Shannon Capacity of Channels with Markov Insertions, Deletions and Substitutions

Ruslan Morozov and Tolga M. Duman
Bilkent University, Ankara, Turkey
Email: ruslan.morozov@bilkent.edu.tr, duman@ee.bilkent.edu.tr

Abstract

We consider channels with synchronization errors modeled as insertions and deletions. A classical result for such channels is their information stability, hence the existence of the Shannon capacity, when the synchronization errors are memoryless. In this paper, we extend this result to the case where the insertions and deletions have memory. Specifically, we assume that the synchronization errors are governed by a stationary and ergodic finite state Markov chain, and prove that such channel is information-stable, which implies the existence of a coding scheme which achieves the limit of mutual information. This result implies the existence of the Shannon capacity for a wide range of channels with synchronization errors, with different applications including DNA storage. The methods developed may also be useful to prove other coding theorems for non-trivial channel sequences.

I. Introduction

A mathematical model of a physical channel is defined by conditional distributions $W(y|x)$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, which means that if a symbol $x \in \mathcal{X}$ is input, output $y \in \mathcal{Y}$ is observed with probability[1] $W(y|x)$. The fact that W is a channel with the input set $\mathcal{X}$ and the output set $\mathcal{Y}$ will be denoted by $W : \mathcal{X} \rightsquigarrow \mathcal{Y}$. Shannon's theorem [1] considers a memoryless channel $W : \mathcal{X} \rightsquigarrow \mathcal{Y}$ that is used n times independently. The equivalent channel that transmits n symbols simultaneously is the channel $W^n : \mathcal{X}^n \rightsquigarrow \mathcal{Y}^n$, where the conditional probabilities are given by $W^n(y_1^n|x_1^n) = \prod_{i=1}^n W(y_i|x_i)$, with the use of the vector notation $a_b^c = (a_b, a_{b+1}, \ldots, a_c)$.

In this paper, our focus is on channels with synchronization errors, modeled as insertions, deletions and channel errors. As an example, consider an insertion channel V_n with n input symbols $x \in \mathcal{X}^n$. In general, for each input *symbol* $x_i \in \mathcal{X}$, an output *vector* y_i over $\mathcal{X}$ of length ≥ 1 is produced. In contrast with the channel W^n, the resulting output of V_n is a concatenation of output vectors, but not the vector y_1^n of output vectors. For example, if input bits are $(1, 0, 0)$ and the corresponding outputs for each bit are $(1, 0)$, (0) and $(1, 0)$, then the overall channel output would be $(1, 0, 0, 1, 0)$, not $((1, 0), (0), (1, 0))$.

In general, a communication setup, besides conditional probabilities of the original channel, also includes some underlying method of combining the outputs of multiple channel uses. In the case of simple memoryless channels, the combining method is just stacking together output symbols into the output vector; in the case of insertion (and/or deletion) channel, the combining method is concatenation. For this reason, in the non-trivial setup a general case of *a channel sequence* $(W_n : \mathcal{X}_n \rightsquigarrow \mathcal{Y}_n)_{n\in\mathbb{N}}$ is usually considered [2], [3], where sets $\mathcal{X}_n$, $\mathcal{Y}_n$ can vary for different n. For simplicity, we consider the case when input set for W_n is $\mathcal{X}_n = \mathcal{X}^n$ and $\mathcal{X}$ is fixed for all n.

For a channel sequence the *mutual information capacity* can be defined as $\lim_{n\to\infty} \mathbf{I}(W_n)/n$, where $\mathbf{I}(W_n)$ denotes the maximum mutual information between the input and the output of a channel W_n over all possible input distributions. The *coding capacity* (or operational capacity) is the (maximum) asymptotic rate of coding schemes which achieve arbitrary small error probability as $n \to \infty$. The *capacity theorem* is a statement that the mutual information capacity is equal to the coding capacity.

There are many capacity theorems that generalize the Shannon's theorem on non-trivial channel sequences. Dobrushin's work [4] considers the most general case, where input and output alphabets can be continuous, and the metric of closeness of decoder's output to the original signal can be arbitrary. It is shown that the so-called "*information stability*" of a channel sequence is sufficient for the capacity theorem to hold. Later, in [2] it was shown that information stability is necessary and sufficient for the capacity theorem. In [5], Dobrushin proves the

This work was funded by the European Union through the ERC Advanced Grant 101054904: TRANCIDS. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

[1]This is valid only for a countable $\mathcal{Y}$; for uncountable case, there should be output PDFs instead; in this paper uncountable $\mathcal{Y}$ are not considered.

capacity theorem for channels with independent synchronization errors. In that model, for each input symbol, an output vector of finite expected length is produced, and all the output vectors are concatenated to form the resulting output. Since this channel model includes any insertion/deletion/substitution errors, we call such model an IDS channel. The generalization of the capacity theorem to the case of IDS channels with continuous input alphabet is studied in [6]. Capacity theorems are also proven for finite-state Markov channels (FSMC) [7], for Gilbert-Elliott channels [8], for deletion channels concatenated with FSMC [9], for IDS channels with intersymbol interference [10].

IDS channels are used to model various communication channels, including bit patterned media recording [11] and DNA storage channels [12]. For the case of memoryless insertion/deletion channels, capacity upper bounds and lower bounds are also obtained in [13], [14], [15], and the asymptotic expansion of the independent identically distributed deletion channel capacity is found in [16]. See [17], [18] for a detailed review of progress in this area. While most existing results focus on the scenario with independent synchronization errors, it is also important to study the case with insertions/deletions with memory. For instance, current reading and writing technologies in DNA sequencing introduce memory to the channel errors [19], [20].

In this paper, we consider IDS channels with Markov memory, i.e., the underlying insertions/deletions are governed by a Markov chain. This model is referred to as a Markov-IDS channel. That is, each state of the Markov chain is an IDS channel, called component channel, and the overall output of Markov-IDS channel is obtained by concatenating the individual component channel outputs. The Markov-IDS channel is a generalization of both IDS and FSMC: it is more general than IDS since statistics of the synchronization errors change over time according to Markov chain rule; it is more general then FSMC since it concatenates the outputs of each component IDS channel (which can be translated to non-concatenated version via inserting a special symbol at the end of each component channel's output). We note that while this model generalizes the synchronization errors as studied by Dobrushin, it is not sufficiently general to encompass some DNA channel models with insertions/deletions with memory as in [19], [20] since the channel states do not depend on the channel input in our model. Alternatively, in [10], the capacity theorem is proven for IDS channels with intersymbol interference (ISI). The difference between the ISI-IDS channel model in [10] and Markov-IDS channel model in this paper is that the state of ISI-IDS channel is defined by a few recent input bits, and the channel state in the Markov-IDS model is random (though the distribution is defined by the previous state).

In this paper, we also introduce a unified notation, which helps to elucidate the idea behind Dobrushin's methods, which can be seen as *applying functions to channel sequences*. Namely, if a function is applied on a channel's output, and this function is not "doing much" in a certain sense, then the modified channel has the same mutual information capacity, coding capacity and information stability, as the original channel. This means that one can apply any finite number of such functions and still have the same capacities.

The paper is organized as follows. In Section II, notations and information-theoretic definitions are introduced. In Section III, functions on channels and their properties are described. In Section IV, the Markov-IDS channel sequence is defined. In Section V, the capacity theorem for Markov-IDS channel sequence is proven. In Section VI, conclusions are drawn.

II. Background

A. Notations

We use the following notation. Sets are written in calligraphic letters $\mathcal{X}, \mathcal{Y}, \overline{\mathcal{Y}}$; operators are written in bold letters $\mathbf{C}, \mathbf{E}, \mathbf{I}$. For $b, c \in \mathbb{N}$, we denote a vector $a_b^c = (a_b, a_{b+1}, \dots, a_c)$. Here, elements a_i can be of any type: a number, a distribution, a channel, etc. Also, we denote infinite sequences as $a^\infty = (a_1, a_2, \dots)$. For a set $\mathcal{A}$, we denote a set of all finite sequences with elements from $\mathcal{A}$ by $\overline{\mathcal{A}} = \{()\} \cup \bigcup_{i=1}^{\infty} \mathcal{A}^i$. For a vector $a = a_1^n \in \overline{\mathcal{A}}$ we denote by $\mathbf{N}(a) = n$ the vector length. We denote the concatenation (gluing) operator $\mathbf{g}$ as $\mathbf{g}(a_1^n, b_1^m) = (a_1, a_2, ..., a_n, b_1, ..., b_m)$, which is also generalized onto any number of input vectors. Also, define the function $\mathbf{g}_n : \overline{\mathcal{A}}^n \to \overline{\mathcal{A}}$, which glues n vectors $\overline{a}_i \in \overline{\mathcal{A}}$ as $\mathbf{g}_n(\overline{a}_1^n) = \mathbf{g}(\overline{a}_1, \overline{a}_2, ..., \overline{a}_n)$. The inverse image $\mathbf{g}_n^{-1}(a)$ is the set of all possible ways of cutting $a \in \overline{\mathcal{A}}$ into n subvectors, i.e., $\mathbf{g}_n^{-1}(a) = \{\overline{a}_1^n | \mathbf{g}_n(\overline{a}_1^n) = a\}$. For a set $\mathcal{S}$, we denote by $\mathbb{D}_{\mathcal{S}}$ the set of all distributions over $\mathcal{S}$, i.e.

$$\mathbb{D}_{\mathcal{S}} = \left\{ f : \mathcal{S} \to [0,1] \,\middle|\, \sum_{x \in \mathcal{S}} f(x) = 1 \right\}.$$

For $n \in \mathbb{N}$, denote by $\mathbb{D}_n$ the set of distributions over the set $\{1, 2, \ldots, n\}$ in a form of row vectors ρ_1^n, $\sum_{i=1}^n \rho_i = 1$, $\rho_i \geq 0$.

In general, probabilities are usually defined using sigma-algebras. In this paper, however, all probabilities are defined on the countable sets, so the probability measure is always completely defined by the probabilities of the elementary events $\Pr[X = x]$.

