbol{Y}} \| \hat{\boldsymbol{p}})\right]\right] \tag{104}$$

$$\leq \frac{1}{\Delta_K}, \tag{105}$$

which converges to zero as $K \to \infty$ since we assume that $\{\Delta_K\}_{K=1}^{\infty}$ is a monotonically increasing sequence with $\lim_{K\to\infty} \Delta_K = \infty$.

We continue by writing,

$$\varepsilon_M(\hat{\boldsymbol{p}}) \leq |\mathcal{C}_M| \cdot B(n, M, K) \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K}, \tag{106}$$

which is independent of the realization of $\boldsymbol{P}_1$, and thus

$$\varepsilon_M \leq |\mathcal{C}_M| \cdot B(n, M, K) \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K}. \tag{107}$$

C. Bounding the Factor $B(n, M, K)$

Recall that

$$B(n,M,K) = (2\omega\sqrt{\pi})^n \cdot \left(\frac{M}{M-n}\right)^K \times \frac{\Gamma(n)}{\Gamma(n+K)} \cdot K^K \cdot e^{-K} \cdot \left(\frac{K}{n}\right)^{\frac{n}{2}}. \tag{108}$$

The second factor in (108) is bounded as

$$\left(\frac{M}{M-n}\right)^K = \exp\left\{K\log\left(\frac{M}{M-n}\right)\right\} \tag{109}$$

$$= \exp\left\{K\log\left(1+\frac{n}{M-n}\right)\right\} \tag{110}$$

$$\le \exp\left\{\frac{nK}{M-n}\right\} \tag{111}$$

$$\le \exp\left\{\frac{nK}{M-\frac{M}{2}}\right\} \tag{112}$$

$$= \exp\{2\xi n\}, \tag{113}$$

where (111) is due to $\log(1+t) \le t$, (112) follows because $n = M^{\beta\log|\mathcal{A}|}$ for some $\beta \in \left(0, \frac{1}{\log|\mathcal{A}|}\right)$, such that $M^{\beta\log|\mathcal{A}|} \le \frac{M}{2}$ for all M sufficiently large, and in (113) we used the definition $\xi = \frac{K}{M}$.

We invoke the following double-sided Stirling's inequality from [59, Theorem 5]. For any $t > 0$, it holds that

$$\sqrt{2\pi}t^{t-1/2}e^{-t} \le \Gamma(t) \le \sqrt{2\pi}t^{t-1/2}e^{-t}e^{\frac{1}{12t}}, \tag{114}$$

and thus, the third factor in (108) is bounded as follows.

$$\frac{\Gamma(n)}{\Gamma(n+K)} \le \frac{\sqrt{2\pi}n^{n-1/2}e^{-n}e^{\frac{1}{12n}}}{\sqrt{2\pi}(n+K)^{n+K-1/2}e^{-(n+K)}} \tag{115}$$

$$= \sqrt{1+\frac{K}{n}} \cdot \frac{n^n}{(n+K)^n} \cdot \frac{1}{(n+K)^K} \cdot e^K e^{\frac{1}{12n}} \tag{116}$$

$$\le 2\sqrt{1+\frac{K}{n}} \cdot \left(\frac{n}{K}\right)^n \cdot \frac{1}{(n+K)^K} \cdot e^K, \tag{117}$$

where in (117) we upper-bounded $e^{\frac{1}{12n}} \le 2$, which holds for any $n \in \{1,2,\ldots\}$.

Upper-bounding (108) with (113) and (117) yields that for all M sufficiently large

$$B(n,M,K) = (2\omega\sqrt{\pi})^n \cdot \left(\frac{M}{M-n}\right)^K \times \frac{\Gamma(n)}{\Gamma(n+K)} \cdot K^K \cdot e^{-K} \cdot \left(\frac{K}{n}\right)^{\frac{n}{2}} \tag{118}$$

$$\le (2\omega\sqrt{\pi})^n \cdot \exp\{2\xi n\} \cdot 2\sqrt{1+\frac{K}{n}} \cdot \left(\frac{n}{K}\right)^n \times \frac{1}{(n+K)^K} \cdot e^K \cdot K^K \cdot e^{-K} \cdot \left(\frac{K}{n}\right)^{\frac{n}{2}} \tag{119}$$

$$= 2\sqrt{1+\frac{K}{n}} \cdot (2\omega\sqrt{\pi})^n \cdot \exp\{2\xi n\} \times \frac{K^K}{(n+K)^K} \cdot \left(\frac{n}{K}\right)^{\frac{n}{2}} \tag{120}$$

$$\le 2\sqrt{1+\frac{K}{n}} \cdot \exp\left\{n\log\left(2e\sqrt{\frac{3}{7}}\right)\right\} \times \exp\{2\xi n\} \cdot \left(\frac{n}{K}\right)^{\frac{n}{2}} \tag{121}$$

$$\le 2\sqrt{1+\frac{K}{n}} \cdot \exp\{2(1+\xi)n\} \cdot \left(\frac{n}{K}\right)^{\frac{n}{2}}, \tag{122}$$

where (121) follows by substituting $\omega = e\sqrt{\frac{3}{7\pi}}$.

D. Wrapping-up the Proof

Upper-bounding (107) with (122) provides that for all M sufficiently large

$$\varepsilon_M \le |\mathcal{C}_M| \cdot B(n,M,K) \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K} \tag{123}$$

$$\le |\mathcal{C}_M| \cdot 2\sqrt{1+\frac{K}{n}} \cdot \exp\{2(1+\xi)n\} \times \left(\frac{n}{K}\right)^{\frac{n}{2}} \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K} \tag{124}$$

$$= |\mathcal{C}_M| \cdot 2\sqrt{1+\frac{K}{n}} \cdot \exp\{2(1+\xi)n\} \times \exp\left\{-\frac{n}{2}\log\left(\frac{K}{n}\right)\right\} \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K} \tag{125}$$

$$= |\mathcal{C}_M| \cdot 2\sqrt{1+\frac{K}{n}} \cdot \exp\left\{2(1+\xi)n - \frac{n}{2}\log(\xi)\right\} \times \exp\left\{-\frac{n}{2}\log\left(\frac{M}{n}\right)\right\} \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K}, \tag{126}$$

where (126) is due to the fact that $K = \xi M$.

