Achievable Rates of Concatenated Codes in DNA Storage under Substitution Errors

Andreas Lenz[*], **Lorenz Welter**[*], and **Sven Puchinger**[†],

[*]Technical University of Munich, andreas.lenz@mytum.de, lorenz.welter@tum.de

[†]Technical University of Denmark, svepu@dtu.dk

Abstract—In this paper, we study achievable rates of concatenated coding schemes over a deoxyribonucleic acid (DNA) storage channel. Our channel model incorporates the main features of DNA-based data storage. First, information is stored on many, short DNA strands. Second, the strands are stored in an unordered fashion inside the storage medium and each strand is replicated many times. Third, the data is accessed in an uncontrollable manner, i.e., random strands are drawn from the medium and received, possibly with errors. As one of our results, we show that there is a significant gap between the channel capacity and the achievable rate of a standard concatenated code in which one strand corresponds to an inner block. This is in fact surprising as for other channels, such as q-ary symmetric channels, concatenated codes are known to achieve the capacity. We further propose a modified concatenated coding scheme by combining several strands into one inner block, which allows to narrow the gap and achieve rates that are close to the capacity.

I. INTRODUCTION

DNA has evolved to a competitive medium for long-term archival storage. Recent experiments [1]–[10] have shown that it is possible to store large amounts of data using these macromolecules. These studies have addressed different important aspects of long-term data storage, such as portability [8], random-access [4], [9], durability [3], reliability and scalability [9]. A typical DNA storage system hereby consists of the three following entities. First, a synthesis machine, which artificially creates DNA strands. Such a device allows to produce strands by chaining nucleotides (adenine [A], cytosine [C], guanine [G] or thymine [T]) to arbitrary vectors. Note that due to practical limitations, it is only possible to synthesize DNA strands of length in the order of 10^2 and 10^3. Second, once the strands have been created, they are stored in an adequate environment and replicated many times. Third, in order to retrieve the stored archive, the data is accessed via a sequencing machine that allows to read the stored strands. The sequencer hereby reads the strand in an uncontrollable manner, i.e., it is not directly possible to select which strand to read.

Such a storage system inherently introduces a novel channel model due to its following properties. Due to the limited strand-length induced by the synthesis machine, the data needs to be split into many strands and then encoded. Depending on the amount of data stored, the number of strands is significantly larger than the length of the strands, and typical values range up to 10^7 for large-scale archives [11]. Further, the strands are stored and received in an unordered fashion, as the sequencing machine cannot control which strand should be sequenced[1]. Finally, the synthesis and sequencing, or in other words, reading and writing process of DNA molecules is prone to errors, and therefore it is necessary to develop forward-error-correcting schemes that allow to cope with the inherent disorder of the read strands and possible distortions. In view of these unique properties of DNA storage systems,

it is immediate that a profound understanding of the channel and associated error-correcting codes is necessary to design cost-efficient DNA-storage systems. This becomes even more apparent as one of the most prominent bottlenecks of todays DNA storage systems is the cost associated with the synthesis and sequencing [12].

The channel induced by DNA-based archival storage has been discussed from different perspectives. Constrained codes restricting long runs of homopolymers or forcing a balanced GC content have, among others, been studied in [13], [14]. Most related to our study is the work discussing channels with unordered input and output strands. Such channels have been discussed from a worst-case point of view in [11], [15], [16]. On the other hand, probabilistic channels have been discussed in [17], [18] for the case where there are no errors inside the strands [19] and for the case, where each strand is drawn exactly once [18]. A lower bound on the capacity of the channel under discussion was derived in [20] and its achievability has been proven using a typicality-like decoder together with a random coding argument in [21]. However, until now, it is unclear how to construct efficiently encodable and decodable capacity-achieving codes for this channel.

In this work, we approach the question of how to design efficiently encodable and decodable codes for the DNA-storage channel model discussed in [17], [20], [21]. We focus on concatenated codes in this paper as they inherently match the channel model and have been used in several experiments [3], [7], [10]. Furthermore, concatenated coding schemes achieve the capacity of other channels, in particular the q-ary symmetric channel [22]. We show, however, that due to the nature of the discussed DNA storage channel, standard concatenated codes with inner block corresponding to one DNA strand are not able to achieve the capacity with a standard hard-decision outer decoder, independent of the selected inner and outer codes. The main result of the paper is that concatenated coding schemes *can* achieve rates close to the capacity if several DNA strands constitute an inner block. This extends the work in [21] to efficiently encodable and decodable coding schemes.

II. PRELIMINARIES

For any integer $n \in \mathbb{N}$, we write $[n] = \{1, 2, \ldots, n\}$ as the set of integers up to n. Throughout the paper, vectors are highlighted by bold letters, such as $\boldsymbol{x} = (x_1, \ldots, x_n)$. Concerning asymptotic statements, whenever we refer to a statement to be true *for large enough* n, we mean that there exists some n_0 such that the statement is fulfilled for all $n \geq n_0$. For a summary of parameters and relevant random variables, we refer the reader to Tables I and II. For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_2^n$, we denote $d(\boldsymbol{x}, \boldsymbol{y})$ as the Hamming distance of $\boldsymbol{x}$ and $\boldsymbol{y}$.

[1]It has been shown in [4] that it is possible to select specific strands for sequencing by appending carefully designed primers to the DNA molecules. Here we restrict our attention to strands with the same or similar primers that do not allow for random selection.

A. Channel Model

We start by introducing the channel model as in [17], [20], [21]. The input of the channel are M strands $\boldsymbol{x}_i \in \mathbb{Z}_2^L$, $i \in [M]$, each of length L. Note that for simplicity we assume binary strands here, however our results can be directly extended to q-ary symmetric channels. From these strands, N strands are drawn uniformly at randomly with replacement and perturbed by errors, resulting in output strands $\boldsymbol{y}_j$, $j \in [N]$ with

$$\boldsymbol{y}_j = \boldsymbol{x}_{I_j} + \boldsymbol{e}_j,$$

where $\boldsymbol{e}_j \sim \mathsf{Ber}(p)^L$ are random error vectors and $I_j \in [M]$ are i.i.d. uniform random draws. For our asymptotic statements, we assume $M = 2^{\beta L}$ and $N = cM$ for some fixed $0 < \beta < 1$ and $c \in \mathbb{R}$ and then let $M \to \infty$. This choice is due to the fact that this is practically and theoretically the most interesting and relevant region, c.f. [17].

Remark 1. *In this model we do not incorporate deletion or insertion errors for two reasons. First, the capacity of channels with insertions or deletions is unknown and therefore also the capacity of the overall channel is unknown. Second, our main focus in this work is to analyze the effect of the disorder of the output strands. However, we believe qualitatively similar results hold also for channels with insertions and deletions.*

B. Channel Capacity and Related Quantities

The capacity of the channel is given by [20], [21]

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

where $p_c(d) = e^{-c}c^d/d!$ is the probability mass function (pmf) of the Poisson distribution and C_d is the capacity of the d-multi-draw channel

$$C_d = 1 + \sum_{i=0}^{d} B_{d,p}(i) \log_2 \left(\frac{B_{d,p}(i)}{B_{d,p}(i) + B_{d,p}(d-i)} \right), \quad (1)$$

c.f. [23], where $B_{d,p}(i) = \binom{d}{i} p^i (1-p)^{d-i}$ is the pmf of the binomial distribution. Hereby, the d-multi-draw channel is a discrete memoryless channel $\mathbb{Z}_2 \to \mathbb{Z}_2^d$, where a single input bit is received over d independent binary symmetric channels with crossover probability p, see [20], [23].

