

SPECTRUM INVARIANCY UNDER OUTPUT APPROXIMATION FOR FULL-RANK DISCRETE MEMORYLESS CHANNELS

T. S. Han and S. Verdú

UDC 621.391.1:519.27

Given a channel, the resolvability of an input process is the minimum randomness of those input processes whose output statistics approximate the original output statistics with arbitrary accuracy. We give a formula for the resolvability of any input process when the channel is full-rank discrete memoryless. When the input process is stationary and ergodic, its resolvability is equal to its mutual information rate. This result is obtained as a corollary of a theorem that shows that if two input processes result in approximately the same output statistics, then their corresponding information spectra (distributions of normalized information density) are almost identical.

Keywords: Shannon Theory, Output Statistics, Resolvability, Discrete Memoryless Channels, Method of Types

1. Introduction

The recent work of Han and Verdú [1] has formulated the following new problem: given a channel and an input process, what is the minimum complexity of an alternate input process which results in almost the same output distributions as the original input? One application elaborated in [1] is the randomness necessary to simulate the input to a random system. Although this question does not involve the coding or decoding of information, it turns out to be Shannon theoretic in nature, and to have strong connections with three important problems in information theory: noiseless source coding, channel coding and identification via channels [2]. For example, the results in [1] lead to a simple proof of the strong converse to the identification coding theorem that holds in wide generality. This is in contrast to the much more involved approaches of [2] and [3] which hold only for discrete memoryless channels.

When the approximation of output statistics is measured in variational distance and the complexity is the worst-case number of bits necessary to generate every realization of the input, the foregoing new concept is called resolvability. The maximum resolvability for all input processes is called the resolvability of the channel. It quantifies the minimum number of bits per input symbol that are necessary to reproduce the output statistics with arbitrary accuracy, regardless of the specific input. The main achievement of [1] is a formula for the resolvability of a channel that holds in complete generality. This formula turns out to coincide with the Shannon capacity for any channel that satisfies the strong converse to the channel coding theorem. The direct part of this result consists of showing that the resolvability of any input process is upper bounded by its sup-information rate (in most cases of interest, this coincides with the conventional mutual information rate). However, a counterexample is given in [1] to show that that bound is not satisfied with equality for certain discrete memoryless channels, thereby leaving the question of finding the resolvability of specific inputs as an open problem. The main purpose of this paper is to solve this open problem for an important class of channels: discrete-memoryless channels with full rank. We show here that the upper bound given by the general direct part of the resolvability theorem is satisfied with equality.

One of the contributions of [1] was to highlight the importance of the information spectrum (distribution of the normalized information density), and to show that the limits in probability (rather than the averages) of the information spectrum are the quantities of interest when dealing with general (nonergodic/nonstationary/unstable) sources and channels. These conclusions give justice to the pioneering works of Dobrushin [4] and Pinsker [5] which championed the use of information densities, albeit in the context of information stable channels. In this paper, our conclusions on resolvability are achieved as simple consequences of a stronger result: in full-rank DMCs, output approximation implies information spectrum

Translated from *Problemy Peredachi Informatsii*, Vol. 29, No. 2, pp. 9-27, April-June, 1993. Original article submitted September 2, 1992.

approximation, which means that if an alternate input process results asymptotically in, essentially, the same output statistics, then it is necessary that the information spectrum it induces is essentially the same as the original one.

The organization of the rest of the paper is: Section 2 presents the necessary notation and definitions. For completeness, the relevant definitions of [1] are reproduced, although some degree of familiarity with [1] is beneficial for the reader of this paper. Section 3 states the main results of the paper. The invariance of the information spectrum under output approximation is proved in Section 4. Finally, Section 5 is devoted to the proof of a converse result for the associated concept of mean-resolvability.

2. Notation and Definitions

DEFINITION 1. A channel W with input and output alphabets, A and B , respectively, is a sequence of conditional distributions

$$W = \{W^n(y^n|x^n) = P_{Y^n|X^n}(y^n|x^n); (x^n, y^n) \in A^n \times B^n\}_{n=1}^{\infty}.$$

In particular, a discrete memoryless channel (DMC) is defined by

$$W^n(y^n|x^n) = \prod_{i=1}^n W(y_i|x_i), \quad x^n \in A^n, y^n \in B^n$$

where A and B are finite sets, $W: A \rightarrow B$ is a stochastic matrix and $x^n = (x_1, \dots, x_n)$, $y^n = (y_1, \dots, y_n)$. A DMC is sometimes denoted simply by W instead of W .

DEFINITION 2. A full-rank discrete memoryless channel (FRDMC) $W = \{W(y|x)\}_{x \in A, y \in B}$ is a DMC such that the transition vectors $\{W(\cdot|x)\}_{x \in A}$ are linearly independent. For example, a binary symmetric channel (BSC) with cross-over probability $\alpha \neq 1/2$ is a FRDMC. Note that every FRDMC satisfies $|A| \leq |B|$.

In order to describe the statistics of input/output processes, we will use the sequence of finite-dimensional distributions $\{X^n = (X_1^{(n)}, \dots, X_n^{(n)})\}_{n=1}^{\infty}$, which is abbreviated as X . The following notation will be used for the output and joint distributions when the input is distributed according to Q^n :

$$Q^n W^n(y^n) = \sum_{x^n \in A^n} W^n(y^n|x^n) Q^n(x^n).$$

$$Q^n \circ W^n(x^n, y^n) = Q^n(x^n) W^n(y^n|x^n), \quad x^n \in A^n, y^n \in B^n.$$

DEFINITION 3 [5]. Given a joint distribution $P_{X^n Y^n}(x^n, y^n) = P_{X^n}(x^n) W^n(y^n|x^n)$, the information density is the function defined on $A^n \times B^n$:

$$i_{X^n W^n}(a^n, b^n) = \log \frac{W^n(b^n|a^n)}{P_{Y^n}(b^n)}.$$

The distribution of the random variable $\frac{1}{n} i_{X^n W^n}(X^n, Y^n)$ where X^n and Y^n have joint distribution $P_{X^n Y^n}$ will be referred to as the information spectrum. The expected value of the information spectrum is the normalized mutual information $\frac{1}{n} I(X^n; Y^n)$.

DEFINITION 4. The limsup in probability of a sequence of random variables $\{A_n\}$ is defined as the smallest extended real number β such that for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P[A_n \geq \beta + \epsilon] = 0.$$

Analogously, the liminf in probability is the largest extended real number α such that for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P[A_n \leq \alpha - \epsilon] = 0.$$

Note that a sequence of random variables converges in probability to a constant if and only if its limsup in probability is equal to its liminf in probability. The limsup in probability [resp. liminf in probability] of the sequence of random variables $\{\frac{1}{n}i_{X^n W^n}(X^n, Y^n)\}_{n=1}^{\infty}$ will be referred to as the sup-information rate [resp. inf-information rate] of the pair (X, Y) and will be denoted as $\bar{I}(X; Y)$ [resp. $\underline{I}(X; Y)$]. The mutual information rate of (X, Y) , if it exists, is the limit

$$I(X; Y) = \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n; Y^n).$$

DEFINITION 5. For any positive integer M , a probability distribution P on Ω is said to be M -type if

$$P(\omega) \in \left\{ 0, \frac{1}{M}, \frac{2}{M}, \dots, \frac{M}{M} \right\} \text{ for all } \omega \in \Omega.$$

DEFINITION 6. The resolution $R(P)$ at a probability distribution of P is the minimum log M such that P is M -type. (If P is not M -type for any integer M , we set $R(P) = +\infty$.)¹

The concept of resolution has first been introduced in [1] as a new kind of measure of complexity/randomness of distributions other than Shannon entropy, R enyi α -entropy etc.

An immediate consequence of the definition of resolution is

Lemma 1 [1].

$$P [i_{X^n W^n}(X^n, Y^n) \leq R(X^n)] = 1. \quad (2.1)$$

DEFINITION 7. The variational distance (or l_1 -distance) between two distributions P and Q defined on the same measurable space (Ω, F) is

$$d(P, Q) = \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| = 2 \sup_{E \in F} |P(E) - Q(E)|.$$

DEFINITION 8. Fix a channel W . Let $\varepsilon \geq 0$. R is an ε -achievable resolution rate for input X if for all $\gamma > 0$, there exists \tilde{X} whose resolution satisfies

$$\frac{1}{n} R(\tilde{X}^n) < R + \gamma \quad (2.2)$$

and

$$d(Y^n, \tilde{Y}^n) < \varepsilon \quad (2.3)$$

for all sufficiently large n , where Y and \tilde{Y} are the output statistics due to input processes X and \tilde{X} , respectively, i.e.

$$P_{Y^n} = P_{X^n} W^n$$

$$P_{\tilde{Y}^n} = P_{\tilde{X}^n} W^n$$

If R is an ε -achievable resolution rate, for every $\varepsilon > 0$, then, we say that R is an achievable resolution rate. By definition, the set of (ε -)achievable resolution rates is either empty or a closed interval. The minimum ε -achievable resolution rate [resp. achievable resolution rate] is called the ε -resolvability [resp. resolvability] of the process X , and it is denoted by $S_\varepsilon(X)$ [resp. $S(X)$]. Note that $S_\varepsilon(X)$ is monotonically nonincreasing in ε and

$$\sup_{\varepsilon > 0} S_\varepsilon(X) = S(X). \quad (2.4)$$

¹For distributions of random variables we shall use the common abbreviation $R(P_X) = R(X)$.