For some $\{\delta_M\}_{M=1}^{\infty}$ with $\lim_{M\to\infty}\delta_M = 0$, let the codebook size be

$$|\mathcal{C}_M| = \exp\left\{\left(\frac{1}{2}-\delta_M\right)n\log\left(\frac{M}{n}\right)\right\}, \tag{127}$$

which implies that for all M sufficiently large

$$\varepsilon_M \le 2\sqrt{1+\frac{K}{n}} \cdot \exp\left\{\left[2+2\xi-\frac{1}{2}\log(\xi)\right]n\right\} \times \exp\left\{-\delta_M n\log\left(\frac{M}{n}\right)\right\} \cdot \exp\{n\Delta_K\} + \frac{1}{\Delta_K}. \tag{128}$$

Finally, choosing $\delta_M = \log^{-\eta/2}(M)$ and $\Delta_K = \log^{1-\eta}(K)$ for some $\eta \in (0,1)$ and substituting $n = M^{\beta\log|\mathcal{A}|}$ and $K = \xi M$ yields that for all M sufficiently large

$$\varepsilon_M \le 2\sqrt{1+\xi M^{1-\beta\log|\mathcal{A}|}} \times \exp\left\{\left[2+2\xi-\frac{1}{2}\log(\xi)\right]M^{\beta\log|\mathcal{A}|}\right\} \times \exp\left\{-\log^{-\eta/2}(M)M^{\beta\log|\mathcal{A}|}\log(M^{1-\beta\log|\mathcal{A}|})\right\} \times \exp\left\{M^{\beta\log|\mathcal{A}|}\log^{1-\eta}(\xi M)\right\} + \frac{1}{\log^{1-\eta}(\xi M)}. \tag{129}$$

To clearly see why (129) converges to zero when $M \to \infty$, we define

$$c_0 = 1 - \beta\log|\mathcal{A}|, \tag{130}$$

$$c_1 = \beta\log|\mathcal{A}|, \tag{131}$$

$$c_2 = 2 + 2\xi - \frac{1}{2}\log(\xi). \tag{132}$$

Using these constants, (129) has the form:

$$\varepsilon_M \le 2\sqrt{1+\xi M^{c_0}} \times \exp\left\{\left[c_2 + \log^{1-\eta}(\xi M) - c_0 \log^{1-\eta/2}(M)\right]\cdot M^{c_1}\right\} + \frac{1}{\log^{1-\eta}(\xi M)}. \tag{133}$$

Since the right-hand side of (133) converges to zero as $M \to \infty$ for any $\eta \in (0,1)$, the codebook-size construction above yields a sequence of codes $\{\mathcal{C}_M\}_{M\ge 1}$ with $\varepsilon_M \to 0$ and cardinality

$$|\mathcal{C}_M| = \exp\left\{\left(\frac{1}{2} - \delta_M\right)(1-\beta\log|\mathcal{A}|)M^{\beta\log|\mathcal{A}|}\log(M)\right\}. \tag{134}$$

Taking the limit $M \to \infty$ then establishes

$$\lim_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}\log(M)} = \frac{1-\beta\log|\mathcal{A}|}{2}, \tag{135}$$

which completes the proof of Theorem 1.

VI. Proof of Theorem 2

We assume without loss of generality that the true message is the partition

$$m = \left\{\left\{1,\ldots,\lfloor n^{1-\rho}\rfloor\right\},\ldots, \ldots,\left\{(\lfloor n^{\rho}\rfloor - 1)\lfloor n^{1-\rho}\rfloor + 1,\ldots,n_{\text{eff}}\right\}\right\}. \tag{136}$$

Let the sets $\{\mathcal{A}_i\}_{i\in[\lfloor n^\rho\rfloor]}$ be defined as in (30). The correct decoding event is given by

$$\mathcal{G} = \left\{\bigcap_{i=1}^{n_{\text{eff}}}\{U_{\boldsymbol{y}}(i)>0\}\right\} \cap \left\{\bigcap_{i=1}^{\lfloor n^\rho\rfloor}\{\hat{\mathcal{S}}_{\boldsymbol{y}}(i) = \mathcal{S}(i)\}\right\} \tag{137}$$

$$= \left\{\bigcap_{i=1}^{\lfloor n^\rho\rfloor}\bigcap_{j\in\mathcal{A}_i}\{U_{\boldsymbol{y}}(j)>0\}\right\} \cap \left\{\bigcap_{i=1}^{\lfloor n^\rho\rfloor-1}\bigcap_{i'=i+1}^{\lfloor n^\rho\rfloor}\bigcap_{j\in\mathcal{A}_i}\bigcap_{j'\in\mathcal{A}_{i'}}\{U_{\boldsymbol{y}}(j) > U_{\boldsymbol{y}}(j')\}\right\} \tag{138}$$

$$= \left\{\bigcap_{i=1}^{\lfloor n^\rho\rfloor}\bigcap_{j\in\mathcal{A}_i}\{U_{\boldsymbol{y}}(j)>0\}\right\} \cap \left\{\bigcap_{i=1}^{\lfloor n^\rho\rfloor-1}\bigcap_{j\in\mathcal{A}_i}\bigcap_{j'\in\mathcal{A}_{i+1}}\{U_{\boldsymbol{y}}(j) > U_{\boldsymbol{y}}(j')\}\right\}. \tag{139}$$

Then, the error event is given by

$$\mathcal{B} = \left\{\bigcup_{i=1}^{\lfloor n^\rho\rfloor}\bigcup_{j\in\mathcal{A}_i}\{U_{\boldsymbol{y}}(j)=0\}\right\} \cup \left\{\bigcup_{i=1}^{\lfloor n^\rho\rfloor-1}\bigcup_{j\in\mathcal{A}_i}\bigcup_{j'\in\mathcal{A}_{i+1}}\{U_{\boldsymbol{y}}(j') \ge U_{\boldsymbol{y}}(j)\}\right\} \tag{140}$$

$$\triangleq \mathcal{B}_1 \cup \mathcal{B}_2, \tag{141}$$

which by the union bound implies that

$$\varepsilon_{\text{PC}}(M,\beta,\xi,\rho) \le \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2]. \tag{142}$$