C. Concatenated Coding Scheme

We present our concatenated coding scheme that is discussed in this paper. For a visualization, see Fig. 1. Assume, we want to encode a data archive of B bits. First, the data is split into k_{out} strands, each of length B/k_{out}. The k_{out} strands are then encoded via a maximum distance separable (MDS) outer code $[M, k_{\mathrm{out}}]$ over $\mathbb{F}_{2^{B/k_{\mathrm{out}}}}$ of length M, dimension

k_{out} and rate $R_{\mathrm{out}} = {}^{k_{\mathrm{out}}}\!/M$. Now, every K consecutive strands are combined to one block each and encoded via an $[KB/(k_{\mathrm{out}}R_{\mathrm{in}}), KB/k_{\mathrm{out}}]$ inner code to be described in detail in Lemma 3. Abbreviating $\ell = B/(k_{\mathrm{out}}R_{\mathrm{in}})$, each inner codeblock is again split into K strands, each of length ℓ. Then, each strand is appended with an index of $\log M = \beta L$ bits and individually protected via a multi-draw capacity-achieving index code of rate R_{ix}. Since the total strand length is L, we have $\ell = L(1 - {}^{\beta}\!/R_{\mathrm{ix}})$ and the overall rate is $R = R_{\mathrm{out}}R_{\mathrm{in}}(1 - {}^{\beta}\!/R_{\mathrm{ix}})$. Note that for $K = 1$ we obtain standard concatenated codes, where each strand is an inner codeword.

D. Decoding Procedure

The decoder acts in four stages: 1) clustering 2) index decoding 3) inner decoding 4) outer decoding. First, the output strands $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are grouped to maximal clusters $\mathcal{Y}_1, \ldots, \mathcal{Y}_{\widehat{M}}$, such that for all $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}_i$, $d(\boldsymbol{y}, \boldsymbol{y}') \leq \rho L$, where ρ is the clustering diameter. This can, for example, be achieved by a greedy algorithm that selects an arbitrary strand $\mathcal{Y}_1 = \{\boldsymbol{y}_j\}$ and then successively adds strands to $\mathcal{Y}_1$ which have distance at most ρL from all sequences in $\mathcal{Y}_1$. This procedure goes on until there are no more output sequences that can be put into the cluster. The algorithm then continues by creating a new cluster $\mathcal{Y}_2$ from the remaining strands and in the same fashion. This procedure continues until each strand $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is assigned exactly one cluster. We will choose $\rho \approx 2p$, as the expected number of errors from the channel are pL and thus two strands that stem from the same origin have an expected distance of at most $2pL$.

Each cluster is now decoded with the index code if $C_{|\mathcal{Y}_i|} > R_{\mathrm{ix}}$ and discarded otherwise. Then, each cluster is assigned to its position i of the decoded index. If there are two or more clusters which have the same index, then discard both. In the following, let $\widehat{d}_i$ be the size of the cluster with decoded index i. Further abbreviate $\widehat{\boldsymbol{d}}^{(b)} = (\widehat{d}_{bK+1}, \ldots, \widehat{d}_{bK+K}) \in \mathbb{N}_0^K$ as the draws of one block $b \in [\frac{M}{K}]$. Note that $\widehat{d}_i \neq d_i$ is possible due to clustering- or index-decoding errors. Now, given the decoded indices, the K strands of an inner block b are results of transmission over K parallel multi-draw channels with draws $\widehat{\boldsymbol{d}}^{(b)}$ and the capacity of these channels is given by $C_{\widehat{\boldsymbol{d}}^{(b)}} = {}^1\!/K \sum_{j=1}^{K} C_{\widehat{d}_{bK+j}}$, as we will see in Lemma 3.[2] The inner code therefore decodes an inner block, if $C_{\widehat{\boldsymbol{d}}^{(b)}} > R_{\mathrm{in}}$

---

[2]Note that the exact drawing realization $\widehat{\boldsymbol{d}}^{(b)}$ is random and thus unknown to the encoder for a specific inner block. This implies that by choosing a fixed inner rate, the rate is too large for weak channels and too small for stronger channels with many draws. As we will show later in Lemma 4, however, for large K, the drawing distribution (and thus also the sum capacity) inside the blocks converges for a large fraction of blocks, which allows to choose the inner rate accordingly.
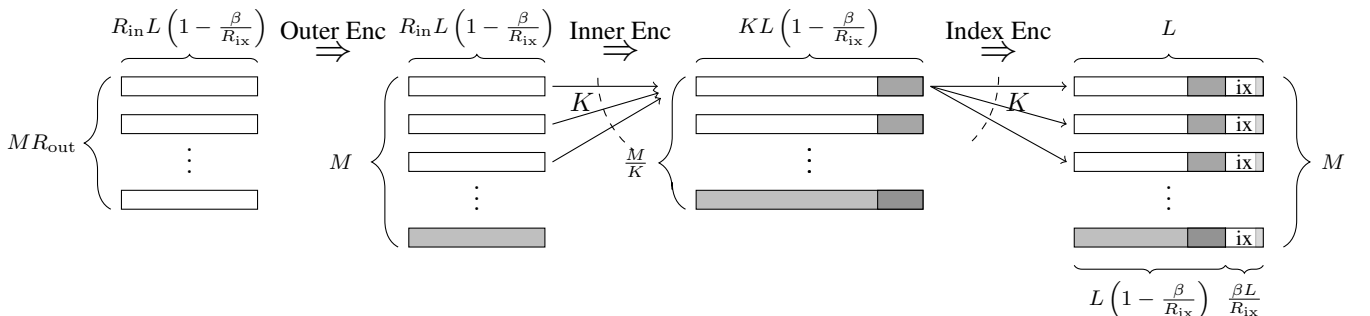

Figure 1. Illustration of the Encoder. k_{out} strands are encoded via a MDS outer code with rate R_{out} by representing them each as symbols over a large alphabet. Then a block of K strands are encoded via an inner encoder by a code with rate R_{in} described in Lemma 3. Each inner codeblock is then split again into K strands. Each strand gets an index appended which is protected via an index code with rate R_{ix}.

Table I
SUMMARY OF RELEVANT PARAMETERS

Param.	Description	Relations
c	Reading rate/Reading rate	$c = N/M$
β	Strand density	$\beta = \log M / L$
p	Error probability of the channel	
R_{out}	Rate of the outer code	$R_{\text{out}} = k_{\text{out}}/M$
K	Block size of the inner code	
R_{in}	Rate of the inner code	$R = R_{\text{out}} R_{\text{in}}(1 - \beta/R_{\text{ix}})$
R_{ix}	Rate of the index code	

Table II
SUMMARY OF IMPORTANT RANDOM VARIABLES

RV	Description	Relations
I_j	Index of the j-th drawn strand	$\boldsymbol{y}_j = \boldsymbol{x}_{I_j} + \boldsymbol{e}_j$
d_i	Num. draws of i-th input strand	$d_i = \lvert\{j : I_j = i\}\rvert$
$\boldsymbol{d}^{(i)}$	Num. draws of strands in block i	$\boldsymbol{d}^{(i)} = (d_{iK+1}, \ldots, d_{iK+K})$
$\mathcal{Y}_i$	Estimated cluster of strand i	$i \in [\widehat{M}]$
$\hat{d}_i$	Estimated num. draws/cluster sizes	$\hat{d}_i = \lvert\mathcal{Y}_i\rvert$
$Q_{\boldsymbol{d}}$	Num. blocks that are drawn $\boldsymbol{d}$ times	$Q_{\boldsymbol{d}} = \{i : \boldsymbol{d}^{(i)} = \boldsymbol{d}\}$
M_{C}	Number of wrong clusters	
M_{Ix}	Number of wrongly decoded indices	
M_{In}	Num. wrongly decoded inner codewords	

and declares an erasure otherwise as it is only possible to decode reliably if the sum capacity is smaller than the inner rate, see Lemma 3. Finally, the decoded inner codewords are fed into a unique decoding algorithm of the outer MDS code.

We associate the following random variables with the decoding process that will be helpful for analyzing the decoding error probability. First, M_{C} counts the number of wrong clusters $M_{\text{C}} = \lvert\{i : \nexists i' \in [M] \text{ s.t. } \mathcal{Y}_i = \{j : I_j = i'\}\}\rvert$, i.e., the number of clusters that do not consist of exactly all output strands that stem from one input strand. Next, M_{Ix} counts the number of clusters with $C_{d_i} > R_{\text{ix}}$ that are decoded wrongly, assuming that all strands were clustered correctly. Similarly, M_{In} counts the number of inner codewords that are decoded wrongly, assuming correct clustering and index decoding. We further write $Q_{\boldsymbol{d}} = \{i : \boldsymbol{d}^{(i)} = \boldsymbol{d}\}$ and $\boldsymbol{Q} = (Q_{\boldsymbol{d}})_{\boldsymbol{d} \in \mathbb{N}_0^K}$. For the readers convenience, we summarize all relevant parameters and random variables in Tables I and II.

