The Noisy Drawing Channel: Reliable Data Storage in DNA Sequences

Andreas Lenz, Paul H. Siegel, Antonia Wachter-Zeh, and Eitan Yaakobi

***Abstract*—Motivated by recent advances in DNA-based data storage, we study a communication system, where information is conveyed over many sequences in parallel. In this system, the receiver cannot control the access to these sequences and can only draw from these sequences, unaware which sequence has been drawn. Further, the drawn sequences are susceptible to errors. In this paper, a suitable channel model that models this input-output relationship is analyzed and its information capacity is computed for a wide range of parameters and a general class of drawing distributions. This generalizes previous results for the noiseless case and specific drawing distributions. The analysis can guide future DNA-based data storage experiments by establishing theoretical limits on achievable information rates and by proposing decoding techniques that can be useful for practical implementations of decoders.**

I. Introduction

DNA-based data storage is a novel approach for long-term archiving of digital data. It has drawn recent attention due to significant advances in biochemical technologies, such as synthesizing and sequencing of DNA. Manifold experiments [4], [5], [6], [7], [8], [9] have been published in the last decade, addressing many different aspects of digital data storage, such as reliability, lifetime, random-access, and efficiency. At the same time, the unique nature of DNA-based storage systems has fueled theoretical investigations inside a variety of research fields, such as computational biology, coding theory, information theory and signal processing.

The process of writing and reading digital data in DNA-based data storage usually involves three main steps. First, the digital binary data is encoded into many short vectors over the alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, which are then synthesized as DNA strands. In most experiments, each strand is synthesized many times such that multiple copies of each strand are present. Second, those strands are transferred into a storage medium that preserves the chemical structure of DNA and enables robustness over a long period of time. Third and finally, when accessing the data inside the archive, the DNA strands from the storage medium are sequenced. This is often an uncontrollable procedure in the sense that it is not possible to choose which strands are sequenced.[1] Using the sequenced data, a decoder then estimates the original digital data.

These writing and reading processes distinguish DNA-based storage systems from conventional transmission or storage systems in the following aspects. First, the unordered nature of reading is rarely observed in traditional communication systems. Next, due to the uncontrolled drawing of strands, it is possible that some strands are never observed at the output and others might be read multiple times. In this work, we study the noisy drawing channel, which embodies these properties.

A. Related Work

Most information-theoretic studies related to DNA-based data storage discuss insertion and deletion error correction. Classical papers on this topic are those by Gallager [10] and Davey and Mackay [11]. More recently, increased interest towards channel models with multiple transmissions over a channel impaired by insertion and deletion errors arose, due to the existence of multiple copies of each stored strand in DNA-based storage systems. For example, [12], [13], [14] study reconstruction from DNA sequences, where [12], [13] discuss uncoded sequences and [14] focuses on coded sequences. Decoding algorithms and achievable information rates are discussed in [15], [16].

Another related line of research is that on channels that permute several parallel input sequences [17], [18]. In their setup, a given number of parallel sequences is arbitrarily permuted and then transmitted over known constituent channels. In principle such a communication scenario is similar to ours, differs however in the nature of the constituent channels and the knowledge of the receiver about the origin of each sequence. A different type of permutation channels, where the symbols of a single sequence can be permuted, has been discussed in [19], [20].

This work deals with the so-called *noisy drawing* channel that models the pipeline from synthesized to sequenced DNA

This paper was presented in part at the 2020 International Conference on Acoustics, Speech, and Signal Processing [1], at the 2021 Information Theory Workshop [2], and the 2021 Visual Communications and Image Processing [3].

A. Lenz is with the Institute for Communications Engineering, Technische Universität München, Munich 80333, Germany (e-mail: andreas.lenz@mytum.de).

P. H. Siegel is with the Electrical and Computer Engineering Department and the Center for Memory and Recording Research, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: psiegel@ucsd.edu).

A. Wachter-Zeh is with the Institute for Communications Engineering, Technische Universität München, Munich 80333, Germany (e-mail: antonia.wachter-zeh@tum.de).

E. Yaakobi is with the Computer Science Department, Technion – Israel Institute of Technology, Haifa 32000, Israel (e-mail: yaakobi@cs.technion.ac.il).

E. Yaakobi was supported by the Center for Memory and Recording Research, University of California San Diego. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 801434). This work was also supported by NSF Grants CCF-BSF-1619053 and CCF-2212437 and by the United States-Israel BSF grant 2018048.

[1] There are studies [6], [7] that have developed methods for random access and for sequencing of specific strands. This was accomplished by designing primers that are appended to the DNA strand. Here however, we are studying the raw system without the usage of such primers.

strands. The channel has an input of many parallel sequences. Out of these sequences, the receiver draws sequences in a random fashion, oblivious of the origin of the drawn sequences, i.e., the drawing indices. The drawn sequences are observed through an erroneous channel and thus can differ from the input sequences. This channel has previously been studied in several variants. Originally, the channel has been studied for the noiseless case [21], [22] assuming uniform and independent draws of the input sequences. The capacity has been derived for this case and it has been shown [21] that a simple indexing and erasure correction scheme achieves capacity. Later, the capacity for the case where each sequence is drawn exactly once and transmitted over a binary symmetric channel has been derived in [23]. Also in this case, an efficient capacity-achieving scheme has been presented. This scheme indexes each sequence and protects the whole sequence with a capacity-achieving code for the binary symmetric channel. The results in [23] have been extended to the case of transmission over erasure channels [24]. Recently, the capacity has been found for the case where each sequence is drawn according to a Bernoulli distribution and transmitted over a binary symmetric channel [25]. It has also been shown that a concatenated code with an outer erasure code and an inner indexing and error correction code can achieve capacity. In [26], the results from [2], [21] have been generalized to general memoryless channels, together with a derivation of bounds on the decoding error probability. Using novel proof techniques, [26] proved converse and achievability bounds, which hold for arbitrary parameters, which is in contrast to previous work, which focused on a special parameter regime. This established capacity results for a larger set of parameters.

B. Contribution and Outline

In this paper, we study the noisy drawing channel for a broad class of distributions on the drawing indices and for the case of transmission over the q-ary memoryless channel. In particular, we define the notion of *regular* drawing distributions in Definition 1 for which we derive the capacity of the noisy drawing channel in Theorem 4. Our results thus generalize and unify previous results [1], [2], [3], [21], [25] to a much broader class of drawing indices distributions and to arbitrary alphabet sizes. Similar as in previous work, our capacity results hold for the *low-noise* scenario, where the number of sequences is moderate, depending on the channel noise. For a thorough discussion on the parameter range, see Section III-C. Compared to [26], our results are more specific regarding the constituent channel and its parameters, however more general with respect to the drawing distribution. Notice that the proof techniques employed in this manuscript follow the lines of our conference contributions [1], [2], [3], which are different from those presented in [26]. For a more detailed comparison of proof techniques, we refer the reader to [26].

We start by defining the channel model in Section II. This section further contains an equivalent channel model that is both useful for the later derivation and also an intuitive understanding of the channel. We proceed with presenting the result about the capacity of this channel together with definitions regarding codes and reliable transmission over the noisy drawing channel in Section III. The presentation is enriched with a discussion of the parameter range and with thoughts regarding practical code constructions over the channel. Sections IV and V are devoted to proving the capacity result by showing that the capacity is an upper bound on achievable information rates and a proof of the existence of codes with vanishing error probabilities and information rates arbitrarily close to capacity. In Section VI we apply the results to popular drawing distributions, recovering and generalizing the results from [1], [2], [3], [21], [25].

C. Notation

Throughout this paper, we discuss sequences over the finite alphabet $\Sigma_q = \{0, 1, \dots, q-1\}$, where we may write $\Sigma_4 = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ to highlight the DNA alphabet. The set of positive integers up to n is denoted by $[n] = \{1, 2, \dots, n\}$. We use the logarithm $\log_q$ with respect to the base q and $\log$ depicts the binary logarithm. We write random variables in upper case and their realizations in lower case. The probability of an event is denoted as $\mathsf{Pr}\left(X = x\right)$, where we sometimes omit the random variable, e.g., $\mathsf{Pr}\left(x\right)$, when it is clear from the context. The expected value and variance of a random variable X are denoted by $\mathsf{E}\left[X\right]$ and $\mathsf{V}\left[X\right]$. Further, we denote by $H(X) = -\sum_x \mathsf{Pr}\left(x\right) \log_q(\mathsf{Pr}\left(x\right))$ the entropy of a random variable X. The q-ary entropy function is denoted by $H_q(p) = p\log_q(q-1) - p\log_q p - (1-p)\log_q(1-p)$. We highlight vectors with bold font, e.g. $\boldsymbol{x}$, and denote by $|\boldsymbol{x}|$ the number of elements in $\boldsymbol{x}$.

II. Channel Model

The input of the noisy drawing channel is M sequences $\boldsymbol{X}_1, \dots, \boldsymbol{X}_M$ where each $\boldsymbol{X}_i = (X_{i,1}, \dots, X_{i,L}) \in \Sigma_q^L$, $i \in [M]$, is a vector of length L over the alphabet Σ_q. From this input, N sequences are drawn and received with errors. Denote by $\boldsymbol{I} = (I_1, \dots, I_N), I_j \in [M]$ the indices of the draws, i.e., in the j-th draw, the input sequence $\boldsymbol{X}_{I_j}$ is drawn. We consider random drawing indices $\boldsymbol{I}$, which are independent of the input $\boldsymbol{X}_1, \dots, \boldsymbol{X}_M$ and will concretize its distribution in Section II-A. Note that, depending on the distribution of $\boldsymbol{I}$, N may also be random. The output of the channel is then given by N sequences $\boldsymbol{Y}_j = (Y_{j,1}, \dots, Y_{j,L}) \in \Sigma_q^L$, $j \in [N]$, each of length L. Each sequence $\boldsymbol{Y}_j$ is obtained by drawing a random input sequence $\boldsymbol{X}_{I_j}$ and transmitting it over the q-ary symmetric channel with error probability p. That is, the output sequences are given by

$$\boldsymbol{Y}_j = \boldsymbol{X}_{I_j} + \boldsymbol{E}_j,$$

for all $j \in [N]$, where the sum is performed over the finite ring of integers Σ_q, i.e., modulo q. Hereby, $\boldsymbol{E}_j = (E_{j,1}, \dots, E_{j,L})$ are random error vectors with independent and identically distributed entries

$$\mathsf{Pr}\left(E_{j,\ell} = e_{j,\ell}\right) = \begin{cases} 1-p, & \text{if } e_{j,\ell} = 0 \\ \frac{p}{q-1}, & \text{if } e_{j,\ell} \neq 0 \end{cases}$$

for all $j \in [N]$ and $\ell \in [L]$ that are independent of the input $\boldsymbol{X}_1, \dots, \boldsymbol{X}_M$ and the drawing indices $\boldsymbol{I}$. For convenience, we stack all input and output sequences to matrices $\boldsymbol{X} =$

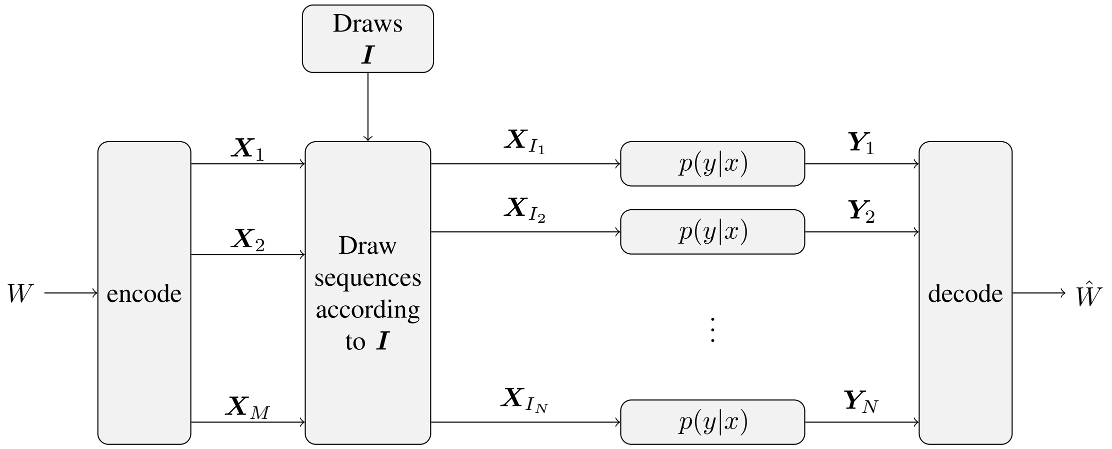

Fig. 1: Visualization of the transmission scheme over the noisy drawing channel. A message W is encoded into a total of M transmit sequences $\boldsymbol{X}_1, \dots, \boldsymbol{X}_M$. Out of these M sequences, N are drawn according to the random drawing indices $\boldsymbol{I}$. The resulting vectors are transmitted over parallel q-ary symmetric channels, resulting in the channel output $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N$.

$(\boldsymbol{X}_1, \dots, \boldsymbol{X}_M) \in \Sigma_q^{M \times L}$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N) \in \Sigma_q^{N \times L}$, such that each sequence is a row of the corresponding matrix. Hence, the input-output relationship can be summarized as

$$\begin{aligned}\mathsf{Pr}\left(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}\right) &= \sum_{\boldsymbol{i}} \mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right) \mathsf{Pr}\left(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{i}\right) \\ &= \sum_{\boldsymbol{i}} \mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right) \prod_{j=1}^{|\boldsymbol{i}|} p(\boldsymbol{y}_j | \boldsymbol{x}_{i_j}),\end{aligned}$$

where $p(\boldsymbol{y}_j | \boldsymbol{x}_{i_j})$ is according to the q-ary symmetric channel described above. Here and in the following, we refer to the sets $\{\boldsymbol{Y}_j : I_j = i\}$, $i \in [M]$ of sequences, which are obtained from the same input sequence as *clusters*. Figure 1 illustrates the transmission scheme over the noisy drawing channel.

A. Drawing Indices, Drawing Composition, and Drawing Frequency

The distribution of the *drawing indices* $\boldsymbol{I}$ is an important aspect of the noisy drawing channel and we continue by defining relevant random variables associated with the drawing indices. Throughout this paper, we use the term *drawing composition* $\boldsymbol{D} = (D_1, \dots, D_M)$ with $D_i = |\{j \in [N] : I_j = i\}|$, $i \in [M]$ for the variables, which count the number of times an input sequence i has been drawn and the term *drawing frequency* $\boldsymbol{N} = (N_0, N_1, \dots)$ with $N_d = |\{i \in [M] : D_i = d\}|$, $d \in \{0, \dots, N\}$ for the variables which count the number of input sequences that have been drawn d times. With this definition,

$$N = \sum_{i=1}^{M} D_i = \sum_{d \geq 0} d N_d$$

and

$$M = \sum_{d \geq 0} N_d.$$

Since $\boldsymbol{I}$ is a random variable, so are the drawing composition $\boldsymbol{D}$ and the drawing frequency $\boldsymbol{N}$. The distributions of $\boldsymbol{N}$ and $\boldsymbol{D}$ can directly be derived from that of $\boldsymbol{I}$.

We impose the following three restrictions on the distribution of $\boldsymbol{I}$ for our capacity result that will both simplify the derivation of the bounds and ensure that the involved quantities are well-defined. The restrictions are as follows.

Definition 1. *Let* $\mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right)$ *be a given family of probability mass functions*[2] *for the drawing indices and denote by* $\boldsymbol{D}$ *and* $\boldsymbol{N}$ *the derived drawing composition and drawing frequency. We say that the distribution* $\mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right)$ *is* regular *if it fulfills the following conditions.*

1) Permutation invariance: *The distributions* $\mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right)$ *and* $\mathsf{Pr}\left(\boldsymbol{D} = \boldsymbol{d}\right)$ *are invariant over permutations of the vectors* $\boldsymbol{i}$ *and* $\boldsymbol{d}$*.*
2) Frequency convergence: *The distribution converges to* $\boldsymbol{\nu} = (\nu_0, \nu_1, \dots)$*,* $\nu_d \in \mathbb{R}$*,* $d \geq 0$ *in frequency, i.e., for every* $\epsilon > 0$*,*
$$\lim_{M \to \infty} \mathsf{Pr}\left(\sum_{d \geq 0} \left|\frac{N_d}{M} - \nu_d\right| > \epsilon\right) = 0.$$
3) Bounded draws: *There exists some constant* $c \in \mathbb{R}$ *such that for all* M*,*
$$\mathsf{Pr}\left(N \leq cM\right) = 1.$$

We concisely highlight the main background and consequences of the constraints imposed in Definition 1.

First, the distributions $\mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right)$ and $\mathsf{Pr}\left(\boldsymbol{D} = \boldsymbol{d}\right)$ are invariant to permutations. Under this constraint, both the index j of an output sequence $\boldsymbol{y}_j$ and the cluster sizes do not reveal anything about the origin of an output sequence. Technically, the uniformity of $\mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right)$ over permutations ensures that the alternative channel model, which will be presented in Section II, is equivalent to the noisy drawing channel. The permutation invariance of $\mathsf{Pr}\left(\boldsymbol{D} = \boldsymbol{d}\right)$ entails the desirable property that the assignment of output sequences with input sequences is independent of the cluster sizes, which will be used in Lemma 7.

[2] We refer here to a *family* of probability mass functions, as the function $\mathsf{Pr}\left(\boldsymbol{I} = \boldsymbol{i}\right)$ may vary with M. This dependence is omitted for reasons of readability.

Both restrictions are essential for the capacity result to hold, as otherwise a receiver could infer information about the origin of output sequences by observing cluster sizes or sequence indices, which may result in a larger channel capacity.

Second, we restrict the relative drawing frequencies $\frac{N_d}{M}$ to converge to a deterministic value. This is a key requirement of our analysis and is reflected through the appearance of the limits $\boldsymbol{\nu}$ in the capacity expression in Theorem 4. The main effect of this property is that the overall channel quality converges to a deterministic value and thus the capacity may be expressed by a weighted sum of channel appearances and their capacity. The lifting of this restriction will likely change the capacity expression and an analysis may require techniques used in probabilistically varying channels.

Third, the total number of draws is deterministically at most cM for a constant $c \in \mathbb{R}$. This is a technical requirement, which simplifies the analysis at several instances, in particular the proof of Lemma 8 and Lemma 14. The useful consequence for Lemma 8 is that we can use trivial bounds on the entropy of some sequences for rare events without an asymptotic rate loss. On the other hand, the achievability proof is simplified, as this restriction forbids the appearance of too many output sequences, which may hinder clustering of sequences. It is conceivable that this restriction is not fundamental to the channel and may be lifted with a different proof technique.

Remark 2. *We note that Definition 1 includes previously studied drawing distributions considered in [21], [23], [24], [25], [26]. Skew sampling distributions, which are not permutation invariant, such as the Poisson sampling distribution, however are not covered in this definition and require further analysis. A generalization to such distributions appears to be possible with the same approach as in this paper. However, it further requires computing the amount of information which may be deduced about the origin of a sequence through observing its draw index and cluster size.*

B. Equivalent Channel using Multinomial Channels

We proceed with introducing an equivalent channel model that will be helpful for both, an intuitive understanding and the derivation of our analytical results. Due to the randomness of the drawing indices, it is possible that the receiver obtains multiple sequences that originate from the same input sequence. To reflect this behavior, consider the following reformulation of the input-output relationship using the drawing composition $\boldsymbol{D}$.

$$\mathsf{Pr}\left(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x}\right) = \sum_{\boldsymbol{i},\boldsymbol{d}} \mathsf{Pr}\left(\boldsymbol{d}\right)\mathsf{Pr}\left(\boldsymbol{i}|\boldsymbol{d}\right)\prod_{i=1}^{M}\prod_{j:i_j=i} p(\boldsymbol{y}_j|\boldsymbol{x}_i).$$

Here, the sets $\{j \in [N] : i_j = i\}$ have size d_i and $\prod_{j:i_j=i} p(\boldsymbol{y}_j|\boldsymbol{x}_i)$ is the conditional probability mass function of a channel that has input $\boldsymbol{X}_i$ and d_i output sequences, each resulting from transmitting the same sequence $\boldsymbol{X}_i$ over a q-ary symmetric channel. This motivates to introduce an equivalent channel presented in Figure 2. In this regard, we denote by $\boldsymbol{Z}_i$ the cluster containing the output sequences that originate from $\boldsymbol{X}_i$, for $i = 1, \ldots, M$. Here $\boldsymbol{Z}_i = (\boldsymbol{Y}_j : I_j = i) \in \Sigma_q^{D_i \times L}$ stems from the D_i-repeated transmission of $\boldsymbol{X}_i$ over a q-ary symmetric channel. The output clusters are permuted by a uniformly random permutation $\boldsymbol{S}$, resulting in the permuted clusters $\boldsymbol{Z}_i'$ and the sequences $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ are then obtained by a permutation of the individual sequences of the clusters. Notice that the permutation of the clusters is immaterial to the input-output relationship due to the uniformly random second permutation of all sequences. However, we introduce this permutation, since the channel from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$ to $\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_M'$ will be used for our converse bound later. The overall channel is equivalent to the noisy drawing channel as derived above due to the fact that $\mathsf{Pr}\left(\boldsymbol{i}|\boldsymbol{d}\right)$ is permutation invariant, as required in Definition 1. Note that the splitting into two separate permutations of clusters and individual sequences has no additional effect on the input-output relationship of this channel, however it will be useful for the derivation of our results later.

