Bounding the Capacity of the Multinomial Channel using KL Divergence Covering and Packing

Jennifer Tang
Laboratory for Information and Decision Systems and Institute for Data, Systems, and Society
Massachusetts Institute of Technology, Cambridge, MA
Email: jstang@mit.edu

***Abstract*—We examine the capacity for the *multinomial channel*, a natural extension of the binomial channel. In the multinomial channel setting, the input is a probability distribution over k entries, and the output of the channel is n items sampled independently with the chosen input probability distribution. Applications of this channel include the use of composite DNA, which is a method for expanding the alphabet set used in DNA storage systems in order to improve the information throughput. In this work, we compute non-asymptotic upper and lower bounds for the information rate of the multinomial channel.**

I. Introduction

Suppose that Alice has n coins where she can control the probability that each coin lands on heads. Specifically, she can choose one parameter $p \in [0, 1]$, and then each of the n coins will be flipped and each independently will have p probability of landing on heads. After p is chosen and all n coins are flipped, Bob observes how many coins ended up as heads and tails (but does not know p). Alice is allowed to repeat this procedure, choosing whatever p she likes in each step, and Bob always observes the outcomes. How much information can Alice communicate to Bob using this method?

This is the problem of finding the information capacity of a channel called the *binomial channel* [1]. A natural extension of the binomial channel, called the *multinomial channel* in [2], is a channel where instead of two outcomes (heads and tails), there are k different outcomes. Alice chooses a probability distribution over k different outcomes at each step; as before, Bob then observes the total outcome counts from n independent draws from Alice's chosen distribution. The binomial and multinomial channels can be used to model biological storage and communication systems. Its capacity is also exactly equivalent to the minimax rates of redundancy for universal prediction problems [3], for which asymptotic results are known.

In this work, our goal is to look at the capacity of the multinomial channel, particularly in the non-asymptotic regime (fixed n, rather than $n \to \infty$), in order to understand the information capacity of systems like the composite DNA storage system.

A. Motivation and Related Work

DNA storage systems are technologies which aim to store digital data on DNA. These methods have a lot of potential due to the natural advantages of DNA, which can store a high density of information with longevity. DNA storage systems have been implemented in practice [4]–[7] but the technology is not efficient enough to be currently viable as a storage solution. Thus, a significant amount of research effort has been put into understanding coding strategies for DNA storage systems in attempt to understand how to design systems to improve the storage rates [8]–[11]. A classic procedure for DNA storage is to synthesize DNA molecules where the desired information is stored by encoding the nucleotides $\{A, C, G, T\}$ (the basic DNA alphabet). The synthesis process creates many copies for each designed strand. Then these short strands of synthesized DNA, called oligos (short for oligonucleotides), can be put in storage. To retrieve the information, the oligos or strands need to be read by sequencing, a procedure which can only randomly access the oglios and has some amount of error.

In order to boost the efficiency of DNA storage systems, a recent development is to increase the size of the DNA alphabet (and thus the stored information content) by using *composite DNA* [12], [13]. Composite DNA increases the alphabet size by allowing mixture of nucleotides (in a predetermined ratio) as one letter. Mixtures are possible due to the oligo copies created during the standard synthesis method.

Let probability vector $\boldsymbol{\pi} = (\pi(A), \pi(C), \pi(G), \pi(T))$, where $\pi(A) + \pi(C) + \pi(G) + \pi(T) = 1$ be such that $\pi(i)$ represents the proportion of nucleotide $i \in \{A, C, G, T\}$ present at some position in the DNA strand (or, rather, collection of strands). As an example, a possible letter in the composite DNA alphabet could be $\boldsymbol{\pi}_1 = (0.25, 0.25, 0.5, 0)$ while another could be $\boldsymbol{\pi}_2 = (0, 0.5, 0.5, 0)$. If $\boldsymbol{\pi}_1$ is chosen for a particular position, then during sequencing, we can expect to see a mixture of 25% A's and C's each and 50% G's in that position. If however, we see roughly 50% C's and 50% G's, then more likely $\boldsymbol{\pi}_2$ is the chosen letter in the position of the strand. The classical model where only nucleotides are used are letters (i.e one of $\{A, C, G, T\}$ is used 100% of the time in each position) can only encode 2 bits per synthesis cycle, whereas composite DNA allows the number of bits encoded per synthesis cycle to be much larger.

What puts a limit on the number of letters available in composite DNA is the number of strands sequenced or read. For example, if the two mixture choices for a position in the strand are $\boldsymbol{\pi}_1 = (0.25, 0.25, 0.5, 0)$ and $\boldsymbol{\pi}_2 = (0.251, 0.249, 0.5, 0)$ and the number of reads is too few, then whether a position was meant to encode $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ cannot be determined as the two mixtures are too similar.