III. ACHIEVABLE RATES USING HARD-DECISION DECODED CONCATENATED CODES

In this section, we derive the main result of this paper: achievable rates for concatenated codes in the DNA storage channel with groups of K strands as inner blocks. In Section III-A, we first formulate the main statement (Theorem 1) and analyze the resulting achievable rate for $K \to \infty$. The remaining sections contain the proof of Theorem 1: first we show three lemmas in Section III-B and then conclude the proof in Section III-C.

A. Main Statement and Discussion

Theorem 1. *Consider a DNA storage channel with parameters p, c, and β as in Section II-A. Let $K \in \mathbb{N}$ and $R_{\text{ix}}, R_{\text{in}} \in (0,1)$ with $\beta < R_{\text{in}}(1 - \beta/R_{\text{ix}})(1 - h(2p))$ be parameters of a concatenated coding scheme as in Sections II-C and II-D. Then there is such a scheme that achieves rate $R = R_{\text{in}} R_{\text{out}}(1 - \beta/R_{\text{ix}})$, for any outer rate R_{out} that satisfies*

$$R_{\text{out}} < \sum_{\boldsymbol{d} \in \mathbb{N}_0^K : C_{\boldsymbol{d}}(R_{\text{ix}}) > R_{\text{in}}} p_c(\boldsymbol{d}), \quad (2)$$

where $p_c(\boldsymbol{d}) = \prod_{i=1}^K p_c(d_i)$, $p_c(d_i) = \frac{c^{d_i}}{d_i!} e^{-c}$ is the pmf of the K-variate Poisson distribution and (using the multi-draw channel capacity C_d, cf. (1))

$$C_{\boldsymbol{d}}(R_{\text{ix}}) = \tfrac{1}{K} \sum_{i=1}^K C_{d_i}(R_{\text{ix}}), \quad C_d(R_{\text{ix}}) = \begin{cases} C_d, & C_d > R_{\text{ix}}, \\ 0, & \text{else.} \end{cases}$$

We prove Theorem 1 in the next subsections, but first we discuss and interpret the result.

Remark 2. *Using Theorem 1, a lower bound on the maximal achievable rate of a concatenated coding scheme with K blocks can be computed by finding the supremum of the right-hand side of (2) over R_{ix} and R_{in}. Since $C_d(R_{\text{ix}})$ attains only discrete values, it is only necessary to consider $R_{\text{ix}} = (1 - \epsilon)C_d$ for $d = 1, 2, 3, \ldots$ and sufficiently small ϵ.*

It is not apparent from Theorem 1 how to compute the sum on the right-hand side of (2) efficiently. We will return to this issue in Section IV, where we show that it can be efficiently estimated using a simple Monte-Carlo simulation.

Remark 3. *It is in fact not possible to achieve higher rates than those given in Theorem 1 with our hard-decision decoded coding scheme. This is because choosing an outer rate violating (2), there will, with high probability, be more erasures than the redundancy of the outer code, which leads to a decoding error.*

For $K \to \infty$, the achievable rate in Theorem 1 converges to the following value.

Corollary 2.A. *Consider the setting of Theorem 1. For large enough K, any rate R is achievable if it fulfills*

$$R < \left(\sum_{d=0}^{\infty} p_c(d) C_d(R_{\text{ix}}) \right) (1 - \beta/R_{\text{ix}}). \quad (3)$$

Proof. We will prove the lemma by slightly rewriting the achievability result from Theorem 1 and the law of large numbers. To start with, we rewrite the bound on R_{out} of Theorem 1 in a probabilistic fashion. Let $\boldsymbol{d} = (d_1, \ldots, d_K)$ be K independent Poisson distributed variables with parameter c. Then, the bound from Theorem 1 becomes

$$R_{\text{out}} < \Pr\left(\tfrac{1}{K} \sum_{i=1}^K C_{d_i}(R_{\text{ix}}) > R_{\text{in}} \right). \quad (4)$$

Note the here the probability is over the artificial random variables $d_1, \ldots, d_K$. Now, by the law of large numbers, for any $\epsilon > 0$ and large enough K,

$$\Pr\left(\left\lvert \tfrac{1}{K} \sum_{i=1}^K C_{d_i}(R_{\text{ix}}) - \mathbb{E}(C_{d_i}(R_{\text{ix}})) \right\rvert > \epsilon \right) < \epsilon.$$

Setting $R_{\text{in}} = \mathbb{E}(C_{d_i}(R_{\text{ix}})) - \epsilon$ this means that any $R_{\text{out}} < 1 - \epsilon$ is achievable for large enough k and we obtain that any

$$R < \mathbb{E}(C_{d_i}(R_{\text{ix}}))(1 - \tfrac{\beta}{R_{\text{ix}}}).$$

is achievable, by choosing ϵ as small as desired. Identifying $\mathbb{E}(C_{d_i}(R_{\text{ix}})) = \sum_{d=0}^{\infty} p_c(d) C_d(R_{\text{ix}})$ proves the lemma. $\square$

For $K \to \infty$, we obtain the achievable rate of a concatenated code by finding the supremum of the right-hand side

of (3) over R_{ix}. Since the $C_d(R_{\mathrm{ix}})$ change only for discrete values of R_{ix}, we have the following:

$$
\begin{aligned}
R_{\max} &= \sup_{R_{\mathrm{ix}} \in (0,1)} \left(\textstyle\sum_{d=0}^{\infty} p_c(d) C_d(R_{\mathrm{ix}}) \right) \left(1 - \beta/R_{\mathrm{ix}}\right) \\
&= \max_{d^* \in \mathbb{Z}_{>0}} \sup_{R_{\mathrm{ix}} \in (C_{d^*-1}, C_{d^*})} \left(\textstyle\sum_{d=d^*}^{\infty} p_c(d) C_d \right) \left(1 - \beta/R_{\mathrm{ix}}\right) \\
&= \max_{d^* \in \mathbb{Z}_{>0}} \left(\textstyle\sum_{d=d^*}^{\infty} p_c(d) C_d \right) \left(1 - \beta/C_{d^*}\right),
\end{aligned}
$$

where the maximization over d^* is indeed a maximization since $1 - \beta/C_{d^*} \to 1$ and the sum converges to 0 for $d^* \to \infty$, hence the supremum is attained a finite d^*. We compare $R_{\max}$ to the capacity of the DNA storage channel [20], [21].

Corollary 2.B. *Let* $C = \sum_{d=1}^{\infty} p_c(d) C_d - \beta(1 - \mathrm{e}^{-c})$ *be the capacity of the DNA storage channel [20], [21]. With notation as above, we have*

$$
R_{\max} - C \to 0
$$

for $p \to 0$ *(and fixed* c*) or* $c \to \infty$ *(and fixed* p*).*

Proof. First consider $p \to 0$: We have

$$
R_{\max} \geq \textstyle\sum_{d=1}^{\infty} p_c(d) C_d - \beta \frac{\sum_{d=1}^{\infty} p_c(d) C_d}{C_1}
$$

Due to C_d converges to 1 for $p \to 0$, we have $C_1 \to 1$ and $\sum_{d=1}^{\infty} p_c(d) C_d \to 1 - p_c(0) = 1 - e^{-c}$.

For $c \to \infty$, for any sequence of $c \to \infty$, there is a sequence of $d^* \in \mathbb{N}$ such that $d^* \to \infty$ (i.e., $C_{d^*} \to 1$) and $\sum_{d=d^*}^{\infty} p_c(d) C_d \to \sum_{d=1}^{\infty} p_c(d) C_d$. This proves the claim. $\qquad\square$

The numerical analysis in Section IV confirms a part of Corollary as for $c = 10$, the maximal achievable rate for $K \to \infty$ is quite close to the actual capacity of the channel.

B. Lemmas

In the following, we present three lemmas that we will need for the proof of the main statement in the next subsection. Recall the notions introduced in Section II-D, in particular the random variables M_C, M_{Ix}, M_{In}, $Q_{\boldsymbol{d}}$ (see also Table II).

