Capacity of Noisy Permutation Channels

Jennifer Tang, *Member, IEEE*, and Yury Polyanskiy, *Senior Member, IEEE*

***Abstract*— We establish the capacity of a class of communication channels introduced by Makur. 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 Makur. The two main ingredients of our proof are: 1) a sharp bound on the Kullback-Leibler divergence of a uniformly sampled vector from a type class and observed through a DMC to an iid vector; 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.**

***Index Terms*— Permutation channel, channel capacity, ϵ-net covering.**

I. Problem Statement and Main Results

The noisy permutation channel, as formally introduced by Makur in [1], 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-B for a motivation of this model. More formally, the channel $P_{Y^n|X^n}$ can be described by the following Markov chain:

$$X^n \to Z^n \to Y^n.$$

Here the channel input 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 $P_{Z|X}$. After the DMC, the sequence Z^n goes through the permutation part of the channel and results in sequence 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 $\hat{W}=g_n(Y^n)$. (See Figure 1 for a diagram depicting the channel.) 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

$$R\triangleq\frac{\log M}{\log n}. \tag{1}$$

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

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

For strictly positive matrices $P_{Z|X}$ (meaning all the transition probabilities are greater than 0), Makur shows a converse bound

$$C_{\mathsf{perm}}(P_{Z|X})\leq\frac{|\mathcal{Y}|-1}{2}.$$

Makur also gives a second converse bound: $C_{\mathsf{perm}}(P_{Z|X})\leq(\mathsf{ext}(P_{Z|X})-1)/2$, 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 (2), resolving Conjecture 1 of [1].

Manuscript received 30 October 2021; revised 27 September 2022; accepted 5 February 2023. Date of publication 22 February 2023; date of current version 16 June 2023. This work was supported in part by the NSF grant CCF-2131115 and 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. An earlier version of this paper was presented in part at the 2022 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT50566.2022.9834509]. *(Corresponding author: Jennifer Tang.)*

The authors are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jstang@mit.edu; yp@mit.edu).

Communicated by A. Somekh-Baruch, Associate Editor for Shannon Theory and Information Measures.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3247812.

Digital Object Identifier 10.1109/TIT.2023.3247812

[1] Notice that rate R for the noisy permutation channel is not the commonly used definition where $R=\frac{\log M}{n}$. The noisy permutation channel would have rate 0 under this commonly used definition. Defining rate as in (1) is appropriate given that we intend to find capacity. Let $M^*(n,\epsilon)=\max\{M:\exists(n,M,\epsilon)-\text{code}\}$ where n is the length (or channel uses), M is the message size, and ϵ is the error probability for a given code (see [2] for more details). Capacity is defined as the limit as $\epsilon\to 0^+$ of the coefficient of the leading term of $\log M^*(n,\epsilon)$. In the case of the noisy permutation channel, the leading term of $\log M^*(n,\epsilon)$ scales as $\log n$.

[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 the effects of δ can make a difference.

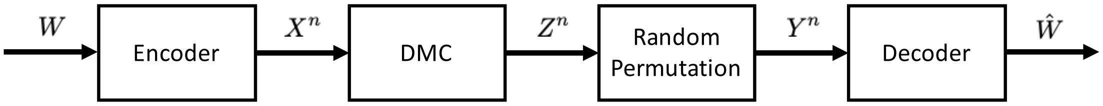

Fig. 1. Diagram of the noisy permutation channel communication system. The key components are a DMC followed by a uniformly random permutation.

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

$$C_{\mathsf{perm}}(P_{Z|X}) = \frac{\mathsf{rank}(P_{Z|X})-1}{2}.$$

Our proof uses the idea of covering the space of distributions via an ε-net under the Kullback-Leibler (KL) divergence as a "distance",[3] 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.

Let k be the alphabet size. We let $\mathcal{P}_n$ be the set of n-types (probabilities which can be written as rational numbers with denominator n), i.e.

$$\mathcal{P}_n = \left\{ P \in \Delta_{k-1} : P = \left(\frac{a_1}{n}, \dots, \frac{a_k}{n}\right) \text{ where } a_1, \dots, a_k \in \mathbb{Z}_{\ge 0} \right\}.$$

(We use Δ_{k-1} for the $k-1$ dimensional probability simplex, see Section I-C.) For $P \in \mathcal{P}_n$, let $T_n(P)$ be the set of sequences of length n in the type class[4] of P, i.e.

$$T_n(P) = \left\{ s_1 \dots s_n : s_t \in [k] \text{ and } \left(\frac{\sum_{t=1}^n \mathbb{I}\{s_t = 1\}}{n}, \dots, \frac{\sum_{t=1}^n \mathbb{I}\{s_t = k\}}{n} \right) = P \right\}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. The notation Q_Y means a distribution on random variable Y. For any distribution Q_Y, we 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-C for more discussion.)

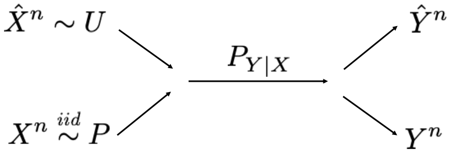

Fig. 2. This diagram illustrates the special case of Theorem 2 where $Q_Y = P_Y$. Variable X^n is distributed iid according to P whereas $\hat{X}^n$ is a sequence uniformly drawn from type class $T_n(P)$ (a distribution represented by U). Variables Y^n and $\hat{Y}^n$ are noisy versions of X^n and $\hat{X}^n$ respectively.

Our next result 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. On the other hand, $\hat{X}^n$ 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 2 for a diagram representing the relations of these variables.) 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 (in which case all the balls are drawn). These observations then 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 $P_{Y|X}$. Our next result will actually show 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.

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

$$nD(P_Y \| Q_Y) \le D(P_{Y|X}^n \circ U \| Q_Y^n) \le nD(P_Y \| Q_Y) + c \quad (3)$$

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

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

where α is a universal constant defined in Theorem 5 (see Section III-C) and if p_{bj} denote the values in matrix $P_{Y|X}$,

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

It is necessary in Theorem 2 that $P_{Y|X}$ is strictly positive. In fact, it is surprising that Theorem 2 can show that the KL divergence of $D(P_{Y|X}^n \circ U \| Q_Y^n)$ when $Q_Y = P_Y$ is constant, considering that this is not the behavior we would expect for transitions $P_{Y|X}$ which are not strictly positive. For example, using our simpler initial example with $X, \hat{X}, Y, \hat{Y}$ (which is depicted by Figure 2), consider when $X, Y \in [2] = \{1, 2\}$ and

$$P_{Y|X} = \begin{cases} 1-\delta & \text{if } X = Y \\ \delta & \text{if } X \neq Y \end{cases} \quad (4)$$

[3] KL divergence is not technically a distance or metric (as it is not symmetric and does not follow triangle inequality), but we choose to use the term *distance* since we are using KL divergence to measure how far two probability distributions are.

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

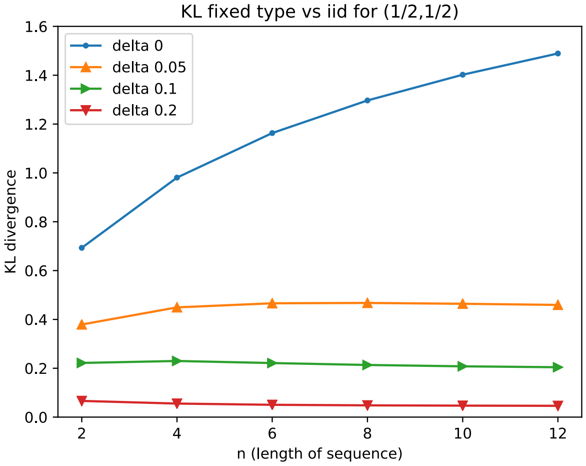

Fig. 3. This plot demonstrates the consequences of Theorem 2. We numerically compute $D(P_{\hat{Y}^n}\|P_{Y^n})$ when $P=(1/2,1/2)$ and $P_{Y|X}$ is given as in (4) for different values of δ. When $\delta=0$, we see that the KL divergence is trending towards infinity (specifically it grows as $\frac{1}{2}\log n$). When any noise is added (δ is positive), this growth to infinity completely disappears. The quantity $D(P_{\hat{Y}^n}\|P_{Y^n})$ becomes constant in n, as Theorem 2 states.

This transition probability represents a binary symmetric channel (BSC) with crossover probability δ. Suppose that $X\in T_n((1/2,1/2))$. If $\delta=0$ (and thus $P_{Y|X}$ is not strictly positive), then we can compute that

$$D(P_{\hat{Y}^n}\|P_{Y^n})\approx\frac{1}{2}\log n. \tag{5}$$

However, if we increase δ slightly (adding any amount of crossover noise), this completely eliminates the growth of $D(P_{\hat{Y}^n}\|P_{Y^n})$ in n. For any positive δ, $D(P_{\hat{Y}^n}\|P_{Y^n})$ is constant as n increases, where the value of the constant depends on δ. An illustration of this example is given in Figure 3.

We note also that Theorem 2 implies that 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.

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 $P_{Y|X}$. See Section A for more details on comparing our bound to that of [5] when $m<n$. We also note that results of [5] as shown in [6] imply the finitary case of de Finetti's theorem.

Other contributions of this work use similar techniques to get converse results in other settings which do not have strictly positive DMC matrices.

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,

$$C_{\mathsf{perm}}(P_{Z|X})=\frac{\mathsf{rank}(P_{Z|X})+\beta-2}{2}. \tag{6}$$

2) For DMC $P_{Z|X}$ which is a q-ary erasure channel for $q\ge 2$ (assuming erasure probabilities are not 0 or 1), then

$$C_{\mathsf{perm}}(P_{Z|X})=\frac{q-1}{2}.$$

3) For DMC $P_{Z|X}$ which is a Z-channel, then

$$C_{\mathsf{perm}}(P_{Z|X})=\frac{1}{2}.$$

The first result in Theorem 3 applies to DMC $P_{Z|X}$ which are block diagonal matrices where each block is strictly positive. As (6) of Theorem 3 implies, we are able to show both the achievability and converse results for block diagonal DMC matrices. We prove both these results in Section C. This result also immediately illustrates that Theorem 1 without the strictly positive condition cannot be true, since block diagonal matrices with 2 or more strictly positive blocks violate the bound in Theorem 1 entirely.

The second result is for binary erasure channels and q-ary erasure channels. The work in [1] determines the capacity for when the DMC matrix is the binary symmetric channel (BSC), but leaves the binary and q-ary erasure channels as open problems. Item 2 of Theorem 3 resolves Conjecture 2 presented in [1] regarding the capacity of binary erasure channels and the conjecture regarding q-ary erasure channels.[5] This result uses (2) which is proved in [1] as the achievability. Our contribution is the tight converse argument.

The third result in Theorem 3 deals with DMC which is the Z-channel. While this is tight, if we generalize to a q-ary Z-channel (or what we call in this work a "zigzag" channel), we are not always able to find tight results with our current covering technique. The erasure channels, Z-channels, and a brief analysis on the zigzag channel are discussed in Section D.

All of these results also use the method of covering. A *covering* is a set of points in a space (we 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

[5] Note that while binary erasure channels and q-ary erasure channels usually have the same erasure probability for each symbol, Item 2 of Theorem 3 is still true even if these erasure probabilities are different. The only requirement is that none of the erasure probabilities are 0 or 1. The capacity when the erasure probabilities are 0 or 1 is discussed in [1].

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.[6]

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 the more complicated 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-B).

a) Paper Organization: We continue this section with the motivation 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 prove Theorem 2 and Theorem 1 in Section III. We prove all the parts of Theorem 3 in the appendix.

B. 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. We briefly describe some of these applications. More details on these applications and other relevant work can be found in [1].

