Information Capacity of BSC and BEC Permutation Channels

Anuran Makur
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA, USA
Email: a_makur@mit.edu

Abstract—In this paper, we describe and study the permutation channel model, which constitutes a discrete memoryless channel (DMC) followed by a random permutation block that reorders the output codeword of the DMC. This model naturally emerges in the context of communication networks, and coding theoretic aspects of such channels have been widely studied. In contrast to the bulk of this literature, we analyze the information theoretic aspects of the model by defining an appropriate notion of permutation channel capacity. We consider two special cases of the permutation channel model: the binary symmetric channel (BSC) and the binary erasure channel (BEC). We establish the permutation channel capacity of the BSC, and prove bounds on the permutation channel capacity of the BEC. Somewhat surprisingly, our results illustrate that permutation channel capacities are generally agnostic to the parameters that define the DMCs. Furthermore, our achievability proof yields a conceptually simple, computationally efficient, and capacity achieving coding scheme for the BSC permutation channel.

I. INTRODUCTION

In this paper, we define and analyze a pertinent notion of information capacity for permutation channels. Permutation channels refer to discrete memoryless channels followed by random permutation transformations that are applied to the entire blocklength of the output codeword. Such channels can be perceived as models of communication links where packets are not delivered in sequence, and hence, the ordering of the packets does not contain any information. The ensuing subsections provide some background literature to motivate our analysis, a formal description of the model, and an outline of our main results, respectively.

A. Related Literature and Motivation

The setting of channel coding with transpositions, where the output codeword undergoes some reordering of its symbols, has been widely studied in both the coding theory and the communication networks communities. For example, in the coding theory literature, one earlier line of work concerned the construction of error-correcting codes that achieve capacity of the random deletion channel. The random deletion channel operates on the codeword space by deleting each input codeword symbol independently with some probability $p \in (0, 1)$, and copying it otherwise. As explained in [1, Section I], with sufficiently large alphabet size $q = 2^b$ (where each symbol can be construed as a packet with $b$ bits), “embedding sequence numbers into the transmitted symbols [turns] the deletion channel [into a memoryless] erasure channel.” Since the erasure channel setting was well-understood, the intriguing question became to construct (nearly) capacity achieving codes for the deletion channel using sufficiently large packet length $b$, but without embedding sequence numbers (see [1], [2], and the references therein). In particular, the author of [1] demonstrated that low density parity check (LDPC) codes with verification-based decoding formed a computationally tractable coding scheme with these properties. This scheme also tolerated transpositions of packets that were not deleted in the process of transmission. Therefore, it was equivalently a coding scheme for a memoryless erasure channel followed by a random permutation block (albeit with very large alphabet size). Several other coding schemes for erasure permutation channels with sufficiently large alphabet size have also been developed in the literature; see e.g. [3], which builds upon the key conceptual ideas in [1], and the references therein.

This discussion regarding the random deletion channel has a patent counterpart in the communication networks literature. Indeed, in the context of the well-known store-and-forward transmission scheme for networks, packet losses (or deletions) were typically corrected using Reed-Solomon codes which assumed that each packet carried a header with a sequence number—see e.g. [4], [5, Section I], and the references therein. Much like in the deletion channel setting, this simplified the error correction problem since packet losses could be treated as erasures. However, “motivated by networks whose topologies change over time, or where several routes with unequal delays are available to transmit the data,” the author of [5] illustrated that packet errors and losses could also be corrected using binary codes under a channel model where the impaired or lost packets were randomly permuted, and the packets were not indexed.

Several other aspects of permutation channels have also been investigated in the communication networks literature. For instance, the permutation channel was a useful model for point-to-point communication between a source and a receiver that used a lower level multipath routed network. Indeed, since packets (or symbols) could take different paths to the receiver in such a network, they would arrive at the
destination out of order due to different delay profiles in the different paths. The authors of [6] established rate-delay tradeoffs for such communication networks, although they neglected to account for packet impairments such as deletions in their analysis for simplicity.

More recently, inspired by packet networks such as mobile ad hoc networks (where the network topology changes over time), and heavily loaded datagram-based networks (where packets are often re-routed for load balancing purposes), the authors of [7], [8], and [9] have considered the general problem of coding in channels where the codeword undergoes a random permutation and is subjected to impairments such as insertions, deletions, substitutions, and erasures. As stated in [7, Section I], the general strategy to communicate across a channel that applies a transformation to its codewords is to “encode the information in an object that is invariant under the given transformation.” In the case of the permutation channel, the appropriate codes are the so-called multiset codes (where the codewords are characterized by their empirical distribution over the underlying alphabet). The existence of certain perfect multiset codes is established in [8], and several other multiset code constructions based on lattices and Sidon sets are analyzed in [9].

An alternative motivation for analyzing permutation channels stems from the study of DNA-based storage systems, cf. [10] and the references therein. The authors of [10] examined the storage capacity of systems where the source is encoded via DNA molecules. These molecules are cached in an unordered fashion akin to the effect of a permutation channel. However, as stated in [9, Section I-B], the model for such systems also differs from our model since the receiver samples the stored codewords with replacement and without errors. We refer readers to the comprehensive bibliography in [9] for other related literature on permutation channels.

As the discussion heretofore reveals, the majority of the literature on permutation channels analyzes its coding theoretic aspects. In contrast, we approach these channels from a purely information theoretic perspective. To our knowledge, there are no known results on the information capacity of the permutation channel model described in the next subsection. (Indeed, while the aforementioned references [2] and [10] have a more information theoretic focus, they analyze different models to ours.) In this paper, we will prove some initial results towards a complete understanding of the information capacity of permutation channels. Rather interestingly, our achievability proofs will automatically yield computationally tractable codes for communication through certain permutation channels, thereby rendering the need for developing sophisticated coding schemes for these channels futile when achieving capacity is the sole objective.