In the concatenated scheme for the DNA storage channel discussed in Section II, there is a fundamental difference to concatenated coding schemes designed for other channels (e.g., the q-ary symmetric channel): inner blocks are transmitted through *different* channels. In this case, an inner block corresponds to a block of K strands and the channel it is transmitted through is a K-parallel multi-draw channel, where the draw vector $\boldsymbol{d}$ varies from block to block. A major difficulty compared to channels with the same "inner channel" is thus to find an inner code that achieves the capacity of a large number of such channels–jointly. We address this question in the following lemma.

Lemma 3. *Let* K*,* $d_{\max}^+$ *be integers and choose a rate* R*. For any* $\epsilon > 0$*, there is a code of rate* R *and sufficiently large length for the family of* K*-parallel multi-draw channels such that the decoding error probability is at most* ϵ *for any realization of the channel with draw vector* $\boldsymbol{d}$ *s. t.* $\sum_{i=1}^{K} d_i \leq d_{\max}^+$ *and*

$$
R < \frac{1}{K} \sum_{i=1}^{K} C_{d_i}.
$$

Proof. We prove this Lemma using a random coding argument and a typical-sequence decoder [24, Chapter 7]. Recall that the K-parallel multi-draw channel with draw vector $\boldsymbol{d} = (d_1, \ldots, d_J)$ is K parallel multidraw channels,

each with d_i draws. It therefore is a memoryless channel with input alphabet $\mathbb{Z}_2^K$ and output alphabet $\mathbb{Z}_2^{d_1} \times \cdots \times \mathbb{Z}_2^{d_K}$. We denote each input symbol by $(X_1, \ldots, X_K)$ and the corresponding output symbol by $\mathcal{Y}_1, \ldots, \mathcal{Y}_K$. For this channel we specify the symbol-wise input distribution as $\Pr\left((X_1, \ldots, X_K) = (x_1, \ldots, x_K)\right) = \prod_{i=1}^{K} p(x_i)$, where $p(x_i)$ is the capacity achieving distribution of the sub-channels, shown in [20, Lemma 3] to be the uniform distribution. Each of the multi-draw channels i, $1 \leq i \leq K$, has the capacity C_{d_i}. Since the transmissions over the K parallel channels are independent of each other, we can conclude for any given, fixed vector $\boldsymbol{d} = (d_1, \ldots, d_K)$

$$
\begin{aligned}
C_{\boldsymbol{d}} &= \frac{1}{K} \max_{p(x_1, \ldots, x_K)} I(X_1, \ldots, X_K; \mathcal{Y}_1, \ldots, \mathcal{Y}_K) \\
&= \frac{1}{K} \sum_{i=1}^{K} \max_{p(x_i)} I(X_i; \mathcal{Y}_i) = \frac{1}{K} \sum_{i=1}^{K} C_{d_i}.
\end{aligned}
$$

This means, that for a fixed and known $\boldsymbol{d}$, $C_{\boldsymbol{d}}$ is clearly achievable. However, here we aim to design a codebook that achieves vanishing error probability for all $\boldsymbol{d} \in \mathcal{D}_{\max}^+ := \{\boldsymbol{d} \in \mathbb{N}_0^K : \sum_{i=1}^{K} d_i \leq d_{\max}^+\}$ and $R < C_{\boldsymbol{d}}$. Note that $\mathcal{D}_{\max}^+$ has finite cardinality resulting in a total number of $|\mathcal{D}_{\max}^+|$ possible channels for vectors $\boldsymbol{d}$.

As a consequence we can use similar arguments as in [25] with a Cover-like random coding argument to show that the capacity expression is achievable. We design the codebook $\mathcal{X} := \{\mathbf{X}(w), w \in [2^{RK\ell}]\}$, where each codeword $\mathbf{X}(w)$ with message index w is a matrix of size $K \times \ell$. Note that the design of this codebook only depends on $\mathcal{D}_{\max}^+$ and not the specific channel realization as desired above. Now, given $\boldsymbol{d} \in \mathcal{D}_{\max}^+$, for each K-parallel multi-draw channel with draws $\boldsymbol{d}$, there exists a decoder $\widehat{w}_{\boldsymbol{d}}$, $\boldsymbol{d} \in \mathcal{D}_{\max}^+$ that has error probability $\Pr(\widehat{w}_{\boldsymbol{d}} \neq w) < \epsilon$, if $R < C_{\boldsymbol{d}} - \epsilon$ by the standard joint typicality argument using the mutual information rate derived above. Therefore, choosing a code $\mathcal{X}$ of rate $R < C_{\boldsymbol{d}} - \epsilon$, for ℓ sufficiently large, by a union bound argument

$$
\bar{\Pr}\left(\bigcup_{\substack{\boldsymbol{d} \in \mathcal{D}_{\max}^+: \\ R < C_{\boldsymbol{d}} - \epsilon}} \widehat{w}_{\boldsymbol{d}} \neq w \right) \leq |\mathcal{D}_{\max}^+| \epsilon.
$$

Since $|\mathcal{D}_{\max}^+|$ is finite we can choose ϵ as small as desired, any rate $R < C_{\boldsymbol{d}}$ is achievable and thus there exist a codebook with vanishing error probability over all channels. $\qquad\square$

Remark 4. *Note that decoding a random code, as in the proof of Lemma 3, still involves a complexity that is only polynomial in* M*. However, more practical capacity-achieving code families for the* K*-parallel multi-draw channel might be obtained by [25]'s rate-matching codes or [26]'s polar code construction.*

In Lemma 3 above, we have seen that there are inner codes that achieve the capacity of several K-parallel multi-draw channels at the same time. A key towards deriving an achievable outer rate for a given inner code is thus to count the number of blocks in which the draw vectors $\boldsymbol{d}$ fall within the set of channels whose capacity is achieved (i.e., how many inner blocks can be recovered with high probability). The following lemma is a generalization of [20, Lemma 2] from $K = 1$ to blocks of size K and states that, asymptotically, the relative number of blocks with draw vector $\boldsymbol{d}$ is with high probability close to the pmf of a K-variate Poisson distribution with parameter c.

Lemma 4. *Let $\epsilon > 0$ and $c \geq 1$ be fixed. Then, for $N = cM$ and M sufficiently large,*

$$\Pr\Big(\sum_{\boldsymbol{d} \in \mathbb{N}_0^K} \big|Q_{\boldsymbol{d}} - \tfrac{M}{K} p_c(\boldsymbol{d})\big| < \epsilon M \Big) > 1 - \epsilon.$$

Proof. In the beginning we need to derive expressions for $\mathbb{E}[Q_{\boldsymbol{d}}]$ and $\mathrm{Var}[Q_{\boldsymbol{d}}]$, which we need later to formally prove the statement of the lemma via Chebyshev's inequality.

Let $\mathcal{D}$ be the set of all possible combinations for a vector $\boldsymbol{d}$ defined as $\mathcal{D} := \{\boldsymbol{d} \in \mathbb{N}_0^K : \sum_{i=1}^K d_i \leq cM\}$ with $M \in \mathbb{Z}$ such that $K \mid M$ and furthermore, $\mathcal{D}^d := \{\boldsymbol{d} \in \mathbb{N}_0^K : \sum_{i=1}^K d_i = d\} \subseteq \mathcal{D}$ for some fixed $d \in \{0, 1, \ldots, N\}$.