The probability distribution for any probability, expectation or entropy is written as a subscript. The values that are not known are interpreted as random variables, and it is assumed that we take summation/expectation over all their possible values. For example, if we have a conditional distribution $W(u, x|v, y)$, where $u \in \mathcal{U}, v \in \mathcal{V}, x \in \mathcal{X}, y \in \mathcal{Y}$, and the values of u and v have been defined previously, then

$$\Pr_{W(u,x|v,y)}[x > 0] = \sum_{\substack{x \in \mathcal{X}: x > 0 \\ y \in \mathcal{Y}}} W(u, x|v, y),$$

$$\mathbf{E}_{W(u,x|v,y)}[f(x, y)] = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} W(u, x|v, y) \cdot f(x, y).$$

We write $\mathbf{H}_A[B] = \mathbf{E}_A[-\log_2 B]$ for entropy-like expressions. If $A = B$, then we omit the subscript for simplicity:

$$\mathbf{H}_{W(x,y)}[W(x|y)] = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} W(x, y) \cdot \log_2 W(x|y),$$

$$\mathbf{H}[W(x, y)] = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} W(x, y) \cdot \log_2 W(x, y).$$

B. Channels and channel sequences

We define a *channel* W as a collection of conditional distributions $W(y|x)$, defined $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$. Throughout the paper, we assume a *finite set $\mathcal{X}$ and a countable set $\mathcal{Y}$*. Formally, a channel W is a function $W : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, such that

$$\forall x \in \mathcal{X} : \sum_{y \in \mathcal{Y}} W(y|x) = 1.$$

The fact that W has input messages from $\mathcal{X}$ and output from $\mathcal{Y}$ is denoted by $W : \mathcal{X} \rightsquigarrow \mathcal{Y}$.

Channel W defines the conditional distribution of the output given its input. If we further define some input distribution $p \in \mathbb{D}_{\mathcal{X}}$, then we can compute the joint distribution of input and output, which is denoted by $p \circ W$:

$$p \circ W(x, y) = p(x)W(y|x) \tag{1a}$$

$$p \circ W(*, y) = \sum_{x \in \mathcal{X}} p(x)W(y|x), \tag{1b}$$

where (1b) defines the probability of output y, given the input distribution p.

For channels $W_i : \mathcal{X}_i \rightsquigarrow \mathcal{Y}_i$, define the channel product $W = \prod_{i=1}^n W_i$ as a channel $W : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \rightsquigarrow \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$, such that

$$W = \prod_{i=1}^n W_i \iff W(y_1^n|x_1^n) = \prod_{i=1}^n W_i(y_i|x_i) \tag{2}$$

A similar notation is used for the n-th power $W^n : \mathcal{X}^n \rightsquigarrow \mathcal{Y}^n$ of a channel $W : \mathcal{X} \rightsquigarrow \mathcal{Y}$.

A *channel sequence* V is defined as an infinite sequence V^∞ of channels $V_i : \mathcal{X}^i \rightsquigarrow \mathcal{Y}_i$. Note that for any i, the output set $\mathcal{Y}_i$ is arbitrary, but the input set for the i-th channel is $\mathcal{X}^i$, so the i-th channel transmits exactly i symbols from $\mathcal{X}$.

C. Mutual information capacity

The mutual information density of a channel W for an input distribution $p \in \mathbb{D}_{\mathcal{X}}$ and given values of an input $x \in \mathcal{X}$ and an output $y \in \mathcal{Y}$ is defined as:

$$\mathbf{i}(W,p,x,y) \stackrel{\triangle}{=} \begin{cases} 0, \text{if } W(y|x)=0 \text{ or } p\circ W(*,y)=0 \\ \log_2 \frac{W(y|x)}{p\circ W(*,y)}, \text{ otherwise} \end{cases} \tag{3}$$

where $p \circ W(*,y)$ is given by (1b). The mutual information of W under input distribution p is the expectation of mutual information density:

$$\mathbf{I}(W,p) \stackrel{\triangle}{=} \mathbf{E}_{p\circ W(x,y)}[\mathbf{i}(W,p,x,y)] = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p\circ W(x,y)\cdot \mathbf{i}(W,p,x,y). \tag{4}$$

The *mutual information capacity/i-capacity* [4] of channel W is the maximum possible mutual information of W under any input distribution, and is given by[2]

$$\mathbf{I}(W) \stackrel{\triangle}{=} \max_{p\in\mathbb{D}_{\mathcal{X}}} \mathbf{I}(W,p). \tag{5}$$

We will call the distribution p, for which the maximum is achieved, as the *optimal distribution* (for W). By the *mutual information capacity/i-capacity* of channel sequence $V = V^{\infty}$ we mean the asymptotic average i-capacity of V_n per input symbol:

$$\mathbf{I}(V) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{\mathbf{I}(V_n)}{n}, \tag{6}$$

if the limit exists. Note that the i-capacity of a *channel* always exists, since $\mathbf{I}(W,p)$ is a concave function of the input distribution p with countable number of coefficients; however, the i-capacity of a *channel sequence* might not exist, since sequence $\mathbf{I}(V_n)/n$ might not converge[3].

D. Coding capacity

For a channel $W : \mathcal{X} \rightsquigarrow \mathcal{Y}$, code size $M \in \mathbb{N}$, $M \le |\mathcal{X}|$ and error probability $\varepsilon \in [0,1]$, we define a (W,M,ε)-*coding scheme* as a collection $(x_1^M, \mathcal{R}_1^M = (\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_M))$, where:

- all x_i are distinct codewords from $\mathcal{X}$, $i = 1 \dots M$;
- $\mathcal{R}_1^M$ is a collection of M disjoint sets $\mathcal{R}_i \subseteq \mathcal{Y}$, which are the decoder's decision regions. When $y \in \mathcal{R}_i$ is received, the decoder outputs x_i;
- the decoding error probability is not greater than ε for any x_i, i.e.

$$\forall i \in \{1,\dots,M\} : \sum_{y\notin\mathcal{R}_i} W(y|x_i) \le \varepsilon.$$

Consider a channel sequence V^{∞}. We call $R \in \mathbb{R}_{\ge 0}$ an achievable rate over V, if

$$\forall \varepsilon > 0, \exists n_0 : \forall n \ge n_0 : \exists (V_n, \lceil 2^{nR} \rceil, \varepsilon)\text{-}\textit{coding scheme}. \tag{7}$$

We denote by $\mathbf{R}(V)$ the set of achievable rates R for channel sequence V. By the *coding capacity/c-capacity* $\mathbf{C}(W)$ of the channel sequence V we mean the supremum of all achievable rates [5]:

$$\mathbf{C}(V) \stackrel{\triangle}{=} \sup \mathbf{R}(V) \tag{8}$$

The existence of a (W,M,ε)-coding scheme obviously implies the existence of a (W,M',ε')-coding scheme for any $M' \le M$ and $\varepsilon' \ge \varepsilon$. So, if $R \in \mathbf{R}(V)$, then $R' \in \mathbf{R}(V)$ for any rate $R' < R$. Thus, the set $\mathbf{R}(V)$ is a connected set, i.e., there always exists $\alpha \in [0, \log_2 |\mathcal{X}|]$, such that $\mathbf{R}(V) \in \{[0,\alpha],[0,\alpha)\}$. This implies that $\mathbf{C}(V) = \alpha$, i.e., any channel sequence has the c-capacity.

In [3], the coding capacity (c-capacity) $\mathbf{C}(V)$ is derived in terms of mutual information density:

$$\mathbf{C}(V) = \sup\left\{ \alpha \,\middle|\, \lim_{n\to\infty} \min_{p_n\in\mathbb{D}_{\mathcal{X}^n}} \Pr_{p_n\circ V_n(x,y)} \left\{ \frac{\mathbf{i}(V_n,p_n,x,y)}{n} \le \alpha \right\} = 0 \right\} \tag{9}$$

[2]In [4], the setup is for the much more general case, and "sup" over the input distributions is used instead of "max". In our setup we assumed (see Section II-B) that $\mathcal{X}$ is finite and $\mathcal{Y}$ is countable, so the maximum is always achieved as a maximum of a continuous function over a compact set.

[3]Consider a channel which is ideal for even n and complete noise for odd n. In this case, the sequence of values of $\mathbf{I}(V_n)/n$ is $(\log_2 |\mathcal{X}|, 0, \log_2 |\mathcal{X}|, 0, \dots)$. Such sequence has no limit for $|\mathcal{X}| > 1$.

E. Information stability

A channel sequence V is called *J-information stable* [2], where $J = J^\infty$ with $J_n \to \infty$, if there exist distributions $p_n^* \in \mathbb{D}_{\mathcal{X}^n}$ and sequence δ with $\delta_n > 0$, $\delta_n \to 0$, such that

$$\Pr_{p_n^* \circ V_n(x,y)} \left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{J_n} - 1 \right| > \delta_n \right] < \delta_n. \tag{10}$$

If $J_n = \mathbf{I}(V_n)$ then V is called information-stable. We denote by $\mathbf{IS}(V)$ the fact that V is information-stable, and, moreover, the i-capacity $\mathbf{I}(V)$ exists. In the case of $\mathbf{IS}(V)$, under our general assumption that $\mathcal{X}$ is finite, (10) can be rewritten as

$$\forall \delta > 0 : \lim_{n\to\infty} \Pr_{p_n^* \circ V_n(x,y)} \left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{n} - \mathbf{I}(V) \right| > \delta \right] = 0. \tag{11}$$

Proposition 1 (Theorems 3.1 and 3.2 in [2])**.** $\mathbf{IS}(V) \iff \mathbf{C}(V) = \mathbf{I}(V)$.

For further derivations we need the following sufficient condition of information stability.

Proposition 2 (A sufficient condition for strict information stability)**.** *If there exists $\mathbf{I}(V)$ and a sequence of distributions p^*, for which*

$$\lim_{n\to\infty} \mathbf{E}_{p_n^* \circ V_n(x,y)} \left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{n} - \mathbf{I}(V) \right| \right] = 0, \tag{12}$$

then $\mathbf{IS}(V)$.