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 which involves the permutation channel is on packet-switched networks. The errors explored in this work include 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, whether it is in synthesizing the DNA molecules or in reading the molecules. 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 [1]. This setting differs from some of the models studied in the works described in the motivations, as it looks at the problem from a purely information theoretic standpoint and does not include assumptions which might be specific to the application.

Among the works relevant to the motivations described, those that have some information theoretical flavors include [13], which deals with 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.

Our results for when the DMC is the erasure channel are particularly interesting to DNA storage applications since the erased symbol can model deletion errors. Permutation channels with deletions are central to the work in [13], where the authors base their code constructions on Sidon sets and determine bounds on optimal codes for a given number of errors. In our work, our errors are not fixed but probabilistic, and hence we find the probabilistic analogue of their bounds.

Also on the topic of deletions, one of the motivations for studying the noisy permutation channel in [1] is the relation between permutation channels with erasures and the random deletion channel [15], [16]. In [16], the author demonstrated a decoding scheme for the random deletion channel based on low density parity check (LDPC) codes. Their scheme can tolerate a reordering of the symbols, allowing it be a viable scheme for the permutation channel with erasures. However, their scheme requires the alphabet size to grow with the blocklength.

C. Notation

The set of all probability distributions on q symbols is defined as the probability simplex

$$\Delta_{q-1} \triangleq \left\{ (\pi_1, \ldots, \pi_q) : \sum_{i=1}^{q} \pi_i = 1, 0 \leq \pi_i \leq 1 \right\}.$$

For a $q \times k$ DMC matrix $P_{Z|X}$, we can express the individual transitions as

$$P_{Z|X} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{q1} & p_{q2} & \cdots & p_{qk} \end{bmatrix}.$$

The values in each row of the matrix sums up to 1 (i.e, the matrix is stochastic). Symbol $b \in \mathcal{X} = [q]$ has probability p_{bj} of becoming symbol $j \in \mathcal{Y} = [k]$. We can also write this probability as $P_{Z|X}(j|b)$. We say that the DMC matrix (or a

[6] A similar notion to covering is packing, which is a set of centers in a space where all the centers in the set are at least distance 2ε from another. Intuitively, covering corresponds to a converse bound while packing corresponds to an achievabaility bound.

submatrix) is *strictly positive* if $p_{bj} > 0$ for all b and j in the matrix (or submatrix).

For example, the DMC matrix for the BSC with crossover probability δ, given in (4), is written as

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

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

Because of the uniform permutation step, the order of the symbols in X^n does not matter. Thus, it is natural to consider the inputs to the channel as types classes rather than sequences. In light of this, we can describe the Markov chain of the noisy permutation channel as

$$\pi \to X^n \to Z^n \to Y^n \tag{7}$$

where each $\pi = (\pi_1, .., \pi_q) \in \Delta_{q-1} \cap \mathcal{P}_n$ describes a type class. For communication, the sender has the freedom to encode messages into any π, and given π, the sequence X^n can be any sequence in $T_n(\pi)$. The value of π_b represents the proportion of positions in sequence X^n which have symbol b.

On the decoding end, the only relevant statistic the receiver would use from Y^n is which type class Y^n belongs to. 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 also use $P_{Y|X}$ to specify the transition matrix, where $P_{Y|X}$ and $P_{Z|X}$ are the same and interchangeable.

Next, we specify a way to parameterize the distributions on Y. We use the notation $Q_{Y|\mu}$ for $\mu = (\mu_1, \dots, \mu_k) \in \Delta_{k-1}$ 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 distribution on sequences Y^n when $\pi \in \mathcal{P}_n$ is the input to the noisy permutation channel on n letters as described in (7). 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 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 $Y^n \sim P^n_{Y|X} \circ U$. This distribution does not depend on the index t since U is uniform on all permutations.

Throughout this work, we use $\log$ to mean the natural logarithm.

II. Covering Converse

Our core method for finding our new results is to use KL divergence covering of the probability simplex. We first show how covering can be applied to the noisy permutation channel and then give the necessary covering results.

A. Covering Basics

The main concept of our proof uses covering ideas similar to [17, 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 *Kullback-Leibler (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 a possible distribution over Y^n generated by the channel. We choose to use multinomial distributions as our set of covering centers.

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

$$I(\pi; Y^n) \le \max_{\pi} D(P_{Y^n|\pi} \| \tilde{Q}_{Y^n}). \tag{8}$$

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

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

The following proposition is the main tool of all our converse results.

Proposition 1 (Covering for Noisy Permutation Channels): Suppose that for the noisy permutation channel with DMC $P_{Y|X}$, we have that for any $\pi \in \mathcal{P}_n$,

$$D(P^n_{Y|X} \circ U \| Q^n_Y) \le nD(P_Y \| Q_Y) + f(n) \tag{9}$$

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

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

In Proposition 1, 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.

For the proof, we need to define for any $\pi \in \Delta_{q-1}$

$$\mu^M(\pi) \triangleq \left(\sum_i \pi_i p_{i1}, \dots, \sum_i \pi_i p_{ik} \right).$$

The vector $\mu^M(\pi)$ is the mean (we use 'M' as short for mean) of the distribution $P_{Y^n|\pi} = P^n_{Y|X} \circ U$. Note that $P_Y(j) = \sum_i \pi_i p_{ij}$. Also if $\mu = \mu^M(\pi)$, we can write $P_Y = Q_{Y|\mu}$.

Proof: Following techniques used in the proof of [17, Theorem 1], we can upper bound the mutual information given in (8) by

$$I(\pi;Y^n) \le \log|\mathcal{N}_n| + \max_{\pi\in\mathcal{P}_n}\min_{\bar\mu\in\mathcal{N}_n} D(P_{Y^n|\pi}\|Q^n_{Y|\bar\mu}). \tag{10}$$

To specify $\mathcal{N}_n$, first define

$$\mathcal{L}(P_{Y|X}) = \bigcup_{\pi\in\Delta_{k-1}} \mu^M(\pi).$$

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, $\mathcal{N}_n = \{\bar\mu^{(1)},\dots,\bar\mu^{(m)}\}$ so that

$$\max_{\mu\in\mathcal{L}(P_{Y|X})}\min_{\bar\mu\in\mathcal{N}_n} D(Q_{Y|\mu}\|Q_{Y|\bar\mu}) \le \frac{1}{n}.$$

Let ℓ be the dimension of $\mathcal{L}(P_{Z|X})$. Using divergence covering results and the result specifically about covering an ℓ-dimensional subspace (see next part Section II-B),

$$|\mathcal{N}_n| \le C(q,\ell)\left(\frac{\ell}{1/n}\right)^{\frac{\ell}{2}} \tag{11}$$

where $C(q,\ell)$ depends on q and ℓ but not on n.

Starting with (10) and substituting in assumption (9) gives

$$\begin{aligned} I(\pi;Y^n) &\le \log\left(C(q,\ell)\left(\frac{\ell}{1/n}\right)^{\frac{\ell}{2}}\right) \\ &\qquad + f(n) + \max_{\pi\in\mathcal{P}_n}\min_{\bar\mu\in\mathcal{N}_n} nD(P_Y\|Q_{Y|\bar\mu}) \\ &\le \frac{\ell}{2}\log n + \log C(q,\ell) + \frac{\ell}{2}\log\ell + f(n) + n\frac{1}{n} \\ &\le \frac{\ell}{2}\log n + c' + f(n) \end{aligned}$$

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

For the noisy permutation channel, recall that the rate is defined as (1). Since asymptotically $\log M \le I(\pi,Y^n) \le \frac{\ell}{2}\log n + c' + f(n)$, we have

$$R \le \frac{\ell}{2} + \frac{c'}{\log n} + \frac{f(n)}{\log n} \to \frac{\ell}{2} + \lim_{n\to\infty}\frac{f(n)}{\log n}. \tag{12}$$

It remains to compute ℓ. Let $r = \mathsf{rank}(P_{Z|X})$. When the domain is any vector in $\mathbb{R}^q$, the image space of this (left) vector multiplied by $P_{Z|X}$ is r dimensional. But since we are restricting the domain and image to probability vectors, this adds an additional constraint to the image space and reduces the dimension by 1, giving that $\ell = \mathsf{rank}(P_{Z|X}) - 1$. Substituting this into (12) gives an upper bound for the capacity of the noisy permutation channel. □

B. Covering Definition and Results

In order to show (11), we have the following definition and results. A *KL divergence covering* is a set of centers in Δ_{k-1} so that every point in Δ_{k-1} is within some KL distance of one of the centers. Let ε be this distance. Since KL divergence is not symmetric, we specify that KL distance is computed where the covering center is placed in the second argument of the KL divergence. This is made explicit in the following definition of a *covering number*.

Definition 1 (Divergence Covering Number):

$$\begin{aligned} M(k,\varepsilon) = \inf\{m : \exists\{Q_1,\dots,Q_m\} \\ \text{s.t} \max_{P\in\Delta_{k-1}}\min_{Q_i} D(P\|Q_i) \le \varepsilon\}. \end{aligned}$$

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

We need upper bounds on the KL divergence covering number in order to get converse results for the permutation channel. One such upper bound is the following:

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

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

for some constant c.

The above result is sufficient for showing our theorems. However, tighter bounds do exist (see [18]). In addition to an upper bound on the KL divergence covering, we also need an additional result which allows us to use covering numbers over the whole simplex to get covering numbers over certain subsets of the simplex.

Proposition 2: 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,

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

The proofs are in Section B. More discussion on KL divergence covering can be found in [18].

III. Divergence Under Fixed Types

For computing our converse bounds, we need to determine the expression (9) for our DMC matrices. This is where we need Theorem 2 which gives the divergence between noisy observations of a fixed type compared to an iid distribution.

We prove Theorem 2 by first showing some relevant intermediate results. The techniques in these intermediate results are also useful for when $P_{Z|X}$ is not strictly positive and therefore relevant for showing some of the results in Theorem 3. Before doing so, we briefly discuss the constant of Theorem 2 and how it is tight.

A. Constant of Theorem 2

Here we show that the constant c in Theorem 2 is sharp (cannot be improved to $o(1)$). One tool we need is the following theorem by Marton [19]:

Lemma 1 (Marton's Transportation Inequality): Let $X^n \sim \prod_{t=1}^n P_{X_t}$ and $\hat X^n \sim P_{\hat X^n}$. Then there exists a joint probability measure $P_{X^n,\hat X^n}$ with these given marginals such that

$$\begin{aligned} \frac{1}{n}\mathbb{E}[d(X^n,\hat X^n)] &= \frac{1}{n}\sum_{t=1}^n \mathbb{P}[X_t \ne \hat X_t] \\ &\le \left(\frac{1}{n}D\left(P_{\hat X^n}\middle\|\prod_{t=1}^n P_{X_t}\right)\right)^{1/2} \end{aligned}$$

where $d(X,Y)$ is the Hamming distance.

Suppose that we are only working with 2 symbols, $\{1,2\}$, for the space of X and Y. Using the notation in Theorem 2, let $Y^n \sim P^n_{Y|X} \circ U$ and $\hat{Y}^n \sim Q^n_Y$, where we set $Q_Y = P_Y$. Choose $P = (1/2, 1/2)$ and $P_{Y|X}$ to be that of a BSC with crossover probability δ. This gives that distribution P_Y is uniform on the two symbols.