The resolvability of the channel \mathbf{W} is defined in [1] as

$$S = \sup_{\mathbf{X}} S(\mathbf{X}).$$

If instead of resolution, the randomness measure in Definition 8 is entropy, then the analogous concept is called mean-resolvability and a bar is used to differentiate it from resolvability: $\bar{S}(\mathbf{X})$, \bar{S} , etc. We refer the reader to [1] for the motivation of these definitions.

3. Theorems

Theorem 6 of [1] states that the resolvability of a finite-input channel is given by

$$S = \sup_{\mathbf{X}} \bar{I}(\mathbf{X}; \mathbf{Y}).$$

If, in addition, the channel satisfies the strong converse to the channel coding theorem, then

$$S = C = \lim_{n \rightarrow \infty} \frac{1}{n} \sup_{X^n} \frac{1}{n} I(X^n; Y^n)$$

where C is the channel capacity. A limited study of mean-resolvability in [1] concluded that

$$\bar{S} = C$$

for a certain class of channels including BSCs.

However, no formula is known for the resolvability and mean-resolvability of individual input \mathbf{X} except in the case of the identity channel, in which case [1]

$$S(\mathbf{X}) = \bar{H}(\mathbf{X}) \tag{3.1}$$

and

$$\bar{S}(\mathbf{X}) = \limsup_{n \rightarrow \infty} \frac{1}{n} H(X^n) \tag{3.2}$$

where $H(X^n)$ denotes the entropy, and $\bar{H}(\mathbf{X})$ is the sup-entropy rate, defined as a special case of the sup-information rate defined in Section 2.

The achievability result in [1] shows that

$$S(\mathbf{X}) \leq \bar{I}(\mathbf{X}; \mathbf{Y}) \tag{3.3}$$

for every input process and channel. This bound of resolvability may not be tight for certain channels as shown in [1]. In this paper we show that (3.3) is satisfied with equality for every input process and every FRDMC. Furthermore, with a FRDMC, an approximation of output statistics turns out to preserve almost invariant a much finer structure of the input/output pair (\mathbf{X}, \mathbf{Y}) , that is, the information spectrum itself, rather than some partial aspects of the information spectrum such as the sup-information rate $\bar{I}(\mathbf{X}; \mathbf{Y})$ and the (normalized) mutual information (i.e., the expected value of the information spectrum).

In this Section, we shall first state our main result (Theorem 1): the invariance of information spectrum under output approximation, and, then, we will prove its consequences on resolvability.

Theorem 1 (Spectrum invariance under output approximation). *Given a full-rank DMC and any $\epsilon > 0$, let \mathbf{X} and $\tilde{\mathbf{X}}$ be such that $d(Y^n, \tilde{Y}^n) < \epsilon$. For every $\tau > 0$, and for every set of reals $D \subset \mathbf{R}$ we have*

$$P \left[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \in D \right] \leq P \left[\frac{1}{n} i_{\tilde{X}^n W^n}(\tilde{X}^n, \tilde{Y}^n) \in D_\tau \right] + \lambda(\epsilon) \tag{3.4}$$

$$P \left[\frac{1}{n} i_{\tilde{X}^n W^n}(\tilde{X}^n, \tilde{Y}^n) \in D \right] \leq P \left[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \in D_\tau \right] + \lambda(\epsilon) \tag{3.5}$$

for all sufficiently large n , where $\lambda(\varepsilon) > 0$ is a continuous function of $\varepsilon > 0$ such that $\lambda(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$ and $D_\tau = \{x \in \mathbf{R} \mid |x - y| \leq \tau \text{ for some } y \in D\}$.

This theorem demonstrates that the information spectrum remains almost invariant under output approximation in the sense of (2.3). The proof of Theorem 1 is given in Section 4.

An immediate consequence (a weaker version of Theorem 1) is the following:

Corollary 1 (Approximation under Lévy distance). *Let $L(U, V)$ be the Lévy distance between the distributions of the random variables U and V , that is,*

$$L(U, V) = \inf \{ \mu > 0 \mid P[U \leq x - \mu] - \mu \leq P[V \leq x] \leq P[U \leq x + \mu] + \mu \}$$

for all real x . Then, every FRDMC satisfies for any $\varepsilon > 0$ and all sufficiently large n

$$L \left(\frac{1}{n} i_{X^n W^n}(X^n, Y^n), \frac{1}{n} i_{\tilde{X}^n W^n}(\tilde{X}^n, \tilde{Y}^n) \right) \leq \lambda(\varepsilon) \quad (3.6)$$

where the relevant quantities are defined in Theorem 1.

Proof. Setting $D = (-\infty, x - \lambda(\varepsilon)]$ in (3.4), $D = (-\infty, x]$ in (3.5) with $\tau = \lambda(\varepsilon)$ yields the required result.

Let us now proceed with the problem of characterizing the resolvability $S(\mathbf{X})$ for individual inputs \mathbf{X} , which is readily solved using Theorem 1.

Theorem 2 (Resolvability for individual inputs). *For every FRDMC and every input \mathbf{X} , we have*

$$S(\mathbf{X}) = \bar{I}(\mathbf{X}; \mathbf{Y}). \quad (3.7)$$

Proof. In view of (3.3), we shall show

$$S(\mathbf{X}) \geq \bar{I}(\mathbf{X}; \mathbf{Y})$$

under the assumption of full-rank discrete memoryless channel.

In order to invoke a contradiction argument, suppose that there exists an achievable resolution rate R for \mathbf{X} such that $R < \bar{I}(\mathbf{X}; \mathbf{Y})$. Fix $\varepsilon > 0, \gamma > 0$, and let $\tilde{\mathbf{X}}$ be as in Definition 8. Then, in view of Lemma 1 we have

$$P \left[\frac{1}{n} i_{\tilde{X}^n W^n}(\tilde{X}^n, \tilde{Y}^n) \leq R + \gamma \right] = 1. \quad (3.8)$$

On the other hand, setting $D = (-\infty, R + \gamma], \tau = \gamma$ in (3.5) of Theorem 1 and using (3.8) yields

$$1 = P \left[\frac{1}{n} i_{\tilde{X}^n W^n}(\tilde{X}^n, \tilde{Y}^n) \leq R + \gamma \right] \leq P \left[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \leq R + 2\gamma \right] + \lambda(\varepsilon).$$

Hence,

$$P \left[\frac{1}{n} i_{X^n W^n}(X^n, Y^n) \leq R + 2\gamma \right] \geq 1 - \lambda(\varepsilon)$$

for all sufficiently large n , which contradicts the definition of $\bar{I}(\mathbf{X}; \mathbf{Y})$, because $\gamma > 0, \varepsilon > 0$ can be chosen arbitrarily small and we have assumed that $R < \bar{I}(\mathbf{X}; \mathbf{Y})$.

If the input process is ergodic and stationary, then the Shannon-MacMillan theorem implies that

$$\bar{I}(\mathbf{X}; \mathbf{Y}) = \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n; Y^n)$$

because the channel is discrete memoryless. Therefore, under those conditions, the resolvability of an input process is equal to its mutual information rate.

Regarding mean-resolvability, the achievability result of [1] implies that

$$\bar{S}(\mathbf{X}) \leq \bar{I}(\mathbf{X}; \mathbf{Y}) \quad (3.9)$$

Even when (3.3) is tight, (3.9) may not be tight, as is evident from (3.1) and (3.2). In Section 5 we prove the following converse

Theorem 3 (Converse for individual mean-resolvability). *For any FRDMC and every input \mathbf{X} , we have*

$$\bar{S}(\mathbf{X}) \geq \limsup_{n \rightarrow \infty} \frac{1}{n} I(X^n; Y^n). \quad (3.10)$$

Corollary 2. *For any full-rank DMC, we have*

$$\bar{S} = S = C,$$

where C is the channel capacity of the DMC.

Proof. Let \bar{X}^n be the distribution that maximizes $I(X^n; Y^n)$, which exists because the channel is a DMC. Consider the following chain of inequalities

$$S \geq \bar{S} \geq \bar{S}(\bar{X}) \geq \limsup_{n \rightarrow \infty} \frac{1}{n} I(\bar{X}^n; \bar{Y}^n) = C = S,$$

where the first inequality follows from the fact that resolution is larger than entropy, the third inequality follows from Theorem 3 and the last relationship follows from the fact that a DMC satisfies the strong converse to the coding theorem [6].