B. Permutation Channel Model

We define the point-to-point permutation channel model in analogy with standard information theoretic definitions, cf. [11, Section 7.5]. Let \( M \in \mathcal{M} \triangleq \{1, \ldots, |\mathcal{M}|\} \) be a message random variable that is drawn uniformly from \( \mathcal{M} \). \( f_n : \mathcal{M} \rightarrow \mathcal{X}^n \) be a (possibly randomized) encoder where \( \mathcal{X} \) is the finite input alphabet of the channel, and \( g_n : \mathcal{Y}^m \rightarrow \mathcal{M} \) be a (possibly randomized) decoder where \( \mathcal{Y} \) is the finite output alphabet of the channel. The message \( M \) is first encoded into a codeword \( X^n_1 = f_n(M) \), where each \( X_i \in \mathcal{X} \), and we use the notation \( X^n_i \triangleq (X_i, \ldots, X_j) \) for \( i < j \). This codeword is transmitted through a discrete memoryless channel (DMC) defined by the conditional probability distributions \( \{P_{Z|X}(\cdot|x) : x \in \mathcal{X}\} \) on \( \mathcal{Y} \) to produce \( Z^n_1 \in \mathcal{Y}^m \), where \( Z_i \in \mathcal{Y} \). The memorylessness property of the DMC implies that:

\[
\forall x^n_i \in \mathcal{X}^n, \ Z^n_i \in \mathcal{Y}^m, \ P_{Z^n_i|X^n}(z^n_i|x^n_i) = \prod_{i=1}^{n} P_{Z_i|X}(z_i|x_i). \tag{1}
\]

The noisy codeword \( Z^n_i \) is then passed through a random permutation transformation to generate \( Y^n_1 \in \mathcal{Y}^nm \), i.e. for a randomly and uniformly chosen permutation \( \pi : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\} \) (which is independent of everything else), each \( Y_i = Z_{\pi(i)} \) for all \( i \in \{1, \ldots, n\} \). Finally, the received codeword \( Y^n_1 \) is decoded to produce an estimate \( M = g_n(Y^n_1) \) of \( M \). Figure 1 illustrates this communication system.

Let the average probability of error in this model be

\[
P_{error} \triangleq \mathbb{P}(M \neq \hat{M}). \tag{2}
\]

The “rate” of the encoder-decoder pair \( (f_n, g_n) \) is defined as:

\[
R \triangleq \frac{\log(|\mathcal{M}|)}{\log(n)}, \tag{2}
\]

where \( \log(\cdot) \) denotes the binary logarithm throughout this paper (and all Shannon entropy, \( H(\cdot) \), and mutual information, \( I(\cdot;\cdot) \), terms are measured in bits). So, we can also write \( |\mathcal{M}| = n^R \). (Strictly speaking, \( n^R \) should be an integer, but we will neglect this detail since it will not affect our results.) We will say that a rate \( R \geq 0 \) is achievable if there is a sequence of encoder-decoder pairs \( \{(f_n, g_n)\}_{n \in \mathbb{N}} \) such that \( \lim_{n \rightarrow \infty} P_{error} = 0 \), where \( \mathbb{N} \triangleq \{1, 2, 3, \ldots\} \). Lastly, we operationally define the permutation channel capacity as:

\[
C_{perm}(P_{Z|X}) \triangleq \sup\{R \geq 0 : R \text{ is achievable}\}. \tag{3}
\]
It is straightforward to verify that the scaling in (2) is indeed $\log(n)$ rather than the standard $n$. As mentioned earlier, due the random permutation in the model, all information embedded in the ordering within codewords is lost. (In fact, canonical fixed composition codes cannot carry more than one message in this setting.) So, the maximum number of decodable messages is upper bounded by the number of possible empirical distributions of $Y^n_1$:

$$n^R = |M| \leq \left( \frac{n + |Y| - 1}{|Y| - 1} \right) \leq (n + 1)^{|Y| - 1} \quad (4)$$

where taking log’s and letting $n \to \infty$ yields $C_{perm}(P_z|x) \leq |Y| - 1$ (at least non-rigorously). This justifies that $\log(n)$ is the correct scaling in (2).

C. Outline

In the remainder of this paper, we will consider two canonical specializations of the aforementioned permutation channel model: the case where the DMC is a binary symmetric channel (BSC), and the case where it is a binary erasure channel (BEC). In the context of [7], [8], and [9], the former case corresponds to permutation channels with substitution errors, and the latter case corresponds to permutation channels with erasures (or equivalently, deletions—see [9, Remark 1]). We will establish the permutation channel capacity of the BSC exactly in section II. In particular, our achievability proof will follow from a binary hypothesis testing result that will be derived using the second moment method. Then, we will prove bounds on the permutation channel capacity of the BEC in section III. Finally, we will conclude our discussion and propose future research directions in section IV.

II. PERMUTATION CHANNEL CAPACITY OF BSC

In this section, we let $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$ within the formalism of subsection I-B. Moreover, we let the DMC be the binary symmetric channel, which is defined by the conditional distributions:

$$\forall z, x \in \{0, 1\}, \ P_{Y|X}(z|x) = \begin{cases} 1 - p, & \text{for } z = x \\ p, & \text{for } z \neq x \end{cases} \quad (5)$$

where $p \in [0, 1]$ is the crossover probability. We denote such a BSC as $BSC(p)$.

A. Auxiliary Lemmata

To derive the permutation channel capacity of BSCs, we first prove a useful lemma.

**Lemma 1** (Testing between Converging Hypotheses). Fix any $n \in \mathbb{N}$, and constants $\epsilon_n \in (0, \frac{1}{2})$ and $p_n \in (0, 1 - (1/n^{\frac{1}{2} - \epsilon_n}))$ that can depend on $n$. Consider a binary hypothesis testing problem with hypothesis random variable $H \sim \text{Ber}(\frac{1}{2})$ (i.e. uniform Bernoulli prior), and likelihoods $P_{X|H=0} = \text{Ber}(p_n)$ and $P_{X|H=1} = \text{Ber}(p_n + (1/n^{\frac{1}{2} - \epsilon_n}))$ on the alphabet $\mathcal{X} = \{0, 1\}$, such that we observe $n$ samples $X^n_1$ that are drawn conditionally i.i.d. given $H$ from the likelihoods:

- Given $H = 0 : X^n_1 \overset{i.i.d.}{\sim} P_{X|H=0} = \text{Ber}(p_n)$,
- Given $H = 1 : X^n_1 \overset{i.i.d.}{\sim} P_{X|H=1} = \text{Ber}(p_n + \frac{1}{n^\frac{1}{2} - \epsilon_n})$.