A. Analysis for the Error Event $\mathcal{B}_1$

Recall that the random variables $\{U_{\boldsymbol{Y}}(1),\ldots,U_{\boldsymbol{Y}}(n_{\text{eff}})\}$ count how many times each type of molecule appears among the K samples. In order to derive the probabilities of $\mathcal{B}_1$ and $\mathcal{B}_2$, we need to handle those random variables, which are related to the various proportions of each molecule type in the joint pool, denoted by $\hat{P}(\ell)$, $\ell \in [n_{\text{eff}}]$. Since each codeword is composed of exactly T molecules, and in each subset $\mathcal{A}_i$, $i \in [\lfloor n^\rho\rfloor]$, each type of molecule has the same number of copies $N(i)$, the various proportions are given by

$$\hat{P}(j) = \frac{N(i)}{T}, \quad \forall j \in \mathcal{A}_i, i \in [\lfloor n^\rho\rfloor]. \tag{143}$$

It follows by the union bound that

$$\mathbb{P}[\mathcal{B}_1] = \mathbb{P}\left[\bigcup_{i=1}^{\lfloor n^\rho\rfloor}\bigcup_{j\in\mathcal{A}_i}\{U_{\boldsymbol{Y}}(j)=0\}\right] \tag{144}$$

$$\le \sum_{i=1}^{\lfloor n^\rho\rfloor}\sum_{j\in\mathcal{A}_i}\mathbb{P}\left[U_{\boldsymbol{Y}}(j)=0\right] \tag{145}$$

$$= \sum_{i=1}^{\lfloor n^\rho\rfloor}\sum_{j\in\mathcal{A}_i}\left(1-\hat{P}(j)\right)^K \tag{146}$$

$$= \sum_{i=1}^{\lfloor n^\rho\rfloor}\sum_{j\in\mathcal{A}_i}\left(1-\frac{N(i)}{T}\right)^K \tag{147}$$

$$= \sum_{i=1}^{\lfloor n^\rho\rfloor}\lfloor n^{1-\rho}\rfloor\left(1-\frac{N(i)}{T}\right)^K \tag{148}$$

$$\le n\cdot\left(1-\frac{N(\lfloor n^\rho\rfloor)}{T}\right)^K \tag{149}$$

$$= n\cdot\exp\left\{K\cdot\log\left(1-\frac{N(\lfloor n^\rho\rfloor)}{T}\right)\right\} \tag{150}$$

$$\le n\cdot\exp\left\{-K\cdot\frac{N(\lfloor n^\rho\rfloor)}{M}\right\}, \tag{151}$$

where (148) follows from the fact that $|\mathcal{A}_i| = \lfloor n^{1-\rho}\rfloor$ for any $i \in [\lfloor n^\rho\rfloor]$, (149) is due to the fact that $N(1) \ge N(2) \ge \cdots \ge N(\lfloor n^\rho\rfloor)$, and (151) follows from the inequalities $\log(1-t) \le -t$ and $T \le M$ (see (41)).

Substituting $K = \xi M$ and the value of $N(\lfloor n^\rho\rfloor)$ from (31) gives

$$\mathbb{P}[\mathcal{B}_1] \le n\cdot\exp\left\{-\xi\cdot\left\lfloor\frac{MR(\lfloor n^\rho\rfloor)}{\lfloor n^{1-\rho}\rfloor}\right\rfloor\right\}, \tag{152}$$

and substituting the value of $R(\lfloor n^\rho\rfloor)$ from (27) provides

$$\mathbb{P}[\mathcal{B}_1] \le n\cdot\exp\left\{-\xi\cdot\left\lfloor\frac{M}{\lfloor n^{1-\rho}\rfloor\lfloor n^\rho\rfloor^2}\right\rfloor\right\} \tag{153}$$

$$\le n\cdot\exp\left\{-\xi\cdot\left\lfloor\frac{M}{n^{1+\rho}}\right\rfloor\right\} \tag{154}$$

$$= M^{\beta\log|\mathcal{A}|}\cdot\exp\left\{-\xi\cdot\left\lfloor M^{1-(1+\rho)\beta\log|\mathcal{A}|}\right\rfloor\right\}, \tag{155}$$

where (155) follows from the fact that $n = M^{\beta\log|\mathcal{A}|}$. The bound in (155) converges to zero as $M \to \infty$ due to the monotonicity condition in (39).

B. Analysis for the Error Event $\mathcal{B}_2$

It follows by the union bound that

$$\mathbb{P}[\mathcal{B}_2] = \mathbb{P}\left[\bigcup_{i=1}^{\lfloor n^\rho \rfloor - 1} \bigcup_{j\in\mathcal{A}_i} \bigcup_{j'\in\mathcal{A}_{i+1}} \{U_{\boldsymbol{Y}}(j') \geq U_{\boldsymbol{Y}}(j)\}\right] \tag{156}$$

$$\leq \sum_{i=1}^{\lfloor n^\rho \rfloor - 1} \sum_{j\in\mathcal{A}_i} \sum_{j'\in\mathcal{A}_{i+1}} \mathbb{P}\left[U_{\boldsymbol{Y}}(j') \geq U_{\boldsymbol{Y}}(j)\right]. \tag{157}$$

Next, for any $i \in \{1,\ldots,\lfloor n^\rho \rfloor - 1\}$, $j \in \mathcal{A}_i$, and $j' \in \mathcal{A}_{i+1}$,

$$\mathbb{P}\left[U_{\boldsymbol{Y}}(j') \geq U_{\boldsymbol{Y}}(j)\right]$$

$$= \mathbb{P}\left[\sum_{k=1}^{K} \mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j')] \geq \sum_{k=1}^{K} \mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j)]\right] \tag{158}$$

$$\leq \mathbb{E}\left[\exp\left\{\theta \cdot \sum_{k=1}^{K} \left(\mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j')] - \mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j)]\right)\right\}\right] \tag{159}$$

$$= \prod_{k=1}^{K} \mathbb{E}\left[\exp\left\{\theta \cdot \left(\mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j')] - \mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j)]\right)\right\}\right], \tag{160}$$

where (158) follows from the definition in (43), (159) from Markov's inequality, and (160) is due to the independence of the samples. The expectation in (160) is given by

$$\mathbb{E}\left[\exp\left\{\theta \cdot \left(\mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j')] - \mathbb{1}[\boldsymbol{Y}_k^L = \boldsymbol{a}(j)]\right)\right\}\right]$$
$$= \hat{P}(j')e^{\theta} + \hat{P}(j)e^{-\theta} + (1 - \hat{P}(j') - \hat{P}(j)). \tag{161}$$

Let us optimize over $\theta \geq 0$. For $a, b \in [0,1]$, denote the function $f(\theta) = ae^{\theta} + be^{-\theta} + (1-a-b)$. Then, $f'(\theta) = 0$ yields

$$0 = f'(\theta) = ae^{\theta} - be^{-\theta} \implies \theta^* = \frac{1}{2}\log\left(\frac{b}{a}\right), \tag{162}$$

as well as

$$f(\theta^*) = 1 - \left(a + b - 2\sqrt{ab}\right), \tag{163}$$

which is strictly smaller than 1, since it always holds that $a \neq b$ under the monotonicity condition, and it follows that the arithmetic mean is strictly greater than the geometric mean.