We choose $\delta = 1/n$, which is non-zero but small enough so that in expectation, X^n and Y^n, as well as $\hat{X}^n$ and $\hat{Y}^n$, will differ by 1. Define function $\#1(\cdot)$ to be mean the number of 1's in a sequence. Then

$$\begin{aligned}\mathbb{E}|\#1(Y^n) - n/2| &= \mathbb{E}|\#1(Y^n) - \#1(X^n)|\\ &\le \mathbb{E}[d(Y^n, X^n)]\\ &= 1\\ \mathbb{E}|\#1(\hat{Y}^n) - \#1(\hat{X}^n)| &\le \mathbb{E}[d(\hat{Y}^n, \hat{X}^n)]\\ &= 1\end{aligned}$$

The number of 1's in $\hat{X}^n$ will differ from its mean by roughly the standard deviation. To be precise, applying a result from [20], we have that

$$\mathbb{E}|\#1(\hat{X}^n) - n/2| \ge \frac{1}{2\sqrt{2}}\sqrt{n}.$$

Using the above and multiple applications of the triangle inequality, we can compute that no matter the coupling chosen

$$\begin{aligned}\mathbb{E}[d(Y^n, \hat{Y}^n)] &\ge \mathbb{E}|\#1(\hat{Y}^n) - \#1(Y^n)|\\ &\ge \mathbb{E}|\#1(\hat{Y}^n) - n/2| - \mathbb{E}|\#1(Y^n) - n/2|\\ &\ge \mathbb{E}|\#1(\hat{X}^n) - n/2| - \mathbb{E}|\#1(\hat{Y}^n) - \#1(\hat{X}^n)|\\ &\quad - \mathbb{E}|\#1(Y^n) - n/2|\\ &= \frac{1}{2\sqrt{2}}\sqrt{n} - 2\end{aligned}$$

We can assume that n is large, so that the Hamming distance can be more simply lower bounded by

$$\mathbb{E}[d(Y^n, \hat{Y}^n)] \ge \frac{1}{4}\sqrt{n}.$$

Combining this Hamming distance with Lemma 1 gets a lower bound

$$\begin{aligned}\frac{1}{4\sqrt{n}} &\le \frac{1}{n}\mathbb{E}[d(Y^n, \hat{Y}^n)]\\ &\le \left(\frac{1}{n} D(P^n_{Y|X} \circ U \| Q^n_Y)\right)^{1/2}\\ &\le \left(\frac{1}{n}(nD(P_Y \| P_Y) + c)\right)^{1/2}\\ &= \frac{\sqrt{c}}{\sqrt{n}}.\end{aligned}$$

Improving c to be $o(1)$ in n would violate this lower bound.

Intuitively, consider what happens when we let $P_{Y|X}$ be the identity matrix where $Y = X$. In such a case, Theorem 2 is not true (in order to get a true statement, the constant c should be replaced with a value that grows logarithmically with n, see Section III-B.1). This is the same setting as the example given above but with $\delta = 0$. It is clear here that $\hat{Y}^n$ likely has $\sqrt{n}$ deviations in the number of 1's from the mean whereas any sequence Y^n has exactly $n/2$ number of 1's. This creates an expected Hamming distance of $\sqrt{n}$. Slightly increasing δ above zero will not change the Hamming distance by much but will make $P_{Z|X}$ strictly positive.

B. Expression for Divergence Under Fixed Types

In order to show Theorem 2, we need some intermediary results about how to work with the quantity $D(P^n_{Y|X} \circ U \| Q^n_Y)$. The next proposition does this and can be used for any $P_{Y|X}$, not just those which are strictly positive.

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,

$$\begin{aligned}D(P^n_{Y|X} \circ U \| Q^n_Y) &= nD(P_Y \| Q_Y)\\ &+ \sum_{y^n \in \mathcal{Y}^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[A = 1 | Y^n = y^n]}{\mathbb{P}[A = 1]} \qquad (13)\end{aligned}$$

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

The second term on the right-hand side of (13) can be written as an expected value:

$$\begin{aligned}&\sum_{y^n \in \mathcal{Y}^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[A = 1 | Y^n = y^n]}{\mathbb{P}[A = 1]}\\ &= \mathbb{E}_{(X,Y)^n \sim (P \times P_{Y|X})}\\ &\quad \left[\log \frac{\mathbb{P}_{(\tilde{X},\tilde{Y})^n \sim (P \times P_{Y|X})}[\tilde{X}^n \in T(P) | \tilde{Y}^n = Y^n]}{\mathbb{P}_{(\tilde{X},\tilde{Y})^n \sim (P \times P_{Y|X})}[\tilde{X}^n \in T(P)]}\right.\\ &\qquad \left.\middle| (X, Y)^n \text{ where } X^n \in T(P)\right]\end{aligned}$$

For ease of notation, we choose to express the term above as

$$\mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n]}{\mathbb{P}[\tilde{A} = 1]} \middle| A = 1\right] \qquad (14)$$

where the $\tilde{\cdot}$ notation emphasizes that the variables are associated with an independent copy $(\tilde{X}, \tilde{Y})^n$ drawn from the same distribution $(P \times P_{Y|X})$ as $(X, Y)^n$ and where $\tilde{A} = \mathbb{I}\{\tilde{X}^n \in T_n(P)\}$.

Proof:

Note that $(P^n_{Y|X} \circ U)(y^n) = \mathbb{P}[Y^n = y^n | A = 1]$.

$$\begin{aligned}D(P^n_{Y|X} \circ U \| Q^n_Y) &= \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1]\\ &\quad \log \frac{\mathbb{P}[Y^n = y^n | A = 1]}{Q^n_Y(y^n)}\\ &= \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1]\\ &\quad \log \frac{\mathbb{P}[A = 1 | Y^n = y^n]\mathbb{P}[Y^n = y^n]}{\mathbb{P}[A = 1] Q^n_Y(y^n)}\\ &= \mathbb{E}\left[\log \frac{P^n_Y(Y^n)}{Q^n_Y(Y^n)} \middle| A = 1\right]\\ &\quad + \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[A = 1 | Y^n = y^n]}{\mathbb{P}[A = 1]}\\ &= \mathbb{E}\left[\log \frac{P^n_Y(Y^n)}{Q^n_Y(Y^n)} \middle| A = 1\right]\\ &\quad + \mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n]}{\mathbb{P}[\tilde{A} = 1]} \middle| A = 1\right].\end{aligned}$$

The marginal distribution $P_Y(a)$ is also the probability that any position t in sequence Y^n takes the value a, i.e. $P_Y(a) = \mathbb{P}[Y_t = a|A = 1]$. This occurs since U is uniform on all permutations of type $T_n(P)$. We get for the first term in the sum,

$$\begin{aligned}&\mathbb{E}\left[\log \frac{P_Y^n(Y^n)}{Q_Y^n(Y^n)}\middle|A=1\right]\\ &= \sum_{y^n} \mathbb{P}[Y^n = y^n|A=1] \log \frac{P_Y^n(y^n)}{Q_Y^n(y^n)}\\ &= \sum_{y^n} \mathbb{P}[Y^n = y^n|A=1] \sum_a n\frac{|\{t : y_t = a\}|}{n} \log \frac{P_Y(a)}{Q_Y(a)}\\ &= n\sum_a P_Y(a) \log \frac{P_Y(a)}{Q_Y(a)}\\ &= nD(P_Y \| Q_Y).\end{aligned}$$

This gives the result (13). □

We can separate (14) into two additive terms due to the logarithm. The next lemma can be used to compute one of these terms.

Lemma 2: Let $P = (p_1, \ldots, p_q) \in \mathcal{P}_n$ and let $A = \mathbb{I}\{X^n \in T_n(P)\}$. If $(X, Y)^n$ is drawn iid from $(P \times P_{Y|X})$, then

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

For this proof, we use a Stirling approximation type bound from [21]: For positive integers n,

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \le n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \quad (15)$$

Proof: We assume that all $p_i > 0$ since we can always reduce P to a shorter vector and decrease q.

The probability that a specific type occurs is given by the multinomial distribution.

$$\begin{aligned}&-\log \mathbb{P}[A=1]\\ &= -\log\left(\frac{n!}{\prod_{i=1}^q (p_i n)!}\prod_{i=1}^q p_i^{p_i n}\right)\\ &= -\log\left(\frac{n!}{n^n}\right) - \log\left(\prod_{i=1}^q \frac{(p_i n)^{p_i n}}{(p_i n)!}\right)\\ &\le n - \frac{1}{2}\log n - \frac{1}{2}\log 2\pi\\ &\quad + \sum_{i=1}^q \left(-p_i n + \frac{1}{2}\log p_i n + \frac{1}{2}\log 2\pi + \frac{q}{12n}\right).\end{aligned}$$

We used (15) in the last inequality (we can do this since each $p_i n$ is an integer greater than 0). Combining terms gives the result. □

1) Theorem 2 Without the Strictly Positive Requirement: To illustrate how to use Proposition 3 and Lemma 2, we compute an upper bound with the same form as the upper bound given in (3) of Theorem 2 for all $P_{Y|X}$. We briefly mentioned above that if we remove the strictly positive requirement for $P_{Y|X}$ in Theorem 2, in the worst case, the constant c would need to be replaced with a logarithmic term. To be exact, if $P_{Y|X}$ is not strictly positive, c needs to be replaced with a poly-logarithmic term in n. Using Proposition 3 and Lemma 2, we can get an upper bound on c with

$$\begin{aligned}c &= \mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A}=1|\tilde{Y}^n = Y^n]}{\mathbb{P}[\tilde{A}=1]}\middle|A=1\right]\\ &\le \frac{q-1}{2}\log n + c' + \mathbb{E}\left[\log \mathbb{P}[\tilde{A}=1|\tilde{Y}^n = Y^n]\middle|A=1\right]\\ &\le \frac{q-1}{2}\log n + c'.\end{aligned}$$

We used the fact that the largest value $\mathbb{P}[\tilde{A}=1|\tilde{Y}^n = Y^n]$ can take is 1 since it is a probability. The inequality is tight when $P_{Y|X}$ is the identity matrix (when $Y = X$). One example when $Y = X$ is the BSC with no noise ($\delta = 0$) example we stated earlier, where (5) holds.

C. Concentration of Sums of Independent Variables

To show Theorem 2, we need to compute (14). We need to determine the probability of $\tilde{A} = 1$, which is the event that $\tilde{X}^n$ has a particular type, under certain conditions. Showing that $\tilde{X}^n$ has a particular type can be equated to the problem of randomly throwing balls into some set of bins and looking at the number of balls which fall into each bin. To help us bound the probability a certain number of balls falls into a particular bin, we make use of the following:

The concentration function $Q(Z;\lambda)$ of random variable Z is defined by

$$Q(Z;\lambda) = \sup_z \mathbb{P}[z \le Z \le z + \lambda]$$

for every $\lambda \ge 0$ [22]. Let $S_n = \sum_{i=1}^n W_i$ where W_i are independent random variables.

Theorem 5 (Petrov [22]): Let the numbers a_i and b_i be such that

$$\mathbb{P}\left[W_i - a_i \le -\frac{\lambda_i}{2}\right] \ge b_i$$
$$\mathbb{P}\left[W_i - a_i \ge \frac{\lambda_i}{2}\right] \ge b_i$$

for $i = 1, \ldots, n$. Then there exists a universal constant α so that

$$Q(S_n;\lambda) \le \alpha\lambda\left(\sum_{i=1}^n \lambda_i^2 b_i\right)^{-1/2}$$

for every positive $\lambda_1, \ldots, \lambda_n$ none of which exceeds λ.

To apply Theorem 5 to our problem, each W_i is a Bernoulli random variable where the probability that $W_i = 1$ is p_i. Let $b_i = \min\{p_i, 1 - p_i\}$. We can fix $a_i = 1/2$. We can also fix $\lambda_i = 1/2$ and $\lambda = 1/2$, though this exact value does not matter so long as $\lambda < 1 - \varepsilon$ for a small $\varepsilon > 0$.

This gives that for any integer z

$$\begin{aligned}\mathbb{P}[S_n = z] &\le Q(S_n;1/2)\\ &\le \frac{\alpha(1/2)}{\sqrt{\sum_{i=1}^n (1/2)^2 \min\{p_i, 1-p_i\}}}\end{aligned}$$

$$\leq \frac{\alpha}{\sqrt{\sum_{i=1}^{n} \min\{p_i, 1-p_i\}}}. \tag{16}$$

We use this in the next lemma which is the key to computing the second term in (13).