Then, the minimum probability of error corresponding to the maximum likelihood (ML) decoder $H^n_{ml} : \{0, 1\}^n \to \{0, 1\}$ for $H$ based on $X^n_1$, $P^n_{ml} \triangleq \mathbb{P}(H^n_{ml}(X^n_1) \neq H)$, satisfies:

$$P^n_{ml} \leq \frac{3}{2n^{2\epsilon_n}}$$

This implies that $\lim_{n \to \infty} P^n_{ml} = 0$ when $\lim_{n \to \infty} n^{\epsilon_n} = +\infty$.

**Proof.** Let $T_n \in T_n = \{ \frac{k}{n} - c_n : k \in \{0, \ldots, n\} \}$ be the translated arithmetic mean of $X^n_1$:

$$T_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i - c_n$$

where the constant $c_n$ (that can depend on $n$) will be chosen later. Moreover, for ease of exposition, let $T^n_n$ and $T^n_0$ be random variables with probability distributions given by the likelihoods $P_{T^n_0} = P_{T^n_0|H=0}$ and $P_{T^n_1} = P_{T^n_1|H=1}$, respectively, such that $P_{T^n_0} = \frac{1}{2} P_{T^n_1} + \frac{1}{2} P_{T^n_1}$. It is straightforward to verify that $T_n$ is a sufficient statistic of $X^n_1$ for $H$. So, the ML decoder of $H$ based on $X^n_1$, $H^n_{ml}(X^n_1)$, is a function of $T_n$ without loss of generality, and we denote it as $H^n_{ml}(T_n)$ (with abuse of notation); in particular, the ML decoder simply thresholds the statistic $T_n$ to detect $H$.

To upper bound $P^n_{ml} = \mathbb{P}(H^n_{ml}(T_n) \neq H)$, recall *Le Cam’s relation* for the ML decoding probability of error, cf. [12, proof of Theorem 2.2(ii)]:

$$P^n_{ml} = \frac{1}{2} \left( 1 - \|P^+_n - P^-_n\|_{TV} \right) \quad (6)$$

where the total variation (TV) distance between two probability measures $P_0$ and $P_1$ on a common measurable space $(\mathcal{S}, \mathcal{F})$ is defined as:

$$\|P_0 - P_1\|_{TV} \triangleq \sup_{A \in \mathcal{F}} |P_0(A) - P_1(A)| = \frac{\|P_0 - P_1\|_1}{2} \quad (7)$$

where $\|\cdot\|_1$ is the $L^1$-norm. Furthermore, recall the so called second moment method lower bound on TV distance—see e.g. [13, Lemma 4.2(iii)]:

$$\|P^+_n - P^-_n\|_{TV} \geq \frac{1}{4} \sum_{t \in T_n} \left( \frac{P_{T^n_1|H}(t|0) - P_{T^n_1|H}(t|1)}{P_{T^n_1}(t)} \right)^2 \quad \text{Vince-Le Cam distance} \quad (8)$$

$$\geq \frac{(E[T^n_1] - E[T^n_0])^2}{4E[T^n_0]^2} \quad (9)$$

where (8) lower bounds $\|P^+_n - P^-_n\|_{TV}$ using the Vince-Le Cam distance between $P^+_n$ and $P^+_n$ (via the observation that $|P_{T^n_1|H}(t|0) - P_{T^n_1|H}(t|1)| \leq P_{T^n_1|H}(t|0) + P_{T^n_1|H}(t|1)$ for all $t \in T_n$), and (9) follows from the Cauchy-Schwarz inequality.
We now select the constant \( c_n \). Since both \( \|P^+_n - P^-_n\|_{TV} \) and the numerator of (9) are invariant to the value of \( c_n \), the best bound of the form (9) is obtained by minimizing the second moment \( E[T^2_n] \). Thus, \( c_n \) is given by:

\[
    c_n = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = p_n + \frac{1}{2n^2} - \epsilon_n
\]

(10)

using the facts that \( X^+_n \overset{i.i.d.}{\sim} \text{Ber}(p_n) \) given \( H = 0 \), and \( X^-_n \overset{i.i.d.}{\sim} \text{Ber}(p_n + (1/n^2 - \epsilon_n)) \) given \( H = 1 \). This ensures that \( E[T_n] = 0 \), and the denominator in (9) is \( E[T^2_n] = \text{VAR}(T_n) \), where \( \text{VAR}(\cdot) \) denotes variance. We remark that with this choice of \( c_n \), (9) can be perceived as a Hammersley-Chapman-Robbins bound [14], [15], where the Vincze-Le Cam distance replaces the usual \( \chi^2 \)-divergence.

