Capacity of Noisy Permutation Channels

Jennifer Tang
EECS (MIT)
Cambridge, MA, USA
jstang@mit.edu

Yury Polyanskiy
EECS (MIT)
Cambridge, MA, USA
yp@mit.edu

***Abstract*—We establish the capacity of a class of communication channels introduced in [2]. The n-letter input from a finite alphabet is passed through a discrete memoryless channel $P_{Z|X}$ and then the output n-letter sequence is uniformly permuted. We show that the maximal communication rate (normalized by $\log n$) equals $\frac{1}{2}(\mathsf{rank}(P_{Z|X}) - 1)$ whenever $P_{Z|X}$ is strictly positive. This is done by establishing a converse bound matching the achievability of [2]. The two main ingredients of our proof are (1) a sharp bound on the entropy of a uniformly sampled vector from a type class and observed through a DMC; and (2) the covering ε-net of a probability simplex with Kullback-Leibler divergence as a metric. In addition to strictly positive DMC we also find the noisy permutation capacity for q-ary erasure channels, the Z-channel and others.**

I. Problem Statement and Main Results

The noisy permutation channel, as formally introduced in [2], is a communication model in which an n-letter input undergoes a concatenation of a discrete memoryless channel (DMC) and a uniform permutation of the n letters. Since the receiver observes a uniformly permuted output, the order of symbols conveys no information. See Section I-C for a motivation of this model. More formally, the channel $P_{Y^n|X^n}$ can be described by the following Markov chain:

$$X^n \to Z^n \to Y^n \,. \tag{1}$$

Here the channel input $X^n = (X_1, \dots X_n)$ is a length n sequence where each position takes a value in $\mathcal{X} = [q]$ (where $[q] = \{1, 2, \dots, q\}$). The sequence X^n goes through the DMC which operates independently and identically on each symbol. This results in a sequence Z^n where each position takes a value in $\mathcal{Y} = [k]$. The DMC transition probabilities can be represented as a $q \times k$ matrix $P_{Z|X}$. Then, the sequence Z^n goes through the permutation part of the channel and results in Y^n which is a uniformly random permutation of symbols on Z^n.

Let f_n and g_n be the channel encoder and decoder respectively. For each message $W \in [M]$, the input to the channel is $X^n = f_n(W)$. The output is Y^n, which the decoder decodes as $\hat{W} = g_n(Y^n)$. The probability of error is given by $P_{\mathsf{error}}^{(n)} \triangleq \mathbb{P}[W \neq \hat{W}]$. The rate[1] for the encoder-decoder pair (f_n, g_n) is defined as

$$R \triangleq \frac{\log M}{\log n} \,. \tag{2}$$

A rate R is *achievable* if there is a sequence of encoder-decoder pairs (f_n, g_n) with rate R such that $\lim_{n\to\infty} P_{\mathsf{error}}^{(n)} = 0$. The capacity for the noisy permutation channel with DMC $P_{Z|X}$ is $C_{\mathsf{perm}}(P_{Z|X}) \triangleq \sup\{R \geq 0 : R \text{ is achievable}\}$.

In [2], the author determined that the noisy permutation channel capacity[2] for DMC $P_{Z|X}$ is bounded by

$$C_{\mathsf{perm}}(P_{Z|X}) \geq \frac{\mathsf{rank}(P_{Z|X}) - 1}{2} \,. \tag{3}$$

For strictly positive matrices $P_{Z|X}$ (meaning all the transition probabilities are greater than 0), the author shows two converse bounds: $C_{\mathsf{perm}}(P_{Z|X}) \leq (|\mathcal{Y}| - 1)/2$ and $C_{\mathsf{perm}}(P_{Z|X}) \leq (\mathsf{ext}(P_{Z|X}) - 1)/2$, where $\mathsf{ext}(P)$ is the number of extreme points of the convex hull of the rows of P. For the case of strictly positive DMC $P_{Z|X}$, these upper and lower bounds do not necessarily match if the rank of matrix $P_{Z|X}$ does not equal to $|\mathcal{Y}|$ or $\mathsf{ext}(P_{Z|X})$.