The multinomial channel discussed above is exactly an abstract channel model for the composite DNA storage system. The number of symbols used to build the mixtures is k (for DNA storage, $k = 4$ since there are 4 nucleotides). The number n corresponds to the number of reads used to determine the letter in each position of the strand. We let the set of composite (or mixture) letters used in each codeword of the channel be $\mathcal{M} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_m\}$ where each $\boldsymbol{\pi}_i$ is a probability vector over k symbols; note that the alphabet for the transmitted codewords is this set of composite (or mixture) letters. To avoid confusion, we refer to the set of nucleotides as the *original alphabet*, and the set of composite letters $\mathcal{M}$ as the *composite alphabet*. Let θ be the random variable representing a selected probability in $\mathcal{M}$. The channel, represented as $\theta \to Y^n$, takes each input letter $\boldsymbol{\pi}_i$ and then outputs $Y^n = (Y_1, \ldots, Y_n)$ which are n random samples independently drawn according to probability vector $\boldsymbol{\pi}_i$. The random samples Y^n models the randomly selected strands for sequencing. Some assumptions of this model are that the DNA strands sequenced are selected with replacement and enough of them are synthesized to represent any mixture. (In practice, the mixtures need to have rational values, for simplicity we allow the probabilities to be any real number.) Other errors that may occur in the synthesis and sequencing processes are not specifically taken into account (though, depending on the type of error, there may be a way to model this directly in the multinomial channel).

Let P_θ be a prior on θ. The capacity of the multinomial channel, is given by

$$C(k, n) = \max_{P_\theta} I(\theta; Y^n)\,. \tag{1}$$

This quantity is intimately connected to the problem of universal prediction [3]. In the universal prediction scenario, samples are independently and identically distributed (iid) according to an unknown (and adversarially selected) distribution, and an estimator is trying to predict the next symbol while minimizing logarithmic loss compared to an oracle who has knowledge of the unknown distribution. This excess loss above the minimum achievable loss is called the minimax redundancy, and due to the structure of the universal prediction problem, it is equivalent to the capacity of the multinomial channel as stated in equation (1). A line of work [14]–[17] has been dedicated to studying this capacity. In [17], it was determined that

$$C(k, n) = \frac{k-1}{2} \log \frac{n}{2\pi e} + \log \frac{\Gamma(1/2)^k}{\Gamma(k/2)} + o(1) \tag{2}$$

where $\Gamma(\cdot)$ is the Gamma function. However, this result is an asymptotic result as $n \to \infty$, leaving open the problem of determining bounds for fixed n. A similar quantity, worst-case minimax redundancy (or regret) which upper bounds $C(k, n)$, is studied in works including [18]–[20]. The formula for regret in [19] is given in terms of sums with multinomial coefficients and [20] approximates these terms asymptotically in its upper bound.

The multinomial channel was studied in the context of DNA storage systems in [21]. Unlike our work, [21] treats the entire DNA strand as a possible letter in their alphabet. Each codeword is a specific collection of strands and their results are stated in terms of the number of possible strands used. There have been some attempts to compute the capacity for the binomial channel (the case when $k = 2$) for fixed values of n. In [1], the authors use a convex optimization problem solved by the ellipsoid method to numerically compute the capacity. More recently, there has been new interest in the binomial channel, since the generalizations of the binomial channel can be used to model the *particle-intensity channel* (PIC) proposed for molecular communication [22]. In molecular communication, transmitters release particles and encode information by controlling the intensity of the particles released. The receiver detects the particles and uses information, such as the number of particles received in a specified time interval, to read the message. However, even given a particular intensity, the number of particles released over an interval is random, and the receiver may also randomly (independently) fail to detect some particles; thus, the binomial channel is an appropriate model for this type of channel. One problem of interest is finding the capacity-achieving input distribution (CAID) for the PIC, which is done in [22] using an algorithm the authors call dynamic assignment Blahut-Arimoto (DAB). A more efficient method for numerically computing the capacity of the binomial channel is given in [23]. In [24], the authors determined that the CAID is unique and give non-asymptotic upper and lower bounds for the binomial channel. They also derive the CAID and exact capacity when $n \leq 3$.

Error correcting codes for composite DNA were studied in [25] for the case where the number of errors is some fixed t. This was done for a few different error types and code constructions were included. In [2], the authors studied the CAID for the multinomial channel in order to understand the capacity of composite DNA storage. Their key technique was to use *Multidimensional DAB* (M-DAB), a variant of DAB algorithm proposed in [22].