Corollary 2 enlarges the class of channels for which it was shown in [1] that mean-resolvability is equal to capacity. We conjecture that the lower bound in Theorem 3 is satisfied with equality for every input process and any FRDMC.

4. Proof of Theorem 1

The proof is performed in several steps. The first step is to modify the given DMC W^n to define the associated clipped channel W_δ^* (cf.[3]).

Given an input distribution $\{Q^n \text{ on } A^n\}_{n=1}^\infty$ and an approximating input distribution $\{\tilde{Q}^n \text{ on } A^n\}_{n=1}^\infty$ in the sense that $d(Q^n W^n, \tilde{Q}^n W^n) < \epsilon$, the second step is to show that the information spectra under the distributions $Q^n \circ W^n$ and $Q^n \circ W_\delta^*$ (resp. the information spectra under the distributions $\tilde{Q}^n \circ W^n$ and $\tilde{Q}^n \circ W_\delta^*$) are close, respectively. The third step is to show that, via $d(Q^n W^n, \tilde{Q}^n W^n) < \epsilon$, the information spectra under $Q^n \circ W_\delta^*$ and $\tilde{Q}^n \circ W_\delta^*$ are close; in this step we invoke the full-rank assumption. Finally, the fourth step is simply to combine the second and third steps to conclude that the information spectra under $Q^n \circ W^n$ and $\tilde{Q}^n \circ W^n$ are close as was stated in Theorem 1.

Step 1. Let us start by giving some notation concerning types of sequences (cf.[6]). For a sequence $z^n = (z_1, \dots, z_n) \in \Omega^n$ (Ω is finite) we define the type of z^n by

$$\text{type}(z^n) = \{n_a/n\}_{a \in \Omega}$$

where n_a is the number of i such that $z_i = a$ ($i = 1, \dots, n$). Similarly, for two sequences $z^n = (z_1, \dots, z_n) \in \Omega^n$ and $w^n = (w_1, \dots, w_n) \in \Phi^n$ (Φ is finite), the conditional type of z^n given w^n is defined by

$$\text{type}(z^n | w^n) = \{n_{ab}/n_b\}_{a \in \Omega, b \in \Phi}$$

where n_{ab} is the number of i such that $(z_i, w_i) = (a, b)$ and n_b is the number of i such that $w_i = b$. It is evident $n_b = \sum_{a \in \Omega} n_{ab}$.

Consider here a DMC with input alphabet A and output alphabet B and let denote by Γ the set of all possible types $\text{type}(\mathbf{x}^n)$ for $\mathbf{x}^n \in A^n$.

Similarly, for each $P \in \Gamma$ let denote by Λ^P the set of all possible conditional types $\text{type}(y^n | x^n)$ for $x^n \in A^n, y^n \in B^n$ such that $P = \text{type}(\mathbf{x}^n)$. For a type $P = \{n_a\}_{a \in A}$ and a conditional type $V = \{n_{ba}/n_a\}_{a \in A, b \in B}$ we indicate by PV the type $\{n_b/n\}_{b \in B}$ for $y^n \in B^n$, where $n_b = \sum_{a \in A} n_{ba}$.

Now, given a DMC W^n , we define the clipped channel W_δ^* of W^n as follows. Let $\delta > 0$ be an arbitrarily small positive number, and for $P \in \Gamma$ set

$$\Lambda_\delta^P = \{V \in \Lambda^P | D(V || W | P) \leq \delta\}, \quad (4.1)$$

where $D(\cdot || \cdot | \cdot)$ is the conditional divergence (cf.[6]). Moreover, for $\mathbf{x}^n \in A^n$ and conditional type V , set

$$T_V(\mathbf{x}^n) = \{y^n \in B^n | \text{type}(y^n | \mathbf{x}^n) = V\}, \quad (4.2)$$

$$T^*(\mathbf{x}^n) = \bigcup_{V \in \Lambda_\delta^P} T_V(\mathbf{x}^n), \quad (4.3)$$

where $P = \text{type}(\mathbf{x}^n)$. It is seen (e.g., cf.[3]) that, if we put

$$W^n(T^*(\mathbf{x}^n) | \mathbf{x}^n) = 1 - \sigma_n(\mathbf{x}^n),$$

then

$$0 \leq \sigma_n(\mathbf{x}^n) \leq (n+1)^{|A||B|} e^{-n\delta} \equiv \gamma_n. \quad (4.4)$$

Clearly, $\gamma_n \rightarrow 0$ as n tends to infinity. The clipped channel $W_\delta^* : A^n \rightarrow B^n$ of W^n (cf.[3]) is given by

$$W_\delta^*(y^n | \mathbf{x}^n) = \begin{cases} \frac{W^n(y^n | \mathbf{x}^n)}{1 - \sigma_n(\mathbf{x}^n)} & \text{for } y^n \in T^*(\mathbf{x}^n), \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

We set

$$\begin{aligned} K &= \{(\mathbf{x}^n, y^n) \in A^n \times B^n | W^n(y^n | \mathbf{x}^n) > 0\}, \\ K^* &= \{(\mathbf{x}^n, y^n) \in A^n \times B^n | W_\delta^*(y^n | \mathbf{x}^n) > 0\}. \end{aligned} \quad (4.6)$$

It is clear that $K^* \subset K$. It follows from (4.4), (4.5) and (4.6) that, for every $(\mathbf{x}^n, y^n) \in K^*$

$$W^n(y^n | \mathbf{x}^n) \leq W_\delta^*(y^n | \mathbf{x}^n) \leq \frac{W^n(y^n | \mathbf{x}^n)}{1 - \gamma_n}. \quad (4.7)$$

Step 2. a). Given any sequence of input distributions $\{Q^n\}_{n=1}^\infty$, we shall show the existence of some subset $\bar{T}_Y \subset B^n$ such that $Q^n W^n(\bar{T}_Y)$ is almost equal to one and $Q^n W_\delta^*(y^n)$ is very close to $Q^n W^n(y^n)$ for every $y^n \in \bar{T}_Y$. To do so, we first observe that, in view of the second inequality in (4.7)

$$\sum_{(\mathbf{x}^n, y^n) \in K^*} W^n(y^n | \mathbf{x}^n) Q^n(\mathbf{x}^n) \geq \sum_{(\mathbf{x}^n, y^n) \in K^*} (1 - \gamma_n) W_\delta^*(y^n | \mathbf{x}^n) Q^n(\mathbf{x}^n) = 1 - \gamma_n, \quad (4.8)$$

which is rewritten as

$$\sum_{(\mathbf{x}^n, y^n) \in K^*} V^n(\mathbf{x}^n | y^n) Q^n W^n(y^n) \geq 1 - \gamma_n, \quad (4.9)$$

where V^n is the inverse channel of W^n defined by

$$V^n(\mathbf{x}^n | y^n) = \frac{W^n(y^n | \mathbf{x}^n) Q^n(\mathbf{x}^n)}{Q^n W^n(y^n)}.$$

We rewrite (4.9) as

$$\sum_{y^n \in T_Y} V^n(T_X(y^n) | y^n) Q^n W^n(y^n) \geq 1 - \gamma_n, \quad (4.10)$$

where

$$T_X(y^n) = \{x^n \in A^n | (x^n, y^n) \in K^*\}, \quad (4.11)$$

$$T_Y = \{y^n \in B^n | (x^n, y^n) \in K^* \text{ for some } x^n \in A^n\}. \quad (4.12)$$

An application of reverse Markov inequality to (4.10) yields

$$Q^n W^n(\bar{T}_Y) \geq 1 - (\gamma_n + \sqrt{\gamma_n}) \geq 1 - 2\sqrt{\gamma_n} \quad (4.13)$$

for all sufficiently large n , where we have put

$$\bar{T}_Y = \{y^n \in B^n | V^n(T_X(y^n)|y^n) \geq 1 - \sqrt{\gamma_n}\}. \quad (4.14)$$

Then, for any $y^n \in \bar{T}_Y$

$$\begin{aligned} \sum_{x^n \in T_X(y^n)} W^n(y^n|x^n) Q^n(x^n) &= \sum_{x^n \in T_X(y^n)} V^n(x^n|y^n) Q^n W^n(y^n) = \\ &= V^n(T_X(y^n)|y^n) Q^n W^n(y^n) \geq (1 - \sqrt{\gamma_n}) Q^n W^n(y^n). \end{aligned} \quad (4.15)$$

On the other hand, for every $y^n \in \bar{T}_Y$

$$Q^n W^*(y^n) = \sum_{x^n \in T_X(y^n)} W^*(y^n|x^n) Q^n(x^n)$$

which combined with (4.7) derives that

$$\sum_{x^n \in T_X(y^n)} W^n(y^n|x^n) Q^n(x^n) \leq Q^n W_\delta^*(y^n) \leq \frac{1}{1 - \gamma_n} \sum_{x^n \in T_X(y^n)} W^n(y^n|x^n) Q^n(x^n) \leq \frac{Q^n W^n(y^n)}{1 - \gamma_n}. \quad (4.16)$$

Therefore, by means of (4.15) and (4.16), it follows that, for every $y^n \in \bar{T}_Y$

$$(1 - \sqrt{\gamma_n}) Q^n W^n(y^n) \leq Q^n W_\delta^*(y^n) \leq \frac{Q^n W^n(y^n)}{1 - \gamma_n}. \quad (4.17)$$

b). We shall show the existence of a subset \bar{K} of K^* such that both $Q^n \circ W^n(\bar{K})$ and $Q^n \circ W_\delta^*(\bar{K})$ are close to one, and such that, for every $(x^n, y^n) \in \bar{K}$, the two information densities under $Q^n \circ W^n$ and $Q^n \circ W_\delta^*$ are close to each other.