The individual channels of repeated transmissions have been discussed first by Mitzenmacher [27] for binary inputs under the name of the *binomial channel*. Here we refer to the channel as the *multinomial channel* as for q-ary input alphabets, the channel law follows a multinomial distribution. The multinomial channel is a discrete memoryless channel with letter-wise input $X \in \Sigma_q$ and output[3] $Z = (Z_1, \ldots, Z_d) \in \Sigma_q^d$, where d is the number of *draws* of the input X. Each Z_i is obtained by transmitting X repeatedly and independently over a q-ary symmetric channel with error probability p. The conditional channel probability of this channel is denoted by $p_d(z|x)$. We will derive the capacity of this channel in Lemma 22 in Appendix A.

III. Capacity of the Noisy Drawing Channel

The *capacity* of a channel characterizes the exact region of code rates for which reliable communication is possible. As introduced by Shannon [28], the capacity of a probabilistic channel is the supremum of code rates for which transmission with vanishing error probability is feasible. In the following, we first specify the notion of code rates and error probabilities over the noisy drawing channel and then proceed with stating our main result about its capacity.

A. Error-Correcting Codes

We start with an introduction of error-correcting codes and their rates. The input of the channel is the sequences $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$, each of length L. Thus, a code is a set $\mathcal{C} \subseteq \Sigma_q^{M \times L}$ such that each codeword consists of M sequences, each of length L over the alphabet Σ_q. Consequently, the *rate* of a code $\mathcal{C} \subseteq \Sigma_q^{M \times L}$ is given by

$$R = \frac{\log_q |\mathcal{C}|}{ML}.$$

Each code $\mathcal{C}$ is equipped with an encoder $\mathsf{enc} : [q^{MLR}] \mapsto \mathcal{C}$ that maps a message $W \in [q^{MLR}]$ to a codeword and a decoder

$$\mathsf{dec} : \bigcup_{n \geq 0} \Sigma_q^{n \times L} \mapsto [q^{MLR}]$$

[3] Note that although strictly speaking, Z is a vector of length d, we view Z as a symbol of the output alphabet Σ_q^d and thus do not highlight it in bold.

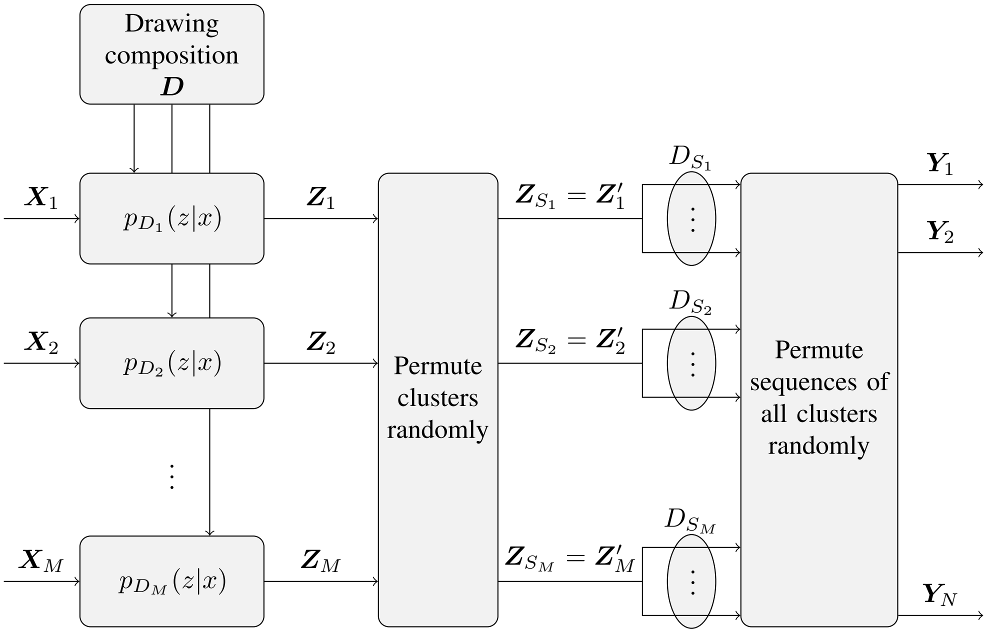

Fig. 2: The noisy drawing channel as the cascading of parallel multinomial channels followed by a permutation of clusters and a joint re-indexing of all sequences within the clusters. Each cluster $\boldsymbol{Z}_i = (\boldsymbol{Y}_j : I_j = i)$ consists of D_i individual sequences. The $N = D_1 + \cdots + D_M$ individual sequences of all clusters are permuted, resulting in the output sequences $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$.

that outputs an estimate $\widehat{W}$ of the original message W given the received sequences $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$. The error probability of a code $\mathcal{C} \subseteq \Sigma_q^{M \times L}$ and a decoder dec is given by

$$\mathsf{Pr}\left(\mathsf{Err}|\mathcal{C}\right) = \frac{1}{q^{MLR}} \sum_{w=1}^{q^{MLR}} \mathsf{Pr}\left(\mathsf{dec}(\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_N) \neq w | W = w\right),$$

where $\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_N$ is the random result of transmitting $\mathsf{enc}(W) = (\boldsymbol{X}_1, \ldots \boldsymbol{X}_M)$ over the noisy drawing channel and Err is the event that $\mathsf{dec}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) \neq W$. Here we assumed that the messages are chosen uniformly from the set of all messages $W \in [q^{MLR}]$, i.e., $\mathsf{Pr}\left(W = w\right) = \frac{1}{q^{MLR}}$.

B. Noisy Channel Coding Theorem

To characterize the channel capacity and achievable rates, it is necessary to specify how the channel parameters scale relative to each other. We consider the regime, where q, p are fixed and $M \to \infty$ and $M = q^{\beta L}$ for some fixed $0 < \beta < 1$. This choice is motivated by the following two facts. First, the case where M is exponential in L is the interesting case, as for $M = q^{\beta L}$ with $\beta > 1$ it has been shown in [21] that no positive rate can be achieved (even in the error-free case) and for the case where M is subexponential in L, the rate loss of indexing is asymptotically vanishing, which essentially removes the unordered nature of the sequencing. Second, this parameter regime is practically relevant for the case, where one wishes to transmit many relatively short sequences, as is the case in DNA-based archival storage. We use the standard notion of achievable rates and channel capacity over the noisy drawing (ND) channel as follows.

Definition 3. *Let $0 < \beta < 1, 0 < p < 1$, $q \in \mathbb{N}$ be fixed and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a regular distribution that converges in frequency to $\boldsymbol{\nu}$. Then, a code rate R is* achievable*, if there exists a family of codes $\mathcal{C}(M, L) \subseteq \Sigma_q^{M \times L}$ with $|\mathcal{C}(M, L)| = q^{RML}$ together with a decoder that has vanishing error probability $\mathsf{Pr}\left(\mathsf{Err}|\mathcal{C}(M, L)\right) \to 0$ as $M \to \infty$, where $M = q^{\beta L}$.*

The Shannon capacity $C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q)$ is the supremum over the set of achievable rates.

With this definition, for any code rate $R < C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q)$ there exists a family of codes with rate R that has vanishing error probability as $M \to \infty$. Conversely, every family of codes with code rate $R > C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q)$ has a non-vanishing error rate. With these prerequisites we are in the position to formulate the main theorem on the capacity of the noisy drawing channel. Recall to this end the definition of regularity for the probability mass function of the drawing indices $\boldsymbol{i}$ from Definition 1, which implies convergence in distribution and a bounded number of draws.

Theorem 4. *Let $\beta > 0, q \in \mathbb{N}$, $0 < p < \frac{q-1}{2q}$ be fixed parameters satisfying $2\beta < 1 - H_q(2p)$ and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a given regular distribution that converges in frequency to $\boldsymbol{\nu}$. Then, the capacity of the noisy drawing channel is given by*

$$C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q) = \sum_{d \geq 0} \nu_d C_{\mathsf{Mul}}(d, p, q) - \beta(1 - \nu_0).$$

We prove Theorem 4 in Sections IV and V.

C. Parameter Range

Theorem 4 holds for the range of the parameters $2\beta < 1 - H_q(2p)$ and $p < \frac{q-1}{2q}$. This parameter restriction appears in the the derivation of the converse and achievability bound and is caused due to the following.

As we will elaborate in Section V, our proof for achievability relies on a decoder that clusters sequences based on their

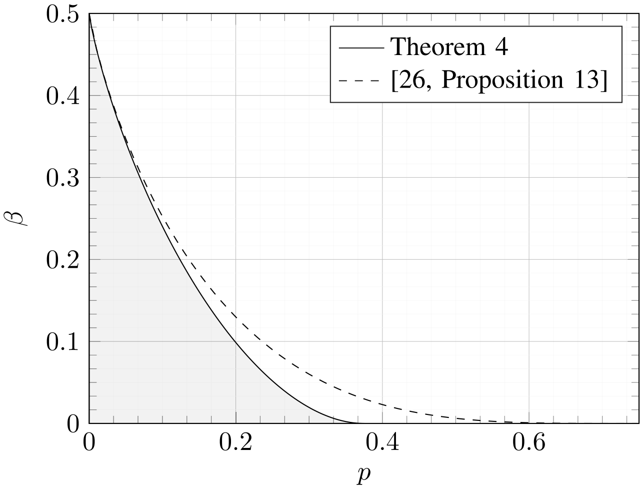

Fig. 3: Region of channel parameters, for which Theorem 4 holds, when $q = 4$. For comparison, we show the region for which the capacity was established in [26]. Common parameters of current DNA-based storage experiments are $\beta < 0.1$ and $p < 0.1$, which lie well within our region.

Hamming distance. This clustering however is only accurate if the sequences form distinguishable clusters, i.e., sequences within a cluster are close as compared to sequences from different clusters. Using sphere-hardening arguments, similar to those used in classical coding theorems [29], the clustering can be performed accurately if the total number sequences is less than the number of input sequences times the potential volume of a cluster. Assuming an expected maximum relative distance of $2p$ between any two sequences within a cluster, and thus a cluster diameter of $2p$, gives the inequalities $\beta < 1 - H_q(2p)$ and $p < \frac{q-1}{2q}$.

On the other hand, the converse bound is derived by finding an upper bound on the mutual information between input and output. In our analysis, we show that under the condition $2\beta < 1 - H_q(2p)$ input distributions that favor well separated clusters maximize mutual information, resulting in matching converse and achievability bounds and thus establishing capacity within this region.

Compared to [25], we obtain the same parameter range. In comparison with [26], where the capacity is derived for $2\beta < 1 - H_q(2p - p^2 \frac{q}{q-1})$ and any p, see [26, Proposition 13], Theorem 4 holds for a smaller range of parameters β, p. The results in [26] are however restricted to independent and uniformly distributed draws. For $q = 4$, we visualize the parameter ranges of Theorem 4 and [26] in Figure 3.

D. Extension to Other Component Channels

The analysis in this paper is focused on the case, where each output sequence is observed through the q-ary symmetric channel with crossover probability p. While the extension to other classes of channels is non-trivial in general[4], one may deduce some insights on more general channels from our analysis. For example the capacity in Theorem 4 may serve as a lower bound on the capacity of an asymmetric memoryless channel, for which a symbol is received correctly with probability $1 - p$. This is because we derive achievable rates with a decoder that uses computations based on the Hamming metric and sequence typicality, which may be conservatively bounded with a symmetric channel. However, since the channel's asymmetry may produce a bias in output sequences, the bound on β needs to be reevaluated for such a channel.

Our analysis further suggests that Theorem 4 may, within an appropriate parameter regime, generalize to channels with memory, such as insertion and deletion channels by replacing the capacity of the multinomial channel with the capacity of the multiple draw insertion and deletion channel in Theorem 4. This conjecture is substantiated by that fact that the main channel property that our analysis requires is the forming of typical sequences and distinguishable clusters. Thus, even without the knowledge of the exact individual capacity expressions, one may prove a corresponding coding theorem.

E. Practical Aspects for Code Design with Rates Approaching Capacity

Interestingly, in contrast to the information-theoretic results, it is still an open problem to find efficiently encodable and decodable schemes that achieve capacity on the noisy drawing channel. This is mainly due to the apriori incertitude how often each input sequence is drawn combined with the loss of ordering of sequences. We will break down these two aspects and existing solutions for each of the aspects in the following.

Channel uncertainty: The amount of information that a receiver may deduce about an input sequence increases with the number of times the input sequence is observed at the output. Since this number is random in the noisy drawing channel, the encoder cannot choose appropriate code rates for each sequence in advance. Thus, in order to operate close to capacity, the input strands must be coded with appropriate cross-correlation such that input sequences with more draws may help in the decoding of those with less (or no) draws. For the case, where the ordering of the output sequences is known to the receiver, rate-matching codes [17], [18] provide a solution to construct this correlation. Explicit constructions of rate-matching codes exist [17], for example based on erasure codes.

Loss of ordering: Through the random drawing of sequences, the receiver has no immediate information about how the output sequences may be associated with input sequences. This loss of ordering can be combat with indexing, i.e., each input sequence $\boldsymbol{X}_i$ is prepended with a field that designates its index i. However, due to channel noise, also these indices require appropriate protection from errors. As explained in the previous paragraph, indices of sequences with more draws are easier to decode, which implies that also the efficient decoding of the indices requires rate-matching techniques.

The crux of the noisy drawing channel is that the combination of these two techniques in an efficient manner is non-trivial. On the one hand, the rate-matching techniques require a correct

[4] The case of general discrete memoryless channels has been treated in [26] using novel techniques in their analysis.

ordering of sequences, on the other hand an efficient decoding of the indices requires rate-matching.

For the case of Bernoulli drawing compositions this code design issue could be elegantly solved [23] using a scheme, which equips each sequence with an index and a capacity-achieving code on the q-ary symmetric channel, together with an outer erasure code. Here, the erasure code takes the roll of the rate-matching code and a rate-matching decoding of the indices is not necessary, as sequences may be drawn at most once. For drawing distributions with more than one draw per sequence, it remains however an open problem to design codes that protect against channel uncertainty and loss of ordering. One possible solution could be the usage of rate-matching techniques that do not require knowledge of the sequence ordering.

IV. Converse Bound

We start with a short overview of the ideas, which will be used in our proof of the converse bound.

The starting point for the bound is Fano's inequality [30], which implies an upper bound on achievable code rates by means of the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$. Using the equivalent channel model from Section II, we simplify the computation of the mutual information using $I(\boldsymbol{X};\boldsymbol{Y}) \leq I(\boldsymbol{X};\boldsymbol{Z}')$. The main difficulty when deriving upper bounds on $I(\boldsymbol{X};\boldsymbol{Z}')$ is the non-trivial dependence of both the output entropy $H(\boldsymbol{Z}')$ and the conditional entropy $H(\boldsymbol{Z}'|\boldsymbol{X})$ on the input distribution $\mathsf{Pr}\left(\boldsymbol{X}\right)$. This dependence arises due to the different effect the permutation of the channel has on sets of input sequences, which are either little or well distributed in Hamming distance. In the subsequent analysis we show that for moderate noise, input distributions that favor well separated input sequences give the largest mutual information. As the main technique, we use an approach similar to that in [23], which introduced a statistic that characterizes the similarity of the output sequences by the means of the largest subset of sequences with a certain minimum Hamming distance. We adapt this statistic for the case of the noisy drawing channel in Definition 6.

The converse bound for the noisy drawing channel is formulated in the following lemma.

Lemma 5. *Let $\beta > 0, q \in \mathbb{N}$, $0 < p < \frac{q-1}{2q}$ be fixed parameters satisfying $2\beta < 1 - H_q(2p)$ and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a regular drawing distribution that converges in frequency to $\boldsymbol{\nu}$. Then, any achievable rate R over the noisy drawing channel satisfies*

$$R \leq C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q).$$

Proof. Let $\mathcal{C} \subseteq \Sigma_q^{M \times L}$ be a code of rate $R = \frac{\log_q |\mathcal{C}|}{ML}$. The code $\mathcal{C}$ has an encoder enc and decoder dec. Denote by $W \in [q^{MLR}]$ a uniformly random message to be transmitted over the channel and $\widehat{W} = \mathsf{dec}(\boldsymbol{Y})$ the output of the decoder, where $\boldsymbol{Y}$ is the result of transmitting $\boldsymbol{X} = \mathsf{enc}(W)$ over the noisy drawing channel. The error probability of this scheme is $\mathsf{Pr}\left(\mathsf{Err}|\mathcal{C}\right) = \mathsf{Pr}\left(W \neq \widehat{W}\right)$ and Fano's inequality implies that

$$R \leq \mathsf{Pr}\left(\mathsf{Err}|\mathcal{C}\right) R + \frac{1 + I(\boldsymbol{X};\boldsymbol{Y})}{ML}.$$

Here we can use Lemma 7, that we will derive in the sequel, which gives an upper bound on the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ to obtain

$$R \leq C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q) + \mathsf{Pr}\left(\mathsf{Err}|\mathcal{C}\right) R + o(1),$$

as $M \to \infty$. By definition of achievable rate, the error probability $\mathsf{Pr}\left(\mathsf{Err}|\mathcal{C}\right)$ has to approach 0, as $M \to \infty$, and we have that any achievable rate R satisfies

$$R \leq C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q).$$

□

A. Bound on Mutual Information

We proceed with bounding the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ from above in a step-by-step fashion. In order to prove our statements, we work with the equivalent channel presented in Section II-B. The following statistic of the output sequences is the key ingredient for deriving an analytically tractable upper bound on the entropy terms of the mutual information.

Definition 6. *Consider the permuted clusters $\boldsymbol{Z}'$ introduced in Section II-B. Denote by $D_i' = D_{S_i}$ the number of draws of the i-th output cluster $\boldsymbol{Z}_i'$. Write each $\boldsymbol{Z}_i'$ as*

$$\boldsymbol{Z}_i' = \begin{pmatrix} \boldsymbol{Z}_{i,1}' \\ \vdots \\ \boldsymbol{Z}_{i,D_i'}' \end{pmatrix},$$

such that each $\boldsymbol{Z}_{i,j}'$ corresponds to one draw of the multinomial channel. We define $\mathcal{U} \subseteq [M]$ to be the largest subset of $[M]$ such that

1) *For all $i \in \mathcal{U}$: $D_i' > 0$.*
2) *For all $i, j \in \mathcal{U}$ with $i \neq j$: $d_{\mathsf{H}}\left(\boldsymbol{Z}_{i,1}', \boldsymbol{Z}_{j,1}'\right) > \alpha L$.*

If the largest subset is not unique, we choose the first according to some (arbitrary) ordering of subsets. We further denote the conditional expectation of $|\mathcal{U}|$ given $\boldsymbol{D}' = \boldsymbol{d}'$ by $U_{\boldsymbol{d}'} \triangleq \mathsf{E}\left[|\mathcal{U}|\ |\boldsymbol{D}' = \boldsymbol{d}'\right]$.

Note that only the size of $\mathcal{U}$ will be of importance later, and we can choose $\mathcal{U}$ arbitrarily (but deterministic) in the case of ties between several subsets. We start with a short explanation of how $\mathcal{U}$ is used to bound the output entropy $H(\boldsymbol{Z}')$. The main idea is the following. Conceptually, in Lemma 8 we will split the output into two types of clusters. Those, which are contained in $\mathcal{U}$ and those, which are not contained in $\mathcal{U}$. Then, with some careful analysis, the entropy of the *free* clusters in $\mathcal{U}$ is bounded simply by the sum of maximum output entropies of the corresponding multinomial channels. On the other hand, the entropy of those clusters, which are not in $\mathcal{U}$ can be bounded more severely, as their first sequence has to be close to at least one of the sequences in the clusters in $\mathcal{U}$, resulting in a smaller entropy as compared to the free clusters. This means that, if the input distribution is chosen such that it favors sequences that are close in Hamming distance, which corresponds to the case of small $\mathcal{U}$, also the bound on the output entropy $H(\boldsymbol{Z}')$ will be smaller. Note that there are a couple of subtleties that need to be overcome when rigorously applying such an argument. One important difference with respect to the derivation of [25] is that, the entropy of the non-free clusters is not trivially bounded by L and we thus apply careful combinatorial arguments using Lemma 23.