Lemma 3: Suppose there are n balls which are thrown into one of q bins. Each ball is thrown independently, and 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

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

where

$$\begin{aligned} B &= c_*^{q-1} \frac{\prod_b \pi_b}{\pi_{max}} \\ c_* &= \min_i \frac{c_-(i)}{c_+(i)} \\ c_-(i) &= \min_b \frac{p_{i,b}}{\pi_b} \\ c_+(i) &= \max_b \frac{p_{i,b}}{\pi_b} \\ \pi_{max} &= \max_b \pi_b \end{aligned}$$

and α is the universal constant used in Theorem 5.

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 $\pi_1 \leq \pi_2 \cdots \leq \pi_q$.

First observe that

$$\begin{aligned} &\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \\ &\quad = \prod_{b=1}^{q} \mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}]. \end{aligned} \tag{17}$$

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

$$\begin{aligned} &\min\left\{\frac{p_{i,b}}{\sum_{a=b}^{q} p_{i,a}}, 1 - \frac{p_{i,b}}{\sum_{a=b}^{q} p_{i,a}}\right\} \\ &\quad = \min\left\{\frac{p_{i,b}}{\sum_{a=b}^{q} p_{i,a}}, \frac{\sum_{a>b}^{q} p_{i,a}}{\sum_{a=b}^{q} p_{i,a}}\right\} \\ &\quad = \min\left\{\frac{\pi_b \frac{p_{i,b}}{\pi_b}}{\sum_{a=b}^{q} \pi_a \frac{p_{i,a}}{\pi_a}}, \frac{\sum_{a>b}^{q} \pi_a \frac{p_{i,a}}{\pi_a}}{\sum_{a=b}^{q} \pi_a \frac{p_{i,a}}{\pi_a}}\right\} \\ &\quad \geq \frac{\min_a \frac{p_{i,a}}{\pi_a}}{\max_a \frac{p_{i,a}}{\pi_a}} \min\left\{\frac{\pi_b}{\sum_{a=b}^{q} \pi_a}, \frac{\sum_{a>b}^{q} \pi_a}{\sum_{a=b}^{q} \pi_a}\right\} \\ &\quad \geq \min_i \frac{c_-(i)}{c_+(i)} \frac{1}{\sum_{a=b}^{q} \pi_a} \min\left\{\pi_b, \sum_{a>b}^{q} \pi_a\right\} \\ &\quad = c_* \frac{\pi_b}{\sum_{a=b}^{q} \pi_a}. \end{aligned}$$

We get the last equality because we have arranged π_b in increasing order. Hence by (16)

$$\begin{aligned} &\mathbb{P}[N_b = n\pi_b | N_1 = n\pi_1, \ldots, N_{b-1} = n\pi_{b-1}] \\ &\quad \leq \frac{\alpha}{\sqrt{\left(n - \sum_{a=1}^{b-1} n\pi_a\right) c_* \frac{\pi_b}{\sum_{a=b}^{q} \pi_a}}} \\ &\quad = \frac{\alpha}{\sqrt{\frac{n - \sum_{a=1}^{b-1} n\pi_a}{n} n c_* \frac{\pi_b}{\sum_{a=b}^{q} \pi_a}}} \\ &\quad = \frac{\alpha}{n^{1/2}\sqrt{c_* \pi_b}} \end{aligned}$$

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 (17), gives

$$\begin{aligned} &\mathbb{P}[N_1 = n\pi_1, \ldots, N_q = n\pi_q] \\ &\quad \leq \prod_{b=1}^{q-1} \frac{\alpha}{n^{1/2}\sqrt{c_* \pi_b}} \\ &\quad = \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{c_*^{q-1} \prod_{b=1}^{q-1} \pi_b}} \\ &\quad = \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{c_*^{q-1} \frac{\prod_b \pi_b}{\pi_q}}} \\ &\quad = \frac{\alpha^{q-1}}{n^{(q-1)/2}\sqrt{B}}. \end{aligned}$$

□

D. Completing Proof of Theorem 2 and Determining Capacity

We can now use Lemma 3 to prove Theorem 2.

Proof: [Proof of Theorem 2] We show the lower bound first, which is easier to show. Using Proposition 3, we need only to show that

$$\mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n]}{\mathbb{P}[\tilde{A} = 1]} \middle| A = 1\right] \geq 0.$$

We do this by

$$\begin{aligned} &\mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n]}{\mathbb{P}[\tilde{A} = 1]} \middle| A = 1\right] \\ &\quad = \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[A = 1 | Y^n = y^n]}{\mathbb{P}[A = 1]} \\ &\quad = \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \\ &\qquad \log \frac{\mathbb{P}[Y^n = y^n | A = 1]\mathbb{P}[A = 1]}{\mathbb{P}[Y^n = y^n]\mathbb{P}[A = 1]} \\ &\quad = \sum_{y^n} \mathbb{P}[Y^n = y^n | A = 1] \log \frac{\mathbb{P}[Y^n = y^n | A = 1]}{\mathbb{P}[Y^n = y^n]} \\ &\quad = D(\mathbb{P}[Y^n | A = 1] \| \mathbb{P}[Y^n]) \\ &\quad \geq 0 \end{aligned}$$

since divergences are always non-negative.

To get the upper bound, we use

$$\mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n]}{\mathbb{P}[A = 1]} \middle| A = 1\right]$$

$$= \mathbb{E}\left[\log \mathbb{P}[\tilde{A}=1|\tilde{Y}^n=Y^n]\middle|\tilde{A}=1\right] - \log \mathbb{P}[A=1] \tag{18}$$

and use Lemma 3 for the first term in the sum and Lemma 2 for the second term in the sum.

Lemma 3 applies to the first term because, given some Y^n, finding the probability that the type of X^n is in $T_n(P)$ is equivalent to finding the number of balls which are randomly thrown into each bin. We want to determine the probability that X^n is in $T_n(P)$ when $(X,Y)^n \sim (P_{Y|X} \times P)$.

Let P of $T_n(P)$ be expressed as $P = (\pi_1, \dots, \pi_q) \in \mathcal{P}_n$. This implies that $\pi_b = \mathbb{P}[X = b]$. Let the balls described in Lemma 3 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 3, N_b is the number of balls in bin b. Then the probability that $X^n \in T_n(P)$ is equivalent to $\mathbb{P}[N_1 = n\pi_1, \dots, N_q = n\pi_q]$. This is computed for a specific value of Y^n, but notice that the expression we derived for $\mathbb{P}[N_1 = n\pi_1, \dots, N_q = n\pi_q]$ in Lemma 3 does not depend on Y^n.

Before computing the rest of the expression, we need to pay particular attention to the case when there exists a b such that $\pi_b = 0$. If $\pi_b = 0$, then $\mathbb{P}[X = b] = 0$, which would also imply that $p_{i,b} = 0$ for all i. In this case, we can remove the symbol b (or bin b in the interpretation of Lemma 3) from consideration and apply Lemma 3 to just the symbols with non-zero probability. We can always reorder the symbols, so that the first q' of the q symbols all have $\pi_b > 0$ and the remaining $b > q'$ are such that $\pi_b = 0$. Like in Lemma 3, we can define

$$\begin{aligned} B &= c_*^{q'-1} \frac{\prod_{b=1}^{q'} \pi_b}{\pi_{max}} \\ c_* &= \min_i \frac{c_-(i)}{c_+(i)} \\ c_-(i) &= \min_{b:b\leq q'} \frac{p_{i,b}}{\pi_b} \\ c_+(i) &= \max_{b:b\leq q'} \frac{p_{i,b}}{\pi_b}. \end{aligned}$$

$$\begin{aligned} &\mathbb{E}\left[\log \mathbb{P}[\tilde{A}=1|\tilde{Y}^n=Y^n]\middle|A=1\right] \\ &= \log \mathbb{P}[N_1 = n\pi_1, \dots, N_{q'} = n\pi_{q'}] \\ &= \log \frac{\alpha^{q'-1}}{n^{(q'-1)/2}\sqrt{B}} \\ &= \log \left(\frac{\alpha^{q'-1}}{n^{(q'-1)/2} c_*^{\frac{q'-1}{2}}} \left(\frac{\pi_{max}}{\prod_b \pi_b} \right)^{1/2} \right) \\ &= \log \left(\frac{\alpha^{q'-1}}{c_*^{\frac{q'-1}{2}}} \left(\frac{n\pi_{max}}{\prod_{b=1}^{q'} n\pi_b} \right)^{1/2} \right) \\ &= \frac{1}{2}\log n\pi_{max} - \sum_{b:\pi_b>0} \frac{1}{2}\log n\pi_b + (q'-1)\log\left(\frac{\alpha}{\sqrt{c_*}}\right) \\ &\leq \frac{1}{2}\log n - \sum_{b:\pi_b>0} \frac{1}{2}\log n\pi_b + c'. \end{aligned} \tag{19}$$

where c' is constant that does not depend on n. Importantly, the value of c' also does not depend on π_b for any b. The quantity c' depends on c_*, which we can compute with:

$$\begin{aligned} \frac{p_{i,b}}{\pi_b} &= \frac{\mathbb{P}[X = b|Y = y_i]}{\pi_b} \\ &= \frac{\mathbb{P}[Y = y_i|X = b]\mathbb{P}[X = b]}{\pi_b \mathbb{P}[Y = y_i]} = \frac{\mathbb{P}[Y = y_i|X = b]}{\mathbb{P}[Y = y_i]} \\ &= \frac{P_{Y|X}(y_i|b)}{\mathbb{P}[Y = y_i]} \end{aligned}$$

and

$$\begin{aligned} c_* &= \min_i \frac{\min_b \frac{p_{i,b}}{\pi_b}}{\max_b \frac{p_{i,b}}{\pi_b}} \\ &= \min_i \frac{\min_b \frac{P_{Y|X}(y_i|b)}{\mathbb{P}[Y=y_i]}}{\max_b \frac{P_{Y|X}(y_i|b)}{\mathbb{P}[Y=y_i]}} \\ &= \min_y \frac{\min_b P_{Y|X}(y|b)}{\max_b P_{Y|X}(y|b)} \end{aligned}$$

So c_* only depends on $P_{Y|X}$.

Combining these terms with those from Lemma 2 gives that the expression in (18) is a constant when $P_{Y|X}$ is strictly positive. This constant depends on q and $P_{Y|X}$ but not on n or π_b for any b. □

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

IV. Conclusion

In summary, our work determines the capacity of the noisy permutation channel for the case of a strictly positive DMC matrix. Our main method is to use KL divergence covering on the probability simplex. A key ingredient necessary to complete this proof is our theorem which computes the KL divergence between noisy observations of a sequence sampled from a fixed type class versus noisy observations of an iid sequence. We expect this key theorem, which is interesting in its own right, to be applicable to other problems as well. We also determine the capacity of the noisy permutation channel for block diagonal DMC matrices with strictly positive blocks, the q-ary erasure channel, and the Z-channel.

Finally, we provide some directions for future research.

1) While we can determine the capacity of the noisy permutation channel for strictly positive DMC matrices and a certain subset of non-strictly positive DMC matrices, we do not know how to compute the capacity for general (non-strictly positive) DMC matrices. We know that the achievability bound (2) will apply in the general case, however strategically placed 0's in the DMC matrix could possibly increase the capacity above the rate specified in (2).
2) Since our converse bound to the capacity is computed using mutual information, this is only a weak converse bound. This leaves open the question of whether we can find strong converse bounds [2, Section 22.1].
3) Capacity gives the (asymptotic) leading coefficient of the $\log n$ term in the expansion of $\log M^*(n, \epsilon)$ as $n \to \infty$.

Finding the next-order terms and their dependence on ϵ would be very interesting.