A. Main Results

Our main result is establishing tightness of the lower bound (3), resolving Conjecture 1 of [2].

Theorem 1 (Strictly Positive DMC)**.** *For strictly positive $P_{Z|X}$,*

$$C_{\mathit{perm}}(P_{Z|X}) = \frac{\mathit{rank}(P_{Z|X}) - 1}{2} \,. \tag{4}$$

Our proof uses the idea of covering the space of distributions via an ε-net under the Kullback-Leibler (KL) divergence distance, following upon our investigations of a similar question in [3]. In order to reduce to the covering question, we first need another result that is, perhaps, of separate interest as well.

This work was supported in part by the NSF grant CCF-2131115 and was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Complete proofs are omitted due to space constraints, see [1] for details.

[1]Notice that rate R for the noisy permutation channel is not the typical definition of information theory rate where $R = \frac{\log M}{n}$. The noisy permutation channel would have rate 0 under this typical definition.

[2]While it might seem that the noisy permutation channel capacity should be a continuous function of the values in $P_{Z|X}$, note that this is not the case due to how capacity is defined. Changing values in $P_{Z|X}$ by a small δ could change the rank of $P_{Z|X}$ by 1, but no matter how small δ is, there exists an n large enough so its effects can make a difference.

978-1-6654-2159-1/22/$31.00 ©2022 IEEE

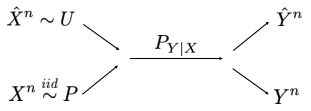

Fig. 1. This diagram illustrates the special case of Theorem 2 where $Q_Y = P_Y$. It shows how $\hat{Y}^n$ relates to Y^n.

We let $\mathcal{P}_n$ be the set of n-types[3] (probabilities which can be written with denominator n). For $P \in \mathcal{P}_n$, let $T_n(P)$ be the set of sequences of length n in the type class of P. The notation Q_Y means a distribution on random variable Y. We will use Q_Y^n to mean the product distribution $Q_Y^n(y^n) = \prod_{t=1}^n Q_Y(y_t)$. For any distribution U on length n sequences, the distribution $P_{Y|X}^n \circ U$ can be understood as the distribution on random sequences derived by first randomly selecting a sequence according to U, then passing each symbol in this sequence through the transition probabilities $P_{Y|X}$ independently. (See Section I-E for more discussion.)

Theorem 2. *Fix channel $P_{Y|X}$ which is strictly positive. Then there exists a constant $c = c(P_{Y|X})$ such that the following holds: For any n-type P, let U be uniform on $T_n(P)$. For all Q_Y we have*

$$nD(P_Y\|Q_Y) \le D(P_{Y|X}^n \circ U\|Q_Y^n) \le nD(P_Y\|Q_Y) + c \tag{5}$$

where P_Y is the marginal distribution of Y under $(P \times P_{Y|X})$.

Remark 1. *It can be shown that the constant c in Theorem 2 is*

$$c \le \frac{q-1}{2}\log\frac{2\pi\alpha^2}{c_*} + \frac{q}{12n} \le \frac{q-1}{2}\log\frac{2\pi\alpha^2}{c_*} + \frac{q}{12}\,. \tag{6}$$

where α is a universal constant (see Section III) and if p_{bj} denote the values in matrix $P_{Y|X}$,

$$c_* = \min_b \frac{\min_j p_{bj}}{\max_j p_{bj}}\,. \tag{7}$$

Theorem 2 deals with the following scenario: Select some $P \in \mathcal{P}_n$ and suppose we have two sequences, X^n and $\hat{X}^n$. The sequence X^n is generated iid using the probability P; whereas $\hat{X}^n$ has uniform probability over all sequences in the type $T_n(P)$. Both sequences X^n and $\hat{X}^n$ undergo the transition $P_{Y|X}$ applied independently on each symbol and respectively results in Y^n and $\hat{Y}^n$. How different are the distributions of Y^n and $\hat{Y}^n$ under KL divergence? See Figure 1 for a diagram. Another interpretation of this scenario is if there are n balls of q colors in an urn. The sequence X^n are n draws from the urn with replacement and $\hat{X}^n$ are n draws without replacement. These observations both go through the same noisy process to produce Y^n and $\hat{Y}^n$.