A similar technique to composite DNA is a method called *combinatorial* DNA, where instead of using probability distributions over nucleotides to expand the alphabet, small sequences of DNA, called *shortmers*, are mixed together to be a new letter, and these shortmers are connected to make the strand [26]–[28]. However, in this method, mixture probabilities of shortmers are not used; rather, each letter is a *set* of shortmers, and for a given shortmer only its presence or absence (rather than exact proportion) in this set is used in the code. In this code, an error occurs when a particular shortmer in the set is missed during the read step. A generalization of this method, where shortmers are mixed in a certain ratio, could be modeled by the multinomial channel, and would represent a way to increase the (original) alphabet size k beyond 4 nucleotides.

B. Contributions

The focus of this work is to determine bounds on the capacity of the multinomial channel, which is given by $C(k, n)$ in (1) and corresponds to the information capacity of the

composite DNA storage system. Primarily, we are interested in determining $C(k,n)$ as a function of the number of reads n, with n is finite rather than asymptotic. This gives a quantitative understanding of how many bits per cycle can be encoded in the composite DNA system with a given number of samples and gives an understanding of how many mixtures are distinguishable when reads are limited.

For our results, we have:

Theorem 1 (Upper Bound on Capacity)**.** *For $k, n \geq 2$,*

$$C(n,k) \leq \frac{k-1}{2}\log n + 1 + (k-1)\log 7\sqrt{k-1}\,.$$

Theorem 2 (Lower Bound on Capacity)**.** *For $k, n \geq 2$,*

$$C(n,k) \geq \frac{k-1}{2}\left(\log n - \log(4(k-1)\log n)\right) - \log 2\,.$$

The proof for Theorem 1 is given in Section II and the proof for Theorem 2 is given in section III. As expected, in the regime of large n, the dominant term in both bounds is $\frac{k-1}{2}\log n$, matching the asymptotic result in (2) as $n \to \infty$. In [24], bounds are only given in the case of the binomial channel (where $k = 2$). We determine bounds for general k.

Our key tool for computing our finite n and k bounds is to use KL divergence covering and Rényi divergence packing. Unlike traditional covering and packing, divergence covering and packing works with the space of probabilities. The reason why these concepts are useful is that the key to being able to communicate over the multinomial channel is to select input probabilities (the letters for the composite alphabet) in a way so that different probabilities can be distinguished at the output. Probabilities can be distinguished if they are sufficiently different from another, which can be measured by KL or Rényi divergences. Finding a set of probabilities with appropriate spacing between them is the key to determining our bounds.

C. Preliminary Notation and Definitions

Let $\mathcal{A}$ be the set representing the original alphabet of the multinomial channel and let $k = |\mathcal{A}|$. Let $\triangle_{k-1}$ be the $(k-1)$-dimensional simplex which represents the set of probability distributions over k objects. Vectors (including probability vectors) are represented as bold (typically lowercase) letters. If $\boldsymbol{p} \in \triangle_{k-1}$, then we write $\boldsymbol{p} = (p(1), ..., p(k))$ so that $p(i)$ is the ith entry of (probability) vector $\boldsymbol{p}$. Notation $X \sim \boldsymbol{p}$ means that random variable X is sampled randomly from probability $\boldsymbol{p}$. The notation $X^n \sim \boldsymbol{p}$ means that random variable $X^n = (X_1, \ldots, X_n)$ is such that each X_i is sampled independently from $\boldsymbol{p}$. Let $[k] = \{1, \ldots, k\}$ and for a set $E \subseteq [k]$, $\boldsymbol{p}(E) = \sum_{x \in E} p(x)$. Logarithms are always base-e unless otherwise specified.

We will use $D(\boldsymbol{\pi}\|\boldsymbol{\mu})$ to denote divergences on two probability distributions $\boldsymbol{\pi}, \boldsymbol{\mu} \in \triangle_{k-1}$. Subscripts will be used to identify what the divergence are. For instance, we denote f-divergences by $D_f(\boldsymbol{\pi}\|\boldsymbol{\mu})$ (see [29]), and, in particular, the Kullback-Leibler (KL) divergence is

$$D_{\mathrm{KL}}(\boldsymbol{\pi}\|\boldsymbol{\mu}) = \sum_{x=1}^{k} \pi(x)\log\frac{\pi(x)}{\mu(x)}$$

and the total variation distance (TV) is

$$D_{\mathrm{TV}}(\boldsymbol{\pi}\|\boldsymbol{\mu}) = \mathrm{TV}(\boldsymbol{\pi}, \boldsymbol{\mu}) = \sup_{E \subseteq [k]} |\boldsymbol{\pi}(E) - \boldsymbol{\mu}(E)|.$$

Notably, Kullback-Leibler divergence is not a metric whereas total variation is (as total variation is half the L_1 norm).