Set

$$\bar{K} = K^* \cap (A^n \times \bar{T}_Y), \quad (4.18)$$

then, from (4.9), (4.13) and (4.14), we have

$$\begin{aligned} Q^n \circ W^n(\bar{K}) &= \sum_{(x^n, y^n) \in \bar{K}} V^n(x^n|y^n) Q^n W^n(y^n) \geq \sum_{(x^n, y^n) \in K^*} V^n(x^n|y^n) Q^n W^n(y^n) + \\ &+ \sum_{(x^n, y^n) \in A^n \times \bar{T}_Y} V^n(x^n|y^n) Q^n W^n(y^n) - 1 \geq (1 - \gamma_n) + (1 - 2\sqrt{\gamma_n}) - 1 \geq 1 - 3\sqrt{\gamma_n} \end{aligned} \quad (4.19)$$

for sufficiently large n . Furthermore, taking account of (4.19) and the first inequality of (4.7), it follows that

$$Q^n \circ W_\delta^*(\bar{K}) = \sum_{(x^n, y^n) \in \bar{K}} W_\delta^*(y^n|x^n) Q^n(x^n) \geq \sum_{(x^n, y^n) \in \bar{K}} W^n(y^n|x^n) Q^n(x^n) = Q^n \circ W^n(\bar{K}) \geq 1 - 3\sqrt{\gamma_n}. \quad (4.20)$$

Now let us derive a relation between two information densities, that is,

$$i(x^n, y^n) = \log \frac{W^n(y^n|x^n)}{Q^n W^n(y^n)}$$

and

$$i^*(x^n, y^n) = \log \frac{W_\delta^*(y^n | x^n)}{Q^n W_\delta^*(y^n)}.$$

An immediate consequence of two inequalities (4.7) and (4.17) is that, for every $(x^n, y^n) \in \bar{K}$

$$(1 - \gamma_n) \frac{W^n(y^n | x^n)}{Q^n W^n(y^n)} \leq \frac{W_\delta^*(y^n | x^n)}{Q^n W_\delta^*(y^n)} \leq \frac{W^n(y^n | x^n)}{(1 - \gamma_n)(1 - \sqrt{\gamma_n})Q^n W^n(y^n)} \leq \frac{1}{1 - 2\sqrt{\gamma_n}} \frac{W^n(y^n | x^n)}{Q^n W^n(y^n)}$$

for all sufficiently large n . Therefore, it follows that, for $(x^n, y^n) \in \bar{K}$

$$i(x^n, y^n) + \log(1 - \gamma_n) \leq i^*(x^n, y^n) \leq i(x^n, y^n) + \log \frac{1}{1 - 2\sqrt{\gamma_n}}$$

and hence

$$i(x^n, y^n) - 2\gamma_n \leq i^*(x^n, y^n) + 3\sqrt{\gamma_n} \quad (4.21)$$

for all sufficiently large n .

c). In this final substep we shall show that the two information spectra under $Q^n \circ W^n$ and $Q^n \circ W_\delta^*$ are close to each other.

For any $D \subset \mathbb{R}$ set

$$\begin{aligned} T_D &= \left\{ (x^n, y^n) \in K : \frac{1}{n} i(x^n, y^n) \in D \right\}, \\ T_D^* &= \left\{ (x^n, y^n) \in K^* : \frac{1}{n} i^*(x^n, y^n) \in D \right\}. \end{aligned}$$

Then, (4.21) implies that, with any small $\delta > 0$ and for all sufficiently large n

$$T_D \cap \bar{K} \subset T_{D_\delta}^*, \quad (4.22)$$

$$T_D^* \cap \bar{K} \subset T_{D_\delta}, \quad (4.23)$$

because $\gamma_n \rightarrow 0$ as $n \rightarrow \infty$, where it should be recalled that $D_\delta = \{x \in \mathbb{R} : |x - y| \leq \delta \text{ for some } y \in D\}$. Then,

$$\begin{aligned} Q^n \circ W_\delta^*(T_D^*) &= \sum_{(x^n, y^n) \in T_D^*} W_\delta^*(y^n | x^n) Q^n(x^n) \leq \\ &\leq \sum_{(x^n, y^n) \in T_D^* \cap \bar{K}} W_\delta^*(y^n | x^n) Q^n(x^n) + \sum_{(x^n, y^n) \notin \bar{K}} W_\delta^*(y^n | x^n) Q^n(x^n) = \\ &= \sum_{(x^n, y^n) \in T_D^* \cap \bar{K}} W_\delta^*(y^n | x^n) Q^n(x^n) + Q^n \circ W_\delta^*(\bar{K}^c). \end{aligned}$$

By virtue of (4.20), (4.23) and the second inequality of (4.7), we have

$$\begin{aligned} Q^n \circ W_\delta^*(T_D^*) &\leq \sum_{(x^n, y^n) \in T_{D_\delta}^*} W_\delta^*(y^n | x^n) Q^n(x^n) + 3\sqrt{\gamma_n} \leq \\ &\leq \frac{1}{1 - \gamma_n} \sum_{(x^n, y^n) \in T_{D_\delta}^*} W^n(y^n | x^n) Q^n(x^n) + 3\sqrt{\gamma_n} = \frac{1}{1 - \gamma_n} Q^n \circ W^n(T_{D_\delta}) + 3\sqrt{\gamma_n}. \end{aligned}$$

Hence, it follows that

$$Q^n \circ W_\delta^*(T_D^*) \leq Q^n \circ W^n(T_{D_\delta}) + 4\sqrt{\gamma_n} \quad (4.24)$$

for all sufficiently large n .

In a similar manner, taking account of (4.19), (4.22) and the first inequality of (4.7), we have

$$\begin{aligned}
Q^n \circ W^n(T_D) &= \sum_{(x^n, y^n) \in T_D} W^n(y^n|x^n)Q^n(x^n) \leq \sum_{(x^n, y^n) \in T_D \cap \bar{K}} W^n(y^n|x^n)Q^n(x^n) + \\
&+ \sum_{(x^n, y^n) \notin \bar{K}} W^n(y^n|x^n)Q^n(x^n) \leq \sum_{(x^n, y^n) \in T_D \cap \bar{K}} W^n(y^n|x^n)Q^n(x^n) + 3\sqrt{\gamma_n} \leq \\
&\leq \sum_{(x^n, y^n) \in T_{D_\delta}^*} W^n(y^n|x^n)Q^n(x^n) + 3\sqrt{\gamma_n} \leq \sum_{(x^n, y^n) \in T_{D_\delta}^*} W_\delta^*(y^n|x^n)Q^n(x^n) + 3\sqrt{\gamma_n} = \\
&= Q^n \circ W_\delta^*(T_{D_\delta}^*) + 3\sqrt{\gamma_n}.
\end{aligned} \tag{4.25}$$

Summarizing (4.24) and (4.25), we have

Lemma 2. *Let $\delta > 0$ be an arbitrary small number. For any $D \subset \mathbb{R}$ it holds that*

$$\begin{aligned}
Q^n \circ W^n \left(\frac{1}{n} i(X^n, Y^n) \in D \right) &\leq Q^n \circ W_\delta^* \left(\frac{1}{n} i^*(X^n, Y^n) \in D_\delta \right) + 3\sqrt{\gamma_n} \quad , \\
Q^n \circ W_\delta^* \left(\frac{1}{n} i^*(X^n, Y^n) \in D \right) &\leq Q^n \circ W^n \left(\frac{1}{n} i(X^n, Y^n) \in D_\delta \right) + 4\sqrt{\gamma_n}
\end{aligned}$$

for all sufficiently large n .

Step 3. a). We shall show that an output approximation for the channel W^n is tantamount to an output approximation for the clipped channel W_δ^* .