Substituting back into (160) and then into (157) yields

$$\mathbb{P}[\mathcal{B}_2]$$

$$\leq \sum_{i=1}^{\lfloor n^\rho \rfloor - 1} \sum_{j\in\mathcal{A}_i} \sum_{j'\in\mathcal{A}_{i+1}} \left[1 - \left(\hat{P}(j) + \hat{P}(j') - 2\sqrt{\hat{P}(j)\hat{P}(j')}\right)\right]^K \tag{164}$$

$$= \sum_{i=1}^{\lfloor n^\rho \rfloor - 1} \sum_{j\in\mathcal{A}_i} \sum_{j'\in\mathcal{A}_{i+1}} \left[1 - \left(\frac{N(i)}{T} + \frac{N(i+1)}{T} - 2\sqrt{\frac{N(i)}{T}\cdot\frac{N(i+1)}{T}}\right)\right]^K \tag{165}$$

$$\leq \sum_{i=1}^{\lfloor n^\rho \rfloor - 1} n^{2(1-\rho)} \left[1 - \frac{2}{T}\left(\frac{N(i)+N(i+1)}{2} - \sqrt{N(i)N(i+1)}\right)\right]^K \tag{166}$$

$$\leq n^{2-\rho} \max_{i\in[\lfloor n^\rho \rfloor - 1]} \left[1 - \frac{2}{T}\left(\frac{N(i)+N(i+1)}{2} - \sqrt{N(i)N(i+1)}\right)\right]^K, \tag{167}$$

where (165) follows from (143) and (166) from the fact that $|\mathcal{A}_i| = \lfloor n^{1-\rho} \rfloor$ for any $i \in [\lfloor n^\rho \rfloor]$.

We now invoke the following refinement of the arithmetic mean–geometric mean inequality, which is a particularization of the main result in [60] to the binary case:

$$\frac{a+b}{2} - \sqrt{ab} \geq \frac{(a-b)^2}{8\max\{a,b\}}. \tag{168}$$

Before we continue, we recall from (35) that

$$N(i) - N(i+1) \geq \frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1, \tag{169}$$

which converges to ∞ as $M \to \infty$ due to the monotonicity condition.

We continue from (167) and arrive at

$$\mathbb{P}[\mathcal{B}_2]$$

$$\leq n^{2-\rho} \max_{i\in[\lfloor n^\rho \rfloor - 1]} \left[1 - \frac{2}{T}\frac{(N(i)-N(i+1))^2}{8\max\{N(i),N(i+1)\}}\right]^K \tag{170}$$

$$= n^{2-\rho} \max_{i\in[\lfloor n^\rho \rfloor - 1]} \left[1 - \frac{(N(i)-N(i+1))^2}{4TN(i)}\right]^K \tag{171}$$

$$\leq n^{2-\rho} \max_{i\in[\lfloor n^\rho \rfloor - 1]} \left[1 - \frac{\left(\frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1\right)^2}{4TN(i)}\right]^K \tag{172}$$

$$= n^{2-\rho} \cdot \left[1 - \frac{\left(\frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1\right)^2}{4TN(1)}\right]^K \tag{173}$$

$$= n^{2-\rho} \cdot \exp\left\{K \cdot \log\left[1 - \frac{\left(\frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1\right)^2}{4TN(1)}\right]\right\} \tag{174}$$

$$\leq n^{2-\rho} \cdot \exp\left\{-\frac{K\left(\frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1\right)^2}{4MN(1)}\right\} \tag{175}$$

$$= n^{2-\rho} \cdot \exp\left\{-\frac{\xi\left(\frac{Md}{\lfloor n^{1-\rho} \rfloor} - 1\right)^2}{4N(1)}\right\}, \tag{176}$$

where (170) is due to (168), (171) and (173) follow from the fact that $\{N(i)\}_{i=1}^{\lfloor n^\rho \rfloor}$ is a non-increasing sequence, in (172) we upper-bounded using (169), (175) is due to $\log(1-t) \leq -t$ and $T \leq M$ (see (41)), and in (176) we used the fact that $K = \xi M$.

Note that

$$N(1) = \left\lfloor \frac{MR(1)}{\lfloor n^{1-\rho} \rfloor} \right\rfloor \tag{177}$$

$$\leq \frac{MR(1)}{\lfloor n^{1-\rho} \rfloor}. \tag{178}$$

Upper-bounding (176) with (178) implies that

$$\mathbb{P}[\mathcal{B}_2] \le n^{2-\rho} \cdot \exp\left\{ -\frac{\xi\left(\frac{Md}{\lfloor n^{1-\rho}\rfloor} - 1\right)^2}{\frac{4MR(1)}{\lfloor n^{1-\rho}\rfloor}} \right\}. \tag{179}$$

The formula in (29) implies that for $\ell = 1$

$$R(1) = R(\lfloor n^\rho \rfloor) + (\lfloor n^\rho \rfloor - 1)d \tag{180}$$

$$= \frac{1}{\lfloor n^\rho \rfloor^2} + \frac{2(\lfloor n^\rho \rfloor - 1)}{\lfloor n^\rho \rfloor^2} \tag{181}$$

$$= \frac{2}{\lfloor n^\rho \rfloor} - \frac{1}{\lfloor n^\rho \rfloor^2}. \tag{182}$$