Lemma 7. *Let $\beta > 0, q \in \mathbb{N}$, $0 < p < \frac{q-1}{2q}$ be fixed parameters satisfying $2\beta < 1 - H_q(2p)$ and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a given regular distribution that converges in frequency to $\boldsymbol{\nu}$. Then,*

$$I(\boldsymbol{X};\boldsymbol{Y}) \leq MLC_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q) + o(ML).$$

Proof. We start by using the data processing inequality to obtain

$$I(\boldsymbol{X};\boldsymbol{Y}) \leq I(\boldsymbol{X};\boldsymbol{Z}').$$

In the next steps, we incorporate the permutation $\boldsymbol{S}$ into the mutual information $I(\boldsymbol{X};\boldsymbol{Z}')$. To start with, we have by the definition of mutual information

$$I(\boldsymbol{X};\boldsymbol{Z}') = H(\boldsymbol{Z}') - H(\boldsymbol{Z}'|\boldsymbol{X}).$$

The draw variable $\boldsymbol{D}'$ is a function of $\boldsymbol{Z}'$, as we can directly infer it from the size of the clusters, and we thus can compute the mutual information by

$$\begin{aligned}I(\boldsymbol{X};\boldsymbol{Z}') &= H(\boldsymbol{Z}',\boldsymbol{D}') - H(\boldsymbol{Z}',\boldsymbol{D}'|\boldsymbol{X})\\ &= H(\boldsymbol{Z}'|\boldsymbol{D}') + H(\boldsymbol{D}') - H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{D}') - H(\boldsymbol{D}'|\boldsymbol{X})\\ &= H(\boldsymbol{Z}'|\boldsymbol{D}') - H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{D}') = I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{D}').\end{aligned}$$

Hence, the condition on $\boldsymbol{D}'$ does not change the mutual information. On the other hand we can express the conditional mutual information as

$$\begin{aligned}I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{D}') &\overset{(a)}{=} H(\boldsymbol{Z}'|\boldsymbol{D}') - H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{S},\boldsymbol{D}')\\ &\quad - H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{D}') + H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\boldsymbol{D}')\\ &\overset{(b)}{=} H(\boldsymbol{Z}'|\boldsymbol{D}') - H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{S},\boldsymbol{D}')\\ &\quad + H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\boldsymbol{D}') - M\log_q M + O(M),\end{aligned}$$

where we applied the chain rule of entropy twice in equality (a). In equality (b), it has been used that $H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{D}') = H(\boldsymbol{S}) = M\log_q M + O(M)$ due to the permutation invariance of $\boldsymbol{D}$. Expanding the condition on $\boldsymbol{D}'$, we obtain

$$\begin{aligned}I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{D}') = &\sum_{\boldsymbol{d}'}\mathsf{Pr}\left(\boldsymbol{d}'\right)(H(\boldsymbol{Z}'|\boldsymbol{d}') - H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{S},\boldsymbol{d}')\\ &+ H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\boldsymbol{d}')) - \beta ML + O(M).\end{aligned}$$

Recall from Definition 6, the notation $U_{\boldsymbol{d}'} \triangleq \mathsf{E}\left[|\mathcal{U}|\,|\boldsymbol{D}' = \boldsymbol{d}'\right]$ for conditional expectation of the size of the random variable $\mathcal{U}$. Plugging in the bounds on the conditional entropy terms $H(\boldsymbol{Z}')$ and $H(\boldsymbol{Z}'|\boldsymbol{X})$ from Lemma 8, Lemma 9 and 10, we obtain

$$\begin{aligned}&H(\boldsymbol{Z}'|\boldsymbol{d}') - H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{S},\boldsymbol{d}') + H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\boldsymbol{d}') - \beta ML\\ &\leq L\sum_{d\geq 0} n_d C_{\mathsf{Mul}}(d,p,q) - U_{\boldsymbol{d}'}\log_q U_{\boldsymbol{d}'} + L\sum_{d\geq D} n_d\\ &\quad + (M - n_0 - U_{\boldsymbol{d}'})(\log_q U_{\boldsymbol{d}'} + L(H_q(\alpha) - 1)) + o(ML), \quad (1)\end{aligned}$$

for any $\alpha > 2p$, $D \in \mathbb{N}$ and large enough M. Denote by $f(U_{\boldsymbol{d}'})$ the terms in the mutual information expression (1) that do not vanish and depend on $U_{\boldsymbol{d}'}$, i.e.,

$$\begin{aligned}f(U_{\boldsymbol{d}'}) = &(M - n_0 - U_{\boldsymbol{d}'})(\log_q U_{\boldsymbol{d}'} + L(H_q(\alpha) - 1))\\ &- U_{\boldsymbol{d}'}\log_q U_{\boldsymbol{d}'}.\end{aligned}$$

Taking the derivative with respect to $U_{\boldsymbol{d}'}$, we see that

$$\begin{aligned}f'(U_{\boldsymbol{d}'}) = &-(\log_q U_{\boldsymbol{d}'} + L(H_q(\alpha) - 1))\\ &+ \log_q(\mathrm{e})\frac{M - n_0 - U_{\boldsymbol{d}'}}{U_{\boldsymbol{d}'}} - \log_q U_{\boldsymbol{d}'} - \log_q(\mathrm{e})\\ &> L(1 - H_q(\alpha))) - 2\log_q U_{\boldsymbol{d}'} - \log_q(\mathrm{e}).\end{aligned}$$

Therefore, $f'(U_{\boldsymbol{d}'}) > 0$ if

$$U_{\boldsymbol{d}'} < \mathrm{e}^{-\frac{1}{2}} q^{\frac{L(1-H_q(\alpha))}{2}} = \mathrm{e}^{-\frac{1}{2}} M^{\frac{1-H_q(\alpha)}{2\beta}}.$$

Hence, if $2\beta < 1 - H_q(\alpha)$, the exponent of M is larger then 1 and $f'(U_{\boldsymbol{d}'}) > 0$ for all $0 \leq U_{\boldsymbol{d}'} \leq M$, provided that M is large enough. This means that $f(U_{\boldsymbol{d}'})$ is strictly increasing and using further $U_{\boldsymbol{d}'} \leq M - n_0$, as $\mathcal{U}$ consists of sequences, which have been drawn at least once, we obtain for $2\beta < 1 - H_q(\alpha)$ and large enough M,

$$f(U_{\boldsymbol{d}'}) \leq f(M - n_0) = -(M - n_0)\log_q(M - n_0).$$

We proceed with introducing the event $\mathcal{N}_\epsilon$ for an arbitrary $\epsilon > 0$ as the event on the random variable $\boldsymbol{D}'$ that $\sum_{d\geq 0}\left|\frac{N_d}{M} - \nu_d\right| \leq \epsilon/4$. Splitting the sum over $\boldsymbol{d}'$ in the computation of $I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{D}')$ according to this event and choosing $D \triangleq D_\epsilon$ as the smallest integer such that $\sum_{d\geq D_\epsilon}\nu_d \leq \epsilon$, we obtain

$$\begin{aligned}I(\boldsymbol{X};&\boldsymbol{Z}'|\boldsymbol{D}') = \sum_{\boldsymbol{d}'}\mathsf{Pr}\left(\boldsymbol{d}'\right) I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{D}' = \boldsymbol{d}')\\ &= \sum_{\boldsymbol{d}'\notin\mathcal{N}_\epsilon}\mathsf{Pr}\left(\boldsymbol{d}'\right) I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{d}') + \sum_{\boldsymbol{d}'\in\mathcal{N}_\epsilon}\mathsf{Pr}\left(\boldsymbol{d}'\right) I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{d}')\\ &\overset{(c)}{\leq} (\mathsf{Pr}\left(\boldsymbol{D}' \notin \mathcal{N}_\epsilon\right) + 3\epsilon)ML - (M - n_0)\log_q(M - n_0)\\ &\quad + L\sum_{\boldsymbol{d}'\in\mathcal{N}_\epsilon}\mathsf{Pr}\left(\boldsymbol{d}'\right)\sum_{d\geq 0} n_d C_{\mathsf{Mul}}(d,p,q) + o(ML),\end{aligned}$$

where we used that for all $\boldsymbol{d}' \in \mathcal{N}_\epsilon$

$$L\sum_{d\geq D_\epsilon} n_d \leq ML\sum_{d\geq D_\epsilon}\nu_d + \epsilon/4ML \leq 2\epsilon ML,$$

and we also used $I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{d}') \leq H(\boldsymbol{X}|\boldsymbol{d}') \leq ML$ to bound the mutual information in the first term in inequality (c). Analyzing the term inside the sum, we find that for all $\boldsymbol{d}' \in \mathcal{N}_\epsilon$, it holds that

$$L\sum_{d\geq 0} n_d C_{\mathsf{Mul}}(d,p,q) \leq ML\sum_{d\geq 0}\nu_d C_{\mathsf{Mul}}(d,p,q) + ML\epsilon/4.$$

On the other hand, we can bound

$$\begin{aligned}-(M - n_0)&\log_q(M - n_0)\\ &\leq -M(1 - \nu_0 - \epsilon/4)\log(M(1 - \nu_0 - \epsilon/4))\\ &\leq -\beta ML(1 - \nu_0) + \beta ML\epsilon/4 + O(M)\end{aligned}$$

for all $\boldsymbol{d}' \in \mathcal{N}_\epsilon$. Using that $\mathsf{Pr}\left(\boldsymbol{D}' \notin \mathcal{N}_\epsilon\right) \to 0$ as $M \to \infty$ by the definition of frequency convergence from Definition 1, we obtain

$$I(\boldsymbol{X};\boldsymbol{Z}'|\boldsymbol{D}') \leq MLC_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q) + o(ML),$$

as we can choose ϵ as small as desired. Under the condition $0 < p < \frac{q-1}{2q}$ and $2\beta < 1 - H_q(2p)$, we can guarantee the existence of an α with $2p < \alpha < \frac{q-1}{q}$ and $2\beta < 1 - H_q(\alpha)$ thus obtain the lemma. □

We proceed with a derivation of the bound on the output entropy $H(\boldsymbol{Z}'|\boldsymbol{D}' = \boldsymbol{d}')$ and will bound the entropy terms $H(\boldsymbol{Z}'|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{D}' = \boldsymbol{d}')$ and $H(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{Z}', \boldsymbol{D}' = \boldsymbol{d}')$ afterwards.

B. Output Entropy Bound

We start with deriving an upper bound on the output entropy, which is given in the following lemma. Recall from Definition 6 that $U_{\boldsymbol{d}'}$ depends on α.

Lemma 8. *Let $0 < \beta < 1, q \in \mathbb{N}$, $0 < p < 1$ be fixed parameters. For any constant $D \in \mathbb{N}$ and any α with $0 < \alpha < \frac{q-1}{q}$, the output entropy satisfies*

$$\begin{aligned} H(\boldsymbol{Z}'|\boldsymbol{D}' = \boldsymbol{d}') \leq L \sum_{d \geq 0} & n_d(C_{\mathsf{Mul}}(d, p, q) + dH_q(p)) \\ & + (M - n_0 - U_{\boldsymbol{d}'})(\log_q U_{\boldsymbol{d}'} + L(H_q(\alpha) - 1)) \\ & + L \sum_{d \geq D} n_d + o(ML). \end{aligned}$$

Proof. We start with the observation that $\boldsymbol{Z}'$ given $\boldsymbol{D}'$ is distributed as M parallel multinomial channels with $\boldsymbol{D}'$ draws, induced by an input distribution, which is shuffled according to $\boldsymbol{S}$. We begin by splitting $H(\boldsymbol{Z}'|\boldsymbol{D}' = \boldsymbol{d}')$ according to $d_i' \geq D$ or $d_i' < D$. To this end, for an arbitrary subset $\mathcal{A} \subseteq [M]$, we introduce the notation $\mathcal{A}^{\mathsf{c}} \triangleq [M] \setminus \mathcal{A}$ and also the notation $\boldsymbol{Z}'_{\mathcal{A}} = (\boldsymbol{Z}'_i : i \in \mathcal{A})$ as the matrix, which contains all output clusters $\boldsymbol{Z}'_i$ with $i \in \mathcal{A}$. The clusters are ordered according to ascending indices such that the matrix is well-defined. Abbreviating $\mathcal{D} = \{i \in [M] : d_i' < D\}$, we obtain

$$H(\boldsymbol{Z}'|\boldsymbol{D}' = \boldsymbol{d}') = H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) + H(\boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}|\boldsymbol{D}' = \boldsymbol{d}'),$$

by the chain rule of entropy. Since the individual entropy of every cluster is trivially bounded by $H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}') \leq L(C_{\mathsf{Mul}}(d_i', p, q) + d_i' H_q(p))$, we may bound the second summand by

$$\begin{aligned} H(\boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}|\boldsymbol{D}' = \boldsymbol{d}') &\leq \sum_{i \in \mathcal{D}^{\mathsf{c}}} H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}') \\ &\leq L \sum_{i \in \mathcal{D}^{\mathsf{c}}} (C_{\mathsf{Mul}}(d_i', p, q) + d_i' H_q(p)). \end{aligned}$$

We proceed with splitting $H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}})$ again into two parts, according to whether $i \in \mathcal{U}$ or not. We obtain

$$\begin{aligned} H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) &\leq H(\boldsymbol{Z}'_{\mathcal{D}}, \mathcal{U}|\boldsymbol{D}' = \boldsymbol{d}', \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) \\ &\leq H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U}, \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) + H(\mathcal{U}|\boldsymbol{D}' = \boldsymbol{d}'). \end{aligned}$$

Since $\mathcal{U}$ is a subset of $[M]$, it has at most 2^M different possible outcomes and, henceforth, the second entropy term is at most $H(\mathcal{U}|\boldsymbol{D}' = \boldsymbol{d}') \leq \log_q(2)M$. We thus have

$$\begin{aligned} H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) &\leq H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U}, \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) + O(M) \\ = \sum_{u \subseteq [M]} & \mathsf{Pr}\left(\mathcal{U} = u|\boldsymbol{D}' = \boldsymbol{d}'\right) H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u, \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) \\ & + O(M). \end{aligned}$$

We are now in the position to use the chain rule of entropy to perform the above mentioned splitting of the remaining clusters according to the partition u and $[M] \setminus u$. By the chain rule of entropy,

$$\begin{aligned} H(\boldsymbol{Z}'_{\mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u, \boldsymbol{Z}'_{\mathcal{D}^{\mathsf{c}}}) &\leq H(\boldsymbol{Z}'_{u \cap \mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u) \\ & + H(\boldsymbol{Z}'_{u^{\mathsf{c}} \cap \mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u, \boldsymbol{Z}'_u). \end{aligned} \tag{2}$$

We proceed with bounding the first term in (2) using the fact that the joint entropy is bounded by the sum of marginal entropies

$$H(\boldsymbol{Z}'_{u \cap \mathcal{D}}|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u) \leq \sum_{i \in u \cap \mathcal{D}} H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u). \tag{3}$$

To simplify the subsequent analysis, we fix an arbitrary $\epsilon > 0$ and introduce the random binary indicator variable F_i, $i \in [M]$, which is equal to 0, if the error vectors of the i-th cluster are ϵ-typical as defined in Lemma 23, and 1, otherwise. We obtain

$$\begin{aligned} H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u) &\leq H(\boldsymbol{Z}'_i, F_i|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u) \\ &\overset{(a)}{\leq} 1 + \sum_{f_i \in \{0,1\}} \mathsf{Pr}\left(F_i = f_i|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u\right) \\ &\qquad H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}', F_i = f_i, \mathcal{U} = u) \\ &\overset{(b)}{\leq} 1 + H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}', F_i = 0, \mathcal{U} = u) \\ &\quad + \mathsf{Pr}\left(F_i = 1|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u\right) d_i' L, \end{aligned} \tag{4}$$

where we used that the entropy of a Bernoulli random variable is at most 1 in inequality (a). Inequality (b) follows from splitting the sum over f_i into two terms and bounding $\mathsf{Pr}\left(F_i = 0|\boldsymbol{D}' = \boldsymbol{d}', \mathcal{U} = u\right) \leq 1$ as well as $H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}', F_i = 1, \mathcal{U} = u) \leq d_i' L$. The latter bound is due to the fact that the cluster $\boldsymbol{Z}'_i$ consists of $d_i' L$ symbols over Σ_q and thus its entropy is directly bounded by $d_i' L$. Denoting the d_i' sequences of the i-th cluster by $\boldsymbol{Z}'_{i,1}, \ldots, \boldsymbol{Z}'_{i,d_i'} \in \Sigma_q^L$ as in Definition 6, we can rewrite the above entropy as

$$\begin{aligned} H(\boldsymbol{Z}'_i|\boldsymbol{D}' = \boldsymbol{d}', F_i = 0, \mathcal{U} = u) & \\ = H\left(\boldsymbol{Z}'_i \middle| \boldsymbol{D}' = \boldsymbol{d}', F_i = 0, \mathcal{U} = u, \boldsymbol{Z}'_{i,1}\right) & \\ + H\left(\boldsymbol{Z}'_{i,1} \middle| \boldsymbol{D}' = \boldsymbol{d}', F_i = 0, \mathcal{U} = u\right). & \end{aligned}$$

The first summand can be bounded using Lemma 23, as follows. To start with, for the sake of the argument, consider the distribution of $\boldsymbol{Z}'_i$ given $\boldsymbol{Z}'_{i,1}$ and $\boldsymbol{D}'$ without the condition on F_i and $\mathcal{U}$ first. To this end, denote by $\boldsymbol{e}_i^{(j)} \triangleq \boldsymbol{Z}'_{i,j} - \boldsymbol{x}_i$ the error vectors of the i-th cluster. As we have remarked in the beginning of the proof, without the condition on F_i and $\mathcal{U}$ those are distributed according to the multinomial channel model, independent from the input. Now, we can express the conditional distribution of $\boldsymbol{Z}'_i$

$$\begin{aligned} \mathsf{Pr}\left(\boldsymbol{Z}'_i = \mathrm{y}_i \middle| \boldsymbol{d}', \boldsymbol{Z}'_{i,1} = \mathrm{y}_i^{(1)}\right) = \mathsf{Pr}\Big(& \boldsymbol{e}_i^{(2)} - \boldsymbol{e}_i^{(1)} = \boldsymbol{y}_i^{(2)} - \boldsymbol{y}_i^{(1)}, \\ \ldots, \boldsymbol{e}_i^{(d_i')} - \boldsymbol{e}_i^{(1)} = \boldsymbol{y}_i^{(d_i')} - \boldsymbol{y}_i^{(1)} & \Big| d_i = d_i', \boldsymbol{Z}'_{i,1} = \boldsymbol{y}_i^{(1)}\Big) \end{aligned}$$

as that of the error vectors $\boldsymbol{e}_i^{(j)} - \boldsymbol{e}_i^{(1)}$. By Lemma 23, given that $F_i = 0$, i.e., the error vectors are ϵ-typical sequences, the number of possible options for the error vectors is at most $q^{L(C_{\mathsf{Mul}}(d'_i,p,q)+d'_iH_q(p)-1+\epsilon)}$. Further conditioning decreases the number of possible options and thus, $H(\boldsymbol{Z}'_i|\boldsymbol{D}'=\boldsymbol{d}',F_i=0,\mathcal{U}=u,\boldsymbol{Z}'_{i,1})\leq L(C_{\mathsf{Mul}}(d'_i,p,q)+d'_iH_q(p)-1+\epsilon)$. Together with the trivial bound $H(\boldsymbol{Z}'_{i,1}|\boldsymbol{D}'=\boldsymbol{d}',F_i=0,\mathcal{U}=u)\leq L$, we obtain for all $i\in u$

$$\begin{aligned}&H(\boldsymbol{Z}'_i|\boldsymbol{D}'=\boldsymbol{d}',F_i=0,\mathcal{U}=u)\\&\qquad\leq L(C_{\mathsf{Mul}}(d'_i,p,q)+d'_iH_q(p)+\epsilon).\end{aligned}\tag{5}$$