Proof. Note that for any $\delta > 0$,

$$\mathbf{E}\left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{n} - \mathbf{I}(V) \right| \right] > \delta \cdot \Pr\left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{n} - \mathbf{I}(V) \right| > \delta \right],$$

so

$$0 \le \lim_{n\to\infty} \Pr\left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{n} - \mathbf{I}(V) \right| > \delta \right] \le \frac{1}{\delta} \cdot \lim_{n\to\infty} \mathbf{E}\left[\left| \frac{\mathbf{i}(V_n, p_n^*, x, y)}{n} - \mathbf{I}(V) \right| \right] = 0,$$

which is exactly the information stability condition (11). □

III. Functions of channels and their properties

For a channel $W : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and function $f : \mathcal{Y} \to \mathcal{Z}$ we write $W' = f(W)$ to refer to a channel $W' : \mathcal{X} \rightsquigarrow \mathcal{Z}$, defined as a composition of channel W and function f. That is,

$$W'(z|x) = \sum_{y \in f^{-1}(z)} W(y|x), \tag{13}$$

where $f^{-1}(z) = \{y : f(y) = z\}$ is the inverse image of f. Channel W' is *degraded* with respect to W, meaning that it can be represented as a pipeline of W and a function f at its output. It is well-known and can be straightforwardly checked that $\mathbf{C}(W') \le \mathbf{C}(W)$, and if both $\mathbf{I}(W)$ and $\mathbf{I}(W')$ exist, then $\mathbf{I}(W') \le \mathbf{I}(W)$. Let $\Phi = \log_2 \max_{z\in\mathcal{Z}} \left| f^{-1}(z) \right|$ be the maximum log-size of a preimage. Then, the following proposition holds.

Proposition 3 (Equation (4.3) in [5])**.** *For any input distribution $p \in \mathbb{D}_{\mathcal{X}}$:*

$$\mathbf{E}_{p\circ W(x,y)} \left[\left| \mathbf{i}(W, p, x, y) - \mathbf{i}(W', p, x, f(y)) \right| \right] \le \Phi. \tag{14}$$

This proposition can be interpreted as follows. The process of forming the (random) output of channel W can be seen as a pipeline of two steps. First, choose the value of z based on probabilities $W'(z|x)$. This is equivalent to channel W'. Second, randomly choose one value among all (at most) 2^Φ possible values of $y \in f^{-1}(y')$. The maximum entropy of the second step is achieved when all preimages are equiprobable, namely, the entropy of the second step is at most Φ.

For a channel sequence $U_n : \mathcal{X}^n \rightsquigarrow \mathcal{Y}_n$ and sequence of functions $f_n : \mathcal{Y}_n \to \mathcal{Z}_n$, we denote by $V = f(U)$ a channel sequence $V_n = f_n(U_n) : \mathcal{X}^n \rightsquigarrow \mathcal{Z}_n$, $n \in \mathbb{N}$. Also, we denote by $\Phi_n = \log_2 \max_{z\in\mathcal{Z}_n} \left| f_n^{-1}(z) \right|$.

Proposition 4 (Function with small preimage)**.** *If $\Phi_n \in o(n)$, then $\mathbf{C}(U) = \mathbf{C}(V)$. If, furthermore, at least one of $\mathbf{I}(U)$, $\mathbf{I}(V)$ exist, then they both exist and $\mathbf{I}(U) = \mathbf{I}(V)$. Moreover, $\mathbf{IS}(U) \iff \mathbf{IS}(V)$.*

Proof. First, note that $\mathbf{E}_{p_n \circ V_n}[\mathbf{i}(V_n, p_n, x, y)] = \mathbf{E}_{p_n \circ U_n}[\mathbf{i}(V_n, p_n, x, f(y))]$. Using Proposition 3 and the fact that V_n is degraded with respect to U_n, for any sequence of input distributions p:

$$0 \le \lim_{n\to\infty} \frac{\mathbf{I}(U_n, p_n) - \mathbf{I}(V_n, p_n)}{n} = \mathbf{E}_{p_n \circ U_n}\left[\frac{\mathbf{i}(U_n, p_n, x, y) - \mathbf{i}(V_n, p_n, x, f_n(y))}{n}\right]$$
$$\le \lim_{n\to\infty} \mathbf{E}_{p_n \circ U_n}\left[\frac{|\mathbf{i}(U_n, p_n, x, y) - \mathbf{i}(V_n, p_n, x, f_n(y))|}{n}\right] \le \lim_{n\to\infty} \frac{\Phi_n}{n} = 0, \tag{15}$$

which for optimal p for U implies that $\mathbf{I}(U) = \mathbf{I}(V)$, if ether $\mathbf{I}(U)$ or $\mathbf{I}(V)$ exists.

Equation (15) implies that for any $p_n \in \mathbb{D}_{\mathcal{X}^n}$,

$$\lim_{n\to\infty} \mathbf{E}_{p_n \circ U_n}\left[\frac{|\mathbf{i}(U_n, p_n, x, y) - \mathbf{i}(V_n, p_n, x, f_n(y))|}{n}\right] = 0.$$

Using the same technique as in Proposition 2, it can be easily shown that for any $\delta > 0$ and $p_n \in \mathbb{D}_{\mathcal{X}^n}$,

$$\lim_{n\to\infty} \Pr_{p_n \circ U_n}\left\{\frac{|\mathbf{i}(U_n, p_n, x, y) - \mathbf{i}(V_n, p_n, x, f_n(y))|}{n} > \delta\right\} = 0. \tag{16}$$

If $\mathbf{IS}(U)$, then for arbitrary $\delta > 0$, both (11) and (16) can be satisfied for $\delta/2$ simultaneously using some sequence of distributions p^*. The equality $\mathbf{I}(U) = \mathbf{I}(V)$, proven above, and the triangle inequality imply

$$\lim_{n\to\infty} \Pr_{p_n^* \circ U_n}\left\{\left|\frac{\mathbf{i}(V_n, p_n^*, x, f_n(y))}{n} - \mathbf{I}(V)\right| > \delta\right\} \le \lim_{n\to\infty} \Pr_{p_n^* \circ U_n}\left\{\frac{|\mathbf{i}(U_n, p_n^*, x, y) - \mathbf{i}(V_n, p_n^*, x, f_n(y))|}{n} > \frac{\delta}{2}\right\}$$
$$+ \lim_{n\to\infty} \Pr_{p_n^* \circ U_n}\left\{\left|\frac{\mathbf{i}(U_n, p_n^*, x, y)}{n} - \mathbf{I}(U)\right| > \frac{\delta}{2}\right\} = 0. \tag{17}$$

This implies $\mathbf{IS}(V)$. The implication $\mathbf{IS}(V) \implies \mathbf{IS}(U)$ can be shown similarly.

To prove that $\mathbf{C}(U) = \mathbf{C}(V)$, we will use the expression of c-capacity given by (9). Assume that $\mathbf{C}(U) = \alpha$ and $\mathbf{C}(V) \neq \alpha$. Then, $\mathbf{C}(V) < \alpha$ and

$$\forall \varepsilon > 0, \delta > 0, \exists n_0^* : \forall n \ge n_0, \exists p_n^* : \quad \Pr_{p^* \circ U_n(x,y)}\left[\frac{\mathbf{i}(U_n, p^*, x, y)}{n} > \alpha - \varepsilon\right] > 1 - \delta. \tag{18}$$

$$\exists \varepsilon^* > 0, \delta^* > 0 : \forall n_0, \exists n^* \ge n_0 : \forall p_n, \quad \Pr_{p \circ U_{n^*}(x,y)}\left[\frac{\mathbf{i}(V_{n^*}, p, x, f_{n^*}(y))}{n^*} \le \alpha - \varepsilon^*\right] \ge \delta^*. \tag{19}$$

Since (18) holds for any $\varepsilon > 0$, $\delta > 0$, we can let $\varepsilon = \varepsilon^*/2$, $\delta = \delta^*/2$, where ε^* and δ^* are from (19), and obtain simultaneously

$$\exists \varepsilon^* > 0, \delta^* > 0 : \begin{cases} \exists n_0^* : \forall n \ge n_0^*, \exists p_n^* : \ \Pr_{p_n^* \circ U_n(x,y)}[\mathbf{i}(U_n, p_n^*, x, y)/n > \alpha - \varepsilon^*/2] > 1 - \delta^*/2 \\ \forall n_0, \exists n^* \ge n_0 : \forall p_{n^*}, \ \Pr_{p_{n^*} \circ U_{n^*}(x,y)}[\mathbf{i}(V_{n^*}, p_{n^*}, x, f_{n^*}(y))/n^* \le \alpha - \varepsilon^*] \ge \delta^* \end{cases}$$

So, the first inequality holds for all n starting from n_0^*, and for each n specific p_n^* should be picked up. The second inequality holds for some n larger than arbitrary n_0, for all p_n. Bringing these two conditions together, we can satisfy both inequalities as follows:

- for any n_0, pick up the specific value of $n^* \ge \max\{n_0, n_0^*\}$, for which the second inequality holds;
- since $n^* \ge n_0^*$, the first equality still holds for some specific selection of $p_{n^*}^*$. The second inequality will hold for $p_{n^*}^*$ as well, since it holds for any p_{n^*}.

The above procedure results in

$$\exists \varepsilon^* > 0, \delta^* > 0 : \forall n_0, \exists n \ge n_0, \exists p_n : \begin{cases} \Pr_{p_n \circ U_n(x,y)}[\mathbf{i}(U_n, p_n, x, y)/n > \alpha - \varepsilon^*/2] > 1 - \delta^*/2 \\ \Pr_{p_n \circ U_n(x,y)}[\mathbf{i}(V_n, p_n, x, f_n(y))/n \le \alpha - \varepsilon^*] \ge \delta^* \end{cases}.$$

Note that both probabilities are defined over the same probability space. The sum of the two probabilities is greater than $1 + \delta^*/2$, which means that

$$\Pr_{p_n \circ U_n(x,y)}\left\{\frac{\mathbf{i}(U_n, p_n, x, y) - \mathbf{i}(V_n, p_n, x, f_n(y))}{n} > \frac{\varepsilon^*}{2}\right\} > \frac{\delta^*}{2},$$

which contradicts (16). □

Thus, one can apply such function sequences to a channel sequence W, changing neither $\mathbf{I}(W)$, nor $\mathbf{C}(W)$, nor information stability property.