Substituting the value of d from (28) yields

$$\left(\frac{Md}{\lfloor n^{1-\rho}\rfloor} - 1\right)^2$$

$$= \left(\frac{2M}{\lfloor n^{1-\rho}\rfloor \lfloor n^\rho \rfloor^2} - 1\right)^2 \tag{183}$$

$$= \left(2 - \frac{\lfloor n^{1-\rho}\rfloor \lfloor n^\rho \rfloor^2}{M}\right)^2 \frac{M^2}{\lfloor n^{1-\rho}\rfloor^2 \lfloor n^\rho \rfloor^4} \tag{184}$$

$$= \left(4 - 4\frac{\lfloor n^{1-\rho}\rfloor \lfloor n^\rho \rfloor^2}{M} + \frac{\lfloor n^{1-\rho}\rfloor^2 \lfloor n^\rho \rfloor^4}{M^2}\right) \frac{M^2}{\lfloor n^{1-\rho}\rfloor^2 \lfloor n^\rho \rfloor^4} \tag{185}$$

$$\ge 4\left(1 - \frac{n^{1+\rho}}{M}\right) \frac{M^2}{\lfloor n^{1-\rho}\rfloor^2 \lfloor n^\rho \rfloor^4}, \tag{186}$$

while substituting the value of $R(1)$ from (180) provides

$$\frac{MR(1)}{\lfloor n^{1-\rho}\rfloor} = \frac{M}{\lfloor n^{1-\rho}\rfloor} \cdot \left(\frac{2}{\lfloor n^\rho \rfloor} - \frac{1}{\lfloor n^\rho \rfloor^2}\right) \tag{187}$$

$$\le \frac{2M}{\lfloor n^{1-\rho}\rfloor \lfloor n^\rho \rfloor}. \tag{188}$$

By upper-bounding (179) with (186) and (188), we arrive at

$$\mathbb{P}[\mathcal{B}_2]$$

$$\le n^{2-\rho} \cdot \exp\left\{ -\xi \cdot \frac{\left(1 - \frac{n^{1+\rho}}{M}\right)}{2} \cdot \frac{M}{\lfloor n^{1-\rho}\rfloor \lfloor n^\rho \rfloor^3} \right\} \tag{189}$$

$$\le n^{2-\rho} \cdot \exp\left\{ -\xi \cdot \frac{\left(1 - \frac{n^{1+\rho}}{M}\right)}{2} \cdot \frac{M}{n^{1+2\rho}} \right\} \tag{190}$$

$$= M^{(2-\rho)\beta \log|\mathcal{A}|} \times \exp\left\{ -\xi \cdot \Phi(M,\beta,\rho) \cdot M^{1-(1+2\rho)\beta\log|\mathcal{A}|} \right\}, \tag{191}$$

where (191) follows from the fact that $n = M^{\beta \log|\mathcal{A}|}$ and by the definition of $\Phi(M,\beta,\rho)$ in (49). This completes the proof of Theorem 2.

VII. Summary and Future Work

In this work, we considered the information density of the DNA storage channel in the short molecule regime. By analyzing a random coding scheme, where each codeword is given by an appropriate quantization of a PMF drawn uniformly from the probability simplex, we were able to complete the proof of a conjecture regarding the largest possible information density.

While in the proposed coding scheme, the number of DNA molecules that make up each codeword is not fixed, an alternative random coding mechanism for generating codewords with a fixed number of molecules is as follows: For any given message, a random PMF $\boldsymbol{P} = (P(1), \ldots, P(n))$ is drawn from $\mathcal{P}_n$ according to the Dirichlet distribution with vector parameters $\boldsymbol{\alpha} = (1, \ldots, 1)$. In a second step, for a given realization $\boldsymbol{P} = \boldsymbol{p}$, the codeword is drawn as $(Z_1, \ldots, Z_n) \sim \text{Multinomial}(M, \boldsymbol{p})$. The analysis of the pairwise error probability of this scheme requires one to find a tight upper bound on the product moments of a Dirichlet-multinomial random variable. Yet another mechanism to generate exactly M molecules in each codeword is as follows: First, we generate $\lfloor MP(\ell) \rfloor$ molecules for each molecule type $\ell \in [n]$. Then, we are missing $q = M - \sum_{\ell=1}^{n} \lfloor MP(\ell) \rfloor$ molecules. To complete these molecules, we simply add these q molecules to arbitrary indices, but at most 1 to each. In this way, it is guaranteed that the number of molecules in the ℓth position is in the range $[MP(\ell) - 1, MP(\ell) + 1]$. The analysis of the pairwise error probability in this coding scheme appears non-trivial, because one can no longer invoke the existing identity for the product moments of a Dirichlet random variable, as given in (21). The problems of finding theoretical guarantees for these two coding schemes are currently open and are of interest for future research.

We also proposed partition coding, a deterministic coding scheme that has a relatively low computational complexity. We proved that partition coding also attains the optimal information density, except in a regime of very short molecules. Without imposing any complexity constraints, short-molecule DNA storage with noisy sequencing was recently studied in [53] and [54]. For low-complexity partition coding, this work addresses only the randomness that stems from the random (multinomial) sampling of the DNA molecules. In future work, it may be interesting to consider the case where also the sequencing process is noisy, e.g., one may assume that each of the M molecules is read through a discrete memoryless channel. A more complicated but realistic setting also includes insertions and deletions while sequencing.

Appendix A
Efficient Encoding of an Integer to a Codeword in Partition Coding

Let $k, \ell \in \mathbb{N}$. Let $P(k,\ell)$ denote the set of all ordered partitions of $\{1, \ldots, k\ell\}$ into k subsets, each one of size ℓ. We note that

$$|P(k,\ell)| = \frac{(k\ell)!}{(\ell!)^k}. \tag{A.1}$$

Given $N \in \{0, 1, 2, \ldots, |P(k,\ell)| - 1\}$, we would like to efficiently generate the Nth partition. The following efficient algorithm consists of two steps. For the first step, we first note that any $N \in \{0, 1, \ldots, |P(k,\ell)| - 1\}$ can be written as a

mixed-radix expansion. For any $i \in \{1, 2, \ldots, k\}$, define the radixes

$$B_i = \binom{(k-i+1)\ell}{\ell} \tag{A.2}$$

and the weights

$$i \in [k-1]: \quad W_i = B_{i+1}B_{i+2} \cdot \ldots \cdot B_k = \frac{((k-i)\ell)!}{(\ell!)^{k-i}} \tag{A.3}$$