It turns out that if $P_{Y|X}$ is strictly positive, then regardless of the sequence length n, $D(P_{\hat{Y}^n}\|P_{Y^n}) \le c$ where c is a constant that only depends on $P_{Y|X}$. Theorem 2 actually shows something more general. The sequence X^n can be generated iid with another distribution Q, and the KL divergence can still be bounded by constant c plus another term which is the KL divergence of the marginals on Y generated by P and Q. In other words, the divergence of (a complicated distribution) $P_{\hat{Y}^n}$ to any iid distribution Q_Y^n can be approximated with $nD(P_Y\|Q_Y)$ and this approximation will only be off by an additive constant. We note also that the constant c in Theorem 2 is sharp (cannot be improved to $o(1)$). This is discussed in [1].

Remark 2. *Note that $D(P_{\hat{X}^m}\|P_X^m)$ describes the difference between sampling m balls from an n-urn with and without replacement. This is a classical question studied in [5]. Our setting studies this question for the particular case when $n = m$ and when the observations are noisy. Bounds for the noiseless case $D(P_{\hat{X}^m}\|P_X^m)$ can still be an upper bound for the noisy case if we apply the data processing inequality. This shows that $D(P_{Y|X}^n \circ U\|P_Y^n) \le D(P_{\hat{X}^n}\|P_X^n) \le \frac{k-1}{2}(\log n + c)$, where the second inequality is shown using Stirling's approximation. Our result removes the $\log n$ term in this bound, but only under the assumption of a strictly positive $P_{Y|X}$. We also note that results of [5] as shown in [6] imply the finitary case of de Finetti's theorem.*

We use similar techniques to get converse results in other settings which do not have strictly positive DMC matrices. These are below and discussed in [1].

Theorem 3. *Other channel results:*

1) *Suppose $P_{Z|X}$ can be written as a block diagonal matrix with β blocks where each block is strictly positive. Then,*
$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{\mathsf{rank}(P_{Z|X}) + \beta - 2}{2}\,. \tag{8}$$
2) *For DMC $P_{Z|X}$ which is a q-ary erasure channel for $q \ge 2$ (assuming non-zero transition probabilities), then*
$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{q-1}{2}\,. \tag{9}$$
3) *For DMC $P_{Z|X}$ which is a Z-channel (assuming non-zero probabilities on the edges), then*
$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{1}{2}\,. \tag{10}$$

B. Paper Organization

We continue this section with the motivation, a high level summary of our covering method, and the notation. In Section II, we discuss how covering is used to determine the capacity of the noisy permutation channel along with some basics in covering. We give the proofs (or proof sketches) of Theorem 2 and Theorem 1 in Section III.

C. Motivation

The motivation for studying the permutation channel is that it captures a setting where codewords get reordered. This occurs in applications such as communication networks and biological storage systems. More details on these applications and other relevant work can be found in [2].

[3]See Section 11.1 of [4] for background on types.

a) Communication Networks: Suppose we have a point-to-point communication network where the information is transmitted through a multipath routed network. Different packets are transmitted through different routes in the network, and each route has its own amount of latency, causing packets traveling on different routes to arrive at different times. The order in which the sender transmits packets is no longer preserved at the receiver end. Such a scenario is studied in [7] where the authors are primarily concerned with reducing delay in their channel. Unlike our work, they do not consider noisy symbols. Another line of work on packet-switched networks deals with the permutation channel along with errors such as insertions, deletions, and substitutions of symbols [8], [9]. Their work primarily focuses on building minimum distance codes and perfect codes for the permutation channel.

b) DNA Storage Systems: DNA-based storage systems are an attractive option for data storage due to its ability to withstand time and encode a very high-density of information [10], [11]. The state-of-the-art technology for storing information on DNA uses nucleotides with relatively small lengths (few hundreds) [12]. Each of these DNA molecules are stored in a pool without any regard to order. The different molecule types can be treated as symbols in the setting of the permutation channel. Noise in this channel models any error that can occur in the process. DNA storage is also the motivation for studying the permutation channel in [13], [14].