Given an input distribution $\{Q^n\}_{n=1}^\infty$, let $\{\tilde{Q}^n\}_{n=1}^\infty$ be an approximating input distribution such that

$$d(Q^n W^n, \tilde{Q}^n W^n) < \varepsilon \tag{4.26}$$

for all sufficiently large n , where $\varepsilon > 0$ is a prescribed arbitrary number. We need here the following lemma:

Lemma 3 [3]. *For every $n, x^n \in A^n, D \subset B^n$ and $\delta > 0$*

$$W^n(D|x^n) \geq (1 - \gamma_n)W_\delta^*(D|x^n) \quad , \tag{4.27}$$

$$W^n(D|x^n) \leq W_\delta^*(D|x^n) + \gamma_n \quad , \tag{4.28}$$

where γ_n is specified in (4.4).

Multiplying both sides of (4.27) and (4.28) by $Q^n(x^n)$ and summing them over all $x^n \in A^n$, we have

$$Q^n W^n(D) \geq (1 - \gamma_n)Q^n W_\delta^*(D) \quad ,$$

$$Q^n W^n(D) \leq Q^n W_\delta^*(D) + \gamma_n \quad .$$

Therefore, for all $D \subset B^n$,

$$|Q^n W^n(D) - Q^n W_\delta^*(D)| \leq \gamma_n$$

and hence

$$d(Q^n W^n, Q^n W_\delta^*) = 2 \sup_{D \subset B^n} |Q^n W^n(D) - Q^n W_\delta^*(D)| \leq 2\gamma_n. \tag{4.29}$$

In an analogous manner, we have

$$d(\tilde{Q}^n W^n, \tilde{Q}^n W_\delta^*) \leq 2\gamma_n. \tag{4.30}$$

Thus, by combining (4.26), (4.29) and (4.30),

$$\begin{aligned}
d(Q^n W_\delta^*, \tilde{Q}^n W_\delta^*) &\leq d(Q^n W_\delta^*, Q^n W^n) + d(Q^n W^n, \tilde{Q}^n W^n) + d(\tilde{Q}^n W^n, \tilde{Q}^n W_\delta^*) \leq \\
&\leq \varepsilon + 4\gamma_n \leq 2\varepsilon
\end{aligned} \tag{4.31}$$

for sufficiently large n .

b). We shall show the existence of a subset F_0 of B^n such that both of $Q^n W_\delta^*(F_0)$ and $\tilde{Q}^n W_\delta^*(F_0)$ are close to one and such that, for every $y^n \in F_0$, $Q^n W_\delta^*(y^n)$ and $\tilde{Q}^n W_\delta^*(y^n)$ are close to each other.

By (4.31) and the definition of the distance d ,

$$\begin{aligned} 2\varepsilon &\geq d(Q^n W_\delta^*, \tilde{Q}^n W_\delta^*) = \sum_{y^n \in B^n} |Q^n W_\delta^*(y^n) - \tilde{Q}^n W_\delta^*(y^n)| = \\ &= \sum_{y^n: Q^n W_\delta^*(y^n) > 0} Q^n W_\delta^*(y^n) \left| 1 - \frac{\tilde{Q}^n W_\delta^*(y^n)}{Q^n W_\delta^*(y^n)} \right| + \sum_{y^n: Q^n W_\delta^*(y^n) = 0} \tilde{Q}^n W_\delta^*(y^n) \geq \\ &\geq \sum_{y^n: Q^n W_\delta^*(y^n) > 0} Q^n W_\delta^*(y^n) \left| 1 - \frac{\tilde{Q}^n W_\delta^*(y^n)}{Q^n W_\delta^*(y^n)} \right|. \end{aligned}$$

Applying here reverse Markov inequality, we have

$$Q^n W_\delta^*(F_1) \geq 1 - \sqrt{2\varepsilon}, \quad (4.32)$$

where

$$F_1 = \left\{ y^n \in B^n \mid Q^n W_\delta^*(y^n) > 0 \text{ and } \left| 1 - \frac{\tilde{Q}^n W_\delta^*(y^n)}{Q^n W_\delta^*(y^n)} \right| \leq \sqrt{2\varepsilon} \right\}.$$

Similarly, exchanging the roles of Q^n and \tilde{Q}^n , we have

$$\tilde{Q}^n W_\delta^*(F_2) \geq 1 - \sqrt{2\varepsilon}, \quad (4.33)$$

where

$$F_2 = \left\{ y^n \in B^n \mid \tilde{Q}^n W_\delta^*(y^n) > 0 \text{ and } \left| 1 - \frac{Q^n W_\delta^*(y^n)}{\tilde{Q}^n W_\delta^*(y^n)} \right| \leq \sqrt{2\varepsilon} \right\}.$$

Therefore, with $F_0 = F_1 \cap F_2$, we have, for any $y^n \in F_0$,

$$1 - \sqrt{2\varepsilon} \leq \frac{\tilde{Q}^n W_\delta^*(y^n)}{Q^n W_\delta^*(y^n)} \leq 1 + \sqrt{2\varepsilon}, \quad (4.34)$$

$$1 - \sqrt{2\varepsilon} \leq \frac{Q^n W_\delta^*(y^n)}{\tilde{Q}^n W_\delta^*(y^n)} \leq 1 + \sqrt{2\varepsilon}. \quad (4.35)$$

On the other hand, by virtue of (4.31)

$$\begin{aligned} Q^n W_\delta^*(F_2) &\geq \tilde{Q}^n W_\delta^*(F_2) - \varepsilon, \\ \tilde{Q}^n W_\delta^*(F_1) &\geq Q^n W_\delta^*(F_1) - \varepsilon. \end{aligned}$$

So that, by using (4.32) and (4.33) it follows that

$$\begin{aligned} Q^n W_\delta^*(F_0) &\geq Q^n W_\delta^*(F_1) + Q^n W_\delta^*(F_2) - 1 \geq Q^n W_\delta^*(F_1) + \tilde{Q}^n W_\delta^*(F_2) - (1 + \varepsilon) \geq \\ &\geq 1 - (\varepsilon + 2\sqrt{2\varepsilon}) = 1 - \lambda_1(\varepsilon), \end{aligned} \quad (4.36)$$

where

$$\lambda_1(\varepsilon) = \varepsilon + 2\sqrt{2\varepsilon}. \quad (4.37)$$

Similarly,

$$\tilde{Q}^n W_\delta^*(F_0) \geq 1 - \lambda_1(\varepsilon). \quad (4.38)$$

c). We shall prepare here some auxiliary propositions which will be needed in the next subsection.

Consider a type $P \in \Gamma$ and define

$$T_P = \{y^n \in B^n | (x^n, y^n) \in K^* \text{ and } \text{type}(x^n) = P \text{ for some } x^n \in A^n\}.$$

Then we have the following result whose proof is relegated to the Appendix. We call the attention to the fact that T_P thus defined actually depends on δ because the definition of T_P contains K^* , which in turn depends on δ (cf.(4.6)).

Lemma 4. For any FRDMC, let $P_1, P_2 \in \Gamma$. If $T_{P_1} \cap T_{P_2} \neq \emptyset$, then

$$d(P_1, P_2) \leq K\sqrt{\delta},$$

where $K > 0$ is a constant depending only on the channel W .

Lemma 5. (see [3]). If $D(V||W|P) \leq \gamma$, then

$$|H(V|P) - H(W|P)| \leq \varrho(\gamma),$$

where ϱ is a continuous function such that $\varrho(\gamma) \rightarrow 0$ as $\gamma \rightarrow 0$.

Take any pair of types $P_1, P_2 \in \Gamma$ and suppose that $T_{P_1} \cap T_{P_2} \neq \emptyset$. Fix a $y^n \in T_{P_1} \cap T_{P_2}$ and let $V_1 \in \Lambda_{P_1}, V_2 \in \Lambda_{P_2}$ be such that $V_1 = \text{type}(y^n|x_1^n), V_2 = \text{type}(y^n|x_2^n)$ for some $x_1^n, x_2^n \in A^n$ with $\text{type}(x_1^n) = P_1, \text{type}(x_2^n) = P_2$. Then, from the definition of K^* (see (4.1), (4.3), (4.5) and (4.6)), it is seen that

$$D(V_1||W|P_1) \leq \delta, \quad (4.39)$$

$$D(V_2||W|P_2) \leq \delta. \quad (4.40)$$

Hence, by virtue of Lemma 4,

$$\begin{aligned} |H(V_1|P_1) - H(W|P_1)| &\leq \varrho(\delta), \\ |H(V_2|P_2) - H(W|P_2)| &\leq \varrho(\delta). \end{aligned}$$