$$i = k: \quad W_k = 1. \tag{A.4}$$

Now, N is expanded as follows

$$N = d_1 \cdot W_1 + d_2 \cdot W_2 + \ldots + d_{k-1} \cdot W_{k-1} + d_k \cdot W_k, \tag{A.5}$$

where $d_i \in \{0, 1, \ldots, B_i - 1\}$ for any $i \in \{1, 2, \ldots, k\}$ can be calculated using the explicit formula

$$d_i = \left\lfloor \frac{N}{W_i} \right\rfloor \mod B_i. \tag{A.6}$$

At this point, we note that $B_k = 1$, which implies that $d_k = 0$. Furthermore, it holds that $W_{k-1} = 1$, such that the actual expansion reduces to

$$N = d_1 \cdot W_1 + d_2 \cdot W_2 + \ldots + d_{k-2} \cdot W_{k-2} + d_{k-1}. \tag{A.7}$$

In this representation, d_1 stands for the choice of the first ℓ indices (out of $k\ell$) to be assigned to the first partition subset, d_2 stands for the choice of the next ℓ indices (out of $(k-1)\ell$) to be assigned to the second partition subset, etc. We note that after assigning $(k-1)\ell$ indices to the first $k-1$ partition subsets, the ℓ indices pertaining to the last subset are already fixed.

In the second step, each number in $(d_1, \ldots, d_{k-1})$ needs to be assigned to a different subset of ℓ indices. Following the notation in (25), we denote those subsets by $\mathcal{S}(1), \mathcal{S}(2), \ldots, \mathcal{S}(k-1)$. The number d_1 is mapped to a specific combination of ℓ distinct numbers from $[k\ell]$. For a general $i \in \{2, \ldots, k-1\}$, the number d_i is mapped to a specific combination of ℓ distinct numbers from $[k\ell] \setminus \cup_{j=1}^{i-1}\mathcal{S}(j)$. As explained above, the final subset $\mathcal{S}(k)$ is just given by $[k\ell] \setminus \cup_{j=1}^{k-1}\mathcal{S}(j)$.

The main idea to efficiently generate each subset $\mathcal{S}(i)$, where $i \in \{1, \ldots, k-1\}$, is to use a methodology known as the *Combinatorial Number System* of degree ℓ. The earliest recorded mathematical proof of this system dates back to 1887; the reader is referred to [61]. The reader is also referred to the modern source [62] and the References therein. The combinatorial number system (also known as Macaulay representation or combinadics) provides a direct, bijective mapping between any integer index (rank) and a specific ℓ-combination from a given set of numbers. The computational complexity of the combinatorial number system algorithm is polynomial in $k\ell$ [63], and therefore the overall proposed scheme is computationally efficient.

Appendix B
Proof of Proposition 2

We first invoke Stirling's bound, which states that for any $n \in \mathbb{N}$,

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi e n}\left(\frac{n}{e}\right)^n. \tag{B.1}$$

Recall that

$$|\mathcal{C}_M| = \frac{n_{\text{eff}}!}{\left(\lfloor n^{1-\rho} \rfloor!\right)^{\lfloor n^\rho \rfloor}}, \tag{B.2}$$

where $n_{\text{eff}} = \lfloor n^\rho \rfloor \cdot \lfloor n^{1-\rho} \rfloor$. We split into two cases. When $\rho = 1$, it holds that $|\mathcal{C}_M| = n!$, and in this case, normalizing by $n\log(n)$, substituting $n = M^{\beta \log|\mathcal{A}|}$, and letting $M \to \infty$, the Stirling bounds yield

$$\lim_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}\log(M)} = \beta \log|\mathcal{A}|. \tag{B.3}$$

Otherwise if $\rho \in [0, 1)$, we upper-bound as follows:

$$|\mathcal{C}_M| = \frac{\left(\lfloor n^\rho \rfloor \cdot \lfloor n^{1-\rho} \rfloor\right)!}{\left(\lfloor n^{1-\rho} \rfloor!\right)^{\lfloor n^\rho \rfloor}} \tag{B.4}$$

$$\leq \frac{\left(n^\rho \cdot n^{1-\rho}\right)!}{\left((n^{1-\rho}-1)!\right)^{n^\rho - 1}} \tag{B.5}$$

$$= \frac{n!}{\left((n^{1-\rho}-1)!\right)^{n^\rho - 1}}, \tag{B.6}$$

where (B.5) holds since $t - 1 \leq \lfloor t \rfloor \leq t$ for any $t \geq 0$.

Taking logarithms implies that

$$\log|\mathcal{C}_M| \leq \log(n!) - (n^\rho - 1)\log\left((n^{1-\rho}-1)!\right). \tag{B.7}$$

It follows from Stirling bounds that

$$\log|\mathcal{C}_M|$$

$$\leq \log\left[\sqrt{2\pi e n}\left(\frac{n}{e}\right)^n\right] - (n^\rho - 1)\log\left[\sqrt{2\pi(n^{1-\rho}-1)}\left(\frac{n^{1-\rho}-1}{e}\right)^{n^{1-\rho}-1}\right] \tag{B.8}$$

$$\leq \log\left[\sqrt{2\pi e n}\left(\frac{n}{e}\right)^n\right] - (n^\rho - 1)(n^{1-\rho}-1)\log\left(\frac{n^{1-\rho}-1}{e}\right) \tag{B.9}$$

$$= n\log(n) - n + \frac{1}{2}\log(2\pi e n) - (n - n^\rho - n^{1-\rho} + 1)\log(n^{1-\rho}-1) + (n - n^\rho - n^{1-\rho} + 1) \tag{B.10}$$

$$\leq n\log(n) + \frac{1}{2}\log(2\pi e n) - (n - n^\rho - n^{1-\rho} + 1)\log(n^{1-\rho}-1) \tag{B.11}$$

$$\leq n\log(n) + \frac{1}{2}\log(2\pi e n) - (n - n^\rho - n^{1-\rho} + 1)\log\left(n^{1-\rho} - \frac{n^{1-\rho}}{2}\right) \tag{B.12}$$

$$\leq n\log(n) + \frac{1}{2}\log(2\pi e n) - (1-\rho)(n - n^\rho - n^{1-\rho} + 1)\log(n) + n\log(2), \tag{B.13}$$

where (B.12) follows since for any $\rho \in [0, 1)$, it holds that $\frac{n^{1-\rho}}{2} \geq 1$ for all n sufficiently large.