Combining (6) and (9) yields the following upper bound on the ML decoding probability of error \( P_{ML}^n \):

\[
    P_{ML}^n \leq \frac{1}{2} \left( 1 - \frac{(E[T^+_n] - E[T^-_n])^2}{4E[T^2_n]} \right) \tag{11}
\]

which we now compute explicitly. Observe using (10) that:

\[
    E[T^+_n] = \frac{1}{n} \sum_{i=1}^{n} X_i - c_n | H = 1 = \frac{1}{2n^2} - \epsilon_n,
\]

\[
    E[T^-_n] = \frac{1}{n} \sum_{i=1}^{n} X_i - c_n | H = 0 = -\frac{1}{2n^2} - \epsilon_n.
\]

Furthermore, using (10), we also get:

\[
    E[T^2_n] = \frac{1}{n^2} \left( \sum_{i=1}^{n} X_i - E[X_i] \right) \left( \sum_{i=1}^{n} X_i - E[X_i] \right) \sum_{i=1}^{n} X_i - E[X_i] \]

\[
    = \frac{1}{n} \text{VAR}(X_1) + \frac{(n-1)}{n} \text{COV}(X_1, X_2)
\]

\[
    = \frac{1}{n} \left( p_n + \frac{1}{2n^2} - \epsilon_n \right) \left( 1 - p_n - \frac{1}{2n^2} - \epsilon_n \right)
\]

\[
    + \left( \frac{n-1}{n} \right) \left( \frac{1}{2} p_n^2 + \frac{1}{2} \left( p_n + \frac{1}{n^2} - \epsilon_n \right) \right)
\]

\[
    - \left( p_n + \frac{1}{2n^2} - \epsilon_n \right)^2
\]

\[
    = p_n(1-p_n) + \frac{1}{2n^2} - \epsilon_n + \frac{n-2}{4n^2 - 2n^2}
\]

\[
    \leq \frac{1}{4n} + \frac{1}{2n^2 - \epsilon_n} + \frac{1}{4n^2 - 2n^2}
\]

where \( \text{COV}(\cdot) \) denotes covariance, and the final inequality follows from the bounds \( p_n(1-p_n) \leq \frac{1}{4}, 1 - 2p_n \leq 1 \), and \( n - 2 \leq n \). Plugging these expressions into (11), we have:

\[
    P_{ML}^n \leq \frac{1}{2} \left( 1 - \frac{\left( \frac{1}{n^2} - \epsilon_n \right)^2}{4 \left( \frac{1}{4n} + \frac{1}{2n^2 - \epsilon_n} + \frac{1}{4n^2 - 2n^2} \right)} \right)
\]

\[
    = \frac{1}{2} \left( 1 - \frac{1}{1 + \frac{1}{n^2} - \epsilon_n} + \frac{2}{1 + \frac{1}{n^2} - \epsilon_n} \right)
\]

\[
    \leq \frac{1}{2} \left( 1 - \frac{1}{1 + \frac{1}{n^2} - \epsilon_n} + \frac{3}{1 + \frac{1}{n^2} - \epsilon_n} \right)
\]

\[
    \leq \frac{3}{2n^2} - \epsilon_n
\]

(12)

where the inequality in the second line holds because \( \epsilon_n < \frac{1}{2} \).

This completes the proof.

This lemma illustrates that as long as the difference between the parameters that define \( P_X|H=0 \) and \( P_X|H=1 \) vanishes slower than \( \Theta(1/\sqrt{n}) \), we can decode the hypothesis \( H \) with vanishing probability of error as \( n \to \infty \). Intuitively, this holds because the standard deviation of the sufficient statistic \( T_n \) is \( O(1/\sqrt{n}) \) (neglecting \( \epsilon_n \)). So, as long as the difference between the two parameters is \( o(1/\sqrt{n}) \), it is possible to distinguish between the two hypotheses. We also remark that tighter upper bounds on \( P_{ML}^n \) can be obtained using standard exponential concentration of measure inequalities. However, the simpler second moment method approach suffices for our purposes.

We will also require the following useful estimate of the entropy of a binomial distribution from the literature in our converse proof of the permutation channel capacity of BSCs.

Lemma 2 (Approximation of Binomial Entropy [16], Equation (7)). Given a binomial random variable \( X \sim \text{Bin}(n, p) \) with \( n \in \mathbb{N} \) and \( p \in (0, 1) \), we have:

\[
    |H(X) - \frac{1}{2} \log(2\pi e n p (1-p))| \leq \frac{c(p)}{n}
\]

for some constant \( c(p) \geq 0 \).

B. Derivation of Permutation Channel Capacity

We next present our first main result, which exploits Lemmata 1 and 2 to derive the permutation channel capacity of BSCs.

Theorem 3 (Permutation Channel Capacity of BSC).

\[
    C_{perm}(\text{BSC}(p)) = \begin{cases} 
    1, & \text{for } p = 0, 1 \\
    \frac{1}{2}, & \text{for } p \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right) \\
    0, & \text{for } p = \frac{1}{2} 
    \end{cases}
\]

Proof.

Converse for \( p \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right) \): Suppose we are given a sequence of encoder-decoder pairs \( \{(f_n, g_n)\} \in \mathcal{E} \) on message sets of size \( |M| = n^R \) such that \( \lim_{n \to \infty} P_{\text{error}}^n = 0 \). Consider the Markov chain \( M \to X^n \to Z^n \to Y^n \to S_n \). Let \( Y^n \) be the sufficient statistic of \( Y^n \) for \( M \) [17, Theorem 3.6]. Then, following the standard argument from [11, Section 7.9], we have:

\[
    R \log(n) = H(M) = H(M|\hat{M}) + I(M; \hat{M})
\]

\[
    \leq 1 + P_{\text{error}}^n R \log(n) + I(M; \hat{Y}^n)
\]

\[
    = 1 + P_{\text{error}}^n R \log(n) + I(M; S_n)
\]

\[
    \leq 1 + P_{\text{error}}^n R \log(n) + I(X^n; S_n)
\]

(13)
where the first equality holds because $M$ is uniformly distributed, the second line follows from Fano’s inequality and the data processing inequality [11, Theorems 2.10.1 and 2.8.1], the third line holds because $S_n$ is a sufficient statistic, cf. [11, Section 2.9], and the last line also follows from the data processing inequality [11, Theorem 2.8.1].