Appendix A
Comparing Theorem 2 to Stam

Here we give some details on how our result compares to that of [5], which we will refer to as Stam's setting or Stam's result. Though similar, our setting is not exactly the same as Stam's setting. The differences are:

1) Stam's result generalizes to m observations, where m can be less than n. Our result Theorem 2 only applies to exactly n observations.
2) Stam's setting has noiseless observations whereas our setting has noisy observations.

In order for our result and Stam's result to be comparable, we apply additional theorems to both our result and Stam's result so that we are in a setting where both results are for all $m \leq n$ and for noisy observations.

Regarding difference 1), we can use a version of Han's inequality for divergence [23, Proposition 5.5] (applies when the second probability argument is independent over the entries of the vector Y^n) on our result Theorem 2, to get the following corollary:

Corollary 1: Let $Y^n \sim P_{Y|X}^n \circ U$, so that P_{Y^n} is the distribution as in Theorem 2. Then for every $m \leq n$ we have:

$$D(P_{Y^m} \| Q_Y^m) \leq mD(P_Y \| Q_Y) + \frac{m}{n}c$$

or when $Q_Y = P_Y$,

$$D(P_{Y^m} \| P_Y^m) \leq \frac{m}{n}c. \tag{20}$$

where Y^m are the first m entries of vector Y^n and c is the same constant as in Theorem 2.

Regarding difference 2), using data processing inequality, we can use Stam's result as an upper bound for the case with noisy observations. Stam's result with data processing gives

$$\begin{aligned} D(P_{Y^m} \| P_Y^m) &\leq D(P_{X^m} \| P_X^m) \\ &\leq \frac{(q-1)}{2} \frac{m(m-1)}{(n-1)(n-m+1)}. \end{aligned} \tag{21}$$

The above equation, which is presented as the final result in [5], is actually not the tightest when m is close to n. For instance, when $m = n$, (21) gives $\frac{q-1}{2}n$ which is far from $\frac{q-1}{2}(\log n + c')$, the actual divergence when computed directly. An improvement on Stam's bound when m is close to n is given in [24]. We show an easier improvement, using an intermediate result in the proof of (21). We can derive for larger m that

$$\begin{aligned} &D(P_{Y^m} \| P_Y^m) \\ &\leq D(P_{X^m} \| P_X^m) \qquad (22) \\ &\leq \frac{q-1}{n-1} \sum_{t=1}^{m-1} \frac{t}{n-t} \\ &= \frac{q-1}{n-1} \sum_{j=n-m+1}^{n-1} \frac{n-j}{j} \\ &= \frac{q-1}{n-1} \left(n \left(\sum_{j=n-m+1}^{n-1} \frac{1}{j} \right) - (m-1) \right) \\ &= \frac{q-1}{n-1} \left(n \left(\log(n-1) - \log(n-m) + c'' \right) - (m-1) \right) \\ &= \frac{q-1}{n-1} \left(n \log \frac{n-1}{n-m} + nc'' - (m-1) \right) \\ &= (q-1) \log \frac{n-1}{n-m} + O(q) \qquad (23) \end{aligned}$$

for $m < n$ and c'' is a constant leftover from approximating the harmonic sum by a logarithm.

We now can compare our result (20) with Stam's result, either (21) and (23), in the setting of $m \leq n$ and noisy observations:

- When $m << n$, (21) is a better bound than (20).
- When m is very close to n, such as when $n - m = o(n)$, (20) is a tighter bound than both (21) and (23).
- When m is linear in n, then it becomes important to compare the constant factors. Let $\gamma = m/n$. To get an estimate on when our bound is tighter, we first assume n is large and ignore the lower order constants which appear in the bounds. Using Remark 1, (20) is tighter than (21) and (23) for large n if
$$\frac{1}{2} \log \frac{2\pi\alpha^2}{c_*} \leq \min \left\{ \frac{\gamma}{1-\gamma}, \frac{1}{\gamma} \log \frac{1}{1-\gamma} \right\}.$$
This can occur for certain values of γ depending on the size of c_*, which is a function of $P_{Y|X}$.

The observations indicate that whether our result is tighter or Stam's result is tighter depends on the value of m and n. This also verifies that our Theorem 2 result cannot be proven as just a corollary of Stam's result and indeed we are offering something new. Our result can have consequences in Stam's setting for noisy observations when n and m are large.

Appendix B
Covering Results

In this section we give the covering results necessary for proving Proposition 1, which include the proofs of Theorem 4 and Proposition 2. Again, the bound in Theorem 4 is sufficient but not the best possible covering bound. Other bounds are explored in [18].

A. Divergence Covering Upper Bound (Proof of Theorem 4)

We need some preliminaries before proving Theorem 4. To define our covering centers for the simplex, we start with a set of scalars. Let

$$\begin{aligned} \Lambda(\varepsilon) \triangleq &\left\{ \varepsilon i^2 : \text{ for } i \in \mathbb{Z}_{>0}, \varepsilon i^2 < \frac{1}{2} \right\} \\ &\cup \left\{ 1 - \varepsilon i^2 : \text{ for } i \in \mathbb{Z}_{>0}, \varepsilon i^2 < \frac{1}{2} \right\} \cup \left\{ \frac{1}{2} \right\} \end{aligned}$$

Define

$$\Lambda_2(\varepsilon) = \{ (\lambda, 1-\lambda) : \lambda \in \Lambda(\varepsilon) \}$$

For each k, let $u_k \in \Delta_{k-1}$ be $u_k = (0, \ldots, 0, 1)$. For each $q \in \Delta_{k-2}$, let $\hat{q}$ be the corresponding $\hat{q} \in \Delta_{k-1}$ where $\hat{q} = (q_1, \ldots, q_{k-1}, 0)$.

For each $q \in \Delta_{k-2}$, define $q^{(\lambda)}$ such that

$$q^{(\lambda)} = \lambda u_k + (1-\lambda)\hat{q}.$$

For $k > 2$, recursively define

$$\Lambda_k(\varepsilon) \triangleq \bigcup_{\lambda \in \Lambda\left(\frac{\varepsilon}{k}\right)} \left\{ q^{(\lambda)} : q \in \Lambda_{k-1}\left(\frac{k-1}{k}\varepsilon\right) \right\}$$

Lemma 4: For any $p \in \Delta_{k-1}$,

$$\min_{q \in \Lambda_k(\varepsilon)} D(p\|q) \leq \gamma\varepsilon$$

where γ is a constant.

Proof: We show this by using induction. First, for any $p \in \Delta_1$, we want to show that

$$\min_{q \in \Lambda_2(\varepsilon)} D(p\|q) \leq \gamma\varepsilon.$$

We use the following fact[7] from [25]. For probabilities P_1 and P_2 on k symbols, we have

$$D(P_1||P_2) \leq \sum_{a=1}^{k} \frac{(P_1(a) - P_2(a))^2}{P_2(a)}.$$

This implies that for any $p, q \in \Delta_1$, where $p = (p_1, 1-p_1)$ and $q = (q_1, 1-q_1)$

$$D(p\|q) \leq \frac{(p_1-q_1)^2}{q_1} + \frac{(1-p_1-1+q_1)^2}{1-q_1} = \frac{(p_1-q_1)^2}{q_1(1-q_1)}.$$

Suppose that $p \in \Delta_1$ and $p_1 < 1/2$. Then $\varepsilon(i-1)^2 < p_1 \leq \varepsilon i^2$ for some positive integer i. Assume for now that $\varepsilon i^2 < 1/2$. Choose $q = (\varepsilon i^2, 1 - \varepsilon i^2) \in \Lambda_2(\varepsilon)$. Note that we must have $1 - \varepsilon i^2 > 1/2$.

$$\begin{aligned} D(p\|q) &\leq \frac{(p_1-q_1)^2}{q_1(1-q_1)} \\ &= \frac{(p_1 - \varepsilon i^2)^2}{(\varepsilon i^2)(1-\varepsilon i^2)} \\ &\leq \frac{(\varepsilon(i-1)^2 - \varepsilon i^2)^2}{(\varepsilon i^2)(1-\varepsilon i^2)} \\ &\leq \varepsilon \frac{(-2i+1)^2}{i^2(1/2)} \\ &\leq \varepsilon \frac{4i^2-4i+1}{i^2/2} \\ &\leq 8\varepsilon \end{aligned}$$

If $\varepsilon i^2 > 1/2$, we can choose $q = (1/2, 1/2)$. For this case, we can also assume $i > 1$, otherwise one center point, $q = (1/2, 1/2)$ is sufficient for covering the whole simplex. Then

$$\begin{aligned} D(p\|q) &\leq \frac{(p_1 - 1/2)^2}{(1/2)(1/2)} \\ &\leq \frac{(\varepsilon(i-1)^2 - \varepsilon i^2)^2}{1/4} \\ &\leq 4\varepsilon^2(-2i+1)^2 \\ &\leq \frac{4}{2}\varepsilon\frac{4i^2-4i+1}{(i-1)^2} \\ &\leq 18\varepsilon \end{aligned}$$

where we used that $\varepsilon < 1/(2(i-1)^2)$. This shows that we can set $\gamma = 18$. By symmetry, $\min_{q \in \Lambda_2(\varepsilon)} D(p\|q) \leq \gamma\varepsilon$ holds for $p_1 > 1/2$ as well.

Suppose in dimension $k-1$, that we have for any $p \in \Delta_{k-2}$,

$$\min_{q \in \Lambda_{k-1}(\varepsilon)} D(p\|q) \leq \gamma\varepsilon$$

For each $p = (p_1, \ldots, p_k) \in \Delta_{k-1}$, we specify a scalar quantity $\lambda_p \in [0,1]$. If $p_k < 1/2$, like above, we can find a positive integer i where

$$\frac{\varepsilon}{k}(i-1)^2 \leq p_k \leq \frac{\varepsilon}{k} i^2$$

and set

$$\lambda_p = \min\left\{\frac{\varepsilon}{k} i^2, \frac{1}{2}\right\} \in \Lambda\left(\frac{\varepsilon}{k}\right).$$

If $p_k > 1/2$, find i such that

$$1 - \frac{\varepsilon}{k} i^2 < p_k \leq 1 - \frac{\varepsilon}{k}(i-1)^2$$

and set

$$\lambda_p = \max\left\{1 - \frac{\varepsilon}{k} i^2, \frac{1}{2}\right\} \in \Lambda\left(\frac{\varepsilon}{k}\right).$$

Define $p_k' = (p_k, 1-p_k)$ and $\lambda_p' = (\lambda_p, 1-\lambda_p)$, then similar to above

$$D(p_k'\|\lambda_p') \leq \gamma\frac{\varepsilon}{k}.$$

$$\begin{aligned} &\min_{q \in \Lambda_k(\varepsilon)} D(p\|q) \\ &\leq \min_{q \in \Lambda_k(\varepsilon): q_k = \lambda_p} p_k \log\frac{p_k}{q_k} + \sum_{i=1}^{k-1} p_i \log\frac{p_i}{q_i} \\ &\leq p_k \log\frac{p_k}{\lambda_p} + \min_{q \in \Lambda_k(\varepsilon): q_k = \lambda_p} \sum_{i=1}^{k-1} p_i \log\frac{p_i}{q_i} \\ &\leq p_k \log\frac{p_k}{\lambda_p} + \min_{q \in \Lambda_k(\varepsilon): q_k = \lambda_p} (1-p_k)\log\frac{1-p_k}{1-\lambda_p} \\ &\qquad + (1-p_k)\sum_{i=1}^{k-1}\frac{p_i}{1-p_k}\log\frac{p_i/(1-p_k)}{q_i/(1-\lambda_p)} \\ &\leq D(p_k'\|\lambda_p') + (1-p_k) \\ &\qquad \min_{q' \in \Lambda_{k-1}\left(\frac{k-1}{k}\varepsilon\right)} \sum_{i=1}^{k-1}\frac{p_i}{1-p_k}\log\frac{p_i/(1-p_k)}{q_i'} \\ &\leq \gamma\frac{\varepsilon}{k} + (1-p_k)\gamma\frac{k-1}{k}\varepsilon \\ &\leq \gamma\varepsilon. \end{aligned}$$

□

Proof: [Proof of Theorem 4] We use $\mathcal{Q}_k(\varepsilon)$ to denote the set of centers we need to cover Δ_{k-1} with radius ε.