Normalizing by $n\log(n)$, substituting $n = M^{\beta\log|\mathcal{A}|}$, and letting $M \to \infty$ provides

$$\lim_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}\log(M)} \leq \rho\beta\log|\mathcal{A}|. \tag{B.14}$$

Using similar algebraic steps, one can derive a matching lower bound. This completes the proof of Proposition 2.

REFERENCES

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

[2] N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 2013.

[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," *Angew. Chem. Int. Ed.*, vol. 54, no. 8, pp. 2552–2555, Feb. 2015.

[4] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, no. 1, p. 14138, Sep. 2015.

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

[6] L. Organick et al., "Random access in large-scale DNA data storage," *Nature Biotechnol.*, vol. 36, no. 3, pp. 242–248, 2018.

[7] P. L. Antkowiak et al., "Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, Oct. 2020.

[8] O. Sabary, H. M. Kiah, P. H. Siegel, and E. Yaakobi, "Survey for a decade of coding for DNA storage," *IEEE Trans. Mol., Biol., Multi-Scale Commun.*, vol. 10, no. 2, pp. 253–271, Jun. 2024.

[9] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Anchor-based correction of substitutions in indexed sets," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 757–761.

[10] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2331–2351, Apr. 2020.

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

[12] A. Lenz, L. Welter, and S. Puchinger, "Achievable rates of concatenated codes in DNA storage under substitution errors," in *Proc. Int. Symp. Inf. Theory Its Appl. (ISITA)*, Oct. 2020, pp. 269–273.

[13] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the DNA storage channel," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8846–8850.

[14] N. Weinberger, "Error probability bounds for coded-index DNA storage systems," *IEEE Trans. Inf. Theory*, vol. 68, no. 11, pp. 7005–7022, Nov. 2022.

[15] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 3675–3689, Jun. 2021.

[16] I. Shomorony and R. Heckel, "Information-theoretic foundations of DNA data storage," *Found. Trends Commun. Inf. Theory*, vol. 19, no. 1, pp. 1–106, Feb. 2022.

[17] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability bounds," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 5657–5700, Sep. 2022.

[18] Y. H. Ling and J. Scarlett, "Exact error exponents of concatenated codes for DNA storage," *IEEE Trans. Inf. Theory*, vol. 71, no. 9, pp. 6566–6585, Sep. 2025.

[19] Y. H. Ling, N. Weinberger, and J. Scarlett, "Error exponents for DNA storage codes with a variable number of reads," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 205–216, 2025.

[20] V. A. Rameshwar and N. Weinberger, "Information rates over multi-view channels," *IEEE Trans. Inf. Theory*, vol. 71, no. 2, pp. 847–861, Feb. 2025.

[21] V. A. Rameshwar and N. Weinberger, "On achievable rates over noisy nanopore channels," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 270–282, 2025.

[22] A. N. Ravi, A. Vahid, and I. Shomorony, "Coded shotgun sequencing," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 1, pp. 147–159, Mar. 2022.

[23] A. Narayan Ravi, A. Vahid, and I. Shomorony, "An information theory for out-of-order media with applications in DNA data storage," *IEEE Trans. Mol., Biol., Multi-Scale Commun.*, vol. 10, no. 2, pp. 334–348, Jun. 2024.

[24] H. Narayanan, P. Krishnan, and N. Parekh, "On achievable rates for the shotgun sequencing channel with erasures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 1730–1735.

[25] D. Bar-Lev, S. Marcovich, E. Yaakobi, and Y. Yehezkeally, "Adversarial torn-paper codes," *IEEE Trans. Inf. Theory*, vol. 69, no. 10, pp. 6414–6427, Oct. 2023.

[26] I. Shomorony and A. Vahid, "Torn-paper coding," *IEEE Trans. Inf. Theory*, vol. 67, no. 12, pp. 7904–7913, Dec. 2021.

[27] B. McBain, E. Viterbo, and J. Saunderson, "Information rates of the noisy nanopore channel," *IEEE Trans. Inf. Theory*, vol. 70, no. 8, pp. 5640–5652, Aug. 2024.

[28] B. McBain and E. Viterbo, "Achievable rates of nanopore-based DNA storage," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 261–269, 2025.

[29] Z. Aharoni and H. D. Pfister, "Neural polar decoders for DNA data storage," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 403–416, 2025.

[30] L. Welter, R. Sokolovskii, T. Heinis, A. Wachter-Zeh, E. Rosnes, and A. G. I. Amat, "An end-to-end coding scheme for DNA-based data storage with nanopore-sequenced reads," *IEEE J. Sel. Areas Inf. Theory*, vol. 7, pp. 17–32, 2026.

[31] A. Kobovich and N. Weinberger, "Input optimization in the composite DNA storage channel," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 248–260, 2025.

[32] D. Bar-Lev, I. Orr, O. Sabary, T. Etzion, and E. Yaakobi, "Scalable and robust DNA-based storage via coding theory and deep learning," *Nature Mach. Intell.*, vol. 7, no. 4, pp. 639–649, Feb. 2025.

[33] P. K. Vippathalla and N. Kashyap, "The secure storage capacity of a DNA wiretap channel model," *IEEE Trans. Inf. Theory*, vol. 69, no. 9, pp. 5550–5569, Sep. 2023.

[34] W. Zhang and Z. Wang, "Secret sharing for DNA probability vectors," in *Proc. ICC - IEEE Int. Conf. Commun.*, Jun. 2024, pp. 4578–4583.

[35] W. Zhang and Z. Wang, "Ramp secret sharing for composite DNA," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 217–231, 2025.

[36] A. Lapidoth and S. M. Moser, "On the capacity of the discrete-time Poisson channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 303–322, Jan. 2009.

[37] Y. Gerzon, I. Shomorony, and N. Weinberger, "Capacity of frequency-based channels: Encoding information in molecular concentrations," *IEEE Trans. Inf. Theory*, vol. 71, no. 8, pp. 5788–5808, Aug. 2025.

[38] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[39] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inf. Theory*, vol. 4, no. 1, pp. 2–22, Jan. 1954.