Rényi divergences (see [30]) with parameters $\lambda \in \mathbb{R} \setminus \{1\}$ are defined as

$$D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \frac{1}{\lambda - 1}\log\sum_{x=1}^{k} p(x)^\lambda q(x)^{1-\lambda}\,.$$

We will call D_λ the λ-Rényi divergence. As stated in [31, page 191], Rényi divergences are not f-divergences, but they are monotone transformation of f-divergences. The Rényi divergence when $\lambda \to 1$ is equivalent to KL divergence.

Definition 1. *Fix a divergence D_f (f-divergence or Rényi divergence) and an alphabet size k with $(k-1)$-dimensional simplex $\triangle_{k-1}$. A set $\mathcal{M} = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_m : \boldsymbol{\mu}_i \in \triangle_{k-1}\}$ is a* covering *of radius R if*

$$\max_{\boldsymbol{\pi} \in \triangle_{k-1}} \min_{\boldsymbol{\mu} \in \mathcal{M}} D_f(\boldsymbol{\pi}||\boldsymbol{\mu}) \leq R$$

and the covering number *for radius R is*

$$M_f(k, R) = \min\{m : \exists \textit{ covering } \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_m\}\}.$$

A set $\mathcal{M} = \{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_m : \boldsymbol{\pi}_i \in \triangle_{k-1}\}$ is a packing *of radius R if*

$$\min_{i,j \in [m]:\ i \neq j} D_f(\boldsymbol{\pi}_i\|\boldsymbol{\pi}_j) \geq 2R$$

and the packing number *for radius R is*

$$m_f(k, R) = \max\{m : \exists \textit{ packing } \{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_m\}\}.$$

We refer to the points in coverings and packings as *centers*. Note that if the divergence D is not symmetric (as is the case for KL divergence), then for $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ in a packing of radius R, we need both $D(\boldsymbol{\pi}_1\|\boldsymbol{\pi}_2) \geq 2R$ and $D(\boldsymbol{\pi}_2\|\boldsymbol{\pi}_1) \geq 2R$.

We let $P_X(\cdot)$ (or $Q_X(\cdot)$) to mean a probability distribution on random variable X. Let $P_{Y|\theta}(\cdot|\boldsymbol{\mu})$ be a distribution where random variable Y is sampled from $\boldsymbol{\mu}$. When $Y^n = y^n$, let

$$P^n_{Y|\theta}(y^n|\boldsymbol{\mu}) = \prod_{t=1}^{n} P_{Y|\theta}(y_t|\boldsymbol{\mu})$$

represent the n-fold product (each y_t is drawn independently). Since in our model Y^n is independent given θ, we also use $P_{Y^n|\theta}(y^n|\boldsymbol{\mu})$ to mean the same quantity as above.

II. Upper Bound

We use covering numbers to derive an upper bound for the capacity of the composite DNA channel. The method is very similar to that used in [32] to find an upper bound for the capacity of the permutation channel. KL divergence covering was also similarly used in [33] to find an upper bound on a minimax regret problem. In this work, we will use the following result on covering numbers:

Theorem 3 (Upper Bound on KL Divergence Covering [32], [34])**.** *For $0 < R \leq 1$,*

$$M_{\mathrm{KL}}(k, R) \leq c^{k-1} \left(\frac{k-1}{R}\right)^{\frac{k-1}{2}}$$

for some constant c.

In [32] it was shown that we can set $c = 7$ (which we will use). Theorem 3 is sufficient for showing our result. However, tighter bounds do exist (see [34]). Additionally, for the special case of $k = 2$ (corresponding to the case of the binomial channel), the proof in Theorem 3 reveals that the bound can be slightly improved to $M_{\mathrm{KL}}(2, R) \leq 6/\sqrt{R} + 1$.