Let

$$\mathcal{Q}_k(\varepsilon) = \Lambda_k\left(\frac{\varepsilon}{\gamma}\right)$$

[7] This fact is that KL divergence is upper-bounded by χ^2-divergence.

where γ is the constant in Lemma 4. Then using Lemma 4, for $p \in \Delta_{k-1}$,

$$\min_{q \in \mathcal{Q}_k(\varepsilon)} D(p||q) \le \varepsilon.$$

Since, $M(k,\varepsilon) \le |\mathcal{Q}_k(\varepsilon)|$, it remains to count the size of each $\mathcal{Q}_k(\varepsilon)$. We show its size by induction. First, we have that

$$\left|\Lambda\left(\frac{\varepsilon}{\gamma}\right)\right| \le 2\sqrt{\frac{\gamma}{2\varepsilon}} + 1 = \sqrt{\frac{2\gamma}{\varepsilon}} + 1 \le \frac{\sqrt{2\gamma}+1}{\sqrt{\varepsilon}}$$

where the last inequality holds if $\varepsilon \le 1$. Therefore we have for some constant c (we can show $c \le 7$),

$$|\mathcal{Q}_2(\varepsilon)| \le c\frac{1}{\sqrt{\varepsilon}}.$$

For the inductive case, given alphabet size k and any $\varepsilon \le 1$, we have $|\Lambda_k\left(\frac{\varepsilon}{\gamma}\right)| \le c^{k-1}\left(\frac{k-1}{\varepsilon}\right)^{\frac{k-1}{2}}$.

Now consider the case of alphabet size $k+1$. The set $\Lambda_{k+1}\left(\frac{\varepsilon}{\gamma}\right)$ is defined as a set of points which is a product of sets $\Lambda\left(\frac{1}{k}\frac{\varepsilon}{\gamma}\right)$ and $\Lambda_{k+1}\left(\frac{k-1}{k}\frac{\varepsilon}{\gamma}\right)$. This gives

$$\begin{aligned}
|\mathcal{Q}_{k+1}(\varepsilon)| &= \left|\Lambda_{k+1}\left(\frac{\varepsilon}{\gamma}\right)\right| \\
&= \left|\Lambda\left(\frac{1}{k}\frac{\varepsilon}{\gamma}\right)\right|\left|\Lambda_k\left(\frac{k-1}{k}\frac{\varepsilon}{\gamma}\right)\right| \\
&\le \left(c\frac{1}{\sqrt{\frac{\varepsilon}{k}}}\right)\left(c^{k-1}\left(\frac{k-1}{\varepsilon\frac{k-1}{k}}\right)^{\frac{k-1}{2}}\right) \\
&= c\frac{\sqrt{k}}{\sqrt{\varepsilon}}c^{k-1}\left(\frac{k}{\varepsilon}\right)^{\frac{k-1}{2}} \\
&= c^k\left(\frac{k}{\varepsilon}\right)^{\frac{k}{2}}
\end{aligned}$$

as the number of centers. □

B. Subspace Covering (Proof of Proposition 2)

To use our covering result for noisy permutation channels, we actually need to cover a lower dimensional subspace of a $(k-1)$-dimensional simplex.

Lemma 5: For $\mathcal{B} \subset \Delta_{k-1}$, suppose there is a stochastic matrix F which maps $\Delta_{\ell-1}$ onto $\mathcal{B}$. Then,

$$M(k,\varepsilon,\mathcal{B}) \le M(\ell,\varepsilon).$$

Proof: Let $\mathcal{N}_c(\ell,\varepsilon)$ be the set of points which are centers for a divergence covering of $\Delta_{\ell-1}$ with covering radius ε. For each $b \in \mathcal{B}$, there exists a $p \in \Delta_{\ell-1}$ such that $pF = b$. For this p, let $r \in \mathcal{N}_c(\ell,\varepsilon)$ be such that $D(p||r) \le \varepsilon$. Let $b^* = rF$. By data processing inequality [2, Theorem 2.2],

$$D(b||b^*) \le D(p||r) \le \varepsilon.$$

Hence the image of the set of centers in $\mathcal{N}_c(\ell,\varepsilon)$ mapped using F, becomes the set of centers for a divergence covering on $\mathcal{B}$ with radius ε. □

Proof: [Proof of Proposition 2] The key to this proof is to divide the space $\mathcal{B}$ into simplices of dimension $\ell - 1$.

We can upper bound the number of simplices needed for a partition of $\mathcal{B}$. The image of F is a convex hull of at most q points (recall q is the size of the input symbols). We call these corner points. Consider all possible choices of ℓ of these q corner points. Let this set of all combinations be S, where

$$|S| = \binom{q}{\ell}.$$

For each $s \in S$, let $\mathcal{B}_s$ be the simplex which is the convex hull of the ℓ corner points in set s.

For each point x in the image of F, since F has rank ℓ, there exists some linear combination of ℓ corner points which results in x. If s is this set of ℓ points, then $x \in \mathcal{B}_s$. Thus for all $x \in \mathcal{B}$, there exists some $s \in S$, so that $x \in \mathcal{B}_s$.

Label each of these simplices as $\mathcal{B}_1, \ldots, \mathcal{B}_{|S|}$. There exists a stochastic matrix F_i which maps from space $\Delta_{\ell-1}$ onto the space $\mathcal{B}_i$. In particular, we can find this map F_i by mapping each of the ℓ corners of $\Delta_{\ell-1}$ into one of the ℓ corner points of $\mathcal{B}_i$. This map covers all of $\mathcal{B}_i$ by linearity.

Hence using Lemma 5, we can find a divergence covering of size $M(\ell,\varepsilon)$ for each $\mathcal{B}_i$. Combining these covering centers together for all i, we get a covering of size

$$\binom{q}{\ell}M(\ell,\varepsilon).$$

□

We are most assuredly over counting the number of simplices $\mathcal{B}$ has to be divided up into. However, this number does not depend on ε, which is sufficient for our application to noisy permutation channels.

APPENDIX C
BLOCK DIAGONAL CASE

With a small modification to the proof of Theorem 1, we can show a converse bound for block diagonal matrices where each block is strictly positive. The key idea is that since each block is independent from all the other blocks, so we can apply the bound for strictly positive matrices separately to each block. We need to show a separate achievability result to match this converse bound.

As a sanity check, the block diagonal case captures the situation where $P_{Z|X}$ is the identity matrix. In which case, it is possible to use all possible permutations of symbols as messages. No errors are allowed so decoding is straightforward. Using an identity matrix of size $q \times q$ for the DMC, for each n,

$$\begin{aligned}
R &= \frac{\log M}{\log n} \\
&= \frac{\log\binom{n}{q-1}}{\log n} \\
&\approx \frac{\log(c^{q-1}n^{q-1}/(q-1)^{q-1})}{\log n} \\
&= q-1+\frac{\log(c^{q-1}/(q-1)^{q-1})}{\log n}
\end{aligned}$$

which goes to $q-1$ asymptotically as n increases. This matches our block diagonal converse result.

A. Converse

Proposition 4 (Block Diagonal Converse): Suppose $P_{Z|X}$ can be written as a block diagonal matrix with β blocks, so that each block is strictly positive. Then,

$$C_{\mathsf{perm}}(P_{Z|X}) \le \frac{\mathsf{rank}(P_{Z|X}) + \beta - 2}{2}.$$

Proof:

We want to use Proposition 1 but we need to show a version of the upper bound in Theorem 2 which applies to block diagonal matrices instead of strictly positive matrices.

Fix $\pi \in \mathcal{P}_n$. Arrange the matrix $P_{Z|X}$ in block diagonal form and let $\mathcal{X}_b$ be the set of symbols in $\mathcal{X}$ which are in the bth block. Let $(X, Y)^n$ be generated iid from $(\pi \times P_{Y|X})$. Let W_i be the number of X which equals i, i.e.

$$W_i = |\{t : X_t = i\}|.$$

Define

$$A_b = \left\{ \bigcap_{i \in \mathcal{X}_b} W_i = \pi_i n \right\}.$$

This is the event that all symbols i associated with block b occur with the count $\pi_i n$. Each block has its own separate set of output symbols in $\mathcal{Y}$. The probability of W_i is independent of what happens in other blocks. Let $Y^n(b)$ (and $y^n(b)$) be notation for the symbol counts restricted to just the output symbols associated with the bth block.

Using the definition in Proposition 3, notice that $\mathbb{I}[A = 1] = \mathbb{I}\left[\bigcap_{b=1}^{\beta} A_b\right]$. (Recall the notation $\tilde{A} = \{\tilde{X}^n \in T(P)\}$ where $(\tilde{X}, \tilde{Y})^n$ is independent copy under the same distribution $(P \times P_{Y|X})$ as $(X, Y)^n$). Then using (13) with Lemma 2,

$$\begin{aligned} &D(P_{Y|X}^n \circ U \| Q_Y^n) \\ &= nD(P_Y \| Q_Y) + \mathbb{E}\left[\log \frac{\mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n]}{\mathbb{P}[\tilde{A} = 1]} \middle| A = 1\right] \\ &\le nD(P_Y \| Q_Y) - \frac{1}{2}\log n + \sum_{i:\pi_i > 0} \frac{1}{2}\log \pi_i n \\ &\quad + c + \mathbb{E}\left[\log \mathbb{P}[\tilde{A} = 1 | \tilde{Y}^n = Y^n] \middle| A = 1\right]. \end{aligned}$$

For any Y^n,

$$\mathbb{P}[A = 1 | Y^n] = \prod_{b=1}^{\beta} \mathbb{P}[A_b = 1 | Y^n(b) = y^n(b)].$$

Each block is a strictly positive matrix. From Lemma 3 and following the same calculations that results in (19), we know that

$$\begin{aligned} &\log \mathbb{P}[A_b = 1 | Y^n(b) = y^n(b)] \\ &\quad \le \frac{1}{2}\log n - \sum_{i \in \mathcal{X}_b : \pi_i > 0} \frac{1}{2}\log n\pi_i + c' \end{aligned}$$

and thus

$$\begin{aligned} &\log \mathbb{P}[A = 1 | Y^n] \\ &\le \sum_{b=1}^{\beta} \left(\frac{1}{2}\log n - \sum_{i \in \mathcal{X}_b : \pi_i > 0} \frac{1}{2}\log n\pi_i + c' \right) \\ &= \frac{\beta}{2}\log n - \sum_{i:\pi_i > 0} \frac{1}{2}\log n\pi_i + \beta c'. \end{aligned}$$

This holds for all Y^n so it automatically gives the expected value. Putting all these terms together, for any π, we get

$$D(P_{Y|X}^n \circ U \| Q_Y^n) = nD(P_Y \| Q_Y) + \frac{\beta - 1}{2}\log n + c''$$

where c'' combines all the constants. Using Proposition 1, gives

$$\begin{aligned} C_{\mathsf{perm}}(P_{Z|X}) &\le \frac{\mathsf{rank}(P_{Z|X}) - 1}{2} + \lim_{n \to \infty} \frac{\frac{\beta - 1}{2}\log n + c''}{\log n} \\ &= \frac{\mathsf{rank}(P_{Z|X}) - 1}{2} + \frac{\beta - 1}{2} \\ &= \frac{\mathsf{rank}(P_{Z|X}) + \beta - 2}{2}. \end{aligned}$$

□

B. Achievability

Proposition 5 (Block Diagonal Achievability): Suppose $P_{Z|X}$ can be written as a block diagonal matrix with β blocks, so that each block is strictly positive. Then,

$$C_{\mathsf{perm}}(P_{Z|X}) \ge \frac{\mathsf{rank}(P_{Z|X}) + \beta - 2}{2}.$$

Proof:

The achievability proof encodes using two steps. The first step is a zero-error code based on which block in the block diagonal matrix the symbols are associated with. Let M_1 denote the total possible messages (or rather message stems) for the first step. The second step operates only on each block independently, and uses the achievability given by (2). Let M_2 denote the total messages (or message tails) possible here.