We now bound the second entropy term in (2) using again the fact that the joint entropy is at most the sum of the individual entropies and obtain

$$\begin{aligned}&H(\boldsymbol{Z}'_{u^{\mathsf{c}}\cap\mathcal{D}}|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u,\boldsymbol{Z}'_u)\\&\qquad\leq\sum_{i\in u^{\mathsf{c}}:0<d'_i<D}H(\boldsymbol{Z}'_i|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u,\boldsymbol{Z}'_u),\end{aligned}\tag{6}$$

where we used that $H(\boldsymbol{Z}'_i|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u,\boldsymbol{Z}'_u)=0$ for all $i\in[M]\setminus u$ with $d'_i=0$. Performing the analogous steps as above to introduce the conditioning on the random variable F_i, we obtain for all $i\in u^{\mathsf{c}}$

$$\begin{aligned}H(\boldsymbol{Z}'_i|\boldsymbol{d}',\mathcal{U}=u,\boldsymbol{Z}'_u)&\leq 1+H(\boldsymbol{Z}'_i|\boldsymbol{d}',F_i=0,\mathcal{U}=u,\boldsymbol{Z}'_u)\\&\quad+\mathsf{Pr}\left(F_i=1|\boldsymbol{d}',\mathcal{U}=u\right)d'_iL.\end{aligned}\tag{7}$$

Using the same notation as in the derivation of the first term, we obtain for all $i\in u^{\mathsf{c}}$

$$\begin{aligned}&H(\boldsymbol{Z}'_i|\boldsymbol{D}'=\boldsymbol{d}',F_i=0,\mathcal{U}=u,\boldsymbol{Z}'_u)\\&=H\left(\boldsymbol{Z}'_i\middle|\boldsymbol{D}'=\boldsymbol{d}',F_i=0,\mathcal{U}=u,\boldsymbol{Z}'_{i,1},\boldsymbol{Z}'_u\right)\\&\quad+H\left(\boldsymbol{Z}'_{i,1}\middle|\boldsymbol{D}'=\boldsymbol{d}',F_i=0,\mathcal{U}=u,\boldsymbol{Z}'_u\right)\\&\overset{(c)}{\leq}L(C_{\mathsf{Mul}}(d'_i,p,q)+d'_iH_q(p)-1+\epsilon+H_q(\alpha))+\log_q|u|,\end{aligned}\tag{8}$$

where the first summand in inequality (c) has been bounded using the same arguments as above. The second summand has been bounded using the fact that, given $\mathcal{U}=u$ and $\boldsymbol{Z}'_u$, there are only $|u|q^{LH_q(\alpha)}$ options for $\boldsymbol{Z}'_{i,1}$, as $\boldsymbol{Z}'_{i,1}$ has to have distance at most αL to one of the sequences in $\boldsymbol{Z}'_u$. Note that this entropy bound on the size of the Hamming ball is only valid if $\alpha<\frac{q-1}{q}$. Plugging (8), (7), (6), and (3), (4), (5) into (2), we conclude that

$$\begin{aligned}&H(\boldsymbol{Z}'|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u)\\&\quad\leq\sum_{i=1}^{M}L(C_{\mathsf{Mul}}(d'_i,p,q)+d'_iH_q(p)+\epsilon)\\&\qquad+\sum_{i\in u^{\mathsf{c}}:0<d'_i<D}\log_q|u|+L(H_q(\alpha)-1)\\&\qquad+\sum_{i\in\mathcal{D}}\mathsf{Pr}\left(F_i=1|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u\right)d'_iL+M\\&\quad\overset{(d)}{=}\sum_{d\geq 0}Ln_d(C_{\mathsf{Mul}}(d,p,q)+dH_q(p)+\epsilon)\\&\qquad+(M-n_0-|u|)(\log_q|u|+L(H_q(\alpha)-1))+L\sum_{d\geq D}n_d\\&\qquad+\sum_{i\in\mathcal{D}}\mathsf{Pr}\left(F_i=1|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u\right)d'_iL+M,\end{aligned}$$

where in equality (d) we replaced the sum over i by a sum over d. We further used that the number of terms in the sum over $i\in[M]\setminus u$ with $d'_i>0$ is precisely $M-n_0-|u|$. As a reminder we note that the above inequality holds for all $0<\epsilon<1$, where F_i is the random variable that depends on ϵ through the ϵ-typical sequences from Lemma 23. We turn to bounding the last summand from above

$$\begin{aligned}&L\sum_{u}\mathsf{Pr}\left(\mathcal{U}=u|\boldsymbol{D}'=\boldsymbol{d}'\right)\sum_{i\in\mathcal{D}}\mathsf{Pr}\left(F_i=1|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u\right)d'_i\\&=L\sum_{i\in\mathcal{D}}\mathsf{Pr}\left(F_i=1|\boldsymbol{D}'=\boldsymbol{d}'\right)d'_i.\end{aligned}$$

We now use that by Lemma 23, $\mathsf{Pr}\left(F_i=1|\boldsymbol{D}'=\boldsymbol{d}'\right)<\epsilon$ for all $L\geq L_{d'_i}(\epsilon)$. Therefore,

$$L\sum_{i\in\mathcal{D}}\mathsf{Pr}\left(F_i=1|\boldsymbol{D}'=\boldsymbol{d}'\right)d'_i\overset{(e)}{\leq}\epsilon L\sum_{i\in\mathcal{D}}d'_i\overset{(f)}{\leq}\epsilon cML,$$

where inequality (e) holds for all $L\geq\max_{0\leq d<D}L_d(\epsilon)$. We further used in inequality (f) that the total number of draws is bounded by cM, according to Definition 1. We are now in the position to compute the overall entropy

$$\begin{aligned}&H(\boldsymbol{Z}'|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U})\\&=\sum_{u\subseteq[M]}\mathsf{Pr}\left(\mathcal{U}=u|\boldsymbol{D}'=\boldsymbol{d}'\right)H(\boldsymbol{Z}'|\boldsymbol{D}'=\boldsymbol{d}',\mathcal{U}=u)\\&\leq L\sum_{d\geq 0}n_d(C_{\mathsf{Mul}}(d,p,q)+dH_q(p))\\&\quad+\mathsf{E}[(M-n_0-|\mathcal{U}|)(\log_q|\mathcal{U}|+L(H_q(\alpha)-1))|\boldsymbol{D}'=\boldsymbol{d}']\\&\quad+L\sum_{d\geq D}n_d+\epsilon cML+O(M)\\&\overset{(g)}{\leq}L\sum_{d\geq 0}n_d(C_{\mathsf{Mul}}(d,p,q)+dH_q(p))+(M-n_0-U_{\boldsymbol{d}'})\cdot\\&\quad(\log_q U_{\boldsymbol{d}'}+L(H_q(\alpha)-1))+L\sum_{d\geq D}n_d+\epsilon cML+O(M),\end{aligned}$$

where inequality (g) is due to Jensen's inequality and the fact that $-|\mathcal{U}|\log_q|\mathcal{U}|$ is a concave function in $|\mathcal{U}|$. Note that this inequality holds for any constant D and large enough $L\geq\max_{0\leq d<D}L_d(\epsilon)$. The claim of the lemma follows as we can choose ϵ arbitrarily small. □

C. Ordered Conditional Entropy Bound

Next, we compute the conditional output entropy, conditioned on the permutation $\boldsymbol{S}$.

Lemma 9. *Let $0<\beta<1, q\in\mathbb{N}$, $0<p<1$ be fixed parameters. Then,*

$$H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{S},\boldsymbol{D}'=\boldsymbol{d}')=\sum_{d\geq 0}n_ddH_q(p).$$

Proof. We can use the fact that, given $\boldsymbol{S}$, there exists a deterministic bijection between $\boldsymbol{Z}$ and $\boldsymbol{Z}'$, since $\boldsymbol{Z}'_i=\boldsymbol{Z}_{S_i}$, to obtain that the conditional output entropy is given by

$$\begin{aligned}H(\boldsymbol{Z}'|\boldsymbol{X},\boldsymbol{S},\boldsymbol{D}'=\boldsymbol{d}')&=H(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{S},\boldsymbol{D}'=\boldsymbol{d}')\\&=\sum_{\boldsymbol{s}}\mathsf{Pr}\left(\boldsymbol{s}|\boldsymbol{d}'\right)H(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{S}=\boldsymbol{s},\boldsymbol{D}'=\boldsymbol{d}').\end{aligned}$$

This allows to compute the entropy by

$$
\begin{aligned}
H(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{S}{=}\boldsymbol{s},\boldsymbol{D}'{=}\boldsymbol{d}') &\overset{(b)}{=} \sum_{i=1}^{M} H(\boldsymbol{Z}_i|\boldsymbol{X},\boldsymbol{S}=\boldsymbol{s},\boldsymbol{D}'=\boldsymbol{d}') \\
&\overset{(c)}{=} \sum_{i=1}^{M} H(\boldsymbol{Z}_i|\boldsymbol{X}_i,\boldsymbol{S}=\boldsymbol{s},\boldsymbol{D}'=\boldsymbol{d}') \\
&\overset{(d)}{=} \sum_{i=1}^{M} d_i H_q(p) = \sum_{d\geq 0} n_d d H_q(p),
\end{aligned}
$$

where equality (b) follows from the independence of the variables $\boldsymbol{Z}_i$ given the input $\boldsymbol{X}$ and drawing composition $\boldsymbol{D}$ (which is implicitly given by the combination of $\boldsymbol{S}$ and $D_i' = D_{S_i}$). In equation (c) we used that, given $\boldsymbol{X}_i$ and D_i, the variable $\boldsymbol{Z}_i'$ is independent of all $\boldsymbol{Z}_j'$ and D_j with $j \neq i$ due to the fact that $\boldsymbol{Z}_i'$ can be expressed as the sum of the D_i-fold repetition of $\boldsymbol{X}_i$ and an error vector that is chosen independently, as presented in Section II-B. Note that $H(\boldsymbol{Z}_i|\boldsymbol{X}_i,\boldsymbol{S}=\boldsymbol{s},\boldsymbol{D}'=\boldsymbol{d}') = d_i H_q(p)$ is precisely the channel entropy of the multinomial channel, which has been shown in Lemma 22 to be independent of the input distribution of $\boldsymbol{X}_i$ and thus is only dependent on D_i, which we used in equality (d). □

D. Permutation Entropy Bound

The last ingredient that is missing to prove Lemma 7 is to bound the entropy of the permutation given the input and output sequences. Note that our proof is motivated by the idea of [25], where a similar statement has been proven for the case where the drawing composition is distributed according to i.i.d. Bernoulli variables.

Lemma 10. *Let $0 < \beta < 1, q \in \mathbb{N}$, $0 < p < 1$ be fixed parameters. Then, for any α with $2p < \alpha < 1$,*

$$H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\boldsymbol{D}'=\boldsymbol{d}') \leq M\log_q M - U_{\boldsymbol{d}'}\log_q U_{\boldsymbol{d}'} + o(ML).$$

Proof. To start with, we observe that

$$H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\boldsymbol{D}'=\boldsymbol{d}') = H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U},\boldsymbol{D}'=\boldsymbol{d}'), \tag{9}$$

as $\mathcal{U}$ is a function of $\boldsymbol{Z}'$ and we thus can introduce the condition without changing the entropy. We can further expand the entropy to

$$
\begin{aligned}
H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U},\boldsymbol{D}'=\boldsymbol{d}') \leq \sum_{u} \mathsf{Pr}\left(\mathcal{U}=u|\boldsymbol{D}'=\boldsymbol{d}'\right)\cdot \\
\sum_{i=1}^{M} H(S_i|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}'). \qquad (10)
\end{aligned}
$$

On the one hand, for each $i \in [M]$ with $d_i' = 0$, we trivially bound $H(S_i|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \leq \log_q M$, as there are at most M options for s_i. On the other hand, for an arbitrary $\delta > p$, for each $i \in [M]$, we introduce the Bernoulli variable E_i, which is equal to one, if $d_i' > 0$ and $d_{\mathsf{H}}\left(\boldsymbol{X}_{S_i},\boldsymbol{Z}_{i,1}'\right) \geq \delta L$ and equal to 0, otherwise. Here $\boldsymbol{Z}_{i,1}' \in \Sigma_q^L$ is the first sequence in the cluster according to the nomenclature of Definition 6. As the Hamming distance between $\boldsymbol{x}_{s_i}$ and $\boldsymbol{Z}_{i,1}'$ is binomial distributed with success probability p and L trials, we know from Lemma 19 that for all $i \in [M]$ with $d_i' > 0$

$$\mathsf{Pr}\left(E_i = 1|\boldsymbol{d}'\right) \leq \mathrm{e}^{-2L(\delta-p)^2}. \tag{11}$$

This allows to derive the following upper bound on the individual entropy terms.

$$
\begin{aligned}
&H(S_i|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \\
&\quad\leq H(S_i,E_i|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \\
&\quad\leq H(E_i|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \\
&\qquad + H(S_i|\boldsymbol{X},\boldsymbol{Z}',E_i,\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \\
&\quad\overset{(a)}{\leq} 1 + \sum_{e_i\in\{0,1\}} \mathsf{Pr}\left(E_i=e_i|\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}'\right)\cdot \\
&\qquad H(S_i|\boldsymbol{X},\boldsymbol{Z}',E_i=e_i,\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \\
&\quad\overset{(b)}{\leq} 1 + \mathsf{Pr}\left(E_i=1|\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}'\right)\log_q M \\
&\qquad + H(S_i|\boldsymbol{X},\boldsymbol{Z}',E_i=0,\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}'), \qquad (12)
\end{aligned}
$$

where we used in inequality (a) that the entropy of a Bernoulli random variable is at most 1 and inequality (b) follows from the fact that $\mathsf{Pr}\left(E_i=0|\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}'\right) \leq 1$ and the fact that we can again trivially bound the entropy of $H(S_i|\boldsymbol{X},\boldsymbol{Z}',E_i=1,\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}')$ by $\log_q M$. It remains to bound $H(S_i|\boldsymbol{X},\boldsymbol{Z}',E_i=0,\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}')$ from above. To this end, we set $\delta = \alpha/2$ and for all $i \in u$, we introduce the set

$$\mathcal{A}_i = \left\{j \in [M] : d_{\mathsf{H}}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{i,1}'\right) < \delta L\right\}$$

of input sequences that have distance less than δL to the first sequence in the i-th output cluster. This set contains all input sequences that could potentially have produced $\boldsymbol{Z}_{i,1}'$, given that $E_i = 0$. Note that by definition of E_i and $\mathcal{A}_i$, we directly have $S_i \in \mathcal{A}_i$, given $E_i = 0$. Further, the sets $\mathcal{A}_i$ are disjoint, as for any $i, k \in u$ and any sequence $j \in \mathcal{A}_i$ it holds by the triangle inequality,

$$
\begin{aligned}
d_{\mathsf{H}}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{k,1}'\right) &\geq d_{\mathsf{H}}\left(\boldsymbol{Z}_{i,1}',\boldsymbol{Z}_{k,1}'\right) - d_{\mathsf{H}}\left(\boldsymbol{X}_j,\boldsymbol{Z}_{i,1}'\right) \\
&> (\alpha-\delta)L = \delta L,
\end{aligned}
$$

implying that each $j \in [M]$ can be contained in at most one set $\mathcal{A}_i$. For all $i \in u$, $S_i \in \mathcal{A}_i$, and S_i can thus assume at most $|\mathcal{A}_i|$ values, limiting its entropy to at most $\log_q|\mathcal{A}_i|$. Bounding the entropy for all other terms $i \notin u$ by $\log_q M$, we obtain

$$
\begin{aligned}
&\sum_{i:d_i'>0} H(S_i|\boldsymbol{X},\boldsymbol{Z}',E_i=0,\mathcal{U}=u,\boldsymbol{D}'=\boldsymbol{d}') \\
&\quad\leq \sum_{i\notin u:d_i'>0} \log_q M + \sum_{i\in u}\log_q|\mathcal{A}_i| \\
&\quad= (M-n_0-|u|)\log M + \sum_{i\in u}\log_q|\mathcal{A}_i| \\
&\quad\overset{(c)}{\leq} (M-n_0-|u|)\log M + |u|\log_q(M/|u|) \\
&\quad= (M-n_0)\log M - |u|\log|u|, \qquad (13)
\end{aligned}
$$

where inequality (c) follows from $\sum_{i\in u}|\mathcal{A}_i| \leq M$ due to the disjointedness of the sets $\mathcal{A}_i$. Thus the sum is bounded from above by setting $|\mathcal{A}_i| = M/|u|$ and using Jensen's inequality and the concavity of the logarithm. Plugging (13) and (12) into (10) and taking also those i with $d_i' = 0$ into account, we obtain

$$\begin{aligned} &H(\boldsymbol{S}|\boldsymbol{X},\boldsymbol{Z}',\mathcal{U},\boldsymbol{D}'=\boldsymbol{d}')\\ &\quad\le M\log_q M + \sum_u \mathsf{Pr}\left(u|\boldsymbol{d}'\right)\cdot\\ &\qquad\left(-|u|\log_q|u| + \sum_{i:d_i'>0}\left(1+\mathsf{Pr}\left(E_i=1|\boldsymbol{d}',u\right)\log_q M\right)\right)\\ &\quad\le M\log_q M - \mathsf{E}\left[|\mathcal{U}|\log_q|\mathcal{U}|\;|\boldsymbol{D}'=\boldsymbol{d}'\right]\\ &\qquad+\log_q M\sum_{i:d_i'>0}^{M}\mathsf{Pr}\left(E_i=1|\boldsymbol{d}'\right)+O(M)\\ &\quad\overset{(d)}{=} M\log_q M - U_{\boldsymbol{d}'}\log_q U_{\boldsymbol{d}'}+o(ML), \end{aligned}\tag{14}$$

where we used Jensen's inequality in inequality (d) together with the bound (11) on the probability $\mathsf{Pr}\left(E_i = 1|\boldsymbol{d}'\right)$, which proves the claim of the lemma with (9). □

V. Achievable Rates

We proceed by deriving achievable rates for the noisy drawing channel. We derive these results using standard random coding techniques [28], [29].

The outline of our arguments is as follows. Choose a random codebook $\mathcal{C}$ of rate R with independently and identically distributed codewords that are drawn from some given input distribution $\mathsf{Pr}\left(\boldsymbol{X}\right)$. Then, we define a suitable decoder and compute its average error probability, averaged over all codebooks. We prove that for a given rate R, the average error probability tends to zero and thus there exists at least one codebook of rate R whose error probability also tends to zero.

Our decoder consists of the two following stages. First, the decoder clusters the output sequences $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N$ to clusters of sequences $\widehat{\boldsymbol{Z}}_1, \dots, \widehat{\boldsymbol{Z}}_M$, according to their Hamming distance. If the channel is not too noisy, we prove that with high probability sequences within a cluster originate from the same input sequence. The second stage then matches the clusters to input sequences of a codeword based on the following measure of typicality. An input sequence $\boldsymbol{X}_i$ and an output cluster $\widehat{\boldsymbol{Z}}_j$ may be matched, if they are jointly typical with respect to the multinomial channel in the sense introduced by Shannon [29], [28]. Given a channel output $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N$, the decoder then decides for a codeword if the number of matching input-cluster pairs is close to M. Analyzing this decoder will show that for any rate R below the capacity of the noisy drawing channel, the probability that the correct transmitted codeword is jointly typical with the received word with high probability and any other codeword is jointly typical with small probability.

We now turn to a rigorous derivation of achievable rates. We will devote the rest of this section to prove the following result about achievable rates in a step-by-step fashion. We will proceed by presenting the final results first and wrap up the necessary ingredients towards the end of the section.

Lemma 11. *Let $0 < \beta < 1, 0 < p < \frac{q-1}{2q}$, $q \in \mathbb{N}$ be fixed parameters satisfying $\beta < 1 - H_q(2p)$ and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a regular distribution that converges in frequency to $\boldsymbol{\nu}$. Then, any rate R with*

$$R < C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q)$$

is achievable over the noisy drawing channel.