As typical in information theory, a question of fundamental interest is to determine the capacity of channels. We determine the capacity of the noisy permutation channel in the strictly positive case, settling the problem introduced in [2]. This setting differs from some of the models studied above. In [13], the authors find asymptotic bounds on rate, but for a fixed number of errors rather than probabilitistic errors. The work in [12] finds the capacity when the symbols are sampled randomly then read, something relevant to DNA models, but not to general permutation channels. The results in [14] are specifically for when the permuted objects are a string of symbols and the noisy process is applied to symbols on a string; the set of strings are permuted but symbols in each string are not.

D. Covering Numbers and Rate

All of our results use the method of covering. A *covering* is a set of points in a space (we will call them centers) for which all other points in the space are within a certain distance ε to (see Definition 1). Using covering as a technique to determine the capacity for the noisy permutation channel is reasonable because the centers which are far apart can intuitively be equated with messages that are distinguishable. When the messages correspond to two distributions Q_1 and Q_2 which are far in KL divergence, it is unlikely that noisy versions of Q_1 will be close to noisy versions of Q_2. If two distributions are close in KL divergence, their noisy versions are likely to be confused. If the messages in our communication are centers of a covering, then we know that if we add another center (or message), it will be close to one of the existing covering centers and thus cause error in determining which of the centers (or messages) was sent. This gives us a limit on the total number of messages which can be sent, creating a converse bound[4].

In order to use this intuition mathematically, we need to overcome the obstacle of computing the KL divergence over the noisy output distributions of the messages. This is difficult to do because these output distributions are not iid. This is where Theorem 2 is useful, as it allows us to use KL divergences over iid distributions in place of the KL divergence over this output distribution (since we can replace a hypergeometric distribution which undergoes noise with a multinomial distribution). Other obstacles include determining the covering number under KL divergence (see Section II).

E. Notation

The set of all probability distributions on q symbols is defined as the probability simplex Δ_{q-1}. The $q \times k$ DMC matrix $P_{Z|X}$ has values in each row which sums up to 1 (i.e, the matrix is stochastic). Symbol $b \in \mathcal{X}$ has probability $P_{Z|X}(j|b)$ (also written as p_{bj}) of becoming symbol $j \in \mathcal{Y}$. We say that the DMC matrix is *strictly positive* if $p_{bj} > 0$ for all b and j in the matrix. For example, we can write the DMC matrix for the binary symmetric channel (BSC) with crossover probability δ as

$$P_{Z|X} = \begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix}. \tag{11}$$

If $0 < \delta < 1$, then this DMC matrix is strictly positive.

Though different from how we described it in the introduction, it is convenient to describe the Markov chain of the noisy permutation channel as $\pi \to X^n \to Z^n \to Y^n$. Each $\pi = (\pi_1, .., \pi_q) \in \Delta_{q-1}$ corresponds to a possible channel input. For each n, we will restrict π to be in $\mathcal{P}_n$. The value of π_b represents the proportion of positions in sequence X^n which have symbol b.

Note that it is entirely equivalent to perform the permutation on the sequence X^n first and then apply the DMC. In this case, we no longer need the random variable Z^n. Because of this, we will also use $P_{Y|X}$ to specify the transition matrix, where $P_{Y|X}$ and $P_{Z|X}$ are the same and interchangeable.

We will specify a way to parameterize the distributions on Y. We use the notation $Q_{Y|\mu}$ for $\mu = (\mu_1, ..., \mu_k) \in \Delta_{k-1}$ to mean a distribution on symbols $\mathcal{Y}$ where the probability of symbol $j \in \mathcal{Y}$ is $Q_{Y|\mu}(j) = \mu_j$. The distribution $Q^n_{Y|\mu}$ is the multinomial distribution with parameters μ and number of independent trials n. These distributions do not (directly) relate the permutation channel; we define them since they are important for our analysis.