Label the β blocks in $P_{Z|X}$ as $B_1, \ldots, B_\beta$. Define the sets of input symbols $\mathcal{X}_1, .., \mathcal{X}_\beta$ and output symbols $\mathcal{Y}_1, \ldots, \mathcal{Y}_\beta$, so that $\mathcal{X}_b$ and $\mathcal{Y}_b$ are the input and output symbols respectively associated with block B_b. (In other words, if $p_{ij} > 0$ and p_{ij} falls into block B_b, then $i \in \mathcal{X}_b$ and $j \in \mathcal{Y}_b$.) These sets $\mathcal{X}_1, .., \mathcal{X}_\beta$ and $\mathcal{Y}_1, \ldots, \mathcal{Y}_\beta$ are both disjoint.

Let $L = \mathsf{rank}(P_{Z|X})$ and let $L_b = \mathsf{rank}(B_b)$. Because of the block diagonal structure, $L = \sum_{b=1}^{\beta} L_b$.

For fixed n, set aside the first $n/2$ input symbol positions so that exactly $n/(2\beta)$ are from set $\mathcal{X}_b$ for each b. These are not used for the first step of the two-step code and are used to make the analysis of the second step easier. The remaining $n/2$ positions can be encoded using symbols from any set and this is used to make the first step of the code. There are

$$\binom{n/2}{\beta - 1} \ge \left(\frac{n/2}{\beta - 1}\right)^{\beta - 1} \tag{24}$$

possible combinations of symbols chosen from β blocks, disregarding order. The DMC maps the symbols in set $\mathcal{X}_b$ to symbols in set $\mathcal{Y}_b$ without any error. Hence, (24) is the number of messages M_1 the first step can encode without any error.

Once it is determined how many symbols of each set will be used, we can determine which symbol in the set will be used for the second step. Suppose there are n_b positions which are designated for symbols in set $\mathcal{X}_b$. This includes the $n/(2\beta)$ we set aside in the beginning and how ever many were chosen to make the first step of the code. Using (2), we know there exists a encoder-decoder pair (f_{n_b}, g_{n_b}) so that the decoding error is vanishingly small as $n_b \to \infty$. Just by choosing which symbol in $\mathcal{X}_b$ to send, for some $\varepsilon_{n_b} > 0$ where $\varepsilon_{n_b} \to 0$, we can encode a set of messages with size M_b satisfying

$$\log M_b \geq \left(\frac{L_b - 1}{2} - \varepsilon_{n_b}\right) \log n_b.$$

The set of messages possible for all the β different sets is

$$\begin{aligned}
\log M_2 &= \log \prod_{b=1}^{\beta} M_b \\
&\geq \sum_{b=1}^{\beta} \left(\frac{L_b - 1}{2} - \varepsilon_{n_b}\right) \log n_b \\
&\geq \sum_{b=1}^{\beta} \left(\frac{L_b - 1}{2} - \varepsilon_{n_b}\right) \log \frac{n}{2\beta} \\
&= \left(\frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b}\right) \log \frac{n}{2\beta}.
\end{aligned}$$

The total number of messages is the product of those available at the first and second steps.

$$\begin{aligned}
\log M &= \log M_1 + \log M_2 \\
&\geq \log \left(\frac{n/2}{\beta - 1}\right)^{\beta - 1} + \left(\frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b}\right) \log \frac{n}{2\beta} \\
&= (\beta - 1) \log \frac{n}{2\beta - 2} + \left(\frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b}\right) \log \frac{n}{2\beta} \\
&\geq \left(\frac{2\beta - 2}{2} + \frac{L - \beta}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b}\right) \log \frac{n}{2\beta} \\
&= \left(\frac{L + \beta - 2}{2} - \sum_{b=1}^{\beta} \varepsilon_{n_b}\right) \log \frac{n}{2\beta}.
\end{aligned}$$

Since each $n_b \geq \frac{n}{2\beta} \to \infty$ as $n \to \infty$, asymptotically the term $\sum_{b=1}^{\beta} \varepsilon_{n_b}$ disappears.

The achievable rate is given by

$$\begin{aligned}
R &= \frac{\log M}{\log n} \\
&\geq \left(\frac{L - \beta - 2}{2} + \sum_{b=1}^{\beta} \varepsilon_{n_b}\right) \frac{\log n - \log 2\beta}{\log n} \to \frac{L - \beta - 2}{2}.
\end{aligned}$$

□

Combining Proposition 4 and Proposition 5 gives the first result in Theorem 3.

Appendix D
Erasure and Z-Channels

A. Concentration Lemma

The following lemma is useful for computing the probability of $A = 1$ (see Proposition 3) when the DMC matrix is not strictly positive. It is a straight-forward concentration bound which is a direct application of Bernstein's inequality. We choose to write the proof anyways for completeness.

Lemma 6: Suppose that Z is a sum of n independent Bernoulli random variables. Let $\mathbb{E}[Z]$ be the expected value of Z.

Fix constant γ. If $\mathbb{E}[Z] > 2\gamma \log n$, then with probability at least $1 - 2/n^{\gamma/4}$, we have that

$$\mathbb{E}[Z] > \mathbb{E}[Z] - \sqrt{\mathbb{E}[Z]\gamma \log n} \geq \frac{1}{5}\mathbb{E}[Z].$$

Proof:

Let $Z = \sum_{i=1}^{n} W_i$ where W_i is the ith Bernoulli random variable. Let $0 < p_i < 1$ be the probability of $W_i = 1$.

$$\sum_{i=1}^{n} \mathbb{E}[(W_i - \mathbb{E}[W_i])^2] = \sum_{i=1}^{n} p_i(1 - p_i) \leq \sum_{i=1}^{n} p_i = \mathbb{E}[Z].$$

Next, we use Bernstein's inequality for bounded variables [26, Theorem 2.8.4].

$$\begin{aligned}
&\mathbb{P}\left[Z - \mathbb{E}[Z] \leq -\sqrt{\mathbb{E}[Z]\gamma \log n}\right] \\
&\quad \leq 2\exp\left(\frac{-\frac{1}{2}\mathbb{E}[Z]\gamma \log n}{\sum_{i=1}^{n} \mathbb{E}[(W_i - \mathbb{E}[W_i])^2] + \frac{1}{3}1\sqrt{\mathbb{E}[Z]\gamma \log n}}\right) \\
&\quad \leq 2\exp\left(\frac{-\frac{1}{2}\mathbb{E}[Z]\gamma \log n}{\mathbb{E}[Z] + \frac{1}{3}\sqrt{\mathbb{E}[Z]\gamma \log n}}\right) \\
&\quad \leq 2\exp\left(\frac{-\frac{1}{2}\gamma \log n}{1 + \frac{1}{3}\frac{\sqrt{\gamma \log n}}{\sqrt{\mathbb{E}Z}}}\right).
\end{aligned}$$

Using that $\mathbb{E}[Z] > 2\gamma \log n$, we have $1 + \frac{1}{3}\frac{\sqrt{\gamma \log n}}{\sqrt{\mathbb{E}[Z]}} \leq 1 + \frac{1}{3}\frac{\sqrt{\gamma \log n}}{\sqrt{2\gamma \log n}} \leq 2$.

$$\begin{aligned}
&\mathbb{P}\left[Z - \mathbb{E}[Z] \leq -\sqrt{\mathbb{E}[Z]\gamma \log n}\right] \\
&\quad \leq 2\exp\left(\frac{-1}{4}\gamma \log n\right) \\
&\quad \leq \frac{2}{n^{\gamma/4}}.
\end{aligned}$$

Hence, with probability $1 - 2/n^{\gamma/4}$,

$$\begin{aligned}
\mathbb{E}[Z] &\geq \mathbb{E}[Z] - \sqrt{\mathbb{E}Z\gamma \log n} \\
&\geq \mathbb{E}[Z] - \sqrt{\mathbb{E}[Z]\frac{1}{2}\mathbb{E}[Z]} \\
&\geq \left(1 - \frac{1}{\sqrt{2}}\right)\mathbb{E}[Z] \\
&\geq \frac{1}{5}\mathbb{E}[Z].
\end{aligned}$$

□

B. The q-Ary Erasure Channel

We now can prove the converse bound for q-ary erasure channels, where q is the number of input symbols. Let $k = q+1$ represent the erased symbol.

The matrix $P_{Z|X}$ for a q-ary erasure channel has the following structure:

$$P_{Z|X} = \begin{bmatrix} p_{11} & 0 & \cdots & 0 & p_{1k} \\ 0 & p_{22} & \cdots & 0 & p_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{qq} & p_{qk} \end{bmatrix}.$$

We assume that $p_{ik} > 0$ for each i.

Proof: [Proof of Item 2 of Theorem 3]

Fix $\pi = (\pi_1, \ldots, \pi_q)$ where $\pi \in \mathcal{P}_n$. (We assume each $\pi_i > 0$, otherwise we can remove it.) Reorder the symbols in $\{1, \ldots, q\}$ so that $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_q$. (Note that $P_{Y|X} = P_{Z|X}$.)

Following Proposition 3, let $(X,Y)^n$ be generated iid according to $(\pi \times P_{Y|X})$. To use Proposition 3 we need to determine $\mathbb{P}[\log \mathbb{P}[\tilde{A} = 1|\tilde{Y}^n = Y^n]|A = 1]$.

Unlike the case of strictly positive $P_{Y|X}$, the value of $\mathbb{P}[A = 1|Y^n]$ depends on Y^n. For instance, it is easy to see that when the erasure symbol k does not appear, then $\mathbb{P}[A = 1|Y^n] = 1$. While Y^n like this can occur under the event $A = 1$, we want to show that these events are rare, this way the expected value of $\mathbb{P}[A = 1|Y^n]$ given that $A = 1$ is much smaller than 1 and close to the value which will give our result. We first show a concentration result on Y^n given that $A = 1$.

Let U_b be the random variable which gives the count of the number of times the symbol b is erased, i.e

$$U_b = \sum_{i=1}^{n} \mathbb{I}\{(X_i, Y_i) = (b, k)\}.$$

Let $v_b(y^n) = \{U_b|A = 1, Y^n = y^n\}$. Note that $v_b(y^n)$ is deterministic. If $A = 1$ and Y^n is known, we can determine exactly what U_b is.

Define $S_b = \sum_{a \geq b} U_a$. Given Y^n, S_1 is deterministic. Given Y^n and $U_1, .., U_{b-1}$, S_b is deterministic.