We define the indicator function $I_{\boldsymbol{d}}(i)$ and an event for all $i = 1, \ldots, M/K$ blocks in the following:

$$I_{\boldsymbol{d}}(i) := \begin{cases} 1, & \text{if } D_{K(i-1)+j} = d_j \text{ for all } j = 1, \ldots, K, \\ 0, & \text{else}; \end{cases}$$

$$\mathcal{P}_d := \{(D_1, \ldots, D_M) : \sum_{j=1}^K D_{K(i-1)+j} = d\}.$$

We can thus compute the probability

$$\Pr\left(I_{\boldsymbol{d}}(i) = 1\right) = \Pr\left((D_1, \ldots, D_K) = (d_1, \ldots, d_K)\right)$$
$$= \Pr\left((D_1, \ldots, D_K) = (d_1, \ldots, d_K)) \mid \mathcal{P}_{d^+}\right) \cdot \Pr\left(\mathcal{P}_{d^+}\right),$$

where we abbreviate $d^+ = \sum_{i=1}^K d_i$. In the last equation the second term can be interpreted as the probability of drawing exactly d^+ strands in one block of length K. Thus, we obtain

$$\Pr\left(\mathcal{P}_{d^+}\right) = \binom{N}{d}\left(\frac{1}{M/K}\right)^{d^+}\left(1 - \frac{1}{M/K}\right)^{N-d^+},$$

since there are in total M/K blocks. For clarity, we abbreviate the first term by

$$B(d^+, \boldsymbol{d}) := \Pr\left(D_j = d_j \mid \mathcal{P}_{d^+}, \forall j = 1, \ldots, K\right).$$

Observe that we can express $Q_{\boldsymbol{d}}$ as the sum of the indicator variables, i.e., $Q_{\boldsymbol{d}} = \sum_{i=1}^{M/K} I_{\boldsymbol{d}}(i)$. We compute

$$\mathbb{E}[Q_{\boldsymbol{d}}] = \sum_{i=1}^{M/K} \mathbb{E}[I_{\boldsymbol{d}}(i)] = \sum_{i=1}^{M/K} \Pr\left(I_{\boldsymbol{d}}(i) = 1\right)$$
$$= \frac{M}{K} \cdot \binom{N}{d^+}\left(\frac{1}{M/K}\right)^{d^+}\left(1 - \frac{1}{M/K}\right)^{N-d^+} \cdot B(d^+, \boldsymbol{d})$$
$$\leq \frac{M}{K}\frac{N^{d^+}}{d^+!}\frac{1}{(M/K)^{d^+}}e^{-\frac{KN}{M}}e^{\frac{d^+K}{M}}B(d^+, \boldsymbol{d})$$
$$= \frac{M}{K}p_{Kc}(d^+)e^{\frac{d^+K}{M}}B(d^+, \boldsymbol{d}),$$

where we used for the inequalities that $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$ and $\binom{N}{d^+} \leq \frac{N^{d^+}}{d^+!}$ for any $N, d^+ \in \mathbb{N}_0$ with $N \geq d^+$. Moreover,

$$\mathbb{E}[Q_{\boldsymbol{d}}^2] = \mathbb{E}[Q_{\boldsymbol{d}}] + B(d^+, \boldsymbol{d})^2$$
$$\cdot \frac{M}{K}\left(\frac{M}{K} - 1\right)\frac{N^{[2d^+]}}{(d^+!)^2}\left(\frac{1}{M/K}\right)^{2d^+}\left(1 - \frac{2}{M/K}\right)^{N-2d^+}.$$

In the last equality of the equation before we used $N^{[2d^+]} = N(N-1)\ldots(N-2d^++1)$. Using the aforementioned bounds and additionally using that $(1 - x)^{d^+} \geq 1 - d^+x$ for any

$0 < x < 1$ and $d^+ \in \mathbb{N}_0$ and $d^+ \cdot p_{Kc}(d^+) \leq Kc$, we obtain the following upper bound after some manipulations.

$$\mathrm{Var}[Q_{\boldsymbol{d}}] \leq \frac{M}{K}p_{Kc}(d^+)e^{\frac{d^+K}{M}}B(d^+, \boldsymbol{d})$$
$$+ \left(\frac{M}{K}\right)^2 p_{Kc}^2(d^+)\left(1 - \left(1 - \frac{2}{M/K}\right)^{d^+}\right)$$
$$\cdot e^{\frac{4d^+K}{M}}B(d^+, \boldsymbol{d})^2$$
$$\leq \frac{M}{K}p_{Kc}(d^+)e^{4Kc}B(d^+, \boldsymbol{d})(1 + 2Kc)$$

Note that this derivation extends the results from [27] to arbitrary block sizes K.

With these valid expressions for $\mathbb{E}[Q_{\boldsymbol{d}}]$ and $\mathrm{Var}[Q_{\boldsymbol{d}}]$ we are prepared to prove the lemma after minor definitions. We abbreviate the statement of the lemma by

$$\mathcal{Q} = \left\{\boldsymbol{Q} : \sum_{\boldsymbol{d} \in \mathcal{D}}\big|Q_{\boldsymbol{d}} - \tfrac{M}{K}p_c(\boldsymbol{d})\big| < \epsilon M\right\}.$$

Let $b_M(\boldsymbol{d})$ be a non-negative real number such that the sum $\sum_{\boldsymbol{d} \in \mathcal{D}} b_M(\boldsymbol{d}) \leq \epsilon M$ for large enough M in the following. Hence, we can derive via the triangular inequality for $|Q_{\boldsymbol{d}} - \tfrac{M}{K}p_c(\boldsymbol{d})| \leq |Q_{\boldsymbol{d}} - \mathbb{E}[Q_{\boldsymbol{d}}]| + |\mathbb{E}[Q_{\boldsymbol{d}}] - \tfrac{M}{K}p_c(\boldsymbol{d})|$ and Chebyshev's inequality

$$\Pr(\mathcal{Q}) = \Pr\Big(\sum_{\boldsymbol{d} \in \mathcal{D}}\big|Q_{\boldsymbol{d}} - \tfrac{M}{K}p_c(\boldsymbol{d})\big| < \epsilon M \Big)$$
$$= 1 - \Pr\Big(\sum_{\boldsymbol{d} \in \mathcal{D}}\big|Q_{\boldsymbol{d}} - \tfrac{M}{K}p_c(\boldsymbol{d})\big| \geq \epsilon M \Big)$$
$$\overset{(a)}{\geq} 1 - \sum_{\boldsymbol{d} \in \mathcal{D}}\Pr\Big(\big|Q_{\boldsymbol{d}} - \tfrac{M}{K}p_c(\boldsymbol{d})\big| \geq b_M(\boldsymbol{d}) \Big)$$
$$\geq 1 - \sum_{\boldsymbol{d} \in \mathcal{D}}\Pr\Big(|Q_{\boldsymbol{d}} - \mathbb{E}[Q_{\boldsymbol{d}}]| \geq b_M(\boldsymbol{d}) - |\mathbb{E}[Q_{\boldsymbol{d}}] - \tfrac{M}{K}p_c(\boldsymbol{d})| \Big)$$
$$> 1 - \sum_{\boldsymbol{d} \in \mathcal{D}}\frac{\mathrm{Var}[Q_{\boldsymbol{d}}]}{\big(b_M(\boldsymbol{d}) - |\mathbb{E}[Q_{\boldsymbol{d}}] - \tfrac{M}{K}p_c(\boldsymbol{d})|\big)^2},$$

where we used a union bound argument for all $\boldsymbol{d} \in \mathcal{D}$ in (a). We now choose $b_M(\boldsymbol{d})$ as

$$b_M(\boldsymbol{d}) = |\mathbb{E}[Q_{\boldsymbol{d}}] - \tfrac{M}{K}p_c(\boldsymbol{d})| + \sqrt{M}M^\epsilon\sqrt{p_{Kc}(d^+)}\, d^+.$$

Observe that one can verify that $p_c(\boldsymbol{d}) = p_{Kc}(d^+) \cdot B(d^+, \boldsymbol{d})$. Inserting this choice of $b_M(\boldsymbol{d})$ into the expression of $\Pr(\mathcal{Q})$ we get

$$\Pr(\mathcal{Q}) > 1 - M^{-2\epsilon}\frac{e^{4Kc}(1 + 2Kc)}{K}\sum_{\boldsymbol{d} \in \mathcal{D}}\frac{B(d^+, \boldsymbol{d})}{d^{+2}}$$
$$= 1 - M^{-2\epsilon}\frac{e^{4Kc}(1 + 2Kc)}{K}\sum_{d=0}^N\frac{1}{d^2}\sum_{\boldsymbol{d} \in \mathcal{D}^d}B(d, \boldsymbol{d})$$

Due to $\sum_{\boldsymbol{d} \in \mathcal{D}^d}B(d, \boldsymbol{d}) = 1$ it follows that for any $\epsilon > 0$ and large enough M this will result to $\Pr(\mathcal{Q}) > 1 - \epsilon$.