Proof of Theorem 1. This proof primarily uses techniques from [35, Theorem 1]. We repeat some of the steps here for completeness. First, for any prior P_θ, we have

$$\begin{aligned}
&I(\theta; Y^n) \\
&\leq \int_{\boldsymbol{\pi}} P_\theta(\boldsymbol{\pi}) \sum_{y^n} P_{Y^n|\theta}(y^n|\boldsymbol{\pi}) \log \frac{P_{Y^n|\boldsymbol{\pi}}(y^n|\theta)}{\tilde{Q}_{Y^n}(y^n)} d(\boldsymbol{\pi}) \\
&\leq \max_{\boldsymbol{\pi} \in \triangle_{k-1}} D_{\mathrm{KL}}(P_{Y^n|\theta}(\cdot\,|\boldsymbol{\pi}) \| \tilde{Q}_{Y^n}(\cdot))\,.
\end{aligned}$$

The above holds for any $\tilde{Q}_{Y^n}$. We will chose a covering for a $k-1$-dimensional simplex for radius R (we specify the value of R later) which we denote as $\mathcal{M}_n$. We then choose

$$\tilde{Q}_{Y^n}(y^n) = \frac{1}{|\mathcal{M}_n|} \sum_{\boldsymbol{\mu} \in \mathcal{M}_n} P^n_{Y|\theta}(y^n|\boldsymbol{\mu})\,.$$

We can bound the mutual information above with

$$\begin{aligned}
I(\theta; Y^n) &\leq \max_{\boldsymbol{\pi} \in \triangle_{k-1}} D_{\mathrm{KL}}(P_{Y^n|\theta}(\cdot\,|\boldsymbol{\pi}) \| \tilde{Q}_{Y^n}(\cdot)) \\
&= \max_{\boldsymbol{\pi} \in \triangle_{k-1}} \sum_{y^n} P_{Y^n|\theta}(y^n|\boldsymbol{\pi}) \log \frac{P_{Y^n|\theta}(y^n|\boldsymbol{\pi})}{\frac{1}{|\mathcal{M}_n|} \sum_{\tilde{\boldsymbol{\mu}} \in \mathcal{M}_n} P^n_{Y|\theta}(y^n|\tilde{\boldsymbol{\mu}})} \\
&\leq \max_{\boldsymbol{\pi} \in \triangle_{k-1}} \min_{\boldsymbol{\mu} \in \mathcal{M}_n} \sum_{y^n} P_{Y^n|\theta}(y^n|\boldsymbol{\pi}) \log \frac{P_{Y^n|\theta}(y^n|\boldsymbol{\pi})}{\frac{1}{|\mathcal{M}_n|} P^n_{Y|\theta}(y^n|\boldsymbol{\mu})} \quad (3) \\
&= \log|\mathcal{M}_n| + \max_{\boldsymbol{\pi} \in \triangle_{k-1}} \min_{\boldsymbol{\mu} \in \mathcal{M}_n} D_{\mathrm{KL}}(P_{Y^n|\theta}(\cdot\,|\boldsymbol{\pi}) \| P^n_{Y|\theta}(\cdot\,|\boldsymbol{\mu})) \\
&= \log|\mathcal{M}_n| + \max_{\boldsymbol{\pi} \in \triangle_{k-1}} \min_{\boldsymbol{\mu} \in \mathcal{M}_n} n D_{\mathrm{KL}}(\boldsymbol{\pi} \| \boldsymbol{\mu}) \\
&= \log|\mathcal{M}_n| + nR\,.
\end{aligned}$$

where in (3) we use that $\sum_{\boldsymbol{\mu} \in \mathcal{M}_n} P^n_{Y|\theta}(y^n|\boldsymbol{\mu}) > P^n_{Y|\theta}(y^n|\boldsymbol{\mu})$ for any $\boldsymbol{\mu} \in \mathcal{M}_n$, including the $\boldsymbol{\mu}$ closest to $\boldsymbol{\pi}$ in KL divergence (though it is sufficient to use any $\boldsymbol{\mu}$ which covers $\boldsymbol{\pi}$).

To compute our result, we use Theorem 3 which gives $|\mathcal{M}_n| = M_{\mathrm{KL}}(k, R)$ and set $R = 1/n$.

$$\begin{aligned}
I(\boldsymbol{\pi}; Y^n) &\leq \log|\mathcal{M}_n| + nR \\
&\leq \frac{k-1}{2} \log \frac{1}{R} + nR + (k-1) \log c\sqrt{k-1} \\
&= \frac{k-1}{2} \log n + 1 + (k-1) \log c\sqrt{k-1}
\end{aligned}$$

and we can use $c = 7$ as stated below Theorem 3. This holds for any prior P_θ which gives the result. □

We will remark that using $k = 2$ with Theorem 1 as stated does not give a tighter bound than the upper bound computed for the binomial channel in [24]. However, if we used the slightly improved covering number bound of $M_{\mathrm{KL}}(2, R) \leq 6/\sqrt{R} + 1$ (see statement below Theorem 3), we can improve our upper bound to have a slight advantage over the upper bound of [24] when $n \geq 1067$.