On the other hand, the distribution $P_{Y^n|\pi}$ refers the the distribution on sequences Y^n when $\pi \in \mathcal{P}_n$ is the input to the noisy permutation channel on n letters. Note that in general $P_{Y^n|\pi}$ is *not* a multinomial distribution. As seen in Theorem 2, $P_{Y^n|\pi} = P^n_{Y|X} \circ U$ where U is a uniform

[4]A similar notion to covering is packing. Intuitively, covering corresponds to a converse bound while packing corresponds to an achievability bound.

distribution on $T_n(\pi)$. Both represent the distribution on the output of the noisy permutation channel. Permuting the input symbols gives a sequence in the support of U, and then each permuted symbol goes through the transition probabilities $P_{Y|X}$ independently.

When it is clear what π is, we use P_Y to mean the marginal distribution for each Y_t in the sequence $Y^n \sim P_{Y|X}^n \circ U$. This distribution does not depend on the index t since U is uniform on all permutations.

II. Covering Converse

Our core method for finding our new results is to use divergence covering of the probability simplex. First, we give some basic definitions and results (which are proved in [1]).

Definition 1 (Divergence Covering Number)**.**

$$M(k,\varepsilon) = \inf\{m : \exists\{Q_1,...,Q_m\} \quad s.t \max_{P\in\Delta_{k-1}} \min_{Q_i} D(P||Q_i) \le \varepsilon\}. \tag{12}$$

Let $M(k,\varepsilon,\mathcal{B})$ be defined like $M(k,\varepsilon)$ except that $P \in \mathcal{B}$ for a subspace $\mathcal{B} \subset \Delta_{k-1}$.

Theorem 4 (Upper Bound on Divergence Covering)**.** *For $0 < \varepsilon \le 1$,*

$$M(k,\varepsilon) \le c^{k-1}\left(\frac{k-1}{\varepsilon}\right)^{\frac{k-1}{2}} \tag{13}$$

for some constant c.

While the above result is sufficient for showing our theorems, stronger bounds do exist (see [15], which also discusses Definition 1 in detail). The next proposition is proved in [1].

Proposition 1. *For $\mathcal{B} \subset \Delta_{k-1}$, suppose there is a stochastic matrix F which maps Δ_{q-1} onto $\mathcal{B}$. Suppose that $\mathcal{B}$ is a space of dimension $\ell - 1$ (or likewise, F has rank ℓ). Then,*

$$M(k,\varepsilon,\mathcal{B}) \le \binom{q}{\ell} M(\ell,\varepsilon). \tag{14}$$

Our main proof, which combines the previous results, uses covering ideas similar to [16, Theorem 1] in order to upper bound the mutual information $I(\pi;Y^n)$. In summary, we need to find a set of covering centers which are close in KL divergence to all the possible distributions on Y^n that can occur as outputs of the noisy permutation channel. Our set of centers need not be possible distributions over Y^n generated by the channel. We will opt for using multinomial distributions as our set of covering centers.

Let $\mathcal{N}_n$ be a discrete set in Δ_{k-1} which we will specify (later) for each n (these will be the covering centers). Mutual information has the property that

$$I(\pi;Y^n) \le \max_{\pi} D(P_{Y^n|\pi}\|\tilde{Q}_{Y^n}). \tag{15}$$

This equation holds for any $\tilde{Q}_{Y^n}$, thus we can choose

$$\tilde{Q}_{Y^n}(y^n) = \frac{1}{|\mathcal{N}_n|}\sum_{\mu\in\mathcal{N}_n} Q_{Y|\mu}^n(y^n) = \frac{1}{|\mathcal{N}_n|}\sum_{\mu\in\mathcal{N}_n}\prod_{t=1}^n Q_{Y|\mu}(y_t). \tag{16}$$

The following proposition is the main work-horse of all our converse results.