We now upper bound $I(X^n_1; S_n)$. Notice that:

$$I(X^n_1; S_n) = H(S_n) - H(S_n|X^n_1)$$

$$\leq \log(n + 1) - \sum_{x^n_1 \in \{0, 1\}^n} P_{X^n_1}(x^n_1) H(S_n|X^n_1 = x^n_1)$$

$$(14)$$

where we use the fact that $S_n \in \{0, \ldots, n\}$. Given $X^n_1 = x^n_1$ for any fixed $x^n_1 \in \{0, 1\}^n$, the third line holds because $S_n$ is a sufficient statistic (almost surely). Hence, we have:

$$H(S_n|X^n_1 = x^n_1) = H\left(\sum_{i=1}^{k} A_i + \sum_{j=k+1}^{n} B_j\right)$$

$$\geq \max\left\{ H\left(\sum_{i=1}^{k} A_i\right), H\left(\sum_{j=k+1}^{n} B_j\right)\right\}$$

$$(15)$$

where $k = |x^n_1|_H$, $A_i \overset{i.i.d.}{\sim} \text{Ber}(p - 1)$, and $B_i \overset{i.i.d.}{\sim} \text{Ber}(p)$ are independent, and the inequality follows from [11, Problem 2.14]. (Note that if $k \in \{0, n\}$, then one of the summations above is trivially 0, and its entropy is also 0.) Since $\max\{k, n - k\} \geq \frac{n}{2}$, we can use Lemma 2 to get:

$$H(S_n|X^n_1 = x^n_1) \geq \frac{1}{2} \log(\pi e p(1 - p)n) - \frac{2c(p)}{n}$$

which we can substitute into (14) and obtain:

$$I(X^n_1; S_n) \leq \log(n + 1) - \frac{1}{2} \log(\pi e p(1 - p)n) + \frac{2c(p)}{n}$$

Combining (13) and (15), and dividing by $\log(n)$, yields:

$$R \leq P^n\text{error} R + \frac{1 + \log(n + 1)}{\log(n)} - \frac{\log(\pi e p(1 - p)n)}{2 \log(n)} + \frac{2c(p)}{n \log(n)}$$

where letting $n \to \infty$ produces $R \leq \frac{1}{2}$. Therefore, we have $C_{\text{perm}}(\text{BSC}(p)) \leq \frac{1}{2}$.

**Converse for $p = \frac{1}{2}$:** The output of the BSC is independent of the input, and $I(X^n_1; S_n) = 0$. So, dividing both sides of (13) by $\log(n)$ yields:

$$R \leq \frac{1}{\log(n)} + P^n\text{error} R$$

where letting $n \to \infty$ produces $R \leq 0$. Therefore, we have $C_{\text{perm}}(\text{BSC}(\frac{1}{2})) = 0$.

**Converse for $p \in \{0, 1\}$:** Starting from (14), we get the inequality $I(X^n_1; S_n) \leq \log(n + 1)$, since $H(S_n|X^n_1) = 0$. As before, combining (13) and this inequality, and dividing by $\log(n)$, yields:

$$R \leq \frac{1}{\log(n)} + P^n\text{error} R + \frac{\log(n + 1)}{\log(n)}$$

where letting $n \to \infty$ produces $R \leq 1$. Therefore, we have $C_{\text{perm}}(\text{BSC}(p)) \leq 1$.

**Achievability for $p \in (0, \frac{1}{2})$:** For any $\epsilon \in (0, \frac{1}{2})$, suppose we have:

1. $|\mathcal{M}| = n^{2^{-\epsilon}}$ messages,
2. a randomized encoder $f_n : \mathcal{M} \to \{0, 1\}^n$ such that:

$$\forall m \in \mathcal{M}, \quad f_n(m) = X^n_1 \overset{i.i.d.}{\sim} \text{Ber}\left(\frac{m}{n^{2^{-\epsilon}}}\right),$$

3. an ML decoder $g_n : \{0, 1\}^n \to \mathcal{M}$ such that:

$$\forall y^n_1 \in \{0, 1\}^n, \quad g_n(y^n_1) = \arg \max_{m \in \mathcal{M}} P_{Y^n_1|M}(y^n_1|m)$$

where the tie-breaking rule (when there are many maximizers) does not affect $P^n\text{error}$.

This completely specifies the communication system model in subsection I-B. We now analyze the average probability of error for this simple encoding and decoding scheme.

Let us condition on the event $M = m \in \mathcal{M}$. Then, $X^n_1 \overset{i.i.d.}{\sim} \text{Ber}(m/n^{2^{\epsilon}})$, and $Z^n_1 \overset{i.i.d.}{\sim} \text{Ber}(p*(m/n^{2^{\epsilon}}))$ since the BSC is memoryless, where $r*s = r(1-s) + s(1-r)$ denotes the convolution of $r, s \in [0, 1]$. Moreover, $Y^n_1 \overset{i.i.d.}{\sim} \text{Ber}(p*(m/n^{2^{\epsilon}}))$ because it is the output of passing $Z^n_1$ through a random permutation. The conditional probability that our ML decoder makes an error is upper bounded by:

$$P(M \neq M|M = m) = P_m(g_n(Y^n_1) \neq m) \leq P_m(\exists i \in \mathcal{M}\{m\}, P_{Y^n_1|M}(Y^n_1|i) \geq P_{Y^n_1|M}(Y^n_1|m))$$

$$\leq \sum_{i \in \mathcal{M}(\{m\})} P_m(P_{Y^n_1|M}(Y^n_1|i) \geq P_{Y^n_1|M}(Y^n_1|m))$$

where $P_m$ represents the probability measure after conditioning on the event $M = m$, the second line is an upper bound because we regard the equality case, $P_{Y^n_1|M}(Y^n_1|i) = P_{Y^n_1|M}(Y^n_1|m)$ for $i \neq m$, as an error (even though the ML decoder may return the correct message in this scenario), and the third line follows from the union bound. To show that this upper bound vanishes, for any message $i \neq m$, consider a binary hypothesis test with likelihoods:

Given $H = 0 : Y^n_1 \overset{i.i.d.}{\sim} \text{Ber}\left(p * \frac{m}{n^{2^{\epsilon}}}\right)$