[40] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2023. [Online]. Available: https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf

[41] F. Bello, Á. Martin, T. Rischewski, and G. Seroussi, "The lattice-input discrete-time Poisson channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 3624–3629.

[42] R. Tamir and N. Weinberger, "Achievable rates and error probability bounds of frequency-based channels of unlimited input resolution," *IEEE J. Sel. Areas Inf. Theory*, vol. 6, pp. 283–295, 2025.

[43] Y. Choi et al., "High information capacity DNA-based data storage with augmented encoding characters using degenerate bases," *Sci. Rep.*, vol. 9, no. 1, p. 6582, Apr. 2019.

[44] O. Sabary, I. Preuss, R. Gabrys, Z. Yakhini, L. Anavy, and E. Yaakobi, "Error-correcting codes for combinatorial composite DNA," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 109–114.

[45] F. Walter, O. Sabary, A. Wachter-Zeh, and E. Yaakobi, "Coding for composite DNA to correct substitutions, strand losses, and deletions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 97–102.

[46] F. Walter and Y. Yehezkeally, "Coding for strand breaks in composite DNA," 2025, *arXiv:2501.15851*.

[47] L. Anavy, I. Vaknin, O. Atar, R. Amit, and Z. Yakhini, "Data storage in DNA with fewer synthesis cycles using composite DNA letters," *Nature Biotechnol.*, vol. 37, no. 10, pp. 1229–1236, Oct. 2019.

[48] M. Kovačević and V. Y. F. Tan, "Coding for the permutation channel with insertions, deletions, substitutions, and erasures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1933–1937.

[49] 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, Jul. 2018.

[50] A. Makur, "Coding theorems for noisy permutation channels," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 6723–6748, Nov. 2020.

[51] J. Tang and Y. Polyanskiy, "Capacity of noisy permutation channels," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4145–4162, Jul. 2023.

[52] W. Lu and A. Makur, "Permutation capacity region of adder multiple-access channels," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 4693–4720, Jul. 2024.

[53] Y. Gerzon, I. Shomorony, and N. Weinberger, "On the capacity of noisy frequency-based channels," 2026, *arXiv:2601.10329*.

[54] R. Tamir, N. Weinberger, and A. Guillén i Fàbregas, "Concatenated codes for short-molecule DNA storage with sequencing channels of positive zero-undetected-error capacity," 2026, *arXiv:2602.12800*.
[55] N. Balakrishnan and V. B. Nevzorov, *A Primer on Statistical Distributions*. Hoboken, NJ, USA: Wiley, 2004.
[56] T. S. Han, *Information-Spectrum Methods in Information Theory*. Cham, Switzerland: Springer, 2002.
[57] C. Mortici, "Sharp inequalities related to Gosper's formula," *Comp. Rendus. Mathématique*, vol. 348, nos. 3–4, pp. 137–140, Feb. 2010.
[58] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Stat. Rev.*, vol. 70, no. 3, pp. 419–435, Dec. 2002.
[59] L. Gordon, "A stochastic approach to the gamma function," *Amer. Math. Monthly*, vol. 101, no. 9, pp. 858–865, Nov. 1994.
[60] D. I. Cartwright and M. J. Field, "A refinement of the arithmetic mean-geometric mean inequality," *Proc. Amer. Math. Soc.*, vol. 71, no. 1, pp. 36–38, Aug. 1978.
[61] E. Pascal, "Sopra una formola di combinazione," *Giornale di Matematiche di Battaglini*, vol. 25, pp. 45–49, May 1887.
[62] D. E. Knuth, "Combinatorial generation, section 7.2.1.3: Generating all combinations," in *The Art of Computer Programming, Volume 4A: Combinatorial Algorithms, Part 1*. Reading, MA, USA: Addison-Wesley, 2011, pp. 355–372.
[63] A. Genitrini and M. Pépin, "Lexicographic unranking of combinations revisited," *Algorithms*, vol. 14, no. 3, p. 97, Mar. 2021.

Ran Tamir (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, in 2010, 2017, and 2021, respectively. From 2021 to 2023, he was a Post-Doctoral Fellow with ETH Zürich. In 2024, he was a Post-Doctoral Fellow with EPFL. He is currently a Post-Doctoral Fellow with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya.

Nir Weinberger (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) from Tel-Aviv University, Tel-Aviv, Israel, in 2006 and 2009, respectively, and the Ph.D. degree from the Technion–Israel Institute of Technology in 2017. From 2017 to 2018, he was a Post-Doctoral Fellow with Tel Aviv University. From 2018 to 2020, he was a Technion-MIT Post-Doctoral Fellow with Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently an Associate Professor with the Viterbi Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology.

Albert Guillén i Fàbregas (Fellow, IEEE) received the Telecommunications Engineering Degree and the Electronics Engineering Degree from Universitat Politècnica de Catalunya and Politecnico di Torino, respectively in 1999, and the Ph.D. in Communication Systems from École Polytechnique Fédérale de Lausanne (EPFL) in 2004.

He is currently a Professor with the Department of Engineering, University of Cambridge; a fellow of Emmanuel College; and a part-time Researcher with the Universitat Politècnica de Catalunya. He has held appointments with New Jersey Institute of Technology; Telecom Italia; European Space Agency (ESA); Institut Eurécom; the University of South Australia; and the Universitat Pompeu Fabra (an ICREA Research Professor). He also held visiting appointments with EPFL; the École Nationale des Télécommunications, Paris; the Universitat Pompeu Fabra; the University of South Australia; the Centrum Wiskunde and Informatica; and Texas A&M University, Qatar. His research interests are in the areas of information theory, communication theory, coding theory, and statistical inference.

Dr. Guillén i Fàbregas is a fellow of the Institute for Mathematics and its Applications (IMA) and a member of the Young Academy of Europe. He received the Starting, Consolidator, Advanced, and Proof of Concept Grants from European Research Council; the Young Authors Award of the 2004 European Signal Processing Conference; the 2004 Best Doctoral Thesis Award from Spanish Institution of Telecommunications Engineers; and the Research Fellowship from Spanish Government to join ESA. Since 2013, he has been an Editor of *Foundations and Trends in Communications and Information Theory*, Now Publishers. He was an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2007 to 2011 and IEEE TRANSACTIONS ON INFORMATION THEORY from 2013 to 2020.