It remains to prove the initial assumption that

$$\sum_{\boldsymbol{d} \in \mathcal{D}} b_M(\boldsymbol{d}) < \epsilon M$$

for large enough M. For the sake of brevity we omit to show that $\sum_{\boldsymbol{d} \in \mathcal{D}}\big|\mathbb{E}[Q_{\boldsymbol{d}}] - \tfrac{M}{K}p_c(\boldsymbol{d})\big| < \tfrac{\epsilon}{2}M$ and refer as a guidance of the proof to [20].

Let us denote $\Delta_M(\boldsymbol{d}) = \sqrt{M}M^\epsilon \sqrt{p_{Kc}(d^+)}\, d^+$ as the remaining summand in $b_M(\boldsymbol{d})$ and investigate its sum as follows.

$$
\begin{aligned}
\sum_{\boldsymbol{d}\in\mathcal{D}} \Delta_M(\boldsymbol{d}) &= M^{\frac{1}{2}+\epsilon} \sum_{d=0}^{N} \sqrt{p_{Kc}(d)}\, d \sum_{\boldsymbol{d}\in\mathcal{D}^d} 1 \\
&\overset{(b)}{\leq} M^{\frac{1}{2}+\epsilon} \sum_{d=0}^{N} e^{\frac{Kc}{2}} \sqrt{Kc}\, p_{Kc}\left(\left\lfloor \frac{d}{2} \right\rfloor\right) d^{(K+1)} \\
&\overset{(c)}{\leq} M^{\frac{1}{2}+\epsilon} e^{\frac{Kc}{2}} \sqrt{Kc}(2\cdot 3^{K+1}) \sum_{d=0}^{\lceil \frac{N}{2}\rceil} p_{Kc}(d)d^{+(K+1)} \\
&\overset{(d)}{\leq} M^{\frac{1}{2}+\epsilon} e^{\frac{Kc}{2}} \sqrt{Kc}(2\cdot 3^{K+1}) \sum_{d=0}^{\infty} p_{Kc}(d)d^{(K+1)} < \frac{\epsilon}{2}M
\end{aligned}
$$

Here we used for (b) that it holds $|\mathcal{D}^d| \leq d^K$. Additionally, the inequality $\sqrt{p_{Kc}(d)} \leq e^{\frac{Kc}{2}}\sqrt{Kc}\,p_{Kc}(\lfloor \frac{d}{2}\rfloor)$ is used, where $\sqrt{d!} > \lfloor \frac{d}{2}\rfloor!$ is applied and due to the flooring operation it holds $(Kc)^{\frac{d}{2}} \leq \sqrt{Kc}(Kc)^{\lfloor \frac{d}{2}\rfloor}$ both for $d\in\mathbb{N}_0$ and $K,c\geq 1$. For (c) we have split the sum in odd and even terms but directly used the upper bound $(2d)^{(K+1)}+(2d+1)^{(K+1)} \leq 2\cdot(3d)^{K+1}$ to combine the split terms again. In the step (d) we transform the expression such that it is resulting to the $K+1$-th moment expression $\sum_{d=0}^{\infty} p_{Kc}(d)d^{(K+1)}$ of the Poisson distribution with parameter Kc and d. Any moment of the Poisson distribution is finite and therefore we can conclude the last inequality for M sufficiently large and small $\epsilon > 0$.

As a consequence both terms of $b_M(\boldsymbol{d})$ are not greater than $\frac{\epsilon}{2}M$, which concludes the proof. $\square$

The previous two lemmas have answered the questions under which conditions an inner block can be recovered given that all of its strands are properly clustered and indices are correctly retrieved. The next and last lemma shows that asymptotically, most strands are indeed correctly clustered and index-recovered with high probability under the assumption of Lemma 4.

Lemma 5. *Let $\mathcal{P}$ be the event from Lemma 4. Fix $\epsilon > 0$, $\beta < R_{\text{in}}(1 - \beta/R_{\text{ix}})(1 - h(2p))$. Then for all $\boldsymbol{Q}\in\mathcal{P}$,*

$$
\begin{aligned}
\Pr(M_{\text{C}} > \epsilon M | \boldsymbol{Q}) &\leq \epsilon, \\
\Pr(M_{\text{Ix}} > \epsilon M | \boldsymbol{Q}) &\leq \epsilon, \\
\Pr(M_{\text{In}} > \epsilon M | \boldsymbol{Q}) &\leq \epsilon
\end{aligned}
$$

for large enough M.

Proof. Let $\epsilon' > 0$ be arbitrary and J_j be the indicator that is 1, if $d(\boldsymbol{y}_j, \boldsymbol{x}_{I_j}) > (p + \frac{\epsilon'}{2})L$ or if there exists an output strand $j' \in [N]$, $I_j \neq I_{j'}$ with distance $d(\boldsymbol{y}_j, \boldsymbol{y}_{j'}) \leq \rho L$, where ρL with $\rho = 2p + \epsilon'$ is the diameter of the clustering algorithm introduced in Section II-D. Then the number of wrong clusters is at most $M_{\text{C}} \leq \sum_{j=1}^{N} J_j$. The individual probabilities can be bounded by

$$
\Pr(d(\boldsymbol{y}_j, \boldsymbol{x}_{I_j}) > (p + \tfrac{\epsilon'}{2})L) = \sum_{i=(p+\frac{\epsilon'}{2})L}^{L} \binom{L}{i} p^i (1-p)^{L-i}
$$
$$
\overset{(a)}{\leq} e^{-LD(p+\frac{\epsilon'}{2}\|p)} \leq \epsilon'/2
$$

for large enough L, where $D(a\|p)$ is the Kullback-Leibler divergence between Bernoulli variables with success probabilities a and p. Here we used the Chernoff bound on the binomial tail in inequality (a). On the other hand, the first $R_{\text{in}}\ell$ bits of each strand $\boldsymbol{y}_j$ form a uniform random vector,

as the marginal distribution of any symbol in a MDS code is uniform. It follows that we can bound the probability of another strand being within the clustering diameter

$$
\Pr(\exists j' : I_j \neq I_{j'}, d(\boldsymbol{y}_j, \boldsymbol{y}_{j'}) \leq \rho L)
$$
$$
\overset{(b)}{\leq} N\Pr(d(\boldsymbol{y}_j, \boldsymbol{y}_{j'}) \leq \rho R_{\text{in}}\ell) \leq N2^{-R_{\text{in}}\ell} \sum_{i=0}^{\rho R_{\text{in}}\ell} \binom{R_{\text{in}}\ell}{i}
$$
$$
\overset{(c)}{\leq} c2^{\beta L} 2^{R_{\text{in}}\ell(h(2p+\epsilon')-1)} \leq \epsilon'/2
$$

for large enough L. Here, (b) follows from the union bound over all output strands and (c) follows from the Chernoff bound on binomial tails, where $h(\bullet)$ denotes the binary entropy function. Thus $\Pr(J_j = 1) \leq 2\epsilon'$ for large enough L. It follows that $\mathbb{E}(\sum_{j=1}^{N} J_j) \leq \epsilon' N$ and $\mathrm{Var}(\sum_{j=1}^{N} J_j) \leq \epsilon' N + \epsilon' N^2$. The statement then follows from Tschebyschev's inequality.

The second claim follows from the fact that for any $\epsilon' > 0$ the probability of erroneous index decoding is at most ϵ' for any cluster with $C_{d_i} \geq R_{\text{ix}}$ and large enough L as we are decoding a capacity achieving code with rate less than the capacity. Given $\boldsymbol{Q}$, the individual decoding processes are statistically independent and thus M_{Ix} can be upper bounded by a binomial variable with success probability ϵ' and N trials. Hence,

$$
\Pr(M_{\text{Ix}} > \epsilon M) \leq \sum_{i=\epsilon M}^{N} \binom{N}{i} (\epsilon')^i (1-\epsilon')^{N-i} \overset{(a)}{\leq} e^{-ND(c\epsilon\|\epsilon')},
$$

where (a) follows from the Chernoff bound. Hence $\Pr(M_{\text{Ix}} > \epsilon M)$ is at most ϵ for large enough M, respectively N.