Algorithm 1 Clustering algorithm

```
Input: $N$ received sequences $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N$; cluster radius $\phi L$
Output: $M$ Clusters $\widehat{\boldsymbol{Z}}_1, \dots, \widehat{\boldsymbol{Z}}_M$
$\widehat{M} \leftarrow 0$
$\mathcal{Y} \leftarrow \{\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N\}\}$
while $\mathcal{Y} \neq \emptyset$ do
    $\widehat{M} \leftarrow \widehat{M} + 1$
    for $\boldsymbol{Y} \in \mathcal{Y}$ do
        if ($\widehat{\boldsymbol{Z}}_{\widehat{M}}$ is empty) or $\left(d_{\mathsf{H}}\left(\boldsymbol{Y}, \widehat{\boldsymbol{Z}}_{\widehat{M},1}\right) < \phi L\right)$ then
            append $\boldsymbol{Y}$ to $\widehat{\boldsymbol{Z}}_{\widehat{M}}$
            $\mathcal{Y} \leftarrow \mathcal{Y} \setminus \{\boldsymbol{Y}\}$
if $\widehat{M} > M$ then discard $\widehat{\boldsymbol{Z}}_{M+1}, \dots, \widehat{\boldsymbol{Z}}_{\widehat{M}}$
if $\widehat{M} < M$ then add empty clusters $\widehat{\boldsymbol{Z}}_{\widehat{M}+1}, \dots, \widehat{\boldsymbol{Z}}_M$
```

We start with setting up the necessary definitions required for the following expositions and assume that the conditions of Lemma 11 are fulfilled throughout the remainder of this section. We will prove the results using the conventional random coding argument. To this end recall the communication setup presented in Section III-A. Let now $\mathcal{C} = \{\boldsymbol{X}(1), \dots, \boldsymbol{X}(q^{MLR})\} \subseteq \Sigma_q^{M \times L}$ be a randomly chosen codebook of code rate R, where each codeword $\boldsymbol{X}(i) \in \Sigma_q^{M \times L}$ is selected independently and uniformly over all possible words in $\Sigma_q^{M \times L}$, i.e., each symbol in $\boldsymbol{X}(i)$ is chosen independently and uniformly over Σ_q. We will write $\boldsymbol{X}(i) = (\boldsymbol{X}_1(i), \dots, \boldsymbol{X}_M(i))$. In order to define the decoder, we fix an $0 < \epsilon < 1$ and a clustering radius $\alpha > 2p$. Consider Algorithm 1, which greedily picks an output sequence and adds other output sequences, such that their Hamming distance with respect to the first pick is less than ϕL. These sequences are combined to a cluster $\widehat{\boldsymbol{Z}}_1$ and all elements in $\widehat{\boldsymbol{Z}}_1$ are removed as candidates for succeeding clusters. The procedure successively continues to form clusters $\widehat{\boldsymbol{Z}}_2, \dots, \widehat{\boldsymbol{Z}}_{\widehat{M}}$ on the remaining sequences with the same procedure until no more sequences are present. Afterwards, the algorithm adds empty clusters or removes excess clusters, such that the total number of estimated clusters is M. It is evident that this clustering algorithm is neither efficient in computational complexity or accuracy, however it is easy to analyze and will suffice for our purposes. Interestingly, under some mild conditions, this naive clustering algorithm produces many correct clusters, as we will see in Section V-A. We proceed with the definition of typicality.

Definition 12. *Consider the q-ary multinomial channel with error probability p, d draws and uniform input $X \in \Sigma_q$ with corresponding output $Z \in \Sigma_q^{d \times L}$. We define the set of ϵ-jointly* *typical sequences $\boldsymbol{x} \in \Sigma_q^L$ and $\boldsymbol{z} \in \Sigma_q^{d \times L}$ by*

$$\mathcal{T}_{\mathsf{Mul}}^{(L,\epsilon)}(d,p,q) \triangleq \Bigg\{ (\boldsymbol{x},\boldsymbol{z}) : \left| -\frac{\log_q \mathsf{Pr}\,(\boldsymbol{z})}{L} - H(Z) \right| < \epsilon, \\ \left| -\frac{\log_q \mathsf{Pr}\,(\boldsymbol{x},\boldsymbol{z})}{L} - H(X,Z) \right| < \epsilon \Bigg\}.$$

Note that usually joint typicality includes also a condition on the input $\mathsf{Pr}\,(\boldsymbol{x})$, however in our case this is trivially fulfilled for all input sequences since we consider uniformly distributed input sequences. We can then define a measure of typicality over parallel multinomial (PM) channels as follows.

Definition 13. *We define the largest typicality matching $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{x},\boldsymbol{z})$ between an input $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_M)$ and output $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_M)$ as the largest integer T such that there exist two sequences of integers $i_1, \dots, i_T$ and $j_1, \dots, j_T$, with $i_t, j_t \in [M]$ for all $1 \leq t \leq T$, each sequence composed of distinct integers, such that $(\boldsymbol{x}_{i_t}, \boldsymbol{z}_{j_t}) \in \mathcal{T}_{\mathsf{Mul}}^{(L,\epsilon)}(d_{j_t}, p, q)$ for all $1 \leq t \leq T$, where d_{j_t} is the size of the cluster $\boldsymbol{z}_{j_t}$.*

In other words, the typicality between an input $\boldsymbol{X}$ and output $\boldsymbol{Z}$ is measured by the number of (distinct) pairs of input sequences and output clusters $(\boldsymbol{X}_i, \boldsymbol{Z}_j)$ that are jointly typical with respect to the multinomial channel. The decoder $\mathsf{dec}(\boldsymbol{Y})$ first estimates the clusters $\widehat{\boldsymbol{Z}}$ using Algorithm 1 and decodes to $\widehat{W}$, if $\boldsymbol{X}(\widehat{W})$ is the unique codeword that satisfies $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(\widehat{W}), \widehat{\boldsymbol{Z}}) \geq M(1-\epsilon)$. If there is none or more than two codewords that are typical in the above sense with $\widehat{\boldsymbol{Z}}$, then the decoder outputs a failure, resulting in a decoding error.

We will proceed with deriving the ingredients used to prove Lemma 11. Roughly speaking, we will first show in Lemma 14 and 15 that with high probability there are many correct clusters and thus the typicality $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \widehat{\boldsymbol{Z}})$ is close to the typicality $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \boldsymbol{Z})$. Afterwards, we will use standard joint typicality results to prove that, $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq M(1-\epsilon)$ with high probability in Lemma 16. Conversely, we show that if the code rate is chosen small enough, then $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \boldsymbol{Z}) < M(1-\epsilon)$ for all $2 \leq i \leq q^{MLR}$ with high probability in Lemma 17. Combining these results with the fact that $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \widehat{\boldsymbol{Z}}) \approx T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \boldsymbol{Z})$ with high probability yields the proof of Lemma 11.

A. Clustering Accuracy

We start with discussing the accuracy of the proposed clustering algorithm. Consider the bipartite graph G_{cluster} with vertices $\boldsymbol{Z}_j$, $j \in [M]$ on the left and $\widehat{\boldsymbol{Z}}_i$, $i \in [M]$ on the right. We draw an edge from $\boldsymbol{Z}_j$ to $\widehat{\boldsymbol{Z}}_i$, if the multiset of sequences in $\boldsymbol{Z}_j$ is equal to the multiset of sequences in $\widehat{\boldsymbol{Z}}_i$. With this graph, we define by G the size of the largest matching[5] of the graph G_{cluster}, which we will refer to by *correct* clusters in the sequel. The following lemma proves that the number of correct clusters is large with high probability and essentially uses two ingredients. First, if $2p < \phi < \frac{q-1}{q}$, the probability that a sequence in a cluster has Hamming distance more than ϕL with respect to the first sequence is small, because the expected number of errors per sequence is pL. Second, if $\beta < 1 - H_q(\phi)$, the probability that a sequence of another cluster is close to the first sequence is small. This is because the probability that a given output sequence has Hamming distance at most ϕL to the first sequence is $q^{-L(1-H_q(\phi))}$ and then, under the condition on β, the union bound over all output sequences is small.

Lemma 14. *Let $\beta > 0, q \in \mathbb{N}$, $0 < p < 1$ be fixed parameters and $\mathsf{Pr}\,(\boldsymbol{i})$ be a given regular distribution. Then, for any ϕ with $2p < \phi < \frac{q-1}{q}$, $\beta < 1 - H_q(\phi)$, the probability of having at least $M(1-\epsilon)$ correct clusters satisfies*

$$\lim_{M \to \infty} \mathsf{Pr}\,(G \geq M(1-\epsilon)) = 1.$$

Proof. Denote by G_i, $i \in [M]$ a binary indicator variable that is equal to 1, if $D_i > 0$ and $\boldsymbol{Z}_i$ has been clustered correctly and 0, otherwise. Further, let $\widehat{M}$ be the number of non-empty clusters produced by Algorithm 1, before removing clusters or adding empty clusters. To start with, it holds that $G \geq \sum_{i=1}^{M} G_i + \min\{N_0, M - \widehat{M}\}$, since we can construct a matching, where we arbitrarily match the $M - \widehat{M}$ empty clusters and we match each cluster $\boldsymbol{y}_i$ with $G_i = 1$ to the correct cluster produced by the algorithm. Note that the edges of this matching share no common vertices, since the matching of empty clusters is arbitrary and the non-empty clusters, produced by Algorithm 1, are disjoint by construction of the algorithm. Notably, the bound further covers the case, where $\widehat{M} > M$ and we possibly remove some of the correct clusters. Thus, by the union bound, the probability on the number of correct clusters is at least

$$\mathsf{Pr}\,(G \geq M(1-\epsilon)) \geq 1 - \mathsf{Pr}\left(\sum_{i=1}^{M} G_i \leq M - N_0 - M\epsilon/2 \right) \\ - \mathsf{Pr}\left(M - \widehat{M} \leq N_0 - M\epsilon/2 \right). \quad (15)$$

We proceed with showing that the sum over the variables G_i in (15) is close to $M - N_0$ with high probability and $M - \widehat{M}$ is close to N_0 with high probability.

We start with the second term. To this end, let $2p < \phi < \frac{q-1}{q}$ and denote by F_j, $j \in \mathbb{N}$ the binary indicator, which is equal to 1, if $d_{\mathsf{H}}\left(\boldsymbol{X}_{I_j}, \boldsymbol{Y}_j\right) > \phi L/2$ and 0, otherwise, where I_j is the original input sequence that corresponds to $\boldsymbol{Y}_j$. With this definition, we observe that $\widehat{M} \leq M - N_0 + \sum_{j=1}^{N} F_j$. This is because, whenever the clustering algorithm selects a sequence $\boldsymbol{Y}_j$ with $F_j = 0$, the remaining sequences j' from this cluster with $F_{j'} = 0$, that have not been clustered yet, will be contained in the estimated cluster. Thus, each sequence $\boldsymbol{Y}_j$ with $F_j = 1$ can produce at most one extra cluster. Since F_j, $j \in [N]$ are independent and identical Bernoulli random variables with success probability at most $\mathrm{e}^{-2L(\phi/2-p)^2}$ (see

[5] A matching of a bipartite graph is a set of edges such that no two edges share common end points.

Lemma 19), it holds that

$$
\begin{aligned}
\mathsf{Pr}\left(M-\widehat{M}\leq N_0-\epsilon M/2\right) &\leq \mathsf{Pr}\left(\sum_{j=1}^{N}F_j\geq M\epsilon/2\right)\\
&=\sum_{n}\mathsf{Pr}\left(N=n\right)\mathsf{Pr}\left(\sum_{j=1}^{n}F_j\geq M\epsilon/2\middle|N=n\right)\\
&\overset{(a)}{\leq}\mathsf{Pr}\left(\sum_{j=1}^{cM}F_j\geq M\epsilon/2\middle|N=cM\right)\\
&\overset{(b)}{\leq}\mathrm{e}^{-2cM\left(\epsilon/(2c)-\mathrm{e}^{-2L(\phi/2-p)^2}\right)^2}=o(1),
\end{aligned}
$$

for all $\epsilon>0$, as $M\to\infty$. We used in inequality (a) that $N\leq cM$ with probability 1 by definition of the regular drawing distribution together with the fact that the second probability term is monotonically increasing in n. In inequality (b), we employed Lemma 19 on the binomial tail.

We turn towards the first summand in (15). Recall that according to Definition 6, we denote by $\boldsymbol{Z}_{i,1},\dots,\boldsymbol{Z}_{i,D_i}$ the sequences of a cluster $\boldsymbol{Z}_i$. We can bound the probability $\mathsf{Pr}\left(G_i=1\right)$ for all i with $D_i>0$ as follows. A cluster $\boldsymbol{Z}_i$ is guaranteed to be estimated correctly, if $d_{\mathsf{H}}\left(\boldsymbol{X}_i,\boldsymbol{Z}_{i,j}\right)\leq\phi L/2$ for all sequences $j\in[d_i]$ and also if there exists no other output sequence $\boldsymbol{Z}_{i',j'}$ from another cluster $i'\neq i$ that has distance less than ϕL to one of the sequences in the cluster $\boldsymbol{y}_i$. Demarginalizing with respect to the drawing composition, we obtain

$$
\begin{aligned}
\mathsf{Pr}(G_i=1)&=\sum_{d_i}\mathsf{Pr}\left(d_i\right)\mathsf{Pr}\left(G_i=1|d_i\right)\\
&\overset{(c)}{\geq}\sum_{d_i\geq 1}\mathsf{Pr}\left(d_i\right)\left(1-d_i\mathrm{e}^{-2L(\phi/2-p)^2}-cMd_iq^{-L(1-H_q(\phi))}\right)\\
&\overset{(d)}{=}\mathsf{Pr}\left(D_i\geq 1\right)-\mathsf{E}\left[D_i\right]\left(\mathrm{e}^{-2L(\phi/2-p)^2}+cq^{-L(1-H_q(\phi)-\beta)}\right)
\end{aligned}
$$

where in inequality (c) we used the union bound and, assuming $\phi<\frac{q-1}{q}$, we used Lemma 19 on the probability of the event $d_{\mathsf{H}}\left(\boldsymbol{X}_i,\boldsymbol{Z}_{i,j}\right)\leq\phi L/2$. We further used Corollary 20 together with a union bound that at least one of at most $N\leq cM$ other i.i.d. uniform output sequences has distance at most ϕL to $\boldsymbol{Z}_{i,1}$. In inequality (d), we used that $\sum_{d_i\geq 1}\mathsf{Pr}\left(d_i\right)d_i=\mathsf{E}\left[d_i\right]$ and $M=q^{\beta L}$. Next, we compute

$$
\mathsf{E}\left[M-N_0-\sum_{i=1}^{M}G_i\right]\leq M\left(c\mathrm{e}^{-2L(\phi/2-p)^2}+c^2q^{-L(1-H_q(\phi)-\beta)}\right),
$$

where we used that $\mathsf{E}\left[N_0\right]=\sum_{i=1}^{M}\mathsf{Pr}\left(D_i=0\right)$ and $\sum_{i=1}^{M}\mathsf{E}\left[D_i\right]\leq cM$. Using Markov's inequality, we conclude that the probability of the first summand in (15) is at most

$$
\begin{aligned}
&\mathsf{Pr}\left(M-N_0-\sum_{i=1}^{M}G_i\geq M\epsilon/2\right)\\
&\qquad\leq\frac{2c\mathrm{e}^{-2L(\phi/2-p)^2}+2c^2q^{-L(1-H_q(\phi)-\beta)}}{\epsilon},
\end{aligned}
$$

which approaches 0 as $M\to\infty$ for any $\beta<1-H_q(\phi)$ and $\phi>2p$ and thus the lemma statement follows. □

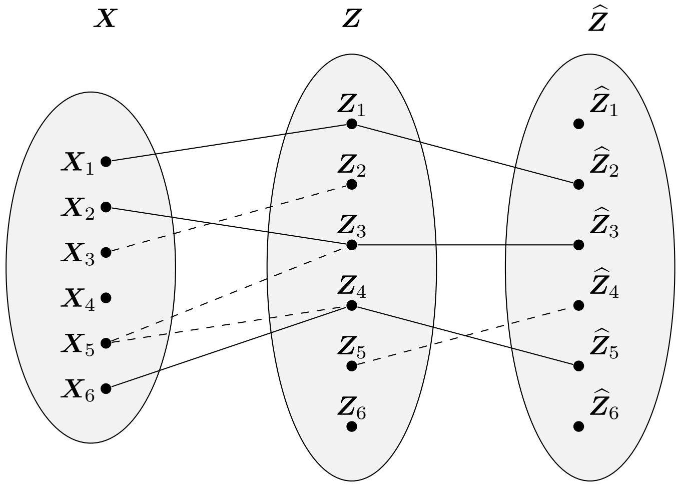

Fig. 4: Illustration of the tripartite matching graph in the proof of Lemma 15. The solid lines highlight edges, which contribute to the joint typicality $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\widehat{\boldsymbol{Z}})$.

We continue with a relationship between the number of correct clusters and the typicality. Loosely speaking, the next lemma shows that the sizes of the matching between $\boldsymbol{X}$, $\boldsymbol{Z}$, and between $\boldsymbol{X}$, $\widehat{\boldsymbol{Z}}$ are close, if the number of correct clusters is close to M.

Lemma 15. *The joint typicality of $\boldsymbol{X}$ and $\widehat{\boldsymbol{Z}}$ satisfies*

$$
|T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\widehat{\boldsymbol{Z}})-T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\boldsymbol{Z})|\leq M-G.
$$

Proof. Consider a tripartite graph G_{tri} with vertices $\boldsymbol{X}_i$, $i\in[M]$ on the left, $\boldsymbol{Z}_i$, $i\in[M]$ in the middle and $\widehat{\boldsymbol{Z}}_i$, $i\in[M]$ on the right. We connect two vertices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_j$, if $(\boldsymbol{X}_i,\boldsymbol{Z}_j)\in\mathcal{T}^{(L,\epsilon)}_{\mathsf{Mul}}(d,p,q)$. We further draw an edge from $\boldsymbol{Z}_j$ to $\widehat{\boldsymbol{Z}}_i$, if the multiset of sequences in $\boldsymbol{Z}_j$ is equal to the multiset of sequences in $\widehat{\boldsymbol{Z}}_i$. This tripartite graph is illustrated in Figure 4.

Let $\mathcal{M}_{\mathsf{g}}\subseteq[M]$ be the vertices in the middle which belong to the largest matching between the middle and right vertices, i.e., that correspond to the correct clusters, and let $\mathcal{M}_{\mathsf{t}}\subseteq[M]$ be the vertices in the middle which belong to a matching between the middle and left vertices. With this definition,

$$
\begin{aligned}
T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\widehat{\boldsymbol{Z}})&\geq|\mathcal{M}_{\mathsf{t}}\cap\mathcal{M}_{\mathsf{g}}|=|\mathcal{M}_{\mathsf{t}}|+|\mathcal{M}_{\mathsf{g}}|-|\mathcal{M}_{\mathsf{t}}\cup\mathcal{M}_{\mathsf{g}}|\\
&\overset{(a)}{\geq}|\mathcal{M}_{\mathsf{t}}|+|\mathcal{M}_{\mathsf{g}}|-M=|\mathcal{M}_{\mathsf{t}}|+G-M,
\end{aligned}
$$

where in inequality (a) we used that both $\mathcal{M}_{\mathsf{t}}$ and $\mathcal{M}_{\mathsf{g}}$ are subsets of $[M]$. Choosing $|\mathcal{M}_{\mathsf{t}}|=T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\boldsymbol{Z})$ as the largest matching between the left and middle vertices, yields an upper bound on the difference. On the other hand,

$$
T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\widehat{\boldsymbol{Z}})\leq|\mathcal{M}_{\mathsf{t}}\cap\mathcal{M}_{\mathsf{g}}|+|[M]\backslash\mathcal{M}_{\mathsf{g}}|\leq T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\boldsymbol{Z})+M-G,
$$

since the number of correct clusters which can be matched to an input sequence is at most the size of the largest matching on the left $|\mathcal{M}_{\mathsf{t}}\cap\mathcal{M}_{\mathsf{g}}|\leq T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X},\boldsymbol{Z})$ and the $|[M]\backslash\mathcal{M}_{\mathsf{g}}|$ incorrect clusters could potentially also add to the joint typicality. □

B. Decoding Probability for the Correct Codeword

We continue with bounding the probability that the correct codeword $\boldsymbol{X}(1)$ has many typical matches with $\boldsymbol{Z}$.