Proposition 2 (Covering for Noisy Permutation Channels)**.** *Suppose that for the noisy permutation channel with DMC $P_{Y|X}$, we have that for any $\pi \in \mathcal{P}_n$,*

$$D(P_{Y|X}^n \circ U\|Q_Y^n) \le nD(P_Y\|Q_Y) + f(n) \tag{17}$$

where U is uniform on the type $T_n(\pi)$, P_Y is the marginal distribution of $P_{Y|X}^n \circ U$ and f is only a function of n and $P_{Y|X}$. Then

$$C_{perm}(P_{Y|X}) \le \frac{\mathsf{rank}(P_{Y|X})-1}{2} + \lim_{n\to\infty}\frac{f(n)}{\log n}. \tag{18}$$

In Proposition 2, when the DMC is strictly positive, the $f(n)$ term is constant in n (which is shown via Theorem 2 and gives the proof for Theorem 1). However, when the DMC is not strictly positive, $f(n)$ is not necessarily constant in n. Non-constant values of $f(n)$ are used in deriving some of the results in Theorem 3.

Proof. Following techniques used in the proof of Theorem 1 from [16], we can upper bound the mutual information given in (15) by

$$I(\pi;Y^n) \le \log|\mathcal{N}_n| + \max_{\pi\in\mathcal{P}_n}\min_{\bar{\mu}\in\mathcal{N}_n} D(P_{Y^n|\pi}\|Q_{Y|\bar{\mu}}^n). \tag{19}$$

To specify $\mathcal{N}_n$, first define $\mathcal{L}(P_{Y|X}) = \bigcup_{\pi\in\Delta_{k-1}}\mu^M(\pi)$ where $\mu^M(\pi) \triangleq (\sum_i \pi_i p_{i1},...,\sum_i \pi_i p_{ik})$ for any $\pi \in \Delta_{q-1}$. This is the space of all possible marginals P_Y.

Let $\mathcal{N}_n$ be a covering of $\mathcal{L}(P_{Y|X})$ under KL divergence with covering radius $1/n$. In other words, $\mathcal{N}_n = \{\bar{\mu}^{(1)},...,\bar{\mu}^{(m)}\}$ so that $\max_{\mu\in\mathcal{L}(P_{Y|X})}\min_{\bar{\mu}\in\mathcal{N}_n} D(Q_{Y|\mu}\|Q_{Y|\bar{\mu}}) \le 1/n$.

Let ℓ be the dimension of $\mathcal{L}(P_{Z|X})$. Using Proposition 1, $|\mathcal{N}_n| \le C(q,\ell)\left(\frac{\ell}{1/n}\right)^{\frac{\ell}{2}}$ where $C(q,\ell)$ depends on q and ℓ but not n.

Using assumption (17) and putting this into (19), gives

$$I(\pi;Y^n) \le \log|\mathcal{N}_n| + \max_{\pi\in\mathcal{P}_n}\min_{\bar{\mu}\in\mathcal{N}_n} D(P_{Y^n|\pi}\|Q_{Y|\bar{\mu}}^n) \tag{20}$$

$$\le \log\left(C(q,\ell)\left(\frac{\ell}{1/n}\right)^{\frac{\ell}{2}}\right) + f(n) + \max_{\pi\in\mathcal{P}_n}\min_{\bar{\mu}\in\mathcal{N}_n} nD(P_Y\|Q_{Y|\bar{\mu}}) \tag{21}$$

$$\le \frac{\ell}{2}\log n + c' + f(n) \tag{22}$$

where c' is a constant which does not depend on n.

For the noisy permutation channel, recall that the rate is defined as (2). Since asymptotically $\log M \le I(\pi,Y^n) \le \frac{\ell}{2}\log n + c' + f(n)$, we have

$$R \le \frac{\ell}{2} + \frac{c'}{\log n} + \frac{f(n)}{\log n} \to \frac{\ell}{2} + \lim_{n\to\infty}\frac{f(n)}{\log n}. \tag{23}$$

Since $\ell = \mathsf{rank}(P_{Z|X}) - 1$, we have an upper bound for the capacity of the noisy permutation channel. □

III. Divergence Under Fixed Types

For computing our converse bounds, we need to determine the expression (17) for strictly positive DMC matrices. This is where we need Theorem 2. The following results are also needed and both are proved in [1].