Given $H = 1 : Y^n_1 \overset{i.i.d.}{\sim} \text{Ber}\left(p * \frac{i}{n^{2^{\epsilon}}}\right)$

where the hypotheses $H = 0$ and $H = 1$ correspond to the messages $M = m$ and $M = i$, respectively. The magnitude of the difference between the two Bernoulli parameters is:

$$\left| p * \frac{m}{n^{2^{\epsilon}}} - p * \frac{i}{n^{2^{\epsilon}}} \right| = \frac{1 - 2p |m - i|}{n^{2^{\epsilon}}} = \frac{1}{n^{2^{\epsilon} - \epsilon_n}}$$

where $\epsilon_n = \epsilon + (\log((1 - 2p)|m - i|)/\log(n)) \in (0, \frac{1}{2})$ (for sufficiently large $n$ depending on $p$). Using Lemma 1, if $H \sim \text{Ber}(\frac{1}{2})$, then $P^n_{\text{ML}} = P(H^n_{\text{ML}}(Y^n_1) \neq H)$ satisfies:

$$P^n_{\text{ML}} = \frac{1}{2} P_m(H^n_{\text{ML}}(Y^n_1) = 1) + \frac{1}{2} P_m(H^n_{\text{ML}}(Y^n_1) = 0) \leq \frac{3}{2n^{2^{\epsilon}n}}$$
where we get $P_{Y_1^n|M}(Y_1^n|i) = P_{Y_1^n|M}(Y_1^n|m)$, by assigning $H_{\text{ML}}^n(Y_1^n) = 1$ (which does not affect the analysis of $P_{\text{ML}}^n$ in Lemma 1).

Combining (16) and (17) yields:

$$\mathbb{P}(\hat{M} \neq M | M = m) \leq \sum_{i \in \mathcal{M}\setminus\{m\}} \frac{3}{n^{2\epsilon_n}}$$

where the second inequality holds because $k = m - i$ ranges over a subset of all non-zero integers. Finally, taking expectations with respect to $M$ in (18) produces:

$$P_{\text{error}}^n \leq \frac{\pi^2}{(1 - 2p)^2n^{2\epsilon}}$$

which implies that $\lim_{n \to \infty} P_{\text{error}}^n = 0$. Therefore, the rate $R = \frac{1}{2} - \epsilon$ is achievable for every $\epsilon \in (0, \frac{1}{2})$, and $C_{\text{perm}}(\text{BSC}(p)) \geq \frac{1}{2}$.

**Achievability for $p \in \{0, 1\}$:** Assume without loss of generality that $p = 0$ since a similar argument holds for $p = 1$. In this case, the BSC is just the deterministic identity channel, and we can use the obvious encoder-decoder pair:

1. $|M| = n + 1$ messages,
2. encoder $f_n : \mathcal{M} \to \{0, 1\}^n$ such that:
   $$\forall m \in \mathcal{M}, \quad f_n(m) = (1, \ldots, 1, 0, \ldots, 0),$$
   $$m = 1 \text{'s} \quad n-m + 1 \text{'s}$$
3. decoder $g_n : \{0, 1\}^n \to \mathcal{M}$ such that:
   $$\forall y_1^n \in \{0, 1\}^n, \quad g_n(y_1^n) = 1 + \sum_{i=1}^{n} y_i$$

which achieves $P_{\text{error}}^n = 0$. Hence, the rate:

$$R = \lim_{n \to \infty} \frac{\log(n + 1)}{\log(n)} = 1$$

is achievable, and $C_{\text{perm}}(\text{BSC}(p)) \geq 1$.

We make a few pertinent remarks regarding Theorem 3. Firstly, in the $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ regime, the randomized encoder and ML decoder presented in the achievability proof constitute a computationally tractable coding scheme. Indeed, unlike traditional channel coding, the ML decoder requires at most $O(n)$ likelihood ratio tests in our setup, which means that the decoder operates in polynomial time in $n$. More precisely, the interval $[0, n]$ can be partitioned into subintervals so that each sub-interval is the decoding region for a message in $M$. The ML decoding can be shown to generate the message $\hat{M}$ that corresponds to the decoding region that contains the sufficient statistic $S_n$. Therefore, communication via permutation channels does not require the development of sophisticated coding schemes to achieve capacity. Furthermore, our achievability proof also implies the existence of a good deterministic code using the probabilistic method.

Secondly, although we have presented Theorem 3 under an average probability of error criterion, our proof establishes the permutation channel capacity of a BSC under a maximal probability of error criterion as well; see e.g. (18). More generally, the permutation channel capacity of a DMC remains the same under a maximal probability of error criterion. This follows from a straightforward expurgation argument similar to [18, Theorem 18.3, Corollary 18.1] or [11, Section 7.7, p.204].

Thirdly, in the $p \in \{0, \frac{1}{2}\} \cup (\frac{1}{2}, 1)$ regime, we intuitively expect the rate of decay of $P_{\text{error}}^n$ to be dominated by the rate of decay of the probability of error in distinguishing between two consecutive messages. Although we do not derive precise rates in this paper, Lemma 1 and (19) indicate that this intuition is accurate.

Fourthly, the proof of Lemma 1 (and the discussion following it) portrays that the distinguishability between two consecutive messages is determined by a careful comparison of the difference between means and the variance. This suggests that the central limit theorem (CLT) can be used to obtain (at least informally) the $|M| \approx \sqrt{n}$ scaling in the $p \in \{0, \frac{1}{2}\} \cup (\frac{1}{2}, 1)$ regime. The CLT is in fact implicitly used in our converse proof when we apply Lemma 2, because estimates for the entropy of a binomial distribution can be obtained using the CLT.

Lastly, Theorem 3 illustrates a few somewhat surprising facts about permutation channel capacity. While traditional channel capacity is convex as a function of the channel (with fixed dimensions), permutation channel capacity is clearly non-convex and discontinuous as a function of the channel. Moreover, for the most part, the permutation channel capacity of a BSC does not depend on $p$. This is because the scaling (with $n$) of the difference between the Bernoulli parameters of two encoded messages does not change after passing through the memoryless BSC. However, (19) suggests that $p$ does affect the rate of decay of $P_{\text{error}}^n$.