Lemma 16. *Let $0 < \beta < 1, 0 < p < 1$, $q \in \mathbb{N}$ be fixed and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a regular distribution. For any $0 < \epsilon < 1$, as $M \to \infty$, it holds that*

$$\mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq M(1-\epsilon)|W = 1\right) \to 1.$$

Proof. We bound $\mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq M(1-\epsilon)|W = 1\right)$ from below. For an arbitrary $\epsilon > 0$ denote by $\mathcal{N}_\epsilon$ the event on the random variable $\boldsymbol{N}$ that $\sum_{d\geq 0}\left|\frac{n_d}{M} - \nu_d\right| \leq \epsilon/4$. We demarginalize with respect to the drawing composition $\boldsymbol{D}$ and obtain

$$\begin{aligned}
&\mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq M(1-\epsilon)|W = 1\right) \\
&= \sum_{\boldsymbol{d}} \mathsf{Pr}\left(\boldsymbol{d}\right) \mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq M(1-\epsilon)|W = 1, \boldsymbol{d}\right) \\
&\geq \sum_{\boldsymbol{d}\in\mathcal{N}_\epsilon} \mathsf{Pr}\left(\boldsymbol{d}\right) \mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq M(1-\epsilon)|W = 1, \boldsymbol{d}\right),
\end{aligned}$$

where we used that the drawing composition $\boldsymbol{D}$ is independent of the message W. Note that the event $\mathcal{N}_\epsilon$ is defined as an event on the drawing frequency $\boldsymbol{N}$, however since $\boldsymbol{N}$ is a function of $\boldsymbol{D}$, one can also view it as an event on the drawing composition $\boldsymbol{D}$. In the sequel, we will analyze the number

$$T^{\epsilon}_{\mathsf{OPM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \triangleq \left|\left\{ i \in [M] : (\boldsymbol{X}_i(1), \boldsymbol{Z}_i) \in \mathcal{T}^{(L,\epsilon)}_{\mathsf{Mul}}(d_i, p, q)\right\}\right|$$

of *ordered* jointly typical pairs over the parallel multinomial (OPM) channel. This is because $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq T^{\epsilon}_{\mathsf{OPM}}(\boldsymbol{X}(1), \boldsymbol{Z})$ due to the fact that each matching $\boldsymbol{Z}_i$ to $\boldsymbol{X}_i(1)$ for all pairs of sequences that contribute to $T^{\epsilon}_{\mathsf{OPM}}(\boldsymbol{X}(1), \boldsymbol{Z})$ also gives a matching for $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z})$. Note that this matching could also be larger, however this bound is sufficient for our analysis as this implies that

$$\begin{aligned}
&\mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq (1-\epsilon)M|W = 1, \boldsymbol{D} = \boldsymbol{d}\right) \\
&\geq \mathsf{Pr}\left(T^{\epsilon}_{\mathsf{OPM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq (1-\epsilon)M|W = 1, \boldsymbol{D} = \boldsymbol{d}\right).
\end{aligned}$$

For a given number of draws $\boldsymbol{D} = \boldsymbol{d}$ the size of the largest typical matching $T^{\epsilon}_{\mathsf{OPM}}(\boldsymbol{Z}(1), \boldsymbol{Y})$ is the sum of M independent random Bernoulli random variables with success probabilities $\pi_i \triangleq \mathsf{Pr}\left((\boldsymbol{X}_i(1), \boldsymbol{Z}_i) \in \mathcal{T}^{(L,\epsilon)}_{\mathsf{Mul}}(d_i, p, q)|W = 1, D_i = d_i\right)$. From the results about jointly typical sequences [29, Thm. 7.6.1] we know that for all $\epsilon > 0$ and $i \in [M]$, it holds that $\pi_i > 1 - \epsilon/2$ for all $L \geq L_{d_i}$, as $\boldsymbol{Z}_i$ is the result of transmitting $\boldsymbol{X}_i(1)$ over the multinomial channel. As $\max_{i\in[M]} L_{d_i}$ might increase with M, we focus our attention to a subset of multinomial channels whose number of draws is bounded from above by a large, but finite quantity. To this end, let D_ϵ be the smallest integer such that $\sum_{d\geq D_\epsilon} \nu_d < \epsilon/4$. We have that for all $\boldsymbol{d} \in \mathcal{N}_\epsilon$, the number of positions $i \in [M]$ with $d_i < D_\epsilon$ is at least

$$\sum_{d=0}^{D_\epsilon - 1} n_d \geq M \sum_{d=0}^{D_\epsilon - 1} \nu_d - \frac{M\epsilon}{4} > M\left(1 - \frac{\epsilon}{2}\right).$$

Thus, at least $M(1-\epsilon/2)$ Bernoulli variables have success probability at least $\pi_i > 1 - \epsilon/2$ for all $L \geq \max_{0\leq d<D_\epsilon} L_d$ (which is finite) and we obtain

$$\begin{aligned}
&\mathsf{Pr}\left(T^{\epsilon}_{\mathsf{OPM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq (1-\epsilon)M|W = 1, \boldsymbol{D} = \boldsymbol{d}\right) \\
&\geq \sum_{i=M-M\epsilon}^{M-\frac{M\epsilon}{2}} \binom{M - \frac{M\epsilon}{2}}{i} \left(1 - \frac{\epsilon}{2}\right)^i \left(\frac{\epsilon}{2}\right)^{M-\frac{M\epsilon}{2}-i} \\
&= \sum_{i=0}^{\frac{M\epsilon}{2}} \binom{M - \frac{M\epsilon}{2}}{i} \left(1 - \frac{\epsilon}{2}\right)^{M-\frac{M\epsilon}{2}-i} \left(\frac{\epsilon}{2}\right)^{i} \\
&= 1 - \sum_{i=\frac{M\epsilon}{2}+1}^{M-\frac{M\epsilon}{2}} \binom{M - \frac{M\epsilon}{2}}{i} \left(1 - \frac{\epsilon}{2}\right)^{M-\frac{M\epsilon}{2}-i} \left(\frac{\epsilon}{2}\right)^{i} \\
&\overset{(a)}{\geq} 1 - \mathrm{e}^{-2\left(M-\frac{M\epsilon}{2}\right)\left(\frac{\epsilon^2}{4-2\epsilon}\right)^2},
\end{aligned}$$

for all $0 < \epsilon < 1$ and large enough L. Here we used Lemma 19 to bound the binomial tail in inequality (a). Thus, finally, for any $0 < \epsilon < 1$ and large enough L,

$$\begin{aligned}
&\mathsf{Pr}\left(T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(1), \boldsymbol{Z}) \geq (1-\epsilon)M|W = 1\right) \\
&\geq \left(1 - \mathrm{e}^{-2\left(M-\frac{M\epsilon}{2}\right)\left(\frac{\epsilon^2}{2-\epsilon}\right)^2}\right) \mathsf{Pr}\left(\boldsymbol{D} \in \mathcal{N}_\epsilon\right),
\end{aligned}$$

where the first term approaches 1 as $M \to \infty$ for any $1 < \epsilon < 0$ and the second term approaches 1 as well by assumption of convergence in frequency on the drawing distribution. The claim of the lemma follows. □

C. Decoding Probability for the Wrong Codewords

The next lemma proves that the probability that any other codeword $\boldsymbol{X}(i)$, $i \geq 2$ has many typical matches with $\boldsymbol{Z}$ is small.

Lemma 17. *Let $0 < \beta < 1, 0 < p < 1$, $q \in \mathbb{N}$ be fixed and $\mathsf{Pr}\left(\boldsymbol{i}\right)$ be a regular distribution that converges in frequency to $\boldsymbol{\nu}$. For any $0 < \epsilon < 1$, and any $R < C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q) - 5\epsilon$, as $M \to \infty$,*

$$\mathsf{Pr}\left(\exists i : 2 \leq i \leq q^{MLR}, T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \boldsymbol{Z}) \geq M(1-\epsilon)|W = 1\right) \to 0.$$

Proof. Denote by $\mathcal{J}'_i$ the event that $T^{\epsilon}_{\mathsf{PM}}(\boldsymbol{X}(i), \boldsymbol{Z}) \geq M(1-\epsilon)$. We again denote by $\mathcal{N}_\epsilon$ the event for $\boldsymbol{N}$ that $\sum_{d\geq 0}\left|\frac{n_d}{M} - \nu_d\right| \leq \epsilon/4$. Similar as in the proof of Lemma 16 we demarginalize with respect to the drawing composition $\boldsymbol{D}$ and obtain

$$\begin{aligned}
&\mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}'_i \middle| W = 1\right) \\
&\leq \mathsf{Pr}\left(\boldsymbol{D} \notin \mathcal{N}_\epsilon\right) + \sum_{\boldsymbol{d}\in\mathcal{N}_\epsilon} \mathsf{Pr}\left(\boldsymbol{d}\right) \mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}'_i \middle| W = 1, \boldsymbol{D} = \boldsymbol{d}\right) \\
&\overset{(a)}{\leq} \mathsf{Pr}\left(\boldsymbol{D} \notin \mathcal{N}_\epsilon\right) + q^{MLR} \sum_{\boldsymbol{d}\in\mathcal{N}_\epsilon} \mathsf{Pr}\left(\boldsymbol{d}\right) \mathsf{Pr}\left(\mathcal{J}'_2|W = 1, \boldsymbol{D} = \boldsymbol{d}\right),
\end{aligned} \tag{16}$$

where in inequality (a) we used the union bound together with the fact that $\mathsf{Pr}\left(\mathcal{J}'_i|W = 1, \boldsymbol{D} = \boldsymbol{d}\right)$ is invariant over all $2 \leq$

$i \leq q^{MLR}$ due to the independent and identical choice of codewords. To start with, denote $\boldsymbol{X}(2) = (\boldsymbol{X}_1(2), \dots, \boldsymbol{X}_M(2))$ as the random codeword $\boldsymbol{X}(2)$. For an arbitrary $h \in [M]$ denote by $\mathcal{P}(M,h) = \{\boldsymbol{m} = (m_1, \dots, m_h) \in [M]^h : m_i \neq m_j \ \forall \ i \neq j\}$ the set of length-h partial permutations of the set $[M]$. Denote further by $T_{m,j}$ a Bernoulli random variable, which is equal to 1, if $(\boldsymbol{X}_m(2), \boldsymbol{Z}_j) \in \mathcal{T}^{(L,\epsilon)}_{\mathsf{Mul}}(d_j, p, q)$ and 0, otherwise. This allows to rewrite the above probability as

$$\begin{aligned}&\mathsf{Pr}\left(\mathcal{J}_2' \middle| W=1, \boldsymbol{D}=\boldsymbol{d}\right)\\ &= \mathsf{Pr}\left(\exists \boldsymbol{m} \in \mathcal{P}(M,M) : \sum_{j=1}^{M} T_{m_j,j} \geq M(1-\epsilon) \middle| W=1, \boldsymbol{D}=\boldsymbol{d}\right).\end{aligned}$$

We next use the fact that $T_{m,j} = 1$ with probability 1 for all empty clusters, i.e., $j \in [M]$ with $d_j = 0$. Denote by $j_1, \dots, j_{M-n_0}$ those indices with $d_{j_t} > 0$ for all $1 \leq t \leq M - n_0$. Then, we can simplify the above expression by

$$\begin{aligned}&\mathsf{Pr}\left(\mathcal{J}_2' \middle| W=1, \boldsymbol{D}=\boldsymbol{d}\right)\\ &\overset{(b)}{\leq} \sum_{\boldsymbol{m}' \in \mathcal{P}(M, M-n_0)} \mathsf{Pr}\left(\sum_{t=1}^{M-n_0} T_{m_t', j_t} \geq M(1-\epsilon) - n_0 \middle| W=1, \boldsymbol{D}=\boldsymbol{d}\right),\end{aligned}$$

where inequality (b) is due to an application of the union bound. We will bound the above probability as follows. To start with, since $\boldsymbol{X}(2)$ is chosen independently from $\boldsymbol{Z}$, given $\boldsymbol{D} = \boldsymbol{d}$, for all $i,j \in [M]$, $\pi_j \triangleq \mathsf{Pr}\left(T_{i,j} = 1 | W = 1, \boldsymbol{D} = \boldsymbol{d}\right) < q^{-L(C_{\mathsf{Mul}}(d_j,p,q)-\epsilon)}$ for $L \geq L_{d_j}$ [29, Thm. 7.6.1], where L_{d_j} are large enough integers that depend on ϵ and the channel d_j. Since at least $M(1-\epsilon) - n_0$ of the Bernoulli variables T_{m_t', j_t} must be equal to 1, we can use these definitions to bound the above probability

$$\begin{aligned}&\mathsf{Pr}\left(\sum_{t=1}^{M-n_0} T_{m_t', j_t} \geq M(1-\epsilon) - n_0 \middle| W=1, \boldsymbol{D}=\boldsymbol{d}\right)\\ &\leq \sum_{\mathcal{I} \subseteq [M-n_0] : |\mathcal{I}| = M(1-\epsilon) - n_0} \prod_{t \in \mathcal{I}} \pi_{j_t} \overset{(c)}{\leq} \sum_{\mathcal{I} \subseteq [M] : |\mathcal{I}| \leq M(1-\epsilon)} \prod_{j \in \mathcal{I}} \pi_j\\ &\leq \binom{M}{M(1-\epsilon)} \max_{\mathcal{I} \subseteq [M] : |\mathcal{I}| = M(1-\epsilon)} \prod_{j \in \mathcal{I}} \pi_j.\end{aligned}$$

Note that in inequality (c) we factored those j with $d_j = 0$ into the product, which will simplify the subsequent notation and analysis. Mathematically, inequality (c) holds, as each set $\mathcal{I}_1$ in the first sum is contained in some set $\mathcal{I}_2$ of the second sum such that the positions $j \in \mathcal{I}_2 \setminus \mathcal{I}_1$ are exactly those positions with $d_j = 0$ and $\pi_j = 1$ and thus each term in the first sum is accounted for by at least one term in the second sum. In order to use the above bound on π_j, we restrict our attention to those channels j with at most a finite but large number of draws, such that the maximum over L_{d_j} is guaranteed to be constant in M. To this end, let D_ϵ be the smallest integer such that $\sum_{d \geq D_\epsilon} \nu_d < \epsilon/4$ and abbreviate $\mathcal{D}(\epsilon) = \{j \in [M] : d_j < D_\epsilon\}$. We can then bound the product over π_j to

$$\begin{aligned}\prod_{j \in \mathcal{I}} \pi_j \leq \prod_{j \in \mathcal{I} \cap \mathcal{D}(\epsilon)} \pi_j < \prod_{j \in \mathcal{I} \cap \mathcal{D}(\epsilon)} q^{-L(C_{\mathsf{Mul}}(d_j,p,q)-\epsilon)}\\ = q^{-L \sum_{j \in \mathcal{I} \cap \mathcal{D}(\epsilon)} (C_{\mathsf{Mul}}(d_i,p,q)-\epsilon)}\end{aligned}$$

for all $L \geq \max_{0 \leq d < D_\epsilon} L_d$, which is constant and not a function of M or L, as desired. Analyzing the exponent of the error probability expression above, we find that

$$\begin{aligned}&\sum_{j \in \mathcal{I} \cap \mathcal{D}(\epsilon)} (C_{\mathsf{Mul}}(d_j,p,q) - \epsilon)\\ &= \left(\sum_{j=1}^{M} (C_{\mathsf{Mul}}(d_j,p,q) - \epsilon) - \sum_{j \notin \mathcal{I} \cap \mathcal{D}(\epsilon)} (C_{\mathsf{Mul}}(d_j,p,q) - \epsilon)\right)\\ &\overset{(d)}{\geq} \left(\sum_{j=1}^{M} (C_{\mathsf{Mul}}(d_j,p,q) - \epsilon) - \frac{3M\epsilon}{2}\right)\\ &= \sum_{d \geq 0} n_d C_{\mathsf{Mul}}(d,p,q) - \frac{5M\epsilon}{2} \geq M \sum_{d \geq 0} \nu_d C_{\mathsf{Mul}}(d,p,q) - \frac{7M\epsilon}{2},\end{aligned}$$

where in inequality (d) we bounded the second sum using $C_{\mathsf{Mul}}(d_j,p,q) \leq 1$ together with the fact that by definition of D_ϵ and for all $\boldsymbol{d} \in \mathcal{N}_\epsilon$,

$$|\mathcal{D}(\epsilon)| = \sum_{d=0}^{D_\epsilon - 1} n_d \geq M \sum_{d=0}^{D_\epsilon - 1} \nu_d - \frac{M\epsilon}{4} > M\left(1 - \frac{\epsilon}{2}\right),$$

and thus

$$|\{j \in [M] : j \notin \mathcal{I} \cap \mathcal{D}(\epsilon)\}| \leq M - |\mathcal{I}| + M - |\mathcal{D}(\epsilon)| = \frac{3M\epsilon}{2}.$$

For any $0 < \epsilon < 1$ and large enough M, and any $\boldsymbol{d} \in \mathcal{N}_\epsilon$, the resulting upper bound $\mathsf{Pr}\left(\mathcal{J}_2' | W = 1, \boldsymbol{D} = \boldsymbol{d}\right)$ is henceforth

$$\begin{aligned}&\mathsf{Pr}\left(\mathcal{J}_2' \middle| W=1, \boldsymbol{D}=\boldsymbol{d}\right)\\ &\leq |\mathcal{P}(M, M-n_0)| \binom{M}{M(1-\epsilon)} q^{-ML(\sum_{d \geq 0} \nu_d C_{\mathsf{Mul}}(d,p,q) - 7\epsilon/2)}\\ &\overset{(e)}{\leq} 2^M q^{-ML(\sum_{d \geq 0} \nu_d C_{\mathsf{Mul}}(d,p,q) - \beta(1-\nu_0) - 9\epsilon/2)}, \qquad (17)\end{aligned}$$

where we used $\binom{M}{M(1-\epsilon)} \leq 2^M$ and $|\mathcal{P}(M, M - n_0)| = \frac{M!}{n_0!} \leq M^{M-n_0} \leq q^{\beta LM(1-\nu_0+\epsilon)}$ for all $\boldsymbol{d} \in \mathcal{N}_\epsilon$ in inequality (e). Plugging (17) into the average code ensemble error probability (16), we obtain

$$\begin{aligned}&\mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i' \middle| W = 1\right)\\ &\leq \mathsf{Pr}\left(\boldsymbol{D} \notin \mathcal{N}_\epsilon\right) + q^{-ML(-R + C_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q) - 9\epsilon/2 - \log_q(2)/L)}.\end{aligned}$$

In the above expression, it holds that $\mathsf{Pr}\left(\boldsymbol{D} \notin \mathcal{N}_\epsilon\right) \to 0$ as $M \to \infty$ by assumption of a drawing distribution that converges in frequency. Further, $\log_q(2)/L \to 0$ as $M \to \infty$ and thus, for any $R < C_{\mathsf{ND}}(\boldsymbol{\nu}, \beta, p, q) - 9\epsilon/2$, the sought-after error probability converges to 0 as $M \to \infty$, which proves the claim of the lemma. □

D. *Proof of Lemma 11*

We are now in the position to prove Lemma 11 using the ingredients derived above.