Proposition 3. *Let U be uniform on the type $T_n(P)$ and $(X,Y)^n$ be iid from $(P \times P_{Y|X})$. Let P_Y be the marginal distribution of Y under $(P \times P_{Y|X})$. Then for all Q_Y,*

$$D(P_{Y|X}^n \circ U \| Q_Y^n) = nD(P_Y \| Q_Y) + \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[A=1|Y^n=y^n]}{\mathbb{P}[A=1]} \quad (24)$$

where $A = \mathbb{I}\{X^n \in T_n(P)\}$ (and we use $\mathbb{P}$ to mean under the probability where $(X,Y)^n$ is iid from $(P \times P_{Y|X})$).

Lemma 1. *Let $P = (p_1, ..., p_q) \in \mathcal{P}_n$ and let $A = \mathbb{I}\{X^n \in T_n(P)\}$. Then if $(X,Y)^n$ is drawn iid from $(P \times P_{Y|X})$,*

$$\log \frac{1}{\mathbb{P}[A=1]} \le -\frac{1}{2}\log n + \sum_{i:p_i>0} \frac{1}{2}\log p_i n + \frac{q-1}{2}\log 2\pi + \frac{1}{12n}. \quad (25)$$

Let $S_n = \sum_{i=1}^n W_i$ where the W_i are independent. Each W_i is distributed Bernoulli with probability p_i. Using concentration results from [17] (details in [1]), we can show that for any integer z (and universal constant α)

$$\mathbb{P}[S_n = z] \le \frac{\alpha}{\sqrt{\sum_{i=1}^n \min\{p_i, 1-p_i\}}}. \quad (26)$$

We will use this in the next lemma which is key to computing the second term in (24).

Lemma 2. *There are n balls thrown in q bins independently, so that for the i-th ball, the probability of landing in bin b is $p_{i,b}$. Let N_b be the ball count of the b-th bin. Then if $\pi_b > 0$ for all b and $\sum_b \pi_b = 1$, we have*

$$\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \le \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{B}} \quad (27)$$

where $\pi_{max} = \max_b \pi_b$,

$$B = c_*^{q-1}\frac{\prod_b \pi_b}{\pi_{max}}, \qquad c_* = \min_i \frac{c_-(i)}{c_+(i)}, \quad (28)$$

$$c_-(i) = \min_b \frac{p_{i,b}}{\pi_b}, \qquad c_+(i) = \max_b \frac{p_{i,b}}{\pi_b}, \quad (29)$$

and α is the universal constant used in [17].

Proof. For notation, let $W_{i,b}$ be the indicator variable of whether ball i was thrown into bin b. We can express $N_b = \sum_{i=1}^n W_{i,b}$. Arrange the indices so that $\pi_1 \le \pi_2 \cdots \le \pi_q$.

First observe that

$$\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \quad (30)$$

$$= \prod_{b=1}^q \mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}]. \quad (31)$$

For $b = q$, $\mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}] = 1$.

For $b < q$, we can compute for any i that

$$\min\left\{\frac{p_{i,b}}{\sum_{a=b}^q p_{i,a}}, 1 - \frac{p_{i,b}}{\sum_{a=b}^q p_{i,a}}\right\} \quad (32)$$

$$\ge \frac{\min_a \frac{p_{i,a}}{\pi_a}}{\max_a \frac{p_{i,a}}{\pi_a}} \min\left\{\frac{\pi_b}{\sum_{a=b}^q \pi_a}, \frac{\sum_{a>b}^q \pi_a}{\sum_{a=b}^q \pi_a}\right\} \quad (33)$$

$$\ge \min_i \frac{c_-(i)}{c_+(i)} \frac{1}{\sum_{a=b}^q \pi_a} \min\left\{\pi_b, \sum_{a>b}^q \pi_a\right\} \quad (34)$$

$$= c_* \frac{\pi_b}{\sum_{a=b}^q \pi_a}. \quad (35)$$

We get the last equality because we have arranged π_b in increasing order. Hence by (26)

$$\mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}] \quad (36)$$

$$\le \frac{\alpha}{\sqrt{\left(n - \sum_{a=1}^{b-1} n\pi_a\right) c_* \frac{\pi_b}{\sum_{a=b}^q \pi_a}}} \quad (37)$$

$$= \frac{\alpha}{n^{1/2}\sqrt{c_*\pi_b}} \quad (38)$$

where we used that $n - \sum_{a=1}^{b-1} n\pi_a = n\sum_{a=b}^q \pi_a$ to get the last inequality. Taking a product of all terms in (30), gives

$$\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \le \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{B}}. \quad (39)$$

□

Proof Sketch of Theorem 2. To show the lower bound, using Proposition 3, we need only to show the last term in (24) is positive, which we can do by expressing it as a KL divergence.