Using Lemma 6, since $\mathbb{E}[S_b] = n \sum_{a \geq b} \pi_a p_{ak} \geq n\pi_q p_{qk} > 2\gamma \log n$ for some γ (chosen later) and all b for large enough n, we have

$$\mathbb{P}\left[S_b > \frac{1}{5} n \sum_{a \geq b} \pi_a p_{ak}\right] \geq 1 - 2/n^{\gamma/4}.$$

Using the union bound,

$$\mathbb{P}\left[\bigcap_{b=1}^{q} \left\{S_b > \frac{1}{5} n \sum_{a \geq b} \pi_a p_{ak}\right\}\right] \geq 1 - 2q/n^{\gamma/4}. \tag{25}$$

Next, for any y^n which has positive probability given $A = 1$,

$$\begin{aligned} &\mathbb{P}[A = 1|Y^n = y^n] \\ &\quad = \mathbb{P}\left[\bigcap_{b=1}^{q} U_b = v_b(y^n) \middle| Y^n = y^n\right] \\ &\quad = \prod_{b=1}^{q-1} \mathbb{P}\left[U_b = v_b(y^n) \middle| \bigcap_{a=1}^{b-1} U_a = v_a(y^n), Y^n = y^n\right]. \end{aligned}$$

We compute the following which is like the proof Lemma 3 of but with appropriate adjustments. Using (16),

$$\begin{aligned} &\mathbb{P}\left[U_b = v_b(y^n) \middle| \bigcap_{a=1}^{b-1} U_a = v_a(y^n), Y^n = y^n\right] \\ &\qquad \leq \frac{\alpha}{\sqrt{\sum_{i=1}^{S_b} \min\left\{\frac{\pi_b p_{bk}}{\sum_{a \geq b} \pi_a p_{ak}}, \frac{\sum_{a > b} \pi_a p_{ak}}{\sum_{a \geq b} \pi_a p_{ak}}\right\}}}. \end{aligned}$$

Like in Lemma 3, define $c_- = \min_i p_{ik}$.

$$\begin{aligned} &\min\left\{\frac{\pi_b p_{bk}}{\sum_{a \geq b} \pi_a p_{ak}}, \frac{\sum_{a > b} \pi_a p_{ak}}{\sum_{a \geq b} \pi_a p_{ak}}\right\} \\ &\quad = (\min_i p_{ik}) \min\left\{\frac{\pi_b}{\sum_{a \geq b} \pi_a p_{ak}}, \frac{\sum_{a > b} \pi_a}{\sum_{a \geq b} \pi_a p_{ak}}\right\} \\ &\quad = \frac{c_- \pi_b}{\sum_{a \geq b} \pi_a p_{ak}}. \end{aligned}$$

We get the last equality since π_i is in increasing order. Hence

$$\begin{aligned} &\mathbb{P}\left[U_b = v_b(y^n) \middle| \bigcap_{a=1}^{b-1} U_a = v_a(y^n), Y^n = y^n\right] \\ &\qquad \leq \frac{\alpha}{\sqrt{S_b \frac{c_- \pi_b}{\sum_{a \geq b} \pi_a p_{ak}}}}. \end{aligned}$$

We can now compute

$$\begin{aligned} &\mathbb{E}[\log \mathbb{P}[A = 1|Y^n]|A = 1] \\ &\quad = \sum_{y^n} \mathbb{P}[Y^n = y^n|A = 1] \log \mathbb{P}[A = 1|Y^n = y^n] \\ &\quad \leq \log \sum_{y^n} \mathbb{P}[Y^n = y^n|A = 1]\mathbb{P}[A = 1|Y^n = y^n] \\ &\quad \leq \log \sum_{y^n} \mathbb{P}[Y^n = y^n|A = 1] \prod_{b=1}^{q-1} \frac{\alpha}{\sqrt{S_b \frac{c_- \pi_b}{\sum_{a \geq b} \pi_a p_{ak}}}} \\ &\quad \leq \log \Bigg((2q/n^{\gamma/4}) + (1 - 2q/n^{\gamma/4}) \\ &\qquad \prod_{b=1}^{q-1} \frac{\alpha}{\sqrt{\left(\frac{1}{5} n \sum_{a \geq b} \pi_a p_{ak}\right) \frac{c_- \pi_b}{\sum_{a \geq b} \pi_a p_{ak}}}} \Bigg) \qquad (26) \\ &\quad \leq \log \left((2q/n^{\gamma/4}) + \frac{\alpha^q}{\left(\frac{c_-}{5}\right)^{\frac{q-1}{2}} n^{\frac{q-1}{2}} \sqrt{\frac{\prod_{b=1}^{q} \pi_b}{\pi_{max}}}} \right) \end{aligned}$$

where in (26) we used (25). We can pick γ large enough[8] so that the first term in the logarithm is negligible compared to the second term for large n.

This gives

$$\mathbb{P}[\log \mathbb{P}[\tilde{A} = 1|\tilde{Y}^n = Y^n]|A = 1]$$

[8] For instance, we can pick $\gamma = 40q$. For large enough n, we still get that $\mathbb{E}[S_b] = n \sum_{a \geq b} \pi_a p_{ak} > 2\gamma \log n$ is true for all b.

$$\leq \log\left(\frac{2\alpha^q}{\left(\frac{c_-}{5}\right)^{\frac{q-1}{2}} n^{\frac{q-1}{2}} \sqrt{\frac{\prod_{b=1}^q \pi_b}{\pi_{max}}}}\right)$$
$$\leq \frac{1}{2}\log n - \sum_{b=1}^{q}\frac{1}{2}\log \pi_b n + c'.$$

The value c' collects all the constants. Combining with Lemma 2, we get that for the q-ary erasure channel and sufficiently large n that

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

where c does not depend on n or π. Using Proposition 1 completes the converse bound for the proof.

The matching achievability bound needed to get the final capacity result is given in [1]. □

C. Z-Channel

The matrix for the Z-channel [4, p 225] is

$$\begin{bmatrix} 1 & 0 \\ p_{21} & p_{22} \end{bmatrix}$$

where we require that $p_{ij} > 0$. (Typically, $p_{21} = p_{22} = 1/2$, but we consider a more general case here.)

We can actually get the capacity of the noisy permutation channel with the Z-channel without altering the proof for the q-ary erasure channels. The transition matrix for the Z-channel can be written as

$$\begin{bmatrix} p_{11} & 0 & p_{13} \\ p_{21} & p_{22} & 0 \end{bmatrix}$$

setting $p_{13} = 0$. This does not change the rank of the matrix or the analysis in the proof.

Corollary 2: Let $P_{Z|X}$ be a stochastic matrix for the Z-channel, then

$$C_{\text{perm}}(P_{Z|X}) = \frac{1}{2}.$$

This is Item 3 of Theorem 3.

D. "Zigzag" Channel

In this section, we explore the limits of our approach. We have a particular DMC matrix which is similar to the q-ary erasure channel, but our method is not known to give a tight converse. We use a matrix which could be considered a q-ary Z-channel and call it a "zigzag" channel since its edges in a transition diagram form a zigzag.

The matrix has the form:

$$\begin{bmatrix} p_{11} & p_{12} & 0 & \cdots & 0 & 0 \\ 0 & p_{22} & p_{23} & \cdots & 0 & 0 \\ 0 & 0 & p_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{q-1,q-1} & p_{q-1,q} \\ 0 & 0 & 0 & \cdots & 0 & p_{qq} \end{bmatrix}$$

where each $p_{ij} > 0$. This matrix has rank q.

Suppose that q is odd and that π is such that π_i is 0 for all even values of i. Following the notation and method in Proposition 3, $\mathbb{P}[A = 1|Y^n] = 1$, since any output symbol can be decoded to exactly one input symbol. For any π of this choice,

$$\begin{aligned} &D(P_{Y|X}\|Q_Y^n) \\ &\quad = -nD(P_Y\|Q_Y) - \frac{1}{2}\log n + \sum_{i:\pi_i>0}\frac{1}{2}\log \pi_i n \\ &\qquad + c + \mathbb{E}[\log \mathbb{P}[A=1|Y^n]|A=1] \\ &\quad = -nD(P_Y\|Q_Y) - \frac{1}{2}\log n + \sum_{i:\pi_i>0}\frac{1}{2}\log \pi_i n + c \\ &\quad \leq -nD(P_Y\|Q_Y) - \frac{1}{2}\log n + \frac{q+1}{2}\frac{1}{2}\log n + c \\ &\quad \leq -nD(P_Y\|Q_Y) + \frac{q-1}{4}\log n + c. \end{aligned}$$

If π of this form is the worst case π to use, meaning it gives the largest possible value of $D(P_{Y|X} \circ U\|Q_{Y^n})$ for any Q_Y, then we get that

$$C_{\text{perm}}(P_{Z|X}) \leq \frac{q-1}{2} + \frac{q-1}{4} = \frac{3(q-1)}{4}. \quad (27)$$

If there is another π which is the worst, then our upper bound on the capacity is larger than the value on the right-hand side of (27). In either case, there is a gap between our upper bound for the capacity and the lower bound of $(q-1)/2$ given by (2). Exploring this gap is an opportunity for future work.

Acknowledgment

The authors would like to thank A. Makur for inspiring this project and for his discussions with them.

References

[1] A. Makur, "Coding theorems for noisy permutation channels," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 672–6748, Nov. 2020.

[2] Y. Polyanskiy and Y. Wu, *Lecture Notes on Information Theory*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf

[3] A. Adler, J. Tang, and Y. Polyanskiy, "Quantization of random distributions under KL divergence," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2762–2767.

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

[5] A. J. Stam, "Distance between sampling with and without replacement," *Statistica Neerlandica*, vol. 32, no. 2, pp. 81–91, Jun. 1978.

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

[7] J. M. Walsh, S. Weber, and C. W. Maina, "Optimal rate delay tradeoffs for multipath routed and network coded networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 682–686.

[8] M. Kovacevic and D. Vukobratovic, "Subset codes for packet networks," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 729–732, Apr. 2013.

[9] M. Kovacevic and D. Vukobratović, "Perfect codes in the discrete simplex," *Des., Codes Cryptogr.*, vol. 75, no. 1, pp. 81–95, Nov. 2013.

[10] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 1, no. 3, pp. 230–248, Sep. 2015.

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

[12] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental limits of DNA storage systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 3130–3134.

[13] M. Kovacevic and V. Y. F. Tan, "Codes in the space of multisets—Coding for permutation channels with impairments," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5156–5169, Jul. 2018.

[14] I. Shomorony and R. Heckel, "Capacity results for the noisy shuffling channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 762–766.

[15] S. N. Diggavi and M. Grossglauser, "On transmission over deletion channels," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 2001, vol. 39. no. 1, pp. 573–582.

[16] M. Mitzenmacher, "Polynomial time low-density parity-check codes with rates very close to the capacity of the q-ary random deletion channel for large q," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5496–5501, Dec. 2006.

[17] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[18] J. Tang, "Divergence Covering," Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2021.

[19] K. Marton, "A simple proof of the blowing-up lemma (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 3, pp. 445–446, May 1986.

[20] D. Berend and A. Kontorovich, "A sharp estimate of the binomial mean absolute deviation with applications," *Statist. Probab. Lett.*, vol. 83, no. 4, pp. 1254–1259, Apr. 2013.

[21] H. Robbins, "A remark on Stirling's formula," *Amer. Math. Monthly*, vol. 62, no. 1, pp. 26–29, Jan. 1955.

[22] V. V. Petrov, *Sums of Independent Random Variables* (Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge/A Series of Modern Surveys in Mathematics). Berlin, Germany: Springer, 2012.

[23] J. Duchi, "Lecture notes for statistics 311/electrical engineering 377," Stanford Statistics 311/EE 377, 2021.

[24] F. Matus, "Urns and entropies revisited," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1451–1454.

[25] I. Csiszar and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1007–1016, Mar. 2006.

[26] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2018.

Jennifer Tang (Member, IEEE) received the B.S.E. degree in electrical engineering from Princeton University and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT). She received the Ph.D. degree while working at LIDS. She is currently a Post-Doctoral Associate with the Institute for Data, Systems, and Society (IDSS), MIT. Her research interests include information theory, prediction and learning theory, quantization and data compression, high-dimensional statistics, data analytics, defect tolerance, and models for social dynamics and inference.

Yury Polyanskiy (Senior Member, IEEE) received the M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 2005, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2010. He is currently a Professor of electrical engineering and computer science and a member of the Laboratory for Information and Decision Systems (LIDS), the Center of Statistics, and Institute for Data, Systems, and Society (IDSS), MIT. His research interests include information theory, statistical machine learning, error-correcting codes, wireless communication, and fault tolerance. He was a recipient of the 2011 IEEE Information Theory Society Paper Award, the 2013 NSF CAREER Award, and the 2020 IEEE Information Theory Society James Massey Award.