Proof of Lemma 11. The average probability of a decoding error, averaged over all codebooks, is given by

$$\mathsf{Pr}\left(\mathsf{Err}\right) = \sum_{\mathcal{C}} \mathsf{Pr}\left(\mathcal{C}\right) \mathsf{Pr}\left(\mathsf{Err} | \mathcal{C}\right) = \mathsf{Pr}\left(\mathsf{Err} | W = 1\right),$$

where the last equality is due to the symmetry of the choice of random codebooks, see, e.g., [29, Ch. 7.7]. The two possible error events are that either $\boldsymbol{X}(1)$ is not jointly typical with $\widehat{\boldsymbol{Z}}$ or that one of the other codewords is jointly typical with respect to $\widehat{\boldsymbol{Z}}$. Denote by $\mathcal{J}_i$ the event that the i-th codeword $\boldsymbol{X}(i)$ is jointly typical with $\widehat{\boldsymbol{Z}}$, i.e., $T_{\mathsf{PM}}^{\epsilon}(\boldsymbol{X}(i), \widehat{\boldsymbol{Z}}) \geq M(1-\epsilon)$ and by $\mathcal{J}_i^{\mathsf{c}}$ the complement event. By the union bound we obtain

$$\begin{aligned}\mathsf{Pr}\left(\mathsf{Err}|W=1\right) &\leq \mathsf{Pr}\left(\mathcal{J}_1^{\mathsf{c}} \cup \bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i \middle| W=1\right)\\ &\leq \mathsf{Pr}\left(\mathcal{J}_1^{\mathsf{c}}|W=1\right) + \mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i \middle| W=1\right).\end{aligned}$$

The first summand can be bounded as follows. Denote by $\mathcal{G}$ the event that $G \geq M(1-\epsilon/2)$. We obtain

$$\begin{aligned}\mathsf{Pr}(\mathcal{J}_1^{\mathsf{c}}|W=1) &\overset{(a)}{\leq} \mathsf{Pr}\left(\mathcal{J}_1^{\mathsf{c}}|W=1,\mathcal{G}\right) + 1 - \mathsf{Pr}\left(\mathcal{G}\right)\\ &\overset{(b)}{\leq} \mathsf{Pr}\left(T_{\mathsf{PM}}^{\epsilon}(\boldsymbol{X}(1),\boldsymbol{Z}) < M(1-\epsilon/2)|W=1,\mathcal{G}\right)+1-\mathsf{Pr}\left(\mathcal{G}\right)\\ &\overset{(c)}{\leq} \frac{\mathsf{Pr}\left(T_{\mathsf{PM}}^{\epsilon}(\boldsymbol{X}(1),\boldsymbol{Z}) < M(1-\epsilon/2)|W=1\right)}{\mathsf{Pr}\left(\mathcal{G}\right)}+1-\mathsf{Pr}\left(\mathcal{G}\right)\\ &\overset{(d)}{=} o(1),\end{aligned}$$

where we used Lemma 21 in inequalities (a) and (c) and Lemma 15 in inequality (b). Inequality (d) was proven by applying Lemma 14 on the probability of having many good clusters, where we require $\beta < 1 - H_q(2p)$ and $p < \frac{q-1}{2q}$ to guarantee the existence of a suitable ϕ. We further used Lemma 16 on the probability of having a large matching between $\boldsymbol{X}$ and $\boldsymbol{Z}$. Conversely, we bound the second summand

$$\mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i \middle| W=1\right) \overset{(e)}{\leq} \mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i \middle| W=1,\mathcal{G}\right) + 1 - \mathsf{Pr}\left(\mathcal{G}\right).$$

Abbreviate further by $\mathcal{J}_i^*$ the event that $T_{\mathsf{PM}}^{\epsilon}(\boldsymbol{X}(i),\boldsymbol{Z}) \geq M(1-3\epsilon/2)$, we obtain

$$\begin{aligned}\mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i \middle| W=1,\mathcal{G}\right) &\overset{(f)}{\leq} \mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i^* \middle| W=1,\mathcal{G}\right)\\ &\overset{(g)}{\leq} \frac{\mathsf{Pr}\left(\bigcup_{i=2}^{q^{MLR}} \mathcal{J}_i^* \middle| W=1\right)}{\mathsf{Pr}\left(\mathcal{G}\right)} \overset{(h)}{=} o(1),\end{aligned}$$

where we used Lemma 21 in inequalities (e) and (g) and Lemma 15 in inequality (f). Inequality (h) follows from Lemma 14 on the probability of having many good clusters and Lemma 17 on the probability of having a large matching between $\boldsymbol{X}$ and $\boldsymbol{Z}$ under the assumption that $R < C_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q) - 5\epsilon$. It follows that for any $\epsilon > 0$ and $R < C_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q) - 5\epsilon$, the error probability vanishes, $\mathsf{Pr}\left(\mathsf{Err}|W=1\right) \to 0$ as $M \to \infty$.

As we can choose ϵ arbitrarily small, it follows that for each $R < C_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q)$, the error probability $\mathsf{Pr}\left(\mathsf{Err}|W=1\right) \to 0$ vanishes as $M \to \infty$. Since the average error probability over all codebooks vanishes, for all code rates $R < C_{\mathsf{ND}}(\boldsymbol{\nu},\beta,p,q)$, there exists at least one codebook of rate R that has vanishing error rate and thus R is an achievable rate. □

VI. Application to Popular Drawing Distributions

We continue with applying Theorem 4 to two examples of popular drawing distributions.

A. Bernoulli Draws

We start with the case of independent Bernoulli draws, i.e., $\mathsf{Pr}\left(\boldsymbol{D}=\boldsymbol{d}\right) = \prod_{i=1}^{M}\mathsf{Pr}\left(D_i=d_i\right)$ and $\mathsf{Pr}\left(D_i=0\right) = 1 - \mathsf{Pr}\left(D_i=1\right) = r$. This case has been investigated in [25]. It is easy to check that this distribution is regular in the sense of Definition 1 as the distribution converges in frequency to $\nu_0 = 1-\nu_1 = r$ by the weak law of large numbers. Further, the number of draws is limited to M. It follows that the capacity in this case is given by

$$C_{\mathsf{Ber}}(r,\beta,p,q) = (1-r)(1-H_q(p)-\beta).$$

B. Independent and Uniform Draws

We now consider the case, where $N = cM$ is fixed for some $c > 0$ and $\boldsymbol{I}$ are i.i.d. variables with $\mathsf{Pr}\left(I_j = i\right) = \frac{1}{M}$ for all $i \in [M]$ and $j \in [N]$. This is the distribution discussed in [1], [2], [3], [21], [26]. We will show that the capacity with such a drawing distribution is precisely

$$C_{\mathsf{DNA}}(c,\beta,p,q) = \sum_{d\geq 0}\mathsf{Poi}_c(d)C_{\mathsf{Mul}}(d,p,q) - \beta(1-\mathrm{e}^{-c}).$$

In view of Theorem 4, we merely have to prove that this distribution is regular.

Lemma 18. *The distribution with i.i.d. uniform draws* $\mathsf{Pr}\left(I_j=i\right) = \frac{1}{M}$ *for all* $j \in [N]$ *and* $i \in [M]$ *is a regular distribution that converges to* $\nu_d = \mathsf{Poi}_c(d) \triangleq \mathrm{e}^{-c}c^d/d!$ *in frequency.*

Proof. We start by showing that the drawing frequency $N_d = |\{i \in [M] : D_i = d\}|$ converges to $\nu_d = \mathsf{Poi}_c(d)$ in frequency. To this end, for an arbitrary $\epsilon > 0$ denote by $\mathcal{N}_\epsilon$ the event that $\sum_{d\geq 0}|N_d - M\nu_d| \leq \epsilon M$. We will use the effect of *Poissonization* [31] of the drawing composition. Denote by $\widetilde{D}_i$, $i \in [M]$ independent and identically distributed random variables with mean c, i.e., $\mathsf{Pr}\left(\widetilde{D}_i = d\right) = \nu_d$. It has been shown [31, Corollary 2.12] that any event on the exact distribution has probability at most $\sqrt{2\pi\mathrm{e}N}$ times the probability of the event for the case of i.i.d. Poisson variables. It follows that

$$\mathsf{Pr}\left(\boldsymbol{D} \notin \mathcal{N}_\epsilon\right) \leq \sqrt{2\pi\mathrm{e}N}\,\mathsf{Pr}\left(\widetilde{\boldsymbol{D}} \notin \widetilde{\mathcal{N}}_\epsilon\right), \tag{18}$$

where $\widetilde{\mathcal{N}}_\epsilon$ is the event that $\sum_{d\geq 0}|\widetilde{N}_d - M\nu_d| \leq \epsilon M$, where $\widetilde{N}_d = |\{i \in [M] : \widetilde{D}_i = d\}|$ is the drawing frequency derived from the i.i.d. Poisson variables. We will split the probability in sequences with many draws and few draws,

$$\mathsf{Pr}\left(\widetilde{\boldsymbol{D}}\notin\mathcal{N}_\epsilon\right) \leq \mathsf{Pr}\left(\sum_{d=0}^{M^{1/3}-1} |\widetilde{N}_d - M\nu_d| \geq \frac{\epsilon M}{2}\right) + \mathsf{Pr}\left(\sum_{d\geq M^{1/3}} |\widetilde{N}_d - M\nu_d| \geq \frac{\epsilon M}{2}\right). \quad (19)$$

We will now use that $\widetilde{N}_d$ is binomial distributed with M trials and success probability ν_d and thus

$$\mathsf{Pr}\left(|\widetilde{N}_d - M\nu_d| \geq \frac{\epsilon M^{2/3}}{2}\right) \overset{(a)}{\leq} 2\mathrm{e}^{-\frac{\epsilon^2}{2}M^{1/3}},$$

where we used Lemma 19 on the two-sided binomial tail distribution in inequality (a). Therefore, by the union bound, the first summand in (19) is at most

$$\mathsf{Pr}\left(\sum_{d=0}^{M^{1/3}-1} |\widetilde{N}_d - M\nu_d| \geq \frac{\epsilon M}{2}\right) \leq \sum_{d=0}^{M^{1/3}-1} \mathsf{Pr}\left(|\widetilde{N}_d - M\nu_d| \geq \frac{\epsilon M^{2/3}}{2}\right) \leq 2M^{\frac{1}{3}}\mathrm{e}^{-\frac{\epsilon^2}{2}M^{\frac{1}{3}}}. \quad (20)$$

Next, we will bound the second summand in (19). By the triangle inequality, $|\widetilde{N}_d - M\nu_d| \leq \widetilde{N}_d + M\nu_d$ and we thus bound the second summand by

$$\mathsf{Pr}\left(\sum_{d\geq M^{1/3}} |\widetilde{N}_d - M\nu_d| \geq \frac{\epsilon M}{2}\right) \leq \mathsf{Pr}\left(\sum_{d\geq M^{1/3}} \widetilde{N}_d \geq M\left(\frac{\epsilon}{2} - \sum_{d\geq M^{1/3}} \nu_d\right)\right).$$

To begin with, we see that $\sum_{d\geq M^{1/3}} \widetilde{N}_d$ is a binomial distribution with M trials and success probability $\sum_{d\geq M^{1/3}} \nu_d$. The tail of the Poisson distribution has exponential decay, see, e.g., [32, Theorem 2.1], or, more precisely,

$$\sum_{d\geq M^{1/3}} \nu_d \leq \mathrm{e}^{-M^{1/3}}$$

for all $M \geq (7c)^3$ by the result of [32, Eq. 2.11]. It follows that we can derive the following upper bound on the outage probability

$$\mathsf{Pr}\left(\sum_{d\geq M^{1/3}} \widetilde{N}_d \geq M\left(\frac{\epsilon}{2} - \sum_{d\geq M^{1/3}} \nu_d\right)\right) \leq \mathrm{e}^{-2M\left(\epsilon/2 - 2\mathrm{e}^{-M^{1/3}}\right)^2} \quad (21)$$

where we used the bound on the binomial tail from Lemma 19. Note that this inequality only holds for large enough M, as we require $M \geq (7c)^3$ and also $\epsilon/2 > 2\mathrm{e}^{-M^{1/3}}$ in order for Lemma 19 to apply. Plugging (21), (20), and (19) into (18), we obtain

$$\mathsf{Pr}\left(\boldsymbol{D} \notin \mathcal{N}_\epsilon\right) \leq \sqrt{2\pi\mathrm{e}cM}\left(2M^{1/3}\mathrm{e}^{-\frac{\epsilon^2}{2}M^{1/3}} + \mathrm{e}^{-2M\left(\epsilon/2 - 2\mathrm{e}^{-M^{1/3}}\right)^2}\right) \to 0,$$

as $M \to \infty$, as the first exponent scales at least as $-\epsilon^2 M^{1/3}$ and the second exponent scales as $-\epsilon^2 M$ and consequently the polynomial scaling factors are asymptotically negligible.

Next, the total number of draws is precisely $N = cM$ by the definition of the drawing distribution and thus also the maximum draws condition from Definition 1 applies.

□

VII. Conclusion

This paper deals with the noisy drawing channel. The main contribution of our results is the derivation of Theorem 4, which states the capacity of the channel for moderate noise levels and a broad class of drawing distributions. Theorem 4 generalizes previous results, which have focused on different drawing distributions.

In view of our results and related research [21], [25], [26], there remain several intriguing open problems regarding the noisy drawing of sequences.

- The extension of the parameter range of Theorem 4 to a broader range or, similarly, extending the results of [26], which hold for a larger range of parameters but more specific drawing distributions, to more general drawing distributions is one interesting direction. Further, even for standard drawing distributions, the capacity for the extremely noisy case outside the range of the results from [26] has not yet been established.
- The generalization of the constituent channels to channels with memory, such as insertion-deletion channels is of practical importance in DNA-based data storage. While the capacity for even a single insertion-deletion channel is unknown to date, it seems not unlikely that expressions similar to those in Theorem 4 hold, provided that one can prove sufficient results on the appearance of clusters and on the nature of the capacity-achieving input distribution for multiple draw insertion-deletion channels.
- Another interesting research direction is to design efficiently encodable and decodable codes that achieve capacity. While there exist codes for the case, where each sequence is drawn at most once [25], the codes for the more general case achieve rates, which are relatively close to capacity [33], but do not yet approach capacity.

Appendix A
Auxiliary Lemmas

Lemma 19. *Let $n \in \mathbb{N}$, $0 < p < 1$ and $k \in \mathbb{N}$, $k \geq np$. Then, the binomial tail distribution can be bounded from above by*

$$\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \leq 2^{-nD\left(\frac{k}{n}||p\right)},$$
$$\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \leq \mathrm{e}^{-2n\left(\frac{k}{n}-p\right)^2}$$

where $D(p_1||p_2) = p_1 \log(p_1/p_2) + (1-p_1)\log((1-p_1)/(1-p_2))$ is the Kullback-Leibler divergence between two Bernoulli distributions with probabilities p_1 and p_2.

Proof. The first inequality is a well-known upper bound on the binomial tail, which can be found in, e.g., [34, Lemma 4.7.2]. The second inequality can directly be proven using Hoeffding inequality [35, Theorem 2.2.6], by using the property that the expected value of the binomial distribution is equal to np. □

Corollary 20. *Let $q \in \mathbb{N}$, $\alpha < \frac{q-1}{q}$ and $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_q^n$ be random sequences. If $\boldsymbol{x}$ is i.i.d. uniform and independent from $\boldsymbol{y}$, then*

$$\mathsf{Pr}\left(d_{\mathsf{H}}\left(\boldsymbol{x},\boldsymbol{y}\right) \leq \alpha L\right) \leq q^{-n(1-H_q(\alpha))}.$$

Proof. The statement follows from Lemma 19 as follows.

$$\begin{aligned}\mathsf{Pr}\left(d_{\mathsf{H}}\left(\boldsymbol{x},\boldsymbol{y}\right) \leq \alpha L\right) &= \sum_{i=0}^{\alpha n} \binom{n}{i} \left(\frac{1}{q}\right)^i \left(\frac{q-1}{q}\right)^{n-i} \\ &\overset{(a)}{=} \sum_{i=(1-\alpha)n}^{n} \binom{n}{i} \left(\frac{q-1}{q}\right)^i \left(\frac{1}{q}\right)^{n-i} \\ &\overset{(b)}{\leq} 2^{-nD((1-\alpha)||(1/q))} = q^{-n(1-H_q(\alpha))},\end{aligned}$$

where in inequality (a), we made a change of variable of the summation index $i \to n-i$ and in inequality (b), we used Lemma 19. □

Lemma 21. *For any events $\mathcal{A}, \mathcal{B}$, the conditional probability of $\mathcal{A}$ given $\mathcal{B}$ satisfies*

$$\mathsf{Pr}\left(\mathcal{A}\right) + \mathsf{Pr}\left(\mathcal{B}\right) - 1 \leq \mathsf{Pr}\left(\mathcal{A}|\mathcal{B}\right) \leq \frac{\mathsf{Pr}\left(\mathcal{A}\right)}{\mathsf{Pr}\left(\mathcal{B}\right)}.$$

Proof. The proof follows directly from basic stochastic principles. On the one hand, we have

$$\mathsf{Pr}\left(\mathcal{A}|\mathcal{B}\right) = \frac{\mathsf{Pr}\left(\mathcal{A}\cap\mathcal{B}\right)}{\mathsf{Pr}\left(\mathcal{B}\right)} \leq \frac{\mathsf{Pr}\left(\mathcal{A}\right)}{\mathsf{Pr}\left(\mathcal{B}\right)}.$$

For the lower bound, we denote by $\mathcal{B}^{\mathsf{c}}$ the *complement* event of $\mathcal{B}$ with $\mathsf{Pr}\left(\mathcal{B}^{\mathsf{c}}\right) = 1 - \mathsf{Pr}\left(\mathcal{B}\right)$. The lower bound then follows from the following series of inequalities

$$\begin{aligned}\mathsf{Pr}\left(\mathcal{A}|\mathcal{B}\right) &= \frac{\mathsf{Pr}\left(\mathcal{A}\cap\mathcal{B}\right)}{\mathsf{Pr}\left(\mathcal{B}\right)} \geq \mathsf{Pr}\left(\mathcal{A}\cap\mathcal{B}\right) \\ &= \mathsf{Pr}\left(\mathcal{A}\right) - \mathsf{Pr}\left(\mathcal{A}\cap\mathcal{B}^{\mathsf{c}}\right) \geq \mathsf{Pr}\left(\mathcal{A}\right) - \mathsf{Pr}\left(\mathcal{B}^{\mathsf{c}}\right).\end{aligned}$$

□

Lemma 22. *The capacity of the q-ary multinomial channel with d draws and error probability p is given by*

$$C_{\mathsf{Mul}}(d,p,q) = \frac{1}{q} \sum_{\substack{t_0,\dots,t_{q-1}: \\ t_0+\dots+t_{q-1}=d}} \binom{d}{t_0,\dots,t_{q-1}} \sum_{i=0}^{q-1} (1-p)^{t_i} \cdot$$
$$\left(\frac{p}{q-1}\right)^{d-t_i} \log_q \left(\frac{(1-p)^{t_i}\left(\frac{p}{q-1}\right)^{-t_i}}{\frac{1}{q}\sum_{j=0}^{q-1}(1-p)^{t_j}\left(\frac{p}{q-1}\right)^{-t_j}} \right),$$

where $\binom{d}{t_0,\dots,t_{q-1}} = \frac{d!}{t_0!\cdot t_1!\dots t_{q-1}!}$ is the multinomial coefficient. The capacity achieving input distribution is the uniform distribution.

Proof. We begin by noticing that the multinomial channel is discrete and memoryless channel. Therefore, the capacity can be found by maximizing the symbol-wise mutual information

$$C_{\mathsf{Mul}}(d,p) = \max_{\mathsf{Pr}(x)} I(X;Y).$$

We start by finding the maximizing input distribution $\mathsf{Pr}\left(x\right)$ and then compute the mutual information for this distribution. To this end, we first show that the channel exhibits symmetry, as defined in [36, Chapter 4.5], which allows to use [36, Theorem 4.5.2] to conclude that the uniform input distribution maximizes the mutual information. Using our notation, a channel is called symmetric[6] if there exists a partition $\Sigma_q^d(1), \dots, \Sigma_q^d(P)$ of Σ_q^d such that for each part $j \in [P]$ it holds that the multiset $\{\{p_d(y|x) : x \in \Sigma_q\}\}$ is invariant over all $y \in \Sigma_q^d(j)$ and the multiset $\left\{\left\{p_d(y|x) : y \in \Sigma_q^d(j)\right\}\right\}$ is invariant over all $x \in \Sigma_q$. In our case we will partition Σ_q^d into parts for which the set of all numbers of symbol occurrences is the same. More precisely, let $\mathsf{ct}_x(y) = |\{i \in [d] : y_i = x\}|$ be the number of occurrences of the symbol $x \in \Sigma_q$ in y and we define the count spectrum

$$\mathsf{cs}(y) = \{\{\mathsf{ct}_x(y) : x \in \Sigma_q\}\}$$

as the multiset of the number of occurrences of each symbol in y. We then partition Σ_q^d into $\Sigma_q^d(1), \dots, \Sigma_q^d(P)$ according to $\mathsf{cs}(Y)$ such that for all $y_1 \in \Sigma_q^d(j_1)$, and $y_2 \in \Sigma_q^d(j_2)$, $\mathsf{cs}(y_1) = \mathsf{cs}(y_2)$ holds if and only if $j_1 = j_2$. Therefore $\mathsf{cs}(y)$ is constant over all y in one part $\Sigma_q^d(j)$. Using the fact that

$$p_d(y|x) = (1-p)^{d-t}\left(\frac{p}{q-1}\right)^t,$$

where $t = |\{i \in [d] : y_i \neq x\}|$ is the number of times that a symbol in y is different from x, we have

$$\{\{p_d(y|x) : x \in \Sigma_q\}\} = \left\{\left\{(1-p)^{d-t}\left(\frac{p}{q-1}\right)^t : t \in \mathsf{cs}(y)\right\}\right\},$$

which is invariant over all $y \in \Sigma_q^d(j)$. Further, for all $x \in \Sigma_q$ the number

$$|\{y \in \Sigma_q^d(j) : |\{i \in [d] : y_i \neq x\}| = t\}|$$

[6] In other words, if we view $p_d(y|x)$ as a matrix, whose rows are indexed by x and whose columns are indexed by y, then a channel is symmetric if there exists a partition of the columns of the matrix $p_d(y|x)$ such that each submatrix, obtained by restricting $p_d(y|x)$ to the columns corresponding to a part, has the property that the rows are permutations of each other and the columns are permutation of each other.

of words $y \in \Sigma_q^d(j)$ with a given $\mathsf{cs}(y)$ that have exactly t symbols equal to a given $x \in \Sigma_q$ is only a function of j and t and does not depend on x. it follows that the set $\left\{\left\{p_d(y|x) : y \in \Sigma_q^d(j)\right\}\right\}$ does not depend on x and thus the multinomial channel is symmetric.