III. Lower Bound

Our key inequality for finding the lower bound will be the following, which appears in [36], [37]. We use the version from [31]. For $\lambda \in [0, 1]$,

$$I(X; Y) \geq -\mathbb{E}_X \left[\log \mathbb{E}_{X'}[e^{-(1-\lambda) d_\lambda(X, X')} | X]\right] \quad (4)$$

where X' is an independent variable generated from the same distribution as X and

$$d_\lambda(x, x') = D_\lambda(P_{Y|X=x} \| P_{Y|X=x'})$$

is given as a Rényi divergence. To use this for our problem, we will substitute X with $\boldsymbol{\pi}$ and Y with Y^n.

Suppose we have a λ-Rényi packing of the $k-1$ probability simplex space for radius R. Call this $\mathcal{M} = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m\}$ where $m = |\mathcal{M}|$. We assume this packing is optimal in the sense that $m = m_\lambda(k, R)$. Recall that, based on the definition of a packing, it must be that for all $i \neq j$, unless $i = j$,

$$D_\lambda(\boldsymbol{q}_i \| \boldsymbol{q}_j) \geq 2R\,. \quad (5)$$

Next, following a similar technique as [31, page 625], using (4) and the tensorization of Rényi divergences, we have that

$$\begin{aligned}
I(\boldsymbol{\pi}; Y^n) &\geq -\sum_{i=1}^m \frac{1}{m} \log \left(\sum_{j=1}^m \frac{1}{m} e^{-n(1-\lambda) D_\lambda(\boldsymbol{q}_i \| \boldsymbol{q}_j)}\right) \\
&\geq -\sum_{i=1}^m \frac{1}{m} \log \left(\frac{m-1}{m} e^{-(1-\lambda) 2nR} + \frac{1}{m}\right) \\
&= -\log \left(\frac{m-1}{m} e^{-(1-\lambda) 2nR} + \frac{1}{m}\right) \\
&\geq -\log \left(e^{-(1-\lambda) 2nR} + \frac{1}{m}\right) \quad (6)
\end{aligned}$$

where in the second inequality, we used (5) (notice we have a factor of 2 because of our definition).

We need to compute m which is the λ-Rényi divergence packing number. We will not compute these packing numbers

directly, but instead show that they are bounded by packing numbers for total variation. We use the following to do this.

Lemma 1. *Let h be a non-decreasing function. For divergences D_f and D_g (which can be either f-divergences or Rényi divergences), if we have that*

$$h(D_f(\boldsymbol{p}\|\boldsymbol{q})) \le D_g(\boldsymbol{p}\|\boldsymbol{q})$$

then

$$m_f(k,R) \le m_g(k, h(2R)/2)$$

Proof. Suppose for divergence D_f, we have a packing $\mathcal{M}_f(k,R)$. This implies for all $\boldsymbol{p},\boldsymbol{q} \in \mathcal{M}_f(k,R)$ that

$$D_f(\boldsymbol{p}\|\boldsymbol{q}) \ge 2R$$

which gives that

$$D_g(\boldsymbol{p}\|\boldsymbol{q}) \ge h(D_f(\boldsymbol{p}\|\boldsymbol{q})) \ge h(2R) = 2h(2R)/2\,.$$

Hence, $\mathcal{M}_f(k,R)$ is also a packing for g with distance $h(2R)/2$. Since the packing number is the largest number of centers possible, the packing number for g with distance $h(2R)/2$ must be at least as big as for any packing in f with distance R. Therefore $m_f(k,R) \le m_g(k,h(2R)/2)\,.$ □

Next, we will use Lemma 1 to find a lower bound on the packing number for λ-Rényi divergences. We can readily do so by using the following relation between Rényi divergences and total variation given in [30].

Theorem 4 ([30], Pinsker's Inequality)**.** *For and $\lambda \in (0,1]$,*

$$2\lambda\,\mathrm{TV}^2(\boldsymbol{p},\boldsymbol{q}) \le D_\lambda(\boldsymbol{p}\|\boldsymbol{q})\,.$$

(Note that differences in constants from what is stated in [30] is due to the fact that our definition of total variation is slightly different).

We can now compute a lower bound on the λ-Rényi divergences packing number.

Lemma 2. *For $\lambda \in (0,1]$ and $R \le \lambda/4$,*

$$m_\lambda(k,R) \ge \left(\frac{\lambda}{4R}\right)^{(k-1)/2}.$$

Otherwise for $R > \lambda/4$, $m_{\mathrm{TV}}(k,R) \ge 1$.