### III. Permutation Channel Capacity of BEC

In this section, we let $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, \text{e}\}$, where $\text{e}$ denotes the erasure symbol, within the formalism of subsection I-B. Moreover, we let the DMC be the binary erasure channel, which is defined by the conditional distributions:

$$\forall z \in \mathcal{Y}, \forall x \in \mathcal{X}, \quad P_{Z|X}(z|x) = \begin{cases} 1 - \delta, & \text{for } z = x \\ \delta, & \text{for } z = \text{e} \\ 0, & \text{otherwise} \end{cases}$$
where $\delta \in [0,1]$ is the erasure probability. We denote such a BEC as $\text{BEC}(\delta)$. The ensuing theorem establishes bounds on the permutation channel capacity of BECs.

**Theorem 4** (Bounds on Permutation Channel Capacity of BEC). For $\delta \in (0,1)$, we have:

$$\frac{1}{2} \leq C_{\text{perm}}(\text{BEC}(\delta)) \leq 1.$$

Furthermore, the extremal permutation channel capacities are $C_{\text{perm}}(\text{BEC}(0)) = 1$ and $C_{\text{perm}}(\text{BEC}(1)) = 0$.

**Proof.**

**Converse for $\delta \in (0,1)$:** As in the converse proof for BSCs, suppose we are given a sequence of encoder-decoder pairs $\{(f_n, g_n)\}_{n \in \mathbb{N}}$ on message sets of size $|M| = n^R$ such that $\lim_{n \to \infty} P_{\text{error}}^n = 0$. For any $y \in \{0, 1, \epsilon\}$, define the functions $S_n^y : \{0, 1, \epsilon\}^n \to \{0, \ldots, n\}$, $S_n^{y_1}(y_1^n) \triangleq \sum_{i=1}^{n} \mathbb{I}(y_i = y)$, where $\mathbb{I}(y_i = y)$ is an indicator that equals 1 if $y_i = y$ and 0 otherwise. Consider the Markov chain $M \to X^n_1 \to Z^n_1 \to Y^n_1 = (S^n_1(Y^n_1), S^n_2(Y^n_2))$. Observe that for every $y_1^n \in \{0,1,\epsilon\}^n$ and $m \in M$:

$$P_{Y^n_1|M}(y_1^n|m) = \left(\frac{n}{S^n_1(y_1^n), S^n_2(y_1^n)}\right)^{-1} \cdot P(S^n_1(Z^n_1) = S^n_1(y_1^n), S^n_2(Z^n_1) = S^n_2(y_1^n)|M = m),$$

where the first term on the right hand side is the reciprocal of a multinomial coefficient, and $S^n_1(y_1^n) = n - S^n_2(y_1^n) - S^n_2(y_1^n)$. As before, since $P_{Y^n_1|M}(y_1^n|m)$ depends on $y_1^n$ through $S^n_1(y_1^n)$ and $S^n_2(y_1^n)$, the Fisher-Neyman factorization theorem implies that $(S^n_1(Y^n_1), S^n_2(Y^n_2))$ is a sufficient statistic of $Y^n_1$ for $M$. Then, following the standard Fano’s inequality argument (see the derivation of (13)), we get:

$$R \log(n) \leq 1 + P_{\text{error}}^n R \log(n) + I(X^n_1; S^n_1, S^n_2) \tag{21}$$

where we let $S^n_1 = S^n_1(Y^n_1)$ and $S^n_2 = S^n_2(Y^n_2)$ (with abuse of notation). To upper bound $I(X^n_1; S^n_1, S^n_2)$, notice that:

$$I(X^n_1; S^n_1, S^n_2) = I(X^n_1; S^n_2) + I(X^n_1; S^n_1|S^n_2) = H(S^n_1|S^n_2) - H(S^n_1|X^n_1, S^n_2) \leq \log(n + 1) \tag{22}$$

where the first line follows from the chain rule, the second line holds because the number of erasures, $S^n_2 = \sum_{i=1}^{n} \mathbb{I}(Z_i = \epsilon)$ (almost surely), is independent of $X^n_1$, and the third line uses the facts that $S^n_1 \in \{0, \ldots, n\}$ and $H(S^n_1|X^n_1, S^n_2) \geq 0$. Combining (21) and (22), and dividing by $\log(n)$, yields:

$$R \leq \frac{1}{\log(n)} + P_{\text{error}}^n R + \frac{\log(n + 1)}{\log(n)} \tag{23}$$

where letting $n \to \infty$ produces $R \leq 1$. Therefore, we have $C_{\text{perm}}(\text{BEC}(\delta)) \leq 1$.

**Case $\delta = 1$:** In this case, the BEC erases all its input symbols so that $S^n_1 = 0$ and $S^n_2 = n$ almost surely. This implies that $I(X^n_1; S^n_1, S^n_2) = 0$. Hence, dividing both sides of (21) by $\log(n)$ yields:

$$R \leq \frac{1}{\log(n)} + P_{\text{error}}^n R \leq 1$$

where letting $n \to \infty$ produces $R \leq 1$. Therefore, we have $C_{\text{perm}}(\text{BEC}(1)) = 0$.

**Case $\delta = 0$:** In this case, the BEC is just the deterministic identity channel $\text{BEC}(0)$. Hence, $C_{\text{perm}}(\text{BEC}(0)) = 1$ using Theorem 3.

**Achievability for $\delta \in (0,1)$:** For the achievability proof, we employ a useful representation of BSCs using BECs. Observe that a BSC($\frac{\delta}{2}$) can be equivalently construed as a channel that copies its input bit with probability $1 - \delta$, and generates a completely independent Ber($\frac{\delta}{2}$) output bit with probability $\delta$. Indeed, the following decomposition of the BSC’s stochastic transition probability matrix demonstrates this equivalence:

$$\begin{pmatrix}
1 - \frac{\delta}{2} & \frac{\delta}{2} \\
\frac{\delta}{2} & 1 - \frac{\delta}{2}
\end{pmatrix} = \begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix} + \delta \begin{pmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}
\end{pmatrix}.$$

We remark that this idea originates from Fortuin-Kasteleyn random cluster representations of Ising models in the study of percolation, cf. [19], and has been exploited in various other discrete probability contexts such as broadcasting on trees [13, p.412], and reliable computation [20, p.570]. A consequence of this idea is that a BSC($\frac{\delta}{2}$) is statistically equivalent to a BEC($\delta$) followed by a channel that outputs an independent Ber($\frac{\epsilon}{2}$) bit for every input erasure symbol, and copies all other input symbols.