Due to the symmetry, we know that the uniform input distribution $\mathsf{Pr}\left(x\right) = \frac{1}{q}$ for all $x \in \Sigma_q$ maximizes mutual information. We thus proceed with computing the entropies $H(Y)$ and $H(Y|X)$ for uniform inputs. We obtain for the output distribution

$$\begin{aligned}\mathsf{Pr}\left(y\right) &= \sum_{x \in \Sigma_q} p_d(y|x) \mathsf{Pr}\left(x\right) \\ &= \frac{1}{q} \sum_{a \in \Sigma_q} (1-p)^{\mathsf{ct}_a(y)} \left(\frac{p}{q-1}\right)^{d-\mathsf{ct}_a(y)},\end{aligned}$$

where we used that the multiset $\{\{p_d(y|x) : x \in \Sigma_q\}\}$ does not depend on x as shown above and we thus can express $\mathsf{Pr}\left(y\right)$ only as a function of the number of appearances of each symbol $a \in \Sigma_q$ in y. In order to compute the output entropy, we now use the fact the number of $y \in \Sigma_q^d$ with a given symbol composition $t_0, \dots, t_{q-1} \in \mathbb{N}_0$, is given by

$$|\{y \in \Sigma_q^d : \mathsf{ct}_a(y) = t_a \ \forall \ a \in \Sigma_q\}| = \binom{d}{t_0, \dots, t_{q-1}},$$

where t_i is the number of times the i-th symbol in Σ_q appears in y and $\binom{d}{t_0,\dots,t_{q-1}}$ is the multinomial coefficient. Combining all words y with a given composition in the computation of the output entropy, we obtain

$$H(Y) = -\frac{1}{q} \sum_{\substack{t_0,\dots,t_{q-1}: \\ t_0+\dots+t_{q-1}=d}} \binom{d}{t_0, \dots, t_{q-1}} \sum_{i=0}^{q-1} (1-p)^{t_i} \cdot \left(\frac{p}{q-1}\right)^{d-t_i} \log_q \left(\frac{1}{q} \sum_{j=0}^{q-1} (1-p)^{t_j} \left(\frac{p}{q-1}\right)^{d-t_j} \right),$$

where the sum over $t_0, \dots, t_{q-1}$ is over all possible compositions of a vector in Σ_q^d. Finally, we compute the conditional entropy to

$$\begin{aligned}H(Y|X) &= -\sum_{y \in \Sigma_q^d} \sum_{x \in \Sigma_d} p_d(y|x) \mathsf{Pr}\left(x\right) \log_q p_d(y|x) \\ &= -\frac{1}{q} \sum_{y \in \Sigma_q^d} \sum_{a \in \Sigma_q} (1-p)^{\mathsf{ct}_a(y)} \left(\frac{p}{q-1}\right)^{d-\mathsf{ct}_a(y)} \cdot \\ &\qquad \log_q \left((1-p)^{\mathsf{ct}_a(y)} \left(\frac{p}{q-1}\right)^{d-\mathsf{ct}_a(y)} \right),\end{aligned}$$

where we used that $\{\{p_d(y|x) : x \in \Sigma_q\}\}$ is independent of x, as shown before and could thus replace the sum over x by a sum over the symbol count spectrum of y. Combining those y with the same composition $t_0, \dots, t_{q-1}$, we arrive at

$$H(Y|X) = -\frac{1}{q} \sum_{\substack{t_0,\dots,t_{q-1}: \\ t_0+\dots+t_{q-1}=d}} \binom{d}{t_0, \dots, t_{q-1}} \sum_{i=0}^{q-1} (1-p)^{t_i} \cdot \left(\frac{p}{q-1}\right)^{d-t_i} \log_q \left((1-p)^{t_i} \left(\frac{p}{q-1}\right)^{d-t_i} \right).$$

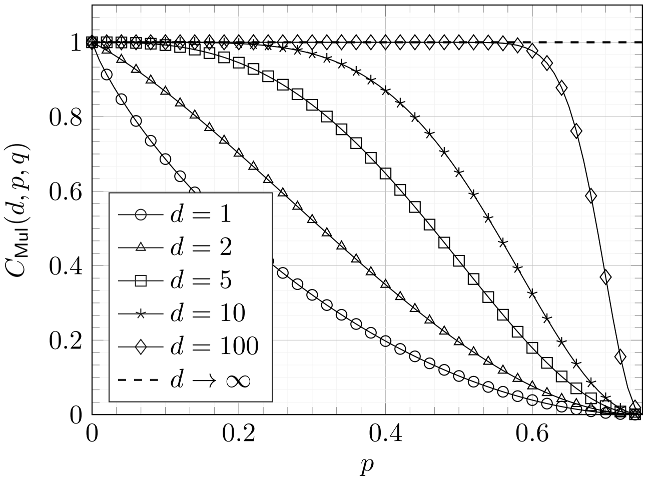

Fig. 5: Capacity of the multinomial channel for $q = 4$ and different number of draws d over the channel error probability p.

Notice that the conditional entropy $H(Y|X) = dH_q(p)$, where $H_q(p)$ is the q-ary entropy function, however here we prefer to express the entropy in the above form as this way it can compactly be combined with $H(Y)$. The lemma then follows from the fact that $I(X;Y) = H(Y) - H(Y|X)$ with uniformly distributed inputs X. □

Figure 5 shows the capacity of the multinomial channel for the DNA alphabet $\Sigma_4 = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ for different numbers of draws over the channel error probability.

Lemma 23. *Let the parameters d, p, q of the multinomial channel be arbitrary and fixed. Further, let $\boldsymbol{E}_i = (E_{i,1}, \dots, E_{i,L}) \in \Sigma_q^L$, $1 \le i \le d$ be random error vectors with independently and identically distributed entries*

$$\mathsf{Pr}\left(E_{i,\ell} = e_{i,\ell}\right) = \begin{cases} 1-p, & \text{if } e_{i,\ell} = 0 \\ \frac{p}{q-1}, & \text{if } e_{i,\ell} \neq 0 \end{cases},$$

for all $1 \le i \le d$ and $1 \le \ell \le L$. For an arbitrary $\epsilon > 0$, let $\mathcal{F}$ be the event that the sequence $\boldsymbol{E}' \triangleq (\boldsymbol{E}_2 - \boldsymbol{E}_1, \dots, \boldsymbol{E}_d - \boldsymbol{E}_1) \in \Sigma_q^{(d-1)\times L}$ is an ϵ-typical sequence[7]. Then, the number of ϵ-typical sequences $\mathbf{e}'$ is at most $q^{L(C_{\mathsf{Mul}}(d,p,q) + dH_q(p) - 1 + \epsilon)}$ and there exists an integer $L_d(\epsilon)$, such that for all $L \ge L_d(\epsilon)$, it holds that $\mathsf{Pr}\left(\mathcal{F}\right) \ge 1 - \epsilon$.

Proof. To start with, define the single letter variable $E' = (E_2 - E_1, \dots, E_d - E_1)$, where $E_i \in \Sigma_q$, $1 \le i \le d$ are independently and identically distributed variables with

$$\mathsf{Pr}\left(E_i = e_i\right) = \begin{cases} 1-p, & \text{if } e_i = 0 \\ \frac{p}{q-1}, & \text{if } e_i \neq 0 \end{cases}.$$

Notice that with this definition $\boldsymbol{E}'$ is a sequence of vectors over Σ_q^{d-1}, where each vector is independently and identically

[7] By *typical*, we refer to typical in the Shannon-sense, i.e., their log-probability is close to the negative entropy. For more details, see [28] and [29, Chapter 3].

distributed according to $\mathsf{Pr}\left(E' = e'\right)$. We can thus define set of ϵ-typical sequences $\mathcal{T}_{\mathsf{QSC}}^{L,\epsilon}(d,p,q)$ as

$$\mathcal{T}_{\mathsf{QSC}}^{(L,\epsilon)}(d,p,q)$$
$$\triangleq \left\{ \boldsymbol{e}' \in \Sigma_q^{(d-1)\times L} : \left| -\frac{\log_q \mathsf{Pr}\left(\boldsymbol{e}'\right)}{L} - H(E') \right| < \epsilon) \right\}.$$

By the standard asymptotic equipartition property [29, Thm 3.1.2], it follows that $\mathsf{Pr}\left(\mathcal{F}\right) \geq 1 - \epsilon$ for all $L \geq L_d(\epsilon)$, where $L_d(\epsilon)$ is a constant that depends only on d, p, q and ϵ. Further, the asymptotic equipartition property implies $|\mathcal{T}_{\mathsf{QSC}}^{(L,\epsilon)}(d,p,q)| \leq |2^{L(H(e')+\epsilon)}|$ for any L.

It remains to compute the entropy $H(e')$. We will do so by relating the entropy $H(E')$ with the output entropy of the multinomial channel. To this end, let $X \in \Sigma_q$ be a random variable, which is uniform and is independent of $E_1, \ldots, E_d$, which will be used as the input of the multinomial channel. Denote further by $Y = (X + E_1, \ldots, X + E_d)$ the corresponding output. Then, as the uniform input distribution maximizes the mutual information between X and Y, on the one hand, we know from Lemma 22 that

$$H(Y) = I(X;Y) + H(Y|X) = C_{\mathsf{Mul}}(d,p,q) + dH_q(p).$$

On the other hand, we show that E' is independent of $E_1 + X$ using the following sequence of equalities,

$$\begin{aligned}
\mathsf{Pr}\left(E' = e' | E_1 + X = y_1\right) &= \frac{\mathsf{Pr}\left(E' = e', E_1 + X = y_1\right)}{\mathsf{Pr}\left(E_1 + X = y_1\right)} \\
&\overset{(a)}{=} q\mathsf{Pr}\left(E' = e', E_1 + X = y_1\right) \\
&\overset{(b)}{=} \sum_{e_1 \in \Sigma_q} q\mathsf{Pr}\left(E' = e', E_1 = e_1, X = y_1 - e_1\right) \\
&\overset{(c)}{=} \sum_{e_1 \in \Sigma_q} \mathsf{Pr}\left(E' = e', E_1 = e_1\right),
\end{aligned}$$

where we used in equality (a) that $\mathsf{Pr}\left(E_1 + X = y_1\right) = 1/q$ for all y_1, as the sum of a uniform distribution x with any independent variable E_1 is again uniformly distributed over Σ_q. In equality (b), we demarginalized with respect to x and in equality (c), we used the independence of x from $E_1, \ldots, E_d$ and $\mathsf{Pr}\left(X = y_1 - e_1\right) = \frac{1}{q}$ due to the uniform distribution of x. This proves that E' is independent of $E_1 + X$. It follows that

$$\begin{aligned}
H(E') = H(E'|E_1 + X) &\overset{(d)}{=} H(E_2 + X, \ldots, E_d + X | E_1 + X) \\
&= H(Y) - H(Y_1) = C_{\mathsf{Mul}}(d,p,q) + dH_q(p) - 1,
\end{aligned}$$

where we used [29, Problem 2.14] on the conditional entropy of a sum in inequality (d). □

References

[1] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *Proc. Inf. Theory Workshop*, Visby, Sweden, Aug. 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8989388/

[2] ——, "Achieving the capacity of the DNA storage channel," in *Proc. Int. Conf. Acoust., Speech, Sig. Process.*, Barcelona, Spain, May 2020, pp. 8846–8850.

[3] ——, "On the capacity of DNA-based data storage under substitution errors," in *Proc. Visual Commun. Image Process.*, Munich, Germany, Dec. 2021.

[4] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, Sep. 2012. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1226355

[5] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 2013. [Online]. Available: http://www.nature.com/articles/nature11875

[6] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific Reports*, vol. 5, no. 1, p. 14138, Nov. 2015. [Online]. Available: http://www.nature.com/articles/srep14138

[7] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific Reports*, vol. 7, no. 1, p. 5011, Dec. 2017. [Online]. Available: http://www.nature.com/articles/s41598-017-05188-1

[8] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, and K. Strauss, "Random access in large-scale DNA data storage," *Nature Biotechnology*, vol. 36, no. 3, pp. 242–248, Mar. 2018. [Online]. Available: http://www.nature.com/articles/nbt.4079

[9] S. Chandak, J. Neu, K. Tatwawadi, J. Mardia, B. Lau, M. Kubit, R. Hulett, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Overcoming high nanopore basecaller error rates for DNA storage via basecaller-decoder integration and convolutional codes," in *Proc. Int. Conf. Acoust., Speech, Sig. Process.*, Barcelona, Spain, May 2020, pp. 8822–8826.

[10] R. G. Gallager, "Sequential decoding for binary channel with noise and synchronization errors," Lincoln Lab Group, Arlington, VA, USA, Tech. Rep., Sep. 1961.

[11] M. C. Davey and D. J. C. Mackay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 687–698, Feb. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/910582/

[12] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2428–2445, Apr. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7809143/

[13] M. A. Sini and E. Yaakobi, "Reconstruction of sequences in DNA storage," in *Proc. Int. Symp. Inf. Theory*, Paris, France, Jul. 2019, pp. 290–294. [Online]. Available: https://ieeexplore.ieee.org/document/8849740/

[14] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6084–6103, Oct. 2020.

[15] A. Lenz, I. Maarouf, L. Welter, A. Wachter-Zeh, E. Rosnes, and A. Graell i Amat, "Concatenated Codes for Recovery From Multiple Reads of DNA Sequences," in *Proc. Inf. Theory Workshop*, Riva del Garda, Italy, Apr. 2021, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9457675/

[16] S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister, and S. Yekhanin, "Trellis BMA: Coded trace reconstruction on IDS channels for DNA storage," in *Proc. Int. Symp. Inf. Theory*, Melbourne, Australia, Jul. 2021.

[17] F. M. J. Willems and A. Gorokhov, "Signaling over arbitrarily permuted parallel channels," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1374–1382, Mar. 2008. [Online]. Available: http://ieeexplore.ieee.org/document/4455764/

[18] E. Hof, I. Sason, S. Shamai, and C. Tian, "Capacity-achieving polar codes for arbitrarily permuted parallel channels," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1505–1516, Mar. 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6399602/

[19] M. Langberg, M. Schwartz, and E. Yaakobi, "Coding for the ℓ_∞-limited permutation channel," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7676–7686, Dec. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8067503/

[20] A. Makur, "Information capacity of BSC and BEC permutation channels," in *Proc. Annu. Allerton Conf. Commun. Control Comp.*, Monticello, IL, USA, Oct. 2018, pp. 1112–1119. [Online]. Available: https://ieeexplore.ieee.org/document/8636070/

[21] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proc. Int. Symp. Inf. Theory*, Aachen, Jun. 2017, pp. 3130–3134. [Online]. Available: http://ieeexplore.ieee.org/document/8007106/

[22] D. J. C. Mackay, J. Sayir, and N. Goldman, "Near-capacity codes for fountain channels with insertions, deletions, and substitutions, with applications to DNA archives," Unpublished Manuscript, Jun. 2015.

[23] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," in *Proc. Int. Symp. Inf. Theory*, Paris, France, Jul. 2019, pp. 762–766.

[24] S. Shin, R. Heckel, and I. Shomorony, "Capacity of the erasure shuffling channel," in *Proc. Int. Conf. Acoust., Speech, Sig. Process.*, May 2020, pp. 8841–8845.

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

[26] N. Weinberger and N. Merhav, "The DNA Storage Channel: Capacity and Error Probability Bounds," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5657–5700, Sep. 2022.

[27] M. Mitzenmacher, "On the theory and practice of data recovery with multiple versions," in *Proc. Int. Symp. Inf. Theory*, Seattle, WA, Jul. 2006, pp. 982–986. [Online]. Available: http://ieeexplore.ieee.org/document/4036111/

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

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[30] R. M. Fano and D. Hawkins, "Transmission of Information: A Statistical Theory of Communications," *American Journal of Physics*, vol. 29, no. 11, pp. 793–794, Nov. 1961. [Online]. Available: http://aapt.scitation.org/doi/10.1119/1.1937609

[31] M. Mitzenmacher, "The power of two choices in randomized load balancing," Ph.D. dissertation, University of California, Berkeley, 1996.

[32] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Feb. 2000. [Online]. Available: http://doi.wiley.com/10.1002/9781118032718

[33] A. Lenz, L. Welter, and S. Puchinger, "Achievable Rates of Concatenated Codes in DNA Storage under Substitution Errors," in *Proc. Int. Symp. Inf. Theory Appl.*, Kapolei, HI, USA, Oct. 2020, pp. 269–273.

[34] R. B. Ash, *Information Theory*. New York: Dover Publications, 1990.

[35] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, 1st ed. Cambridge University Press, Sep. 2018. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781108231596/type/book

[36] R. Gallager, *Information Theory and Reliable Communication*. Vienna: Springer Vienna, 1972. [Online]. Available: http://link.springer.com/10.1007/978-3-7091-2945-6

Andreas Lenz (S'17) received the B.Sc. degree (Hons.), the M.Sc. degree (Hons.), in electrical engineering and information technology from Technische Universität München (TUM), Germany. He received the Ph.D degree (summa cum laude) from TUM in 2022 on the topic of coding for modern memories. His research interests include signal processing, coding theory and information theory.

Antonia Wachter-Zeh (S'10–M'14-SM'20) is an Associate Professor at the Technical University of Munich (TUM), Munich, Germany in the School of Computation, Information and Technology. She received the M.Sc. degree in communications technology in 2009 from Ulm University, Germany. She obtained her Ph.D. degree in 2013 from Ulm University and from Universite de Rennes 1, Rennes, France. From 2013 to 2016, she was a postdoctoral researcher at the Technion—Israel Institute of Technology, Haifa, Israel, and from 2016 to 2020 a Tenure Track Assistant Professor at TUM. She is a recipient of the DFG Heinz Maier-Leibnitz-Preis and of an ERC Starting Grant. She is currently an Associate Editor for the IEEE Transactions on Information Theory. Her research interests are coding theory, cryptography and information theory and their application to storage, communications, privacy, security and machine learning.

Paul H. Siegel (M'82–SM'90–F'97–LF'19) received the S.B. and Ph.D. degrees in mathematics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1975 and 1979, respectively. He held a Chaim Weizmann Postdoctoral Fellowship with the Courant Institute, New York University, New York, NY, USA. He was with the IBM Research Division, San Jose, CA, USA, from 1980 to 1995. He joined the faculty at the University of California San Diego, La Jolla, CA, USA, in 1995, where he is currently a Distinguished Professor of electrical and computer engineering with the Jacobs School of Engineering. He is affiliated with the Center for Memory and Recording Research where he holds an Endowed Chair and served as Director from 2000 to 2011. His research interests include information theory, coding techniques, and machine learning, with applications to digital data storage and transmission. He is a Member of the National Academy of Engineering. He was a Member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996 and from 2009 to 2014. He was the 2015 Padovani Lecturer of the IEEE Information Theory Society. He was a co-recipient of the 1992 IEEE Information Theory Society Paper Award, the 1993 IEEE Communications Society Leonard G. Abraham Prize Paper Award, and the 2007 Best Paper Award in Signal Processing and Coding for Data Storage from the Data Storage Technical Committee of the IEEE Communications Society. He served as an Associate Editor of Coding Techniques of the IEEE Transactions on Information Theory from 1992 to 1995, and as the Editor-in-Chief from 2001 to 2004. He served as a Co-Guest Editor of the 1991 Special Issue on Coding for Storage Devices of the IEEE Transactions on Information Theory. He was also a Co-Guest Editor of the 2001 two-part issue on The Turbo Principle: From Theory to Practice and the 2016 issue on Recent Advances in Capacity Approaching Codes of the IEEE Journal on Selected Areas in Communications.

Eitan Yaakobi (S'07–M'12–SM'17) is an Associate Professor at the Computer Science Department at the Technion — Israel Institute of Technology. He received the B.A. degrees in computer science and mathematics, and the M.Sc. degree in computer science from the Technion — Israel Institute of Technology, Haifa, Israel, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering from the University of California, San Diego, in 2011. Between 2011-2013, he was a postdoctoral researcher in the department of Electrical Engineering at the California Institute of Technology and at the Center for Memory and Recording Research at the University of California, San Diego. His research interests include information and coding theory with applications to non-volatile memories, associative memories, DNA storage, data storage and retrieval, and private information retrieval. He received the Marconi Society Young Scholar in 2009 and the Intel Ph.D. Fellowship in 2010-2011. Since 2020, he serves as an Associate Editor for Coding snd Decoding for the IEEE Transactions on Information Theory.