Proof. Applying Lemma 1, we can let $h(\cdot) = 2\lambda(\cdot)^2$, which gives

$$m_{\mathrm{TV}}(k,R) \le m_\lambda(k, 2\lambda(2R)^2/2)$$
$$m_{\mathrm{TV}}\left(k,\sqrt{\frac{R}{4\lambda}}\right) \le m_\lambda(k,R)\,.$$

We then present the sequence of steps to get this bound and then discuss how each step was derived.

$$\begin{aligned} m_\lambda(k,R) &\ge m_{\mathrm{TV}}\left(k,\sqrt{\frac{R}{4\lambda}}\right) \\ &\ge M_{\mathrm{TV}}\left(k,2\sqrt{\frac{R}{4\lambda}}\right) \qquad (7)\\ &\ge \left(\frac{1}{4}\sqrt{\frac{4\lambda}{R}}\right)^{(k-1)} \qquad (8)\\ &= \left(\frac{\lambda}{4R}\right)^{(k-1)/2}. \end{aligned}$$

To get (7), we use that for packing and covering for a norm D_f, we have

$$m_f(k,R) \le M_f(k,R) \le m_f(k,R/2)\,.$$

(see [38, Theorem 14.1]). For (8), a lower bound on the covering number for TV can be computed using a volume argument (see [38, Theorem 14.2]). This gives that

$$\left(\frac{1}{2R}\right)^{(k-1)} \le M_{\mathrm{TV}}(k,R)$$

which is computed in [34, Section 2.4]. The last step simplifies the terms. The packing number should always be at least 1, so when $R > \lambda/4$, we simply use 1 as the lower bound. □

Proof of Theorem 2. Combining the packing result Lemma 2 with (6), we get that

$$\begin{aligned} I(\boldsymbol{\pi};Y^n) &\ge -\log\left(e^{-(1-\lambda)2nR} + \frac{1}{m}\right) \\ &\ge -\log\left(e^{-(1-\lambda)2nR} + e^{-\frac{k-1}{2}\log\frac{\lambda}{4R}}\right). \end{aligned}$$

We can choose

$$\lambda = 1/2 \text{ and } R = \frac{k-1}{2}\frac{\log n}{n}\,.$$

(The choice for R is not optimal, but is something chosen which is not too difficult or complex to work with.)

This gives that

$$\begin{aligned} &I(\boldsymbol{\pi};Y^n) \\ &\ge -\log\left(e^{-\frac{k-1}{2}\log n} + e^{-\frac{k-1}{2}\log\frac{n}{4(k-1)\log n}}\right) \\ &\ge -\log\left(2e^{-\frac{k-1}{2}\log\frac{n}{4(k-1)\log n}}\right) \\ &= \frac{k-1}{2}\left(\log n - \log(k-1) - \log\log n - \log 4\right) - \log 2\,. \end{aligned}$$

□

Acknowledgments

We would like to thank the Dagstuhl Seminar on Coding Theory and Algorithms for Emerging Technologies in Synthetic Biology (organized by Rawad Bitar, Olgica Milenkovic, Zohar Yakhini, and Yonatan Yehezkeally) for introducing us to this problem. We thank the many participants of the seminar and give a special thanks to Eitan Yaakobi for connecting us to relevant resources.

REFERENCES

[1] C. Komninakis, L. Vandenberghe, and R.D. Wesel, "Capacity of the binomial channel, or minimax redundancy for memoryless sources," in *Proceedings. 2001 IEEE International Symposium on Information Theory (IEEE Cat. No.01CH37252)*, 2001, pp. 127–.

[2] Adir Kobovich, Eitan Yaakobi, and Nir Weinberger, "M-dab: An input-distribution optimization algorithm for composite dna storage by the multinomial channel," .

[3] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[4] George M. Church, Yuan Gao, and Sriram Kosuri, "Next-generation digital information storage in dna," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[5] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos, and Ewan Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized dna," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[6] Meinolf Blawat, Klaus Gaedke, Ingo Hütter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, and George M. Church, "Forward error correction for dna data storage," *Procedia Computer Science*, vol. 80, pp. 1011–1022, 2016, International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

[7] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss, "Random access in large-scale dna data storage," *Nature Biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[8] Yaniv Erlich and Dina Zielinski, "DNA fountain enables a robust and efficient storage architecture," *bioRxiv*, 2016.

[9] Reinhard Heckel, Ilan Shomorony, Kannan Ramchandran, and David N. C. Tse, "Fundamental limits of DNA storage systems," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 3130–3134.

[10] Omer Sabary, Han Mao Kiah, Paul H. Siegel, and Eitan Yaakobi, "Survey for a decade of coding for dna storage," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 253–271, 2024.