On the other hand,

$$\begin{aligned} |H(W|P_1) - H(W|P_2)| &= \left| \sum_{a \in A} H(W|a)(P_1(a) - P_2(a)) \right| \leq \\ &\leq \sum_{a \in A} H(W|a) |P_1(a) - P_2(a)| \leq \log |A| \cdot d(P_1, P_2) \leq K_0\sqrt{\delta}, \end{aligned}$$

where we have used Lemma 3 in the last step, and put $K_0 = K \log |A|$. Thus, we have

$$\begin{aligned} |H(V_1|P_1) - H(V_2|P_2)| &\leq |H(V_1|P_1) - H(W|P_1)| + |H(W|P_1) - H(W|P_2)| + \\ &+ |H(W|P_2) - H(V_2|P_2)| \leq 2\varrho(\delta) + K_0\sqrt{\delta}, \end{aligned}$$

which is summarized as

Lemma 6. Let $(x_1^n, y^n), (x_2^n, y^n) \in K^*$ and $P_1 = \text{type}(x_1^n), P_2 = \text{type}(x_2^n), V_1 = \text{type}(y^n|x_1^n), V_2 = \text{type}(y^n|x_2^n)$. Then

$$|H(V_1|P_1) - H(V_2|P_2)| \leq \varrho_1(\delta),$$

where $\varrho_1(\delta) \equiv 2\varrho(\delta) + K_0\sqrt{\delta} > 0$ is a continuous function of $\delta > 0$ such that $\varrho_1(\delta) \rightarrow 0$ as $\delta \rightarrow 0$.

d). We shall show that the two information spectra for the channels W_δ^* under the input distributions $\{Q^n\}$ and $\{\bar{Q}^n\}$ are close to each other. The argument is similar to that used in Step 2 along with a more subtle evaluation.

First, we observe that the channel transition probabilities $W^n(y^n|x_1^n), W^n(y^n|x_2^n)$ have the expressions (cf.[6]):

$$\begin{aligned} W^n(y^n|x_1^n) &= \exp[-n(H(V_1|P_1) + D(V_1|W|P_1))] \quad , \\ W^n(y^n|x_2^n) &= \exp[-n(H(V_2|P_2) + D(V_2|W|P_2))] \quad , \end{aligned}$$

where the same notation as in Lemma 6 has been used. Take any x_1^n, x_2^n, y^n such that $x_1^n \in T_X(y^n)$ and $x_2^n \in T_X(y^n)$ (cf.(4.11)), then, from (4.7), (4.39), (4.40) and Lemma 6 it follows that

$$(1 - \gamma_n) \exp[-n\varrho_2(\delta)] \leq \frac{W_\delta^*(y^n|x_1^n)}{W_\delta^*(y^n|x_2^n)} \leq \frac{1}{1 - \gamma_n} \exp[n\varrho_2(\delta)] \quad ,$$

where

$$\varrho_2(\delta) = \varrho_1(\delta) + \delta. \quad (4.41)$$

Consequently, by means of (4.34), (4.35) and (4.41) we have, for any $x_1^n, x_2^n \in T_X(y^n), y^n \in F_0$,

$$\frac{(1 - \gamma_n) \exp[-n\varrho_2(\delta)]}{1 - \sqrt{2\varepsilon}} \frac{W_\delta^*(y^n|x_2^n)}{\bar{Q}^n W_\delta^*(y^n)} \leq \frac{W_\delta^*(y^n|x_1^n)}{Q^n W_\delta^*(y^n)} \leq \frac{(1 + \sqrt{2\varepsilon}) \exp[n\varrho_2(\delta)]}{1 - \gamma_n} \frac{W_\delta^*(y^n|x_2^n)}{\bar{Q}^n W_\delta^*(y^n)}.$$

Therefore,

$$\tilde{i}^*(x_2^n, y^n) - n\varrho_2(\delta) + \log \frac{1 - \gamma_n}{1 - \sqrt{2\varepsilon}} \leq i^*(x_1^n, y^n) \leq \tilde{i}^*(x_2^n, y^n) + n\varrho_2(\delta) + \log \frac{1 + \sqrt{2\varepsilon}}{1 - \gamma_n} \quad ,$$

and hence

$$\left| \frac{1}{n} i^*(x_1^n, y^n) - \frac{1}{n} \tilde{i}^*(x_2^n, y^n) \right| \leq \mu(\delta) \quad (4.42)$$

for sufficiently large n , where

$$\mu(\delta) = \varrho_2(\delta) + \delta. \quad (4.43)$$

Define

$$Y_D^* = \{y^n \in B^n | (x^n, y^n) \in T_D^* \text{ for some } x^n \in B^n\} \quad , \quad (4.44)$$

$$T_D^*(y^n) = \{(x^n, y^n) | (x^n, y^n) \in T_D^*\}. \quad (4.45)$$

Since $y^n \in Y_D^* \cap F_0$ means the existence of some $x_1^n \in T_X(y^n)$ such that $\frac{1}{n} i^*(x_1^n, y^n) \in D$, (4.42) implies that

$$\frac{1}{n} \tilde{i}^*(x_2^n, y^n) \in D_{\mu(\delta)} \quad (4.46)$$

for any $x_2^n \in T_X(y^n)$. As a consequence, by virtue of (4.35) we have for any $y^n \in Y_D^* \cap F_0$,

$$\begin{aligned} Q^n \circ W_\delta^*(T_D^*(y^n)) &\leq Q^n \circ W_\delta^*(T_X(y^n)) = Q^n W_\delta^*(y^n) \leq \\ &\leq (1 + \sqrt{2\varepsilon}) \bar{Q}^n W_\delta^*(y^n) = (1 + \sqrt{2\varepsilon}) \bar{Q}^n \circ W_\delta^*(T_X(y^n)). \end{aligned} \quad (4.47)$$

Define here a counterpart of (4.44) and (4.45):

$$\tilde{Y}_D^* = \{y^n \in B^n | (x^n, y^n) \in \tilde{T}_D^* \text{ for some } x^n \in B^n\} \quad ,$$

$$\tilde{T}_D^*(y^n) = \{(x^n, y^n) | (x^n, y^n) \in \tilde{T}_D^*\} \quad ,$$

then, in view of (4.46), the derivation (4.47) can be continued as

$$Q^n \circ W_\delta^*(T_D^*(y^n)) \leq (1 + \sqrt{2\varepsilon}) \bar{Q}^n \circ W_\delta^*(\tilde{T}_{D_{\mu(\delta)}}^*(y^n)) \quad ,$$

from which together with (4.47) it follows that

$$\begin{aligned}
Q^n \circ W_\delta^* (T_D^*) &= \sum_{y^n \in Y_D^*} Q^n \circ W_\delta^* (T_D^*(y^n)) \leq \\
&\leq \sum_{y^n \in Y_D^* \cap F_0} Q^n \circ W_\delta^* (T_D^*(y^n)) + \sum_{y^n \notin F_0} Q^n \circ W_\delta^* (T_D^*(y^n)) \leq \\
&\leq (1 + \sqrt{2\varepsilon}) \sum_{y^n \in Y_D^* \cap F_0} \tilde{Q}^n \circ W_\delta^* (\tilde{T}_{D_{\mu(\delta)}}^*(y^n)) + \sum_{y^n \notin F_0} Q^n \circ W_\delta^* (T_X(y^n)) = \\
&= (1 + \sqrt{2\varepsilon}) \sum_{y^n \in Y_D^* \cap F_0} \tilde{Q}^n \circ W_\delta^* (\tilde{T}_{D_{\mu(\delta)}}^*(y^n)) + Q^n W_\delta^* (F_0^c) \leq \\
&\leq (1 + \sqrt{2\varepsilon}) \sum_{y^n \in Y_D^*} \tilde{Q}^n \circ W_\delta^* (\tilde{T}_{D_{\mu(\delta)}}^*(y^n)) + \lambda_1(\varepsilon) \leq \\
&\leq (1 + \sqrt{2\varepsilon}) \sum_{y^n \in \tilde{Y}_{D_{\mu(\delta)}}^*} \tilde{Q}^n \circ W_\delta^* (\tilde{T}_{D_{\mu(\delta)}}^*(y^n)) + \lambda_1(\varepsilon) = \\
&= (1 + \sqrt{2\varepsilon}) \tilde{Q}^n \circ W_\delta^* (\tilde{T}_{D_{\mu(\delta)}}^*(y^n)) + \lambda_1(\varepsilon), \tag{4.48}
\end{aligned}$$

where we have used (4.36) in the third inequality and $Y_D^* \subset \tilde{Y}_{D_{\mu(\delta)}}^*$ in the last inequality.