The bound for M_{In} is obtained by choosing a $d_{\max}^+$ such that

$$
\sum_{d^+ > d_{\max}^+} p_{Kc}(d^+) < \epsilon'.
$$

Then, for all blocks for which $d_{bK+1} + \cdots + d_{bK+K} \leq d_{\max}^+$ the number of errors can be bounded as for the case of M_{Ix} using the bound on the error probability of inner decoding in Lemma 3. Since the number of blocks with $d_{bK+1} + \cdots + d_{bK+K} > d_{\max}^+$ is given by $Q_{d_{\max}^+ + 1} + Q_{d_{\max}^+ + 1} + \ldots$, which is at most $M(\sum_{d^+ > d_{\max}^+} p_{Kc}(d^+) + \epsilon') \leq 2M\epsilon'$ with high probability (c.f. Lemma 4), the claim follows. $\square$

C. Proof of Main Statement

Proof of Theorem 1. We will prove the theorem by showing that the probability of a decoding error inside the outer codeword goes to zero as $M \to \infty$. In this view, let s, t denote the random variables that count number of erasures, respectively errors inside the outer codeword after clustering, index decoding and inner decoding. Recalling the definitions of the number of wrong clusters M_{C}, and wrongly decoded indices M_{Ix}, respectively inner codewords M_{In} from Section II-D, we obtain

$$
\begin{aligned}
s &\leq \sum_{\boldsymbol{d}: C_{\boldsymbol{d}}(R_{\text{ix}})\leq R_{\text{in}}} Q_{\boldsymbol{d}} + 2M_{\text{C}} + 2M_{\text{Ix}}, \\
t &\leq M_{\text{In}} + 2M_{\text{C}} + 2M_{\text{Ix}}.
\end{aligned}
$$

This can be explained as follows. First, the inner decoder declares an erasure, when $C_{\widehat{\boldsymbol{d}}} \leq R_{\text{in}}$ for a block with $\widehat{\boldsymbol{d}}$ draws. Assuming no clustering errors or index decoding errors, the first term counts the number of erasures declared by the inner decoder. Then, for each wrong cluster or a wrongly decoded index, there can be at most two inner codewords affected - the inner codeword, where the cluster actually belongs to and

the inner codeword that the cluster is mistakenly associated to. The same argument can be used for the upper bound on t. We obtain for the success probability

$$\Pr(\mathcal{S}) = \sum_{\boldsymbol{Q}} \Pr(\mathcal{S}|\boldsymbol{Q}) \Pr(\boldsymbol{Q})$$
$$\overset{(a)}{\geq} \sum_{\boldsymbol{Q} \in \mathcal{P}} \Pr(s + 2t \leq M(1 - R_{\text{out}})|\boldsymbol{Q}) \Pr(\boldsymbol{Q}),$$

where (a) holds by the properties of unique decoding of MDS codes. We further obtain for the conditional success probability

$$\Pr(s + 2t \leq M(1 - R_{\text{out}})|\boldsymbol{Q})$$
$$= 1 - \Pr(s + 2t + 10\epsilon M > M(1 - R_{\text{out}}) + 10\epsilon M|\boldsymbol{Q})$$
$$\geq 1 - \Pr(s > M(1 - R_{\text{out}}) - 10\epsilon M|\boldsymbol{Q}) - \Pr(t > 5\epsilon M|\boldsymbol{Q}).$$

On the one hand, we have

$$\Pr(t > 5\epsilon M|\boldsymbol{Q}) \leq \Pr(M_{\text{In}} + 2M_{\text{C}} + 2M_{\text{Ix}} > 5\epsilon M|\boldsymbol{Q})$$
$$\leq \Pr(M_{\text{C}} > \epsilon M|\boldsymbol{Q}) + \Pr(M_{\text{Ix}} > \epsilon M|\boldsymbol{Q}) + \Pr(M_{\text{In}} > \epsilon M|\boldsymbol{Q}) \overset{(b)}{\leq} 3\epsilon,$$

where (b) follows from Lemma 5. On the other hand, abbreviate $\mathcal{D}_C = \{\boldsymbol{d} \in \mathbb{N}_0^K : C_{\boldsymbol{d}}(R_{\text{ix}}) \leq R_{\text{in}}\}$. Then, for $\boldsymbol{Q} \in \mathcal{P}$

$$\Pr(s > M(1 - R_{\text{out}}) - 10\epsilon M|\boldsymbol{Q})$$
$$\leq \Pr\Big(\sum_{\boldsymbol{d} \in \mathcal{D}_C} Q_{\boldsymbol{d}} + 2M_{\text{C}} + 2M_{\text{Ix}} > M(1 - R_{\text{out}}) - 10\epsilon M|\boldsymbol{Q}\Big)$$
$$\overset{(c)}{\leq} \Pr\Big(2M_{\text{C}} + 2M_{\text{Ix}} > M(1 - R_{\text{out}} - \sum_{\boldsymbol{d} \in \mathcal{D}_C} p_c(\boldsymbol{d}) - 11\epsilon)|\boldsymbol{Q}\Big)$$
$$\overset{(d)}{\leq} \Pr(2M_{\text{C}} + 2M_{\text{Ix}} > 4\epsilon M|\boldsymbol{Q}) \leq 2\epsilon$$

where in (c) we used that $\boldsymbol{Q} \in \mathcal{P}$ and in (d), we chose $R_{\text{out}} = \sum_{\boldsymbol{d}:C_{\boldsymbol{d}}(R_{\text{ix}}) > R_{\text{in}}} p_c(\boldsymbol{d}) - 15\epsilon$. Finally inserting these results into the overall success probability, and using the bound on $\Pr(\mathcal{P})$ from Lemma 4, we obtain

$$\Pr(\mathcal{S}) \geq (1 - 5\epsilon) \sum_{\boldsymbol{Q} \in \mathcal{P}} \Pr(\boldsymbol{Q}) \geq (1 - 5\epsilon)(1 - \epsilon)$$

for large enough M and the claim of the theorem follows, as we can choose ϵ as small as desired. $\square$

IV. NUMERICAL RESULTS

The achievable rate given in the main statement of the paper (Theorem 1) is a semi-closed expression it is not obvious from the formulation how to compute it efficiently.

One way is to approximate it by summing only over a finite subset $\mathcal{A} \subset \mathbb{N}_0^K$ of vectors $\boldsymbol{d}$ for which $\sum_{\boldsymbol{d} \in \mathbb{N}_0^K \setminus \mathcal{A}} p_c(\boldsymbol{d})$ is sufficiently small (e.g., $\mathcal{A} = \{\boldsymbol{d} \in \mathbb{N}_0^K : \sum_{i=1}^K d_i \leq d^+\}$ for a large d^+). A drawback of this approach is that the

cardinality of the set $\mathcal{A}$ might grow exponentially[3] in the channel parameters K, c and the inverse of the sought precision. A second and much more efficient approach is to approximate the achievable rate by a Monte-Carlo simulation. The foundation of this method is that the achievable rate equals the probability $\Pr\left(\frac{1}{K} \sum_{i=1}^K C_{d_i}(R_{\text{ix}}) > R_{\text{in}}\right)$ for a random variable $\boldsymbol{d}$ drawn from a K-variate Poisson distribution with parameter c. Hence, we can repeatedly take samples of $\boldsymbol{d}$ and estimate the achievable rate by the relative number of times that $\frac{1}{K} \sum_{i=1}^K C_{d_i}(R_{\text{ix}}) > R_{\text{in}}$. The complexity of this method grows only linearly in K and the inverse of the sought variance of the estimator.

In the following, we give numerical results obtained from the Monte-Carlo method. All simulations were performed using the computer-algebra system SageMath [28] (version 8.1). As fixed channel parameters, we took $\beta = {}^1\!/_{20}$ and $p = 0.1$, which are common values for nanopore sequencing, see e.g. [11], [19].

Figure 2 presents the achievable overall rate R as a function of the block length K for several channel parameters c. For comparison, it displays the capacity of the channel [20], [21] and the achievable rate for $K \to \infty$ (cf. Corollary 2.A). In all cases, the achievable rate for $K \to \infty$ is close to the capacity. Furthermore, for all simulated parameters, the achievable rate approaches the expected asymptotic value for growing K. There is a significant gap between the asymptotic rate and the rate for $K = 1$, especially for $c = 2, 4, 6$. For small c, the achievable rate first decreases for small K, but then increases until close to the asymptotic value. Using a sufficiently large block length K (e.g., $K = 100$ for $c \geq 2$), the achievable rate is significantly larger than for $K = 1$, and can get very close to the capacity.