[11] Ilan Shomorony and Reinhard Heckel, "Capacity results for the noisy shuffling channel," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 762–766.

[12] Leon Anavy, Inbal Vaknin, Orna Atar, Roee Amit, and Zohar Yakhini, "Data storage in dna with fewer synthesis cycles using composite dna letters," *Nature biotechnology*, vol. 37, no. 10, pp. 1229–1236, 2019.

[13] Yeongjae Choi, Taehoon Ryu, Amos C. Lee, Hansol Choi, Hansaem Lee, Jaejun Park, Suk-Heung Song, Seojoo Kim, Hyeli Kim, Wook Park, and Sunghoon Kwon, "High information capacity dna-based data storage with augmented encoding characters using degenerate bases," *Scientific Reports*, vol. 9, no. 1, pp. 6582, 2019.

[14] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[15] B.S. Clarke and A.R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[16] Bertrand S. Clarke and Andrew R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[17] Q. Xie and A. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 646–657, 1997.

[18] Qun Xie and A.R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[19] Y. Shtarkov, T. Tjalkens, and F. Willems, "Multialphabet universal coding of memoryless sources," *Problems of Information Transmission*, vol. 31, pp. 114–127, 1995.

[20] A. Orlitsky and N.P. Santhanam, "Speaking of infinity [i.i.d. strings]," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2215–2230, 2004.

[21] Yuval Gerzon, Ilan Shomorony, and Nir Weinberger, "Capacity of frequency-based channels: Encoding information in molecular concentrations," in *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 681–686.

[22] Nariman Farsad, Will Chuang, Andrea Goldsmith, Christos Komninakis, Muriel Médard, Christopher Rose, Lieven Vandenberghe, Emily E. Wesel, and Richard D. Wesel, "Capacities and optimal input distributions for particle-intensity channels," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 6, no. 3, pp. 220–232, 2020.

[23] Richard D. Wesel, Emily E. Wesel, Lieven Vandenberghe, Christos Komninakis, and Muriel Medard, "Efficient binomial channel capacity computation with an application to molecular communication," in *2018 Information Theory and Applications Workshop (ITA)*, 2018, pp. 1–5.

[24] Luca Barletta, Ian Zieder, Antonino Favano, and Alex Dytso, "Binomial channel: On the capacity-achieving distribution and bounds on the capacity," in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 711–716.

[25] Frederik Walter, Omer Sabary, Antonia Wachter-Zeh, and Eitan Yaakobi, "Coding for composite dna to correct substitutions, strand losses, and deletions," in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 97–102.

[26] Yiqing Yan, Nimesh Pinnamaneni, Sachin Chalapati, Conor Crosbie, and Raja Appuswamy, "Scaling logical density of dna storage with enzymatically-ligated composite motifs," *Scientific Reports*, vol. 13, no. 1, pp. 15978, 2023.

[27] Inbal Preuss, Ben Galili, Zohar Yakhini, and Leon Anavy, "Sequencing coverage analysis for combinatorial dna-based storage systems," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 297–316, 2024.

[28] Omer Sabary, Inbal Preuss, Ryan Gabrys, Zohar Yakhini, Leon Anavy, and Eitan Yaakobi, "Error-correcting codes for combinatorial composite dna," 2024.

[29] Yury Polyanskiy and Yihong Wu, "Lecture notes on information theory," class notes for MIT 6.441, 2013-2016.

[30] Tim Van Erven and Peter Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[31] Yury Polyanskiy and Yihong Wu, *Information Theory: From Coding to Learning*, Cambridge University Press, 2025.

[32] Jennifer Tang and Yury Polyanskiy, "Capacity of noisy permutation channels," *IEEE Transactions on Information Theory*, vol. 69, no. 7, pp. 4145–4162, 2023.

[33] Jennifer Tang, "Minimax regret on patterns using kullback-leibler divergence covering," in *Annual Conference on Learning Theory (COLT)*, 2022, vol. 178, pp. 1–18.

[34] Jennifer Tang, *Divergence Covering*, Ph.D. thesis, Massachusetts Institute of Technology, 2022.

[35] Yuhong Yang and Andrew Barron, "Information-theoretic determination of minimax rates of convergence," *The Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[36] David Haussler and Manfred Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.

[37] Artemy Kolchinsky and Brendan D. Tracey, "Estimating mixture entropy with pairwise distances," *Entropy*, vol. 19, no. 7, 2017.

[38] Yihong Wu, "Lecture notes on: Information-theoretic methods for high-dimensional statistics," class notes for ECE 598, 016.