In an analogous way, we have

$$\tilde{Q}^n \circ W_\delta^* (\tilde{T}_D^*) \leq (1 + \sqrt{2\varepsilon}) Q^n \circ W_\delta^* (T_{D_{\mu(\delta)}}^*) + \lambda_1(\varepsilon). \tag{4.49}$$

Finally, rewriting (4.48) and (4.49), we have

Lemma 7. *Given any $\varepsilon > 0, \delta > 0$, and any $D \subset B^n$, it holds that*

$$Q^n \circ W_\delta^* \left(\frac{1}{n} i^*(X^n, Y^n) \in D \right) \leq \tilde{Q}^n \circ W_\delta^* \left(\frac{1}{n} \tilde{i}^*(X^n, Y^n) \in D_{\mu(\delta)} \right) + \lambda_2(\varepsilon), \tag{4.50}$$

$$\tilde{Q}^n \circ W_\delta^* \left(\frac{1}{n} \tilde{i}^*(X^n, Y^n) \in D \right) \leq Q^n \circ W_\delta^* \left(\frac{1}{n} i^*(X^n, Y^n) \in D_{\mu(\delta)} \right) + \lambda_2(\varepsilon), \tag{4.51}$$

where $\lambda_2(\varepsilon) = \lambda_1(\varepsilon) + \sqrt{2\varepsilon}$ is a continuous function of $\varepsilon > 0$ such as $\lambda_2(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$; and similarly for $\mu(\delta) > 0$ (cf.(4.43)).

Step 4. In this final step we simply combine Lemmas 2 and 7 to establish Theorem 1. First, consider (4.26) in Lemma 2 and (4.50) in Lemma 7, then

$$\begin{aligned}
P_n \left(\frac{1}{n} i(X^n, Y^n) \in D \right) &= Q^n \circ W^n \left(\frac{1}{n} i(X^n, Y^n) \in D \right) \leq \\
&\leq Q^n \circ W_\delta^* \left(\frac{1}{n} i^*(X^n, Y^n) \in D_\delta \right) + 3\sqrt{\gamma_n} \leq \tilde{Q}^n \circ W_\delta^* \left(\frac{1}{n} \tilde{i}^*(X^n, Y^n) \in D_{\mu(\delta)+\delta} \right) + 3\sqrt{\gamma_n} + \lambda_2(\varepsilon).
\end{aligned}$$

Furthermore, considering (4.27) in Lemma 2 with \tilde{Q}^n instead of Q^n , we have

$$\begin{aligned}
P \left[\frac{1}{n} i(X^n, Y^n) \in D \right] &\leq \tilde{Q}^n \circ W^n \left(\frac{1}{n} \tilde{i}(X^n, Y^n) \in D_{\mu(\delta)+2\delta} \right) + 7\sqrt{\gamma_n} + \lambda_2(\varepsilon) = \\
&= P \left[\frac{1}{n} \tilde{i}(X^n, Y^n) \in D_{\mu(\delta)+2\delta} \right] + 7\sqrt{\gamma_n} + \lambda_2(\varepsilon). \tag{4.52}
\end{aligned}$$

In entirely the same way, we have

$$P \left[\frac{1}{n} \tilde{i}(\tilde{X}^n, \tilde{Y}^n) \in D \right] \leq P \left[\frac{1}{n} i(X^n, Y^n) \in D_{\mu(\delta)+2\delta} \right] + 7\sqrt{\gamma_n} + \lambda_2(\varepsilon). \tag{4.53}$$

For an arbitrary given $\tau > 0$, we can chose $\delta > 0$ so that $\mu(\delta) + 2\delta < \tau$. Moreover, we see that $\lambda(\varepsilon) = \lambda_2(\varepsilon) + \varepsilon > 7\sqrt{\gamma_n} + \lambda_2(\varepsilon)$ for all sufficiently large n , because $\gamma_n \rightarrow 0$ as $n \rightarrow \infty$, establishing Theorem 1.

5. Proof of Theorem 3

Suppose that we are given a DMC W and an input distribution $\{Q^n\}_{n=1}^\infty$. Let us take any approximating input distribution $\{\tilde{Q}^n\}_{n=1}^\infty$ such that

$$d(Q^n W^n, \tilde{Q}^n W^n) < \varepsilon \quad (5.1)$$

for all sufficiently large n , where $\varepsilon > 0$ is any prescribed number. We write the input processes corresponding to $\{Q^n\}_{n=1}^\infty$ and $\{\tilde{Q}^n\}_{n=1}^\infty$ as $\mathbf{X} = \{X^n\}_{n=1}^\infty$ and $\tilde{\mathbf{X}} = \{\tilde{X}^n\}_{n=1}^\infty$, respectively, and denote by $\mathbf{Y} = \{Y^n\}_{n=1}^\infty$ and $\tilde{\mathbf{Y}} = \{\tilde{Y}^n\}_{n=1}^\infty$ the output processes induced by \mathbf{X} and $\tilde{\mathbf{X}}$ via the channel $\{W^n\}_{n=1}^\infty$, respectively.

We are concerned here with evaluating the values of the mutual informations $I(X^n; Y^n)$ and $I(\tilde{X}^n; \tilde{Y}^n)$. To this end, define the information densities:

$$i(x^n, y^n) = \log \frac{W^n(y^n | x^n)}{Q^n W^n(y^n)},$$

$$\tilde{i}(x^n, y^n) = \log \frac{W^n(y^n | x^n)}{\tilde{Q}^n W^n(y^n)},$$

and denote by F_n and \tilde{F}_n the distribution functions of $\frac{1}{n}i(X^n, Y^n)$ and $\frac{1}{n}\tilde{i}(\tilde{X}^n, \tilde{Y}^n)$, that is, for all $x \in \mathbb{R}$,

$$F_n(x) = Q^n \circ W^n \left(\frac{1}{n}i(X^n, Y^n) \leq x \right),$$

$$\tilde{F}_n(x) = \tilde{Q}^n \circ W^n \left(\frac{1}{n}\tilde{i}(\tilde{X}^n, \tilde{Y}^n) \leq x \right).$$

Then, we have the following expressions:

$$\frac{1}{n}I(X^n; Y^n) = \int_{-\infty}^{\infty} x dF_n(x), \quad (5.2)$$

$$\frac{1}{n}I(\tilde{X}^n; \tilde{Y}^n) = \int_{-\infty}^{\infty} x d\tilde{F}_n(x). \quad (5.3)$$

Let us decompose the right-hand side of (5.2) as follows.

$$\int_{-\infty}^{\infty} x dF_n(x) = \int_{-\infty}^0 x dF_n(x) + \int_0^{c+\tau} x dF_n(x) + \int_{c+\tau}^{\infty} x dF_n(x), \quad (5.4)$$

where $c = \log |A|$ and $\tau > 0$ is an arbitrarily small number. The first term of (5.4) is evaluated as

$$\begin{aligned} 0 &\geq \int_{-\infty}^0 x dF_n(x) = \frac{1}{n} \sum_{(x^n, y^n): i(x^n, y^n) \leq 0} Q^n(x^n) W^n(y^n | x^n) \log \frac{W^n(y^n | x^n)}{Q^n W^n(y^n)} \geq \\ &\geq \frac{1}{n} \sum_{(x^n, y^n): i(x^n, y^n) \leq 0} Q^n(x^n) Q^n W^n(y^n) \frac{1}{e} \log \frac{1}{e} \geq \frac{1}{ne} \log \frac{1}{e}. \end{aligned}$$

Therefore, we have

$$\lim_{n \rightarrow \infty} \int_{-\infty}^0 x dF_n(x) = 0. \quad (5.5)$$

In a similar manner, we have

$$\lim_{n \rightarrow \infty} \int_{-\infty}^0 x d\tilde{F}_n(x) = 0. \quad (5.6)$$

In order to evaluate the third term of (5.4) we need the following lemma.

Lemma 8. (see Appendix of [1]). *Let $G > \log |A|$, then*

$$\lim_{n \rightarrow \infty} \int_G^\infty x dF_n(x) = \lim_{n \rightarrow \infty} \int_G^\infty x d\tilde{F}_n(x) = 0. \quad (5.7)$$

In view of this lemma, the third term on the right-hand side of (5.4) vanishes as $n \rightarrow \infty$, that is

$$\lim_{n \rightarrow \infty} \int_{c+\tau}^\infty x dF_n(x) = 0. \quad (5.8)$$

Similarly,

$$\lim_{n \rightarrow \infty} \int_{c+\tau}^\infty x d\tilde{F}_n(x) = 0. \quad (5.9)$$

Consequently, what remains to be evaluated is the second term of (5.4). Set

$$I_n(\tau) = \int_0^{c+\tau} x dF_n(x), \quad (5.10)$$

$$\tilde{I}_n(\tau) = \int_0^{c+\tau} x d\tilde{F}_n(x). \quad (5.11)$$

Then,

$$I_n(\tau) = \int_0^{c+\tau} x dF_n(x) = (c+\tau)F_n(c+\tau) - \int_0^{c+\tau} F_n(x) dx. \quad (5.12)$$