Thus, for our BEC($\delta$) permutation channel, let us use the randomized encoder from the achievability proof for a BSC($\frac{\delta}{2}$) with $\delta \in (0, \frac{\epsilon}{2})$ (in subsection II-B). Furthermore, let us use a randomized decoder which first generates independent Ber($\frac{\epsilon}{2}$) bits to replace every erasure symbol in $Y^n_1$, and then applies the ML decoder from the achievability proof for a BSC($\frac{\epsilon}{2}$) to the resulting codeword (which belongs to $\{0,1\}^n$). By our previous discussion, it is straightforward to verify that the $P_{\text{error}}^n$ for this encoder-decoder pair under the BEC($\delta$) model is equal to the $P_{\text{error}}^n$ analyzed in the achievability proof for a BSC($\frac{\epsilon}{2}$). (We omit the details of this equivalence for brevity.) This portrays that $C_{\text{perm}}(\text{BEC}(\delta)) \geq \frac{1}{2}$ using the achievability result of Theorem 3.

We conjecture that $C_{\text{perm}}(\text{BEC}(\delta)) \geq \frac{1}{2}$ in the $\delta \in (0, 1)$ regime, i.e. the achievability result is tight, and the permutation channel capacities of the BSC and BEC are equal (in the non-trivial regimes of their parameters). Our converse bound, $C_{\text{perm}}(\text{BEC}(\delta)) \leq 1$, is intuitively trivial, since there are only $n+1$ distinct empirical distributions of codewords in $\{0, 1\}^n$ (which non-rigorously shows the upper bound on capacity). So, we believe that this bound can be tightened.

To elucidate the difficulty in improving this bound, consider the expression in (22), which along with the fact that $S^n_1 \in \{0, \ldots, n\}$, produces:

$$I(X^n_1; S^n_1, S^n_2) \leq \log(n + 1) - H(S^n_1|X^n_1, S^n_2). \tag{24}$$

As in the proof of the converse for the BSC, if we can lower bound $H(S^n_1|X^n_1, S^n_2)$ by:

$$H(S^n_1|X^n_1, S^n_2) \geq \frac{1}{2} \log(n) + o(\log(n)) \tag{25}$$
then combining (21), (24), and (25) will yield the desired bound $C_{\text{perm}}(\text{BEC}(\delta)) \leq \frac{1}{2}$. It is straightforward to verify that given $X^*_1 = x^*_n \in \{0,1\}^n$ such that $S_n^R(x^*_1) = m \in \{0, \ldots, n\}$, and $S_n^L = k \in \{0, \ldots, n\}$, $S_n^R$ has a hypergeometric distribution:

$$P_{S_n^R|X^*_1, S_n^L}(r|x^*_1, k) = \binom{m}{r} \binom{n-m}{n-k-r} \binom{n}{k}$$

for every $\max\{0, m-k\} \leq r \leq \min\{m, n-k\}$. Furthermore, $S_n^R \sim \text{bin}(n, \delta)$ because $\{1\{Z_i = e\} : i \in \{1, \ldots, n\}\}$ i.i.d. $\text{Ber}(\delta)$ for a memoryless $\text{BEC}(\delta)$, and $S_n^R$ is independent of $X^*_1$ (as mentioned earlier). However, it is unclear how to use these facts to find an estimate of the form (25) since we cannot immediately apply a CLT based argument (as we might have done to obtain Lemma 2).

IV. CONCLUSION

In closing, we first briefly reiterate our main contributions. Propelled by existing literature on coding for permutation channels, we formulated the information theoretic notion of permutation channel capacity for the problem of communicating through a DMC followed by a random permutation transformation. We then proved that the permutation channel capacity of a BSC is $C_{\text{perm}}(\text{BSC}(p)) = \frac{1}{2}$ for $p \in (0, \frac{1}{2}] \cup [\frac{1}{2}, 1)$ in Theorem 3. Furthermore, we derived bounds on the permutation channel capacity of a BEC, $\frac{1}{2} \leq C_{\text{perm}}(\text{BEC}(\delta)) \leq 1$. for $\delta \in (0, 1)$ in Theorem 4.

We next propose some directions for future research. Firstly, our proof technique for Theorem 3 can be extended to establish the permutation channel capacity of DMCs with entry-wise strictly positive stochastic transition probability matrices. In particular, this will entail employing multivariate versions of the results used (either implicitly or explicitly) in our argument such as the second moment method bound in (9) and the CLT. Secondly, the exact permutation channel capacity of the BEC should be determined. As conveyed in the previous section, this will presumably involve a more careful analysis of the converse proof. Thirdly, our ultimate objective is to establish the permutation channel capacity of general DMCs (whose row stochastic matrices have zero entries). Evidently, achieving this goal will first require us to completely resolve the permutation channel capacity of BECs. Finally, there are several other open problems that parallel aspects of classical information theoretic development such as finding tight bounds on probability of error (like classical error exponent analysis, cf. [21, Chapter 5]), developing strong converse results, cf. [18, Section 22.1], [21, Theorem 5.8.5], establishing exact asymptotics for the maximum achievable value of $|M|$ (akin to finite blocklength analysis, cf. [22], [23, Chapter II.4], and the references therein), and extending the permutation channel model by replacing DMCs with other kinds of channels or networks.

ACKNOWLEDGMENT

The author thanks Professors Yury Polyanskiy and Lihong Zheng, and Mohamed AlHajri for stimulating discussions.

REFERENCES