Proposition 5 (Semi-bijective function)**.** *Consider channels $U : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $V : \mathcal{X} \rightsquigarrow \mathcal{Z}$, such that $V = f(U)$, $f : \mathcal{Y} \to \mathcal{Z}$. Consider partitions $\mathcal{Y} = \mathcal{A} \sqcup \mathcal{B}$, $\mathcal{Z} = \mathcal{A}' \sqcup \mathcal{B}'$, such that $f(\mathcal{A}) = \mathcal{A}'$ and f is bijective between $\mathcal{A}$ and $\mathcal{A}'$. Denote by $\beta(x) = \Pr_{U(y|x)}[y \in \mathcal{B}]$ the probability that the output of U belongs to $\mathcal{B}$ given input $x \in \mathcal{X}$. If $\forall x \in \mathcal{X} : \beta(x) \le \overline{\beta}$, then for any $p \in \mathbb{D}_{\mathcal{X}}$,*

$$\mathbf{I}(U,p) - \mathbf{I}(V,p) \le \overline{\beta} \log_2 |\mathcal{X}| + \frac{1}{e \ln 2}. \tag{20}$$

Proof. Let $U' : \mathcal{X} \rightsquigarrow \mathcal{B}$ be a channel equivalent to U with an additional condition that the output belongs to $\mathcal{B}$, with the normalized transition probabilities

$$U'(y|x) = \frac{U(y|x)}{\beta(x)}.$$

By definition, $\mathbf{I}(U,p) = \mathbf{E}_{p \circ U(x,y)}[\mathbf{i}(U,p,x,y)]$. The expectation can be expressed by two separate summations over $\mathcal{A}$ and $\mathcal{B}$:

$$\begin{aligned}\mathbf{E}_{p\circ U(x,y)}[\mathbf{i}(U,p,x,y)] &= \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}^n} p(x)U(y|x)\mathbf{i}(U,p,x,y) \\ &= \underbrace{\sum_{y\in\mathcal{A}} \sum_{x\in\mathcal{X}} p(x)U(y|x)\mathbf{i}(U,p,x,y)}_{a} + \underbrace{\sum_{y\in\mathcal{B}} \sum_{x\in\mathcal{X}} p(x)U(y|x)\mathbf{i}(U,p,x,y)}_{b}\end{aligned}$$

Let us analyze term b. First, introduce the modified distribution $\widetilde{p}(x) = p(x)\beta(x)/\Delta$, where $\Delta = \sum_x p(x)\beta(x)$. Note that $0 \le \Delta \le \overline{\beta}$. Substituting $p(x)U(y|x) = \Delta \widetilde{p}(x)U'(y|x)$ for $y \in \mathcal{B}$, one obtains

$$\begin{aligned}b &= \sum_{y\in\mathcal{B},x\in\mathcal{X}} p(x)U(y|x)\log_2 \frac{U(y|x)}{\sum_w p(w)U(y|w)} = \Delta \sum_{y\in\mathcal{B},x\in\mathcal{X}} \widetilde{p}(x)U'(y|x)\log_2 \frac{\beta(x)U'(y|x)}{\Delta \sum_w \widetilde{p}(w)U'(y|w)} \\ &= \Delta \cdot \mathbf{I}(U',\widetilde{p}) + \Delta \sum_{y\in\mathcal{B},x\in\mathcal{X}} \widetilde{p}(x)U'(y|x)\log_2 \frac{\beta(x)}{\Delta} = \Delta(\mathbf{I}(U',\widetilde{p}) - \log_2 \Delta) + \sum_{y\in\mathcal{B},x\in\mathcal{X}} p(x)U'(y|x)\beta(x)\log_2 \beta(x)\end{aligned}$$

Using inequality $-\frac{1}{e\ln 2} \le x \log_2 x \le 0$ when $0 < x \le 1$, one obtains

$$-\Delta \log_2 \Delta - \frac{1}{e\ln 2} \le b \le \Delta \log_2 \frac{|\mathcal{X}|}{\Delta} \tag{21}$$

The same can be done for the channel V:

$$\begin{aligned}\mathbf{E}_{p\circ V(x,y)}[\mathbf{i}(V,p,x,y)] &= \sum_{y\in\mathcal{A}'} \sum_{x\in\mathcal{X}} p(x)V(y|x)\mathbf{i}(V,p,x,y) + \underbrace{\sum_{y\in\mathcal{B}'} \sum_{x\in\mathcal{X}} p(x)V(y|x)\mathbf{i}(V,p,x,y)}_{b'} \\ &= \sum_{y\in\mathcal{A}} \sum_{x\in\mathcal{X}} p(x)U(y|x)\mathbf{i}(U,p,x,y) + b' = a + b'\end{aligned}$$

Note that $\beta(x) = \sum_{y\in\mathcal{B}} U(y|x) = \sum_{z\in\mathcal{B}'} V(z|x)$, so, using the auxiliary channel $V'(z|x) : \mathcal{X} \rightsquigarrow \mathcal{B}'$ defined as $V'(z|x) = V(z|x)/\beta(x)$, one obtains for any $z \in \mathcal{B}'$, $p(x)V(z|x) = \Delta\widetilde{p}(x)V'(z|x)$. Similarly to b, the value of b' can be represented as

$$\begin{aligned}b' &= \sum_{z\in\mathcal{B}',x\in\mathcal{X}} p(x)V(z|x)\mathbf{i}(V,p,x,z) = \Delta \sum_{z\in\mathcal{B}',x\in\mathcal{X}} \widetilde{p}(x)V'(z|x)\log_2 \frac{\beta(x)V'(z|x)}{\Delta \sum_w \widetilde{p}(w)V'(z|w)} \\ &= \Delta(\mathbf{I}(V',\widetilde{p}) - \log_2 \Delta) + \sum_{z\in\mathcal{B}',x\in\mathcal{X}} p(x)V'(y|x)\beta(x)\log_2 \beta(x)\end{aligned}$$

so (21) holds for b' as well. This implies

$$\forall p \in \mathbb{D}_{\mathcal{X}} : \mathbf{I}(U,p) - \mathbf{I}(V,p) = b - b' \le \Delta \log_2 \frac{|\mathcal{X}|}{\Delta} + \Delta \log_2 \Delta + \frac{1}{e\ln 2} = \Delta \log_2 |\mathcal{X}| + \frac{1}{e\ln 2}$$

Since $\Delta \le \overline{\beta}$, (20) holds. □

Proposition 6 (Function that does nothing almost surely)**.** *Consider channel sequences $U_n : \mathcal{X}^n \rightsquigarrow \mathcal{Y}_n$ and $V_n : \mathcal{X}^n \rightsquigarrow \mathcal{Z}_n$, such that $V_n = f_n(U_n)$, $f_n : \mathcal{Y}_n \to \mathcal{Z}_n$. Consider partitions $\mathcal{Y}_n = \mathcal{A}_n \sqcup \mathcal{B}_n$, $\mathcal{Z}_n = \mathcal{A}'_n \sqcup \mathcal{B}'_n$, such that $f_n(\mathcal{A}_n) = \mathcal{A}'_n$ and f_n is bijective between $\mathcal{A}_n$ and $\mathcal{A}'_n$. Denote by $\beta_n(x) = \Pr_{U_n(y|x)}[y \in \mathcal{B}_n]$, and $\overline{\beta}_n = \max_x \beta_n(x)$. If the output of channel U_n almost always belongs to $\mathcal{A}_n$, i.e., $\lim_{n\to\infty} \overline{\beta}_n = 0$, then $\mathbf{C}(U) = \mathbf{C}(V)$. If, in addition, one of $\mathbf{I}(U)$, $\mathbf{I}(V)$ exists, then they both exist and $\mathbf{I}(U) = \mathbf{I}(V)$. Moreover, $\mathbf{IS}(U) \iff \mathbf{IS}(V)$.*

Proof. If either $\mathbf{I}(U)$ or $\mathbf{I}(V)$ exists, then, using Proposition 5 and distributions p_n^*, optimal for U_n,

$$0 \le \mathbf{I}(U) - \mathbf{I}(V) \le \lim_{n\to\infty} \frac{\mathbf{I}(U_n, p_n^*) - \mathbf{I}(V_n, p_n^*)}{n} \le \lim_{n\to\infty} \left(\frac{\overline{\beta}_n n \log_2 |\mathcal{X}|}{n} + \frac{1}{ne \ln 2} \right) = 0,$$

so $\mathbf{I}(U) = \mathbf{I}(V)$.

Assume that $\mathbf{IS}(U)$ is true. Note that, for any p_n,

$$\lim_{n\to\infty} \Pr_{p_n \circ U_n(x,y)} \{\mathbf{i}(U_n, p_n, x, y) \ne \mathbf{i}(V_n, p_n, x, f_n(y))\} = \lim_{n\to\infty} \Pr_{p_n \circ U_n(x,y)} \{y \in \mathcal{B}\} \le \lim_{n\to\infty} \overline{\beta}_n = 0. \tag{22}$$

Since $\mathbf{I}(U) = \mathbf{I}(V)$, there exists a sequence of distributions $(p_n^*)_{n\in\mathbb{N}}$, such that for any $\delta > 0$:

$$\lim_{n\to\infty} \Pr_{p_n^* \circ V_n(x,z)} \left\{ \left| \frac{\mathbf{i}(V_n, p_n^*, x, z)}{n} - \mathbf{I}(V) \right| > \delta \right\} = \lim_{n\to\infty} \Pr_{p_n^* \circ U_n(x,y)} \left\{ \left| \frac{\mathbf{i}(V_n, p_n^*, x, f_n(y))}{n} - \mathbf{I}(U) \right| > \delta \right\}$$
$$\le \lim_{n\to\infty} \left[\Pr_{p_n^* \circ U_n(x,y)} \{\mathbf{i}(V_n, p_n^*, x, f_n(y)) \ne \mathbf{i}(U_n, p_n^*, x, y)\} + \Pr_{p_n^* \circ U_n(x,y)} \left\{ \left| \frac{\mathbf{i}(U_n, p_n^*, x, y)}{n} - \mathbf{I}(U) \right| > \delta \right\} \right],$$

and by (22) and information stability of U, the limit is 0. Thus, $\mathbf{IS}(U) \implies \mathbf{IS}(V)$ is information stable. The other implication can be shown similarly.