On the other hand, putting $D = (-\infty, x]$ in (3.4) of Theorem 1 and putting $D = (-\infty, x - \frac{\tau}{2}]$ in (3.5) of Theorem 1 with $\frac{\tau}{2}$ instead of τ , we have

$$F_n(x) \leq \tilde{F}_n(x + \tau) + \lambda(\varepsilon), \quad (5.13)$$

$$F_n(x) \geq \tilde{F}_n\left(x - \frac{\tau}{2}\right) - \lambda(\varepsilon). \quad (5.14)$$

Substituting (5.13) and (5.14) into the right-hand side of (5.12) yields

$$\begin{aligned} I_n(\tau) &\leq (c+\tau) \left(\tilde{F}_n(c+2\tau) + \lambda(\varepsilon) \right) - \int_0^{c+\tau} \tilde{F}_n\left(x - \frac{\tau}{2}\right) dx + \lambda(\varepsilon) = \\ &= (c+\tau)\tilde{F}_n(c+2\tau) - \int_{-\tau/2}^{c+\tau/2} \tilde{F}_n(x) dx + (c+\tau+1)\lambda(\varepsilon) = (c+2\tau)\tilde{F}_n(c+2\tau) - \\ &- \int_0^{c+\tau/2} \tilde{F}_n(x) dx - \int_{-\tau/2}^0 \tilde{F}_n(x) dx + (c+\tau+1)\lambda(\varepsilon) - \tau\tilde{F}_n(c+2\tau) \leq \\ &\leq (c+2\tau)\tilde{F}_n(c+2\tau) - \int_0^{c+\tau/2} \tilde{F}_n(x) dx - \int_{-\tau/2}^0 \tilde{F}_n(x) dx + (c+\tau+1)\lambda(\varepsilon) = \\ &= (c+2\tau)\tilde{F}_n(c+2\tau) - \int_0^{c+2\tau} \tilde{F}_n(x) dx + \int_{c+\tau/2}^{c+2\tau} \tilde{F}_n(x) dx + (c+\tau+1)\lambda(\varepsilon) - \int_{-\tau/2}^0 \tilde{F}_n(x) dx. \end{aligned} \quad (5.15)$$

Since $0 \leq \tilde{F}_n(x) \leq 1$ for all $x \in \mathbb{R}$, the third term on the right-hand side of (5.15) is evaluated as

$$\int_{c+\tau/2}^{c+2\tau} \tilde{F}_n(x) dx \leq \frac{3\tau}{2}.$$

Thus, we have

$$\begin{aligned} I_n(\tau) &\leq (c+2\tau)\bar{F}_n(c+2\tau) - \int_0^{c+2\tau} \bar{F}_n(x)dx + \frac{3\tau}{2} + (c+\tau+1)\lambda(\varepsilon) = \\ &= \int_0^{c+2\tau} x d\bar{F}_n(x) + \varrho(\tau, \varepsilon) = \tilde{I}_n(2\tau) + \varrho(\tau, \varepsilon), \end{aligned} \quad (5.16)$$

where $\varrho(\tau, \varepsilon) = \frac{3\tau}{2} + (c+\tau+1)\lambda(\varepsilon)$ is obviously a continuous function of τ and ε such that $\varrho(\tau, \varepsilon) \rightarrow 0$ as $\tau \rightarrow 0$ and $\varepsilon \rightarrow 0$.

In the entirely same way, we have

$$\tilde{I}_n(\tau) \leq I_n(2\tau) + \varrho(\tau, \varepsilon), \quad (5.17)$$

although this inequality is not used in the sequel. Then, taking account of (5.5), (5.6) and (5.8), (5.9), (5.16), we have

$$\begin{aligned} \frac{1}{n}H(\tilde{X}^n) &\geq \frac{1}{n}I(\tilde{X}^n; \tilde{Y}^n) = \tilde{I}_n(2\tau) + \alpha_n \geq I_n(\tau) + \alpha_n - \varrho(\tau, \varepsilon) = \\ &= \frac{1}{n}I(X^n; Y^n) + \alpha_n + \beta_n - \varrho(\tau, \varepsilon), \end{aligned} \quad (5.18)$$

where α_n, β_n are some sequences such that $\alpha_n \rightarrow 0, \beta_n \rightarrow 0$ as $n \rightarrow \infty$.

On the other hand, if R is an achievable entropy rate for the input process $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and the approximating input process is $\tilde{\mathbf{X}} = \{\tilde{X}^n\}_{n=1}^{\infty}$, it must hold that, for any $\gamma > 0$,

$$\frac{1}{n}H(\tilde{X}^n) \leq R + \gamma \quad (5.19)$$

for all sufficiently large n (cf.(2.12)). From (5.18) and (5.19) it follows that

$$R + \gamma \geq \limsup_{n \rightarrow \infty} \frac{1}{n}H(\tilde{X}^n) \geq \limsup_{n \rightarrow \infty} \frac{1}{n}I(X^n; Y^n) - \varrho(\tau, \varepsilon).$$

Since $\gamma > 0, \tau > 0, \varepsilon > 0$ can be let arbitrarily small, it follows that

$$R \geq \limsup_{n \rightarrow \infty} \frac{1}{n}I(X^n; Y^n),$$

which proves Theorem 3.

Appendix: Proof of Lemma 4

Suppose that there exist two types $P_1, P_2 \in \Gamma$ such that $T_{P_1} \cap T_{P_2} \neq \emptyset$. Then, there exist $x_1^n, x_2^n \in A^n$ and $y^n \in T_{P_1}, y^n \in T_{P_2}$ such that

$$P_1 V_1 = P_2 V_2 = \text{type}(y^n), \quad (A.1)$$

where

$$P_1 = \text{type}(x_1^n), P_2 = \text{type}(x_2^n), \quad (A.2)$$

$$V_1 = \text{type}(y^n | x_1^n), V_2 = \text{type}(y^n | x_2^n). \quad (A.3)$$

By the definition of T_P (cf. c). of Step 3 in Section 4), $y^n \in T_{P_1}$ and $y^n \in T_{P_2}$ imply

$$D(V_1 || W | P_1) \leq \delta,$$

$$D(V_2 || W | P_2) \leq \delta,$$

respectively. Then, from the convexity of divergence it follows that

$$\begin{aligned} D(P_1 V_1 \| P_1 W) &\leq \delta \quad , \\ D(P_2 V_2 \| P_2 W) &\leq \delta \quad . \end{aligned}$$

By virtue of the general relation (cf.[6])

$$d^2(P, Q) \leq \frac{2}{\log e} D(P \| Q) ,$$

we have

$$\begin{aligned} d(P_1 V_1, P_1 W) &\leq \sigma \quad , \\ d(P_2 V_2, P_2 W) &\leq \sigma \quad , \end{aligned}$$

where $\sigma^2 = \frac{2\delta}{\log e}$, from which follows that

$$\begin{aligned} d(P_1 W, P_2 W) &\leq d(P_1 W, P_1 V_1) + d(P_1 V_1, P_2 V_2) + d(P_2 V_2, P_2 W) = \\ &= d(P_1 W, P_1 V_1) + d(P_2 V_2, P_2 W) \leq 2\sigma \quad , \end{aligned} \tag{A.4}$$

because $d(P_1 V_1, P_2 V_2) = 0$ owing to (A.1). Let us denote by Φ, Ψ be the set of all probability distributions on A^n and B^n respectively, and regard the channel $W : A \rightarrow B$ as a linear mapping $W : \Phi \rightarrow \Psi$. Then, since we have assumed that W is with full-rank, W is the one-to-one correspondence between Φ and its image $W(\Phi) \subset \Psi$. Therefore, there must exist the inverse linear mapping W^{-1} of W from $W(\Phi)$ to Φ . As $P_1 W, P_2 W$ in (A.4) belong to $W(\Phi)$, we can write

$$\begin{aligned} P_1 &= W^{-1}(P_1 W) \quad , \\ P_2 &= W^{-1}(P_2 W) \quad , \end{aligned}$$

and hence

$$P_1 - P_2 = W^{-1}(P_1 W - P_2 W) \quad ,$$

which together with (A.4) leads to

$$d(P_1, P_2) \leq c|A|d(P_1 W, P_2 W) \leq 2c|A|\sigma \quad ,$$

where $c > 0$ is the maximum modulus of components of the matrix W^{-1} .

The authors are grateful to V. Prelov for useful suggestions and the translation of the paper.

REFERENCES

1. T. S. Han and S. Verdú, "Approximation Theory of Output Statistics," submitted for publication.
2. R. Ahlswede and G. Dueck, "Identification via Channels," *IEEE Trans. Inform. Theory*, 35, 15–29 (1989).
3. T. S. Han and S. Verdú, "New Results in the Theory and Applications of Identification via Channels," *IEEE Trans. Inform. Theory*, 38, 14–25 (1992).
4. R. L. Dobrushin, "General Formulation of Shannon's Main Theorem in Information Theory," *AMS Translations*, 33, 323–438 (1963).
5. M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Francisco (1964).
6. I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic, New York (1981).