Figure 3 shows both the achievable outer rate R_{out} and the corresponding overall rate R as a function of the inner rate R_{in} for $c = 2$. The figure also illustrates the asymptotic behavior ($K \to \infty$), which follows from the proof of Corollary 2.A: For $K \to \infty$, the achievable outer rate, given by the probability $\Pr\left(\frac{1}{K} \sum_{i=1}^K C_{d_i}(R_{\text{ix}}) > R_{\text{in}}\right)$, approaches 1 for $R_{\text{in}} < \mathbb{E}(C_{d_i}(R_{\text{ix}}))$ and 0 for larger R_{in}. As we have $R = R_{\text{in}} R_{\text{out}}(1 - \beta/R_{\text{ix}})$, for $K \to \infty$, we get a linear increase of the overall rate up to $R_{\text{in}} = \mathbb{E}(C_{d_i}(R_{\text{ix}}))$ and $R = 0$ for larger R_{in}. As expected, the curves for approach the asymptotic behavior for growing K. An important interpretation of the "R over R_{in}" plot is that the positions of maxima of the curves give a design criterion on how to choose R_{in} to maximize R.

---

[3]For instance, in the case $\mathcal{A} = \{\boldsymbol{d} \in \mathbb{N}_0^K : \sum_{i=1}^K d_i \leq d^+\}$, the cardinality is related to the number of partitions of integers $\leq d^+$.

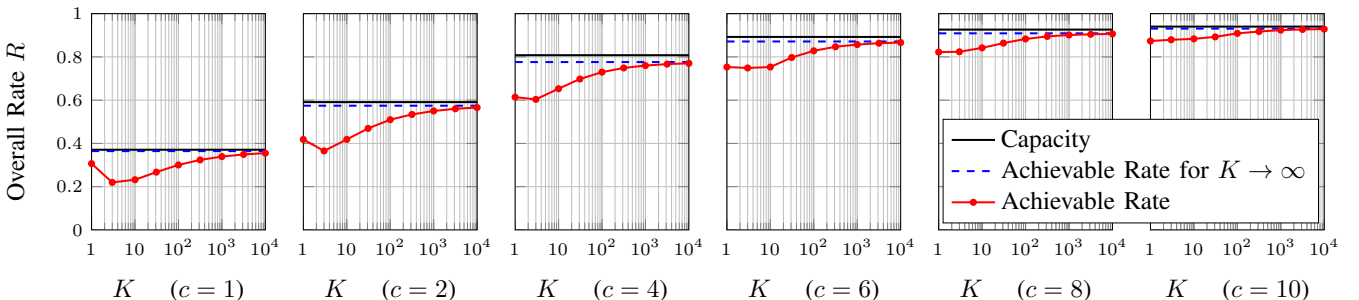Figure 2. Overall rate R over block size K (logarithmic scale) for varying channel parameter $c = 1, 2, 4, 6, 8, 10$. Fixed parameters: $\beta = {}^1\!/_{20}$ and $p = 0.1$. For each c, the index rate is chosen to maximize the achievable rate for $K \to \infty$ given in Corollary 2.A: $R_{\text{ix}} = 0.999 \cdot C_{d^*}$ with $d^* = 1$ for $c = 1, 2, 4$, $d^* = 2$ for $c = 6, 8$, and $d^* = 3$ for $c = 10$. The curves for finite K are obtained by Monte-Carlo simulations with $N_i = 10^4$ samples.
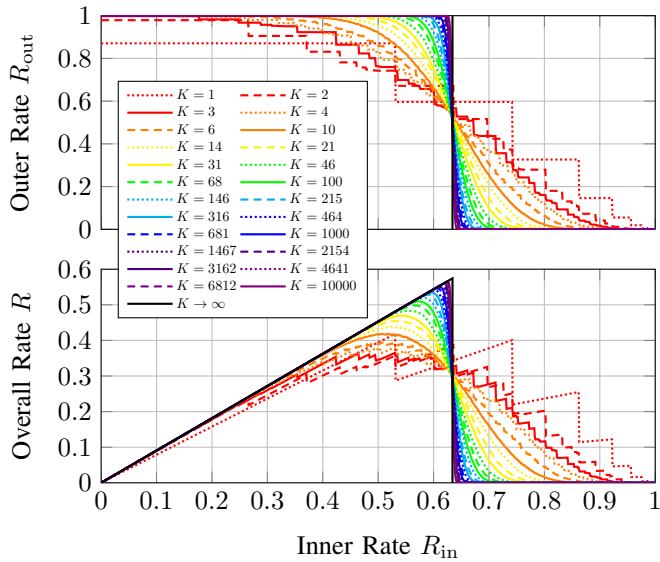
Figure 3. Achievable outer rate R_{out} and overall rate R, respectively, plotted over the inner rate R_{in} for several K (here: $K = \lfloor 10^{i/6} \rfloor$ for $i = 1, \ldots, 24$). Fixed parameters: $c = 2$, $\beta = 1/20$, $p = 0.1$, $R_{\mathrm{ix}} = 0.5304$ ($= 0.999 \cdot C_1$, where C_1 is the capacity of the multi-draw channel with $d = 1$ draw and crossover probability p). The curves for finite K are obtained from Monte-Carlo simulations with $N = 10^4$ samples. The curves for $K \to \infty$ show the theoretical predictions obtained from Lemma 2.A.

REFERENCES

[1] 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.

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

[3] R. N. Grass *et al.*, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, Feb. 2015.

[4] S. M. H. T. Yazdi *et al.*, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, no. 1, p. 14138, Nov. 2015.

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

[6] J. Bornholt *et al.*, "A DNA-based archival storage system," in *Proc. Int. Conf. Architectural Support for Programming Languages and Operating Systems*, Atlanta, Georgia, USA, 2016, pp. 637–649.

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

[19] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, no. 1, p. 9663, Dec. 2019.

[8] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Sci. Rep.*, vol. 7, no. 1, p. 5011, Dec. 2017.

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

[10] S. Chandak *et al.*, "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes," in *Proc. Annu. Allerton Conf. Commun. Control Comp.*, Monticello, IL, Sep. 2019, pp. 147–156.

[11] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Trans. Inf. Theory*, vol. 66, pp. 2331–2351, Apr. 2020.

[12] S. K. Tabatabaei *et al.*, "DNA punch cards: Storing data on native DNA sequences via nicking," Synthetic Biology, preprint, Jun. 2019.

[13] K. A. Schouhamer Immink and K. Cai, "Design of capacity-approaching constrained codes for DNA-based storage systems," *IEEE Communications Letters*, vol. 22, no. 2, pp. 224–227, Feb. 2018.

[14] Y. Wang *et al.*, "Construction of bio-constrained code for DNA data storage," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 963–966, Jun. 2019.

[15] W. Song, K. Cai, and K. A. Schouhamer Immink, "Sequence-subset distance and coding for error control for DNA-based data storage," in *Proc. Int. Symp. Inf. Theory*, Paris, France, Jul. 2019, pp. 86–90.

[16] J. Sima, N. Raviv, and J. Bruck, "On coding over sliced information," in *Proc. Int. Symp. Inf. Theory*, Paris, France, Jul. 2019, pp. 767–771, arXiv: 1809.02716.

[17] 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.

[18] 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, arXiv: 1902.10832.

[20] 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.

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

[22] G. Forney, Jr., *Concatenated Codes*. MIT Press, Cambridge, Massachusetts, 1966.

[23] 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.

[24] T. M. Cover and J. A. Thomas, *Elements of information theory*, second edition ed. Hoboken, NJ: Wiley-Interscience, 2006, oCLC: 255202515.

[25] 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.

[26] 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.

[27] V. F. Kolchin, B. A. Sevastyanov, and V. P. Chistyakov, *Random Allocations*. Washington D.C.: V.H. Winston & Sons, 1978.

[28] The Sage Developers, *SageMath*, https://www.sagemath.org.