According to (9), $\mathbf{C}(U) = \mathbf{C}(V)$ can be proved by analyzing the probability of $\mathbf{i}(V_n, p_n, x, y)$ being less than αn for some value of α. This can be done as follows:

$$\Pr_{p_n \circ V_n(x,y)} [\mathbf{i}(V_n, p_n, x, y) \le \alpha n] = \Pr_{p_n \circ U_n(x,y)} [\mathbf{i}(V_n, p_n, x, f(y)) \le \alpha n]$$
$$\le \Pr_{p_n \circ U_n(x,y)} [\mathbf{i}(V_n, p_n, x, f(y)) \ne \mathbf{i}(U_n, p_n, x, y) \vee \mathbf{i}(V_n, p_n, x, f(y)) = \mathbf{i}(U_n, p_n, x, y) \le \alpha n]$$
$$\le \Pr_{p_n \circ U_n(x,y)} [\mathbf{i}(V_n, p_n, x, f(y)) \ne \mathbf{i}(U_n, p_n, x, y) \vee \mathbf{i}(U_n, p_n, x, y) \le \alpha n]$$
$$\le \Pr_{p_n \circ U_n(x,y)} [\mathbf{i}(V_n, p_n, x, f(y)) \ne \mathbf{i}(U_n, p_n, x, y)] + \Pr_{p_n \circ U_n(x,y)} [\mathbf{i}(U_n, p_n, x, y) \le \alpha n],$$

where the first probability goes to zero according to (22). Thus,

$$\lim_{n\to\infty} \min_{p_n} \Pr_{p_n \circ U_n(x,y)} [\mathbf{i}(U_n, p_n, x, y) \le \alpha n] = 0 \implies \lim_{n\to\infty} \min_{p_n} \Pr_{p_n \circ V_n(x,y)} [\mathbf{i}(V_n, p_n, x, y) \le \alpha n] = 0,$$

which means that the set of such α's for U_n is a subset of such set for V_n, which implies $\mathbf{C}(V) \ge \mathbf{C}(U)$ and thus $\mathbf{C}(V) = \mathbf{C}(U)$. □

If this proposition is combined with Proposition 4, the result is that one can apply any finite number of functions, which either have sufficiently small pre-image or sufficiently small probability of changing the output, without changing neither $\mathbf{I}$ nor $\mathbf{C}$ of the original channel sequence. The $\mathbf{IS}$ property is also saved upon such transformations.

IV. The Markov-IDS channel sequence

A. The IDS channel sequence

In [5] a general channel model, which represents any channel with independent insertions, deletions, and substitutions is considered. Consider a basic channel $W : \mathcal{X} \rightsquigarrow \overline{\mathcal{Y}}$, such that for each $x \in \mathcal{X}$ the expected length $\mathbf{N}(y)$ of the output sequence $y \in \overline{\mathcal{Y}}$ is finite:

$$\exists A \in \mathbb{R} : \forall x \in \mathcal{X} : \mathbf{E}_{W(y|x)}[\mathbf{N}(y)] \le A. \tag{23}$$

We will call such channel *an IDS channel*. The *IDS channel sequence* is $D = D^\infty$, where D_n is channel W^n, followed by a gluing operator on its output:

$$D_n = \mathbf{g}_n(W^n), \tag{24}$$

where $\mathbf{g}_n : \overline{\mathcal{Y}}^n \to \overline{\mathcal{Y}}$ glues together n vectors over $\mathcal{Y}$. Such channel sequence D is information-stable [5], so $\mathbf{C}(D) = \mathbf{I}(D)$.

Proposition 7 (Lemma 9 in [10])**.** *Any IDS channel (with finite expectation of output vector length) has finite output entropy for each input:*

$$\exists \overline{H} \in \mathbb{R} : \forall x \in \mathcal{X} : \mathbf{H}[W(y|x)] \le \overline{H}. \tag{25}$$

Proof. Fix some $x_0 \in \mathcal{X}$. Denote by $w_y = W(y|x_0)$. The entropy is a series

$$\mathbf{H}[w_y] = \sum_{y \in \overline{\mathcal{Y}} : w_y > 0} -w_y \log_2 w_y. \tag{26}$$

Since all the terms of the summation are non-negative, we only need to prove that the sum is constrained. Denote by $p_n = \sum_{y \in \mathcal{Y}^n} w_y$ the probability that the length of output is n. From (23),

$$\sum_{y \in \overline{\mathcal{Y}}} w_y \mathbf{N}(y) = \sum_{n=0}^{\infty} n \cdot \sum_{y \in \mathcal{Y}^n} w_y = \sum_{n=0}^{\infty} n p_n \le A.$$

We obtained a constraint on p_n. Expression $\sum_{y \in \mathcal{Y}^n} -w_y \log_2 w_y$ achieves its maximum (with fixed p_n) with uniform distribution within $\mathcal{Y}^n$, i.e., when $w_y = p_n / |\mathcal{Y}|^n$. Thus,

$$\mathbf{H}[w_y] \le -\sum_{n=0}^{\infty} \sum_{y \in \mathcal{Y}^n} \frac{p_n}{|\mathcal{Y}|^n} \log_2 \frac{p_n}{|\mathcal{Y}|^n} \le \mathbf{H}[p_n] - A \cdot |\mathcal{Y}|.$$

The second term is finite. The first term is the entropy of distribution p_n, corresponding to the length of the output of W, which has finite expectation $\le A$. The maximum possible entropy of discrete distribution over $\mathbb{N} \cup \{0\}$ with given (finite) mean is equal to the entropy of the geometric distribution, which is finite. □

B. The Markov-IDS channel sequence

Consider a discrete-time Markov chain with s states $\mathcal{S} = \{1, \dots, s\}$. Let G be an $s \times s$ matrix which defines the transition probabilities: $G_{\sigma,\tau} \in [0, 1]$ is equal to the conditional probability of the next state being τ if the current state is σ.

Assume that each state $\sigma \in \mathcal{S}$ corresponds to a *state channel* $W_1[\sigma] : \mathcal{X} \rightsquigarrow \overline{\mathcal{Y}}$, which is an IDS channel. If the state channels have different output alphabets $\mathcal{Y}_\sigma$ for each σ, we can easily generalize this case by letting $\mathcal{Y} = \cup_\sigma \mathcal{Y}_\sigma$.

The finite-state Markov chain (FSMC) channel [7] works as follows:

1) The initial state σ_1 is given.
2) The i-th symbol $x_i \in \mathcal{X}$ is transmitted through channel $W_1[\sigma_i]$. The output is $\overline{y}_i \in \overline{\mathcal{Y}}$.
3) After the i-th symbol is transmitted, the system transits from state σ_i to state σ_{i+1} with probability $G_{\sigma_i,\sigma_{i+1}}$.
4) For n input symbols, one obtains $\overline{y}_1^n \in \overline{\mathcal{Y}}^n$.

Let $W_{*,1}[\sigma_1] = (W_{1,1}[\sigma_1], W_{2,1}[\sigma_1], \dots)$ be an FSMC channel sequence with initial state σ_1. Then

$$W_{n,1}[\sigma_1](y_1^n | x_1^n) = W_1[\sigma_1](y_1|x_1) \cdot \sum_{\sigma_2^n \in \mathcal{S}^{n-1}} \prod_{i=2}^{n} G_{\sigma_{i-1},\sigma_i} W_1[\sigma_i](y_i|x_i). \tag{27}$$

Also, introduce channel $W_n[\sigma]$ which outputs $\mathbf{g}_n(\overline{y}_1^n) \in \overline{\mathcal{Y}}$, where $\overline{y}_1^n$ is the output of $W_{n,1}[\sigma]$:

$$W_n[\sigma] = \mathbf{g}_n(W_{n,1}[\sigma]). \tag{28}$$

Assume that the initial state of Markov chain is chosen according to distribution $\rho \in \mathbb{D}_s$. Then, the corresponding FSMC channel sequence is denoted by $V_{*,1}[\rho]$, $V_{n,1}[\rho] : \mathcal{X}^n \rightsquigarrow \overline{\mathcal{Y}}^n$:

$$V_{n,1}[\rho](y|x) = \sum_{\sigma \in \mathcal{S}} \rho_\sigma \cdot W_{n,1}[\sigma](y|x). \tag{29}$$

The Markov-IDS channel sequence $V[\rho]$, $V_n[\rho] = \mathbf{g}_n(V_{n,1}[\rho]) : \mathcal{X}^n \to \overline{\mathcal{Y}}$ is given by

$$V_n[\rho](y|x) = \sum_{\sigma \in \mathcal{S}} \rho_\sigma \cdot W_n[\sigma](y|x). \tag{30}$$

V. Capacity theorem for Markov-IDS channel

Throughout the section, we use notation of the channel sequences $W_{n,1}[\rho]$, $W_n[\rho]$, $V_{n,1}[\rho]$ and $V_n[\rho]$, $n \in \mathbb{N}$, as defined in (27), (28), (29) and (30).

A. Markov chain state distribution convergence

It is well-known (Theorem 8.9 in [21]), that if a Markov chain with a finite state space is aperiodic and irreducible, then

$$\exists B \geq 0, 0 \leq P < 1 : \forall n \in \mathbb{N}, \forall i, j \in \mathcal{S} : |G^n_{i,j} - \pi_j| \leq BP^n = B' e^{-Cn} \stackrel{\triangle}{=} \delta_n, \tag{31}$$

where $G^n_{i,j} = (G^n)_{i,j}$, $\pi \in \mathbb{D}_s$ is the stationary distribution of Markov chain G: $\pi \cdot G = \pi$, constants $B, B' \geq 0$, $0 \leq P < 1$ and $C > 0$ depend only on the Markov chain.