For the upper bound, we express the last term of (24) as a difference of two terms $\sum_{y^n \in \mathcal{Y}^n} \mathbb{P}[Y^n = y^n | A = 1] \log \mathbb{P}[A = 1 | Y^n = y^n] - \log \mathbb{P}[A = 1]$ and use Lemma 2 for the first term and Lemma 1 for the second term. To see why Lemma 2 applies to the first term, note that the first term is trying to calculate given some Y^n, what the probability that the type of X^n is equal to $T_n(P)$. This under a distribution where $(X,Y)^n \sim (P_{Y|X} \times P)$.

We will express the type $T_n(P)$ with $P = (\pi_1, ..., \pi_q)$ where $P \in \mathcal{P}_n$. This implies that $\pi_b = \mathbb{P}[X = b]$. Let the balls described in Lemma 2 be each of the elements of Y^n. If $Y_i = y_i$, then let $p_{i,b} = \mathbb{P}[X_i = b | Y_i = y_i] = \mathbb{P}[X = b | Y = y_i]$ (because the symbols are iid). This way $p_{i,b}$ is appropriately the probability that the ith symbol lands in bin b. As in Lemma 2, N_b is the number of balls in bin b. Then the probability that $X^n \in T_n(P)$ is equivalent to $\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q]$.

The remaining details we give in [1]. □

Proof of Theorem 1. Using Theorem 2 with Proposition 2 completes the proof for strictly positive DMC. □

Acknowledgement

We would like to thank A. Makur for inspiring this project and for his discussions with us.

REFERENCES

[1] Jennifer Tang and Yury Polyanskiy, "Capacity of noisy permutation channels," *arXiv preprint arXiv:2111.00559*, 2021.

[2] Anuran Makur, "Coding theorems for noisy permutation channels," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 672–6748, Nov 2020.

[3] Aviv Adler, Jennifer Tang, and Yury Polyanskiy, "Quantization of random distributions under kl divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 2762–2767.

[4] T.M. Cover and J.A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, USA, 2006.

[5] A. J. Stam, "Distance between sampling with and without replacement," *Statistica Neerlandica*, vol. 32, pp. 81–91, 1978.

[6] P. Diaconis and D. Freedman, "Finite exchangeable sequences," *The Annals of Probability*, vol. 8, no. 4, pp. 745–764, 1980.

[7] John MacLaren Walsh, Steven Weber, and Ciira wa Maina, "Optimal rate delay tradeoffs for multipath routed and network coded networks," in *2008 IEEE International Symposium on Information Theory*, 2008, pp. 682–686.

[8] Mladen Kovacevic and Dejan Vukobratovic, "Subset codes for packet networks," *IEEE Communications Letters*, vol. 17, no. 4, pp. 729–732, Apr 2013.

[9] Mladen Kovacevic and Dejan Vukobratovic, "Perfect codes in the discrete simplex," *Designs, Codes and Cryptography*, vol. 75, no. 1, pp. 81–95, Nov 2013.

[10] S. M. Hossein Tabatabaei Yazdi, Han Mao Kiah, Eva Garcia-Ruiz, Jian Ma, Huimin Zhao, and Olgica Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.

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

[12] 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.

[13] Mladen Kovacevic and Vincent Y. F. Tan, "Codes in the space of multisets-coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, Jul 2018.

[14] 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.

[15] Jennifer Tang, *Divergence Covering*, Ph.D. thesis, Massachusetts Institute of Technology, 2021.

[16] 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.

[17] V.V. Petrov, *Sums of Independent Random Variables*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 2. Folge. Springer Berlin Heidelberg, 2012.