For any starting distribution $\rho \in \mathbb{D}_s$ we have $\sum_i \rho_i = 1$, and so

$$|(\rho G^n)_j - \pi_j| = \left| \sum_i \rho_i G^n_{i,j} - \sum_i \rho_i \cdot \pi_j \right| \leq \sum_i \rho_i \left| G^n_{i,j} - \pi_j \right| \stackrel{(31)}{\leq} \delta_n. \tag{32}$$

Introduce sequence ε_n, such that

$$\delta_n = \varepsilon_n \min_i \pi_i. \tag{33}$$

Then, the small (with $n \to \infty$) additive term can be substituted with small multiplicative term as $\pi_i + \delta_n \leq \pi_i(1+\varepsilon_n)$. Combining this with (31), (32)

$$\begin{aligned} \forall \rho \in \mathbb{D}_s, \forall j : \ & \pi_j(1-\varepsilon_n) \leq (\rho G^n)_j \leq \pi_j(1+\varepsilon_n), \\ & \varepsilon_n \leq De^{-Cn}, C = \text{const}, D = \text{const}, \end{aligned} \tag{34}$$

which also implies

$$\pi_j(1-\varepsilon_n) \leq G^n_{i,j} \leq \pi_j(1+\varepsilon_n).$$

B. Main Theorems

Theorem 1 (Existence of i-capacity)**.** *If a Markov chain G of s states, corresponding to Markov-IDS channel sequence V, is aperiodic and irreducible, then for any $\rho \in \mathbb{D}_s$ the i-capacity $\mathbf{I}(V[\rho])$ exists and does not depend on ρ. In other words, for any $\rho \in \mathbb{D}_s$, $\mathbf{I}(V[\rho])$ exists and*

$$\mathbf{I}(V[\rho]) = \mathbf{I}(V[\pi]), \tag{35}$$

where π is the stationary distribution of the Markov chain.

Proof. We now will prove that the limit $\mathbf{I}(V[\rho]) = \lim_{n\to\infty} \mathbf{I}(V_n[\rho])/n$ exists. For that, we will use the stronger formulation [22] of Fekete's lemma, which says that if function $u : \mathbb{N} \to \mathbb{R}$ is quasi-subadditive in the sense that

$$\forall m, n \in \mathbb{N}, n \leq m \leq 2n : u(m+n) \leq u(m) + u(n) + f(m+n), \tag{36}$$

where term $f(m+n)$ is such that series $\sum_{n=1}^{\infty} f(n)/n^2$ converges, then the limit $\lim_{n\to\infty} u(n)/n$ exists.

Denote by $V_{n_1|n_2|\ldots|n_t}[\rho] : \mathcal{X}^n \rightsquigarrow \overline{\mathcal{Y}}^t$, $n = \sum_i n_i$, a channel, which works as a channel $V_{n,1}[\rho]$ combined with the merging operation. The merging operation concatenates the outputs inside i-th output sub-block for each i, where the corresponding i-th input sub-block has length n_i. So, the receiver knows the output, corresponding to each n_i-block, but does not know the exact output for each input symbol. Also, we write $\widetilde{n}_i$ instead of n_i to denote the fact that the output, corresponding to the i-th input block is erased, i.e., substituted by the zero-length vector. Consider the case of $\rho = \pi$, i.e., the initial state distribution is the stationary distribution of the Markov chain (later we will show that the initial state distribution does not influence the i-capacity). Then, the following sequence of inequalities holds:

$$\mathbf{I}(V_{m+n}[\pi]) \stackrel{1}{\leq} \mathbf{I}(V_{m|n}[\pi]) \stackrel{2}{\leq} \mathbf{I}(V_{m|l|n-l}[\pi]) \stackrel{3}{\leq} \mathbf{I}(V_{m|\widetilde{l}|n-l}[\pi]) + l\overline{H} \stackrel{4}{\leq} \mathbf{I}(V_{m|\widetilde{l}|n-l|l}[\pi]) + l\overline{H} \stackrel{5}{\leq} \mathbf{I}(V_{m|\widetilde{l}|n}[\pi]) + 2l\overline{H} \tag{37}$$

where $\overline{H} = \max_{\sigma\in\mathcal{S},x\in\mathcal{X}} \mathbf{H}[W[\sigma](y|x)]$ is the highest output entropy of a component channel (see Proposition 7). Here, transitions 1 and 2 correspond to providing the receiver with additional information.

In transition 3, let us assume the same input distribution $p \in \mathbb{D}[\mathcal{X}^{m+n}]$ for both $V_{m|l|n-l}[\pi]$ and $V_{m|\widetilde{l}|n-l}[\pi]$ and prove that $\mathbf{I}(V_{m|l|n-l}[\pi], p) \leq \mathbf{I}(V_{m|\widetilde{l}|n-l}[\pi], p) + l\overline{H}$. Since this holds for any p, then we establish that the transition 3 is also valid. Denote by X_1, X_2, X_3 the random variables corresponding to three input blocks of m, l and $(n-l)$ symbols (they are the same for both channels since we take the same input distribution), and denote the random variables corresponding to the outputs by Y_1, Y_2, Y_3. Then, the amount of information about X_1^3 that we have lost by erasing Y_2 is given by

$$\mathbf{I}(V_{m|l|n-l}[\pi], p) - \mathbf{I}(V_{m|\widetilde{l}|n-l}[\pi], p) = I(Y_1^3; X_1^3) - I(Y_1, Y_3; X_1^3) = I(Y_2; X_1^3|Y_1, Y_3) \leq H(Y_2|Y_1, Y_3) \leq H(Y_2),$$

where $H(Y_2)$ is the unconditional entropy of output corresponding to the second block; vector Y_2 consists of l outputs $Y_2^{(i)}$ of component channels, glued together, $i = 1 \dots l$. So, $H(Y_2) \leq H(Y_2^{(1)}, Y_2^{(2)}, \dots, Y_2^{(l)}) \leq \sum_i H(Y_2^{(i)})$. Each $Y_2^{(i)}$ is an unconditioned output of a state channel. It can be expressed as

$$H(Y_2^{(i)}) = \mathbf{E}_{\sigma_i, x_i}\Big[H(Y_2^{(i)}|x_i, \sigma_i)\Big] \leq \max_{\sigma,x} \mathbf{H}[W[\sigma](y|x)] = \overline{H},$$

where distribution of σ_i is given by initial state distribution ρ and Markov chain transition probabilities, and distribution of x_i is given by marginalization of distribution p over all coordinates except the $(m+i)$-th coordinate. The above implies that $H(Y_2) \leq l\overline{H}$ and inequality 3 holds.

In transition 4 after block of $(n-l)$ symbols we further transmit a block of l symbols. This cannot decrease the mutual information, since we can just ignore last block of input and output.

In transition 5 the output of $(n-l)$-block is merged with the output of the l-block. We can additionally tell the receiver the exact values of the output for the last l-block, and then the receiver can know the boundaries of the two outputs. By the same consideration as in inequality 3, the conditional entropy about the last l glued outputs, which we have partially lost by gluing, is not greater than unconditional entropy of l unglued outputs, which is not greater than $l\overline{H}$ bits. This holds for any input distribution, so this holds for maxima over the input distributions.

Now, consider channel $V_{m|\widetilde{l}|n}[\rho] : \mathcal{X}^{m+l+n} \rightsquigarrow \overline{\mathcal{Y}}^2$. We will need auxiliary channel $W_{n,1}[\sigma \to \tau] : \mathcal{X}^n \to \mathcal{S}\times\overline{\mathcal{Y}}^n$, where τ is considered as a part of the channel's output. The channel can be represented as channel $W_{n,1}[\sigma]$ which additionally outputs the state after transmission of the last symbol:

$$W_{n,1}[\sigma_1 \to \tau](y_1^n|x_1^n) = W_1[\sigma_1](y_1|x_1) \cdot \sum_{\sigma_2^n\in\mathcal{S}^{n-1}} G_{\sigma_n,\tau} \cdot \prod_{i=2}^{n} G_{\sigma_{i-1},\sigma_i} W_1[\sigma](y_i|x_i). \tag{38}$$

Also, define the glued version of this channel as

$$W_n[\sigma \to \tau](\overline{y}|x) = \sum_{y\in\mathbf{g}_n^{-1}(\overline{y})} W_{n,1}[\sigma \to \tau](y|x). \tag{39}$$

The channel which does not output the final state and the channel which outputs the final state are related via marginalization over the final state:

$$W_n[\sigma](\overline{y}|x) = \sum_{\tau} W_n[\sigma \to \tau](\overline{y}|x), \;\; W_{n,1}[\sigma](\overline{y}|x) = \sum_{\tau} W_{n,1}[\sigma \to \tau](\overline{y}|x).$$

Now we can define the transition probabilities, corresponding to transmitting an m and an n-block, separated by an l-block, which is then erased:

$$\begin{aligned}
V_{m|\widetilde{l}|n}[\pi](y_1^2|x_1^3) &\triangleq \sum_{\substack{z_1\in\mathbf{g}_m^{-1}(y_1)\\ z_2\in\mathbf{g}_n^{-1}(y_2)}} \sum_{\sigma_1^3\in\mathcal{S}^3} \pi_{\sigma_1} W_m[\sigma_1 \to \sigma_2](z_1|x_1) G^l_{\sigma_2,\sigma_3} W_n[\sigma_3](z_2|x_3) \\
&\overset{1}{\leq} \sum_{z_1^2,\sigma_1^3} \pi_{\sigma_1} W_m[\sigma_1 \to \sigma_2](z_1|x_1) \pi_{\sigma_3}(1+\varepsilon_l) W_n[\sigma_3](z_2|x_3) \\
&\overset{2}{\leq} (1+\varepsilon_l) \sum_{\sigma_1,z_1} \pi_{\sigma_1} W_m[\sigma_1](z_1|x_1) \sum_{\sigma_3,z_2} \pi_{\sigma_3} W_n[\sigma_3](z_2|x_3) = (1+\varepsilon_l) V_m[\pi](y_1|x_1) V_n[\pi](y_2|x_3).
\end{aligned} \tag{40}$$

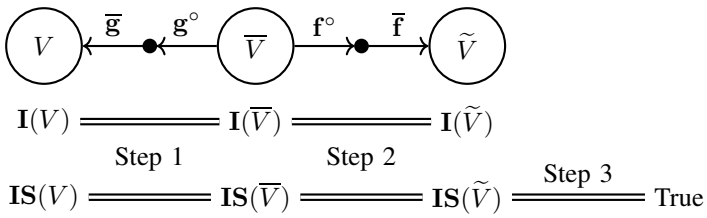

Fig. 1. The outline of proof of Theorem 2

For brevity, introduce some shortcuts: let $x = x_1^3$, $y = y_1^2$. Since $V_{m|\widetilde{l}|n}[\pi](y_1^2|x_1^3)$ does not depend on x_2, consider shortcut $U(y_1^2|x_1^2) = V_{m|\widetilde{l}|n}[\pi](y_1^2|x_1, x', x_2)$, where $x' \in \mathcal{X}^l$ is some fixed vector. Also, consider shortcut $U'(y|x) = V_m[\pi](y_1|x_1)V_n[\pi](y_2|x_2)$. Denote marginalized distribution $\widetilde{p}(x_1, x_2) = \sum_{x'} p(x_1, x', x_2)$. Rewriting (40) in shortcuts, one obtains $U(y|x) \le (1+\varepsilon_l)U'(y|x)$, and one can upper-bound the mutual information by

$$\begin{aligned}
&\mathbf{I}(V_{m|\widetilde{l}|n}[\pi], p) = \mathbf{I}(U, \widetilde{p}) = \sum_{x,y} \widetilde{p} \circ U(x,y)[\log_2 U(y|x) - \log_2 \widetilde{p} \circ U(*,y)] \\
&\overset{1}{\le} \sum_{x,y} (1-\varepsilon_l)\widetilde{p} \circ U'(x,y) \log_2 \left[(1+\varepsilon_l)U'(y|x)\right] - (1+\varepsilon_l)\widetilde{p} \circ U'(x,y) \log_2 \left[(1-\varepsilon_l)\widetilde{p} \circ U'(*,y)\right] \\
&= (1-\varepsilon_l)\mathbf{E}_{\widetilde{p} \circ U'(x,y)}[\log_2 U'(y|x) + \log_2 (1+\varepsilon_l)] - (1+\varepsilon_l)\mathbf{E}_{\widetilde{p} \circ U'(x,y)}[\log_2 \widetilde{p} \circ U'(*,y) + \log_2 (1-\varepsilon_l)] \\
&= \mathbf{I}(U', \widetilde{p}) - \varepsilon_l \mathbf{E}_{\widetilde{p} \circ U'(x,y)}[\log_2 U'(y|x)] - \varepsilon_l \mathbf{E}[\log_2 \widetilde{p} \circ U'(*,y)] + \Gamma_l \\
&= \mathbf{I}(U', \widetilde{p}) + \varepsilon_l \mathbf{H}_{\widetilde{p} \circ U'(x,y)}[U'(y|x)] + \varepsilon_l \mathbf{H}[\widetilde{p} \circ U'(*,y)] + \Gamma_l \\
&\le \mathbf{I}(V_m[\pi] \cdot V_n[\pi], \widetilde{p}) + \varepsilon_l (m+n)\overline{H} + \varepsilon_l (m+n)\overline{H} + \Gamma_l,
\end{aligned} \tag{41}$$

where $\Gamma_l = (1-\varepsilon_l)\log_2(1+\varepsilon_l) - (1+\varepsilon_l)\log_2(1-\varepsilon_l)$. The transition in $\overset{1}{\le}$ is done as follows. Each $\log$ in the first term adds a negative amount, so the multiplicative term before the first logarithm is lower-bounded. The logarithm itself is an increasing function, so its argument is upper-bounded. Note that the upper bound might turn the value of logarithm to a positive value. Then the whole term under $\sum$ sign is upper-bounded, because the original term was non-positive. The second term is taken with a negative sign, so all bounds are the opposite.

Bringing together (37) and (41), for $u_n = \mathbf{I}(V_n[\pi])$ one obtains

$$u_{m+n} \le u_m + u_n + 2\varepsilon_l (m+n)\overline{H} + \Gamma_l + 2l\overline{H}$$

Let $l = \lfloor \sqrt{m+n} \rfloor$ and denote by $f(m+n) = 2\varepsilon_l (m+n)\overline{H} + \Gamma_l + 2l\overline{H}$. Then,

$$0 \le f(n) \le 2\varepsilon_{2\sqrt{2n}} \cdot n + \Gamma_{\lfloor \sqrt{2n} \rfloor} + 2\sqrt{2n} \cdot \overline{H}$$

Note that by (34), the first term tends to zero. The second term $\Gamma_l \to 0$ with $l \to \infty$. Thus, $f(n) \in O(\sqrt{n})$, so the series $\sum_n f(n)/n^2$ converges. This implies the existence of $\lim_{n\to\infty} u_n/n = \mathbf{I}(V[\pi])$.

We can show that $\mathbf{I}(V[\rho]) = \mathbf{I}(V[\pi])$ by the same trick of skipping $\lfloor \sqrt{n} \rfloor$ symbols at the beginning. □

To underline the fact that $\mathbf{I}(V[\rho])$ does not depend on ρ, we will always denote the i-capacity by $\mathbf{I}(V)$.

Theorem 2 (Information stability of the Markov-IDS channel sequence)**.** *For any ρ, the channel sequence $V[\rho]$ is information stable. In particular, the coding capacity $\mathbf{C}(V[\rho])$ achieves the mutual information capacity $\mathbf{I}(V)$, and thus does not depend on ρ:*

$$\mathbf{C}(V[\rho]) = \mathbf{C}(V) = \mathbf{I}(V). \tag{42}$$

Proof. The outline of the proof is depicted in Fig. 1.

Step 1. Transition $V \to \overline{V}$.

Consider channel sequence $\overline{V}_n[\rho] = V_{l_n|m_n|l_n|\ldots|m_n}[\rho] : \mathcal{X}^n \rightsquigarrow \overline{\mathcal{Y}}^{Q_n}$, for which the receiver knows the outputs corresponding to all the input subvectors of lengths l_n and m_n, where

$$m_n = \lfloor \sqrt{n} \rfloor,\ l_n = \lfloor \sqrt[4]{n} \rfloor,\ q_n = \left\lfloor \frac{n}{m_n + l_n} \right\rfloor,\ Q_n = 2q_n \tag{43}$$

and the last $r_n = n - q_n(m_n + l_n)$ input symbols are ignored. Note that with $n \to \infty$, the asymptotics of these variables are

$$Q_n \sim q_n \sim m_n \in O(\sqrt{n}),\ l_n \in O(\sqrt[4]{n}),\ r_n \in O(n^{3/4}).$$

Below, we will omit index $_n$ and write l, m, q and Q meaning that they depend on n by (43).

Note that in the case of channel $\overline{V}_n[\rho]$ the receiver knows all the information that is here in the case of channel $V_n[\rho]$ plus the boundaries of l- and m-blocks. This means that channel $V_n[\rho]$ can be represented as a pipeline of channel $\overline{V}_n[\rho]$ and merging function $\mathbf{g}_Q$, which merges Q vectors into one, i.e., $V_n[\rho] = \mathbf{g}_Q(\overline{V}_n[\rho])$.

We propose to represent the merging function $\mathbf{g}_Q$ as a pipeline of two functions. The first one merges the output of $\overline{V}_n[\rho]$, if the total length of output is $< 2An$. Formally, let

$$\mathbf{g}^\circ_n(\overline{y}_1^Q) = \begin{cases} \overline{y}_1^Q & \text{if } \mathbf{N}(\mathbf{g}_Q(\overline{y}_1^Q)) \le 2An \\ \mathbf{g}_Q(\overline{y}_1^Q) & \text{if } \mathbf{N}(\mathbf{g}_Q(\overline{y}_1^Q)) > 2An \end{cases} \tag{44}$$

Note that, with $n \to \infty$, function $\mathbf{g}^\circ_n(y)$ changes the input with probability 0 , i.e., $\Pr[\mathbf{g}^\circ_n(y) \ne y] \to 0$. Moreover, the image of short sequences does not intersect the image of long sequences, so function sequence $\mathbf{g}^\circ$ satisfy Proposition 6.

Now let $\overline{\mathbf{g}}$ be a sequence of functions, which glue the vectors of vectors that remained unglued after $\mathbf{g}^\circ$:

$$\overline{\mathbf{g}}_n(\overline{Y}) = \begin{cases} \mathbf{g}_Q(\overline{Y}), & \text{if } \overline{Y} \in \overline{\mathcal{Y}}^Q \wedge \mathbf{N}(\mathbf{g}_Q(\overline{Y})) \le 2An \\ \overline{Y}, & \text{if } \overline{Y} \in \overline{\mathcal{Y}} \wedge \mathbf{N}(\overline{Y}) > 2An \end{cases} \tag{45}$$

Note that the log-size of the inverse image of this function is

$$\Phi_n = \log_2 \max_{\overline{Y}} \left|\overline{\mathbf{g}}_n^{-1}(\overline{Y})\right| = \log_2 \binom{2An + Q - 1}{Q - 1} \sim \log_2 (2An)^{\sqrt{n}} \in O(\sqrt{n}\log n) \subset o(n).$$

Thus, function sequence $\overline{\mathbf{g}}$ satisfies Proposition 4. One can see that $\mathbf{g}_Q = \mathbf{g}^\circ_n \circ \overline{\mathbf{g}}_n$, thus, $V[\rho] = \overline{\mathbf{g}}(\mathbf{g}^\circ(\overline{V}[\rho]))$, and we obtain

$$\mathbf{I}(\overline{V}[\rho]) = \mathbf{I}(V),\ \mathbf{IS}(\overline{V}[\rho]) \iff \mathbf{IS}(V[\rho]). \tag{46}$$

Step 2. Transition $\overline{V} \to \widetilde{V}$.

Now, denote by $\widetilde{V}[\rho]$ the sequence of channels $\widetilde{V}_n[\rho] = V_{\widetilde{l}|m|\widetilde{l}|m|\ldots}[\rho]$, where number of l- and m-blocks is q, and $l = l_n$, $m = m_n$, and $q = q_n$ are defined by (43).

The difference between channels $\widetilde{V}_n[\rho]$ and $\overline{V}_n[\rho]$ is that each l-block is erased. Similarly to the previous step, transition from $\overline{V}_n[\rho]$ to $\widetilde{V}_n[\rho]$, i.e., erasing of the l-blocks, can be done by two steps via pipeline of two functions $\mathbf{f}^\circ_n$ and $\overline{\mathbf{f}}_n$, such that $\widetilde{V}_n[\rho] = \overline{\mathbf{f}}_n(\mathbf{f}^\circ_n(\widetilde{V}_n[\rho]))$.

Let function $\mathbf{f}^\circ_n$ substitute all l-blocks with special erasure symbol $\iota \notin \mathcal{Y}$, if their total length is greater than $2Aql$. Similarly to $\mathbf{g}^\circ$, function sequence $\mathbf{f}^\circ$ satisfies conditions for Proposition 6.

Let function $\overline{\mathbf{f}}_n$ erase all l-blocks after applying function $\mathbf{f}^\circ_n$. Then, the number Φ_n of possible inputs, corresponding to the same output, or, alternatively, the number of possible values of erased l-blocks, is upper-bounded by $\sum_{i=0}^{\lfloor 2Aql \rfloor} |\mathcal{Y} \cup \{\iota\}|^i \le (|\mathcal{Y}| + 1)^{2Aql+1}$. Obviously, A and $|\mathcal{Y}|$ are constants, and $ql \in O(n^{3/4})$, thus, $\log_2 \Phi_n \in O(n^{3/4}\log_2 n) \subset o(n)$, which means that function sequence $\overline{\mathbf{f}}$ satisfies Proposition 4.

The above considerations imply

$$\mathbf{I}(\widetilde{V}[\rho]) = \mathbf{I}(\overline{V}[\rho]), \mathbf{IS}(\widetilde{V}[\rho]) \iff \mathbf{IS}(\overline{V}[\rho]). \tag{47}$$

Step 3. Information stability of $\widetilde{V}$.

Similarly to (40), one can obtain

$$(1 - \varepsilon_l)^q V_m^q[\pi](y_{\text{even}}|w) \le \widetilde{V}_n[\rho](y_1^Q|x) \le (1 + \varepsilon_l)^q V_m^q[\pi](y_{\text{even}}|w)$$

where $w = \widetilde{f}_n(x)$ is the input vector x without the l-blocks and the last incomplete block, i.e. $\widetilde{f}_n(x) = (x^{(1)}, x^{(2)}, \ldots, x^{(q)})$, $x^{(i)} = x_{i(m+l)-m+1}^{i(m+l)}$. Denote by $\overline{p}_n \in \mathbb{D}_{\mathcal{X}^{n'}}$, $n' = qm$ the marginalized distribution of $\widetilde{f}_n(x)$, given distribution p_n over x:

$$\overline{p}_n(w) = \sum_{x:\widetilde{f}_n(x)=w} p_n(x)$$

The fact that channel probabilities of $\widetilde{V}_n[\rho]$ and $V_m^q[\pi]$ are close implies that mutual information densities of $\widetilde{V}_n[\rho]$ and $V_m^q[\pi]$ are close too, since

$$\begin{aligned}\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y) &= \log_2 \frac{\widetilde{V}_n[\rho](y|x)}{\sum_{x'} p_n(x')\widetilde{V}_n[\rho](y|x')} \le \log_2 \frac{(1+\varepsilon_l)^q \cdot V_m^q[\pi](y_{\text{even}}|w)}{(1-\varepsilon_l)^q \cdot \overline{p}_n \circ V_m^q[\pi](*, y_{\text{even}})} \\ &= \mathbf{i}(V_m^q[\pi], \overline{p}_n, w, y_{\text{even}}) + q\log_2 \frac{1+\varepsilon_l}{1-\varepsilon_l}. \end{aligned} \tag{48}$$

The last term tends to zero with $n \to \infty$, since $\varepsilon_l \le Ce^{-D\sqrt[4]{n}}$ decays exponentially fast and $q \in O(\sqrt{n})$. Similarly, $\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y) \ge \mathbf{i}(V_m^q[\pi], \overline{p}_n, w, y_{\text{even}}) - q\log_2 \frac{1+\varepsilon_l}{1-\varepsilon_l}$, and thus

$$\lim_{n\to\infty} \mathbf{E}\Big[\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y) - \mathbf{i}(V_m^q[\pi], \overline{p}_n, x, y_{\text{even}})\Big] = \lim_{n\to\infty} \mathbf{E}\Big[\Big|\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y) - \mathbf{i}(V_m^q[\pi], \overline{p}_n, x, y_{\text{even}})\Big|\Big] = 0.$$

Now, let $\overline{p}_n$ be the direct product of q distributions p_m^* which are optimal for $V_m[\pi]$, i.e. let $\overline{p}_n(x_1^q) = \prod_{i=1}^q p_m^*(x_i)$ and let p_n be any distribution such that $\widetilde{f}_n(p_n) = \overline{p}_n$. Then,

$$\begin{aligned}&\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n} \cdot \left|\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y) - \sum_{i=1}^q \mathbf{i}(V_m[\pi], p_m^*, x^{(i)}, y_{\text{even}})\right|\right] \\ &= \lim_{n\to\infty} \mathbf{E}\left[\left|\frac{\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y)}{n} - \frac{q}{n} \cdot \mathbf{I}(V_m[\pi])\right|\right] = \lim_{n\to\infty} \mathbf{E}\left[\left|\frac{\mathbf{i}(\widetilde{V}_n[\rho], p_n, x, y)}{n} - \mathbf{I}(V)\right|\right] = 0. \end{aligned} \tag{49}$$

By (46) and (47), $\mathbf{I}(\widetilde{V}[\rho]) = \mathbf{I}(V)$. By (49), the channel sequence $\widetilde{V}[\rho]$ satisfies Proposition 2, which implies $\mathbf{IS}(\widetilde{V}[\rho])$, which by (46) and (47) is equivalent to $\mathbf{IS}(V)$. □

VI. Conclusion

In this paper, we have focused on channels with synchronization errors, for which the insertions/deletions are governed by a Markov chain. Specifically, we have shown that if the underlying Markov chain is stationary and ergodic, then the information capacity of the corresponding channel sequence exists and it is equal to the coding capacity. This result generalizes the classical result of Dobrushin on the existence of the Shannon capacity for channels with memoryless synchronization errors. To accomplish this goal, we generalized methods used in [4], [5] as separate independent propositions. These propositions state that one can apply a function sequence to a channel sequence, changing neither the coding capacity nor the mutual information capacity of the latter, if the function sequence has asymptotically small preimage or asymptotically small probability of changing the output. By applying such function sequences, one can bring the original channel sequence to another channel sequence, for which the capacity theorem or capacity bounds can be more easily derived, and one can be sure that the coding capacity / mutual information capacity is not changed under such transformation. The methodology may be useful for deriving capacity theorems and capacity bounds for other non-trivial channel sequences. The future work can be done towards generalization of this paper and [10] in order to include simultaneously the state's randomness and its dependence on input symbols.

References

[1] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[2] G. D. Hu, "On Shannon theorem and its converse for sequences of communication schemes in the case of abstract random variables," in *Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, 1962, pp. 285–332.

[3] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, 1994.

[4] R. L. Dobrushin, "A general formulation of the fundamental theorem of Shannon in the theory of information," *Uspekhi Mat. Nauk*, vol. 14, no. 6, pp. 3–104, 1959.

[5] ——, "Shannon's theorems for channels with synchronization errors," *Problems of Information Transmission*, vol. 3, no. 4, pp. 18–36, 1967.

[6] S. Z. Stambler, "Memoryless channels with synchronization errors: the general case," *Probl. Peredachi Inf.*, vol. 6, no. 3, pp. 43–49, 1970.

[7] R. G. Gallager, *Information theory and reliable communication*. Wiley, 1968.

[8] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channels," *IEEE Transactions on Information Theory*, vol. 35, no. 6, pp. 1277–1290, 1989.

[9] Y. Li and V. Y. F. Tan, "On the capacity of channels with deletions and states," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2663–2679, 2021.

[10] W. Mao, S. N. Diggavi, and S. Kannan, "Models and information-theoretic bounds for nanopore sequencing," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 3216–3236, 2018.
[11] J. Hu, T. M. Duman, K. E. M., and E. M. F., "Bit patterned media with written-in errors: Modelling, detection and theoretical limits," *IEEE Transactions on Magnetics*, vol. 43, pp. 3517–3524, 2007.
[12] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3675–3689, 2021.
[13] D. Fertonani and T. M. Duman, "Novel bounds on the capacity of binary deletion channels," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2753–2765, 2010.
[14] D. Fertonani, T. M. Duman, and M. F. Erden, "Upper bounds on the capacity of deletion channels using channel fragmentation," *IEEE Transactions on Communications*, vol. 59, no. 1, pp. 2–6, 2011.
[15] E. Drinea and M. Mitzenmacher, "On lower bounds for the capacity of deletion channels," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4648–4657, 2006.
[16] Y. Kanoria and A. Montanari, "Optimal coding for the binary deletion channel with small deletion probability," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6192–6219, 2013.
[17] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, pp. 1–33, 2009.
[18] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3207–3232, 2021.
[19] L. Deng, Y. Wang, M. Noor-A-Rahim, Y. L. Guan, Z. Shi, E. Gunawan, and C. L. Poh, "Optimized code design for constrained DNA data storage with asymmetric errors," *IEEE Access*, vol. 7, pp. 84 107–84 121, 2019.
[20] B. Hamoum and E. Dupraz, "Channel model and decoder with memory for DNA data storage with nanopore sequencing," *IEEE Access*, vol. 11, pp. 52 075–52 087, 2023.
[21] P. Billingsley, *Probability and Measure*, ser. Wiley Series in Probability and Statistics. Wiley, 2012.
[22] N. Bruijn, de and P. Erdös, "Some linear and some quadratic recursion formulas. II," *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen: Series A: Mathematical Sciences*, vol. 14, pp. 152–163, 1952.