ON COMMON INFORMATION AND RELATED CHARACTERISTICS OF CORRELATED INFORMATION SOURCES

R. Ahlswede and J. Körner

Preface

This is a literal copy of a manuscript from 1974. References have been updated. It contains a critical discussion of in those days recent concepts of "common information" and suggests also alternative definitions. (Compare pages 402–405 in the book by I. Csiszár, J. Körner "Information Theory: Coding Theorems for Discrete Memoryless Systems", Akademiai Kiado, Budapest 1981.) One of our definitions gave rise to the now well–known source coding problem for two helpers (formulated in 2.) on page 7).

More importantly, an extension of one concept to "common information with list knowledge" has recently (R. Ahlswede and V. Balakirsky "Identification under Random Processes" invited paper in honour of Mark Pinsker, Sept. 1995) turned out to play a key role in analysing the contribution of a correlated source to the identification capacity of a channel.

Thus the old ideas have led now to concepts of operational significance and therefore are made accessible here.

0. Introduction

Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of pairs of random variables which are independent, identically distributed and take finitely many different values. $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are to be viewed as two correlated discrete memoryless stationary information sources (DCMSS).

In [1] a notion of "common information" was introduced for those sources. It was meant as the maximal common part of the total amount of information contained individually in each of the two sources $\{X_i\}$ and $\{Y_i\}$ and which can therefore be encoded separately by any of them without knowing the actual outcomes of the other source. It was shown in [1] that common codes of a DCMSS can use only deterministic interdependence of the sources and no further correlation can be exploited in this manner. This result was sharpened later by H.S. Witsenhausen [2].[1]

At a first glance the results may seem unsatisfactory because the common information thus defined depends only on the zeroes of the joint pr. d. matrix and does not involve its actual values. It is therefore natural to look for other notions of common information. Motivated by the work of Gray and Wyner [3], Wyner proposed another notion of common information in [4]. He expresses the believe that he has found the right notion of common information and that the earlier one of Gács and Körner is not the right notion, because of the properties mentioned above. The quantity introduced by Wyner seems to be indeed an interesting characteristic of correlated sources. However, the present authors take the position that his notion does not reflect at all what we would mean intuitively by "common information".

In this paper some arguments are provided which substantiate this position. It is therefore natural to look again for other notions of common information. We proceed systematically and investigate several coding schemes. It will become clear in our discussion that all notions introduced heavily depend on the network used for connecting encoders and decoders. Therefore it seems to us that a question as "what is the right notion of common information of $\{X_i\}$ and $\{Y_i\}$?" is meaningless. However, we shall introduce some concepts which we believe to be natural, because they relate to some basic source coding problems.

The main aim of the present contribution is to stimulate further discussions on the subject.

A word about notation. Throughout the paper "code" shall always mean deterministic block codes and the r.v. $\tilde{X}^n$ will be said to ε–reproduce X^n if $P(\tilde{X}^n \neq X^n) < \varepsilon$. All the r.v.'s have finite ranges. The unexplained basic notation is that of Gallager [9]. For the random variables (r.v) X and Y, $H(X)$ stands for the entropy of X, $\|X\|$ denotes the cardinality of the (finite) range of X, $H(X|Y)$ is the average conditional entropy of X given Y and $I(X \wedge Y)$ denotes the mutual information of X and Y. Exp's and log's are to the base 2, $h(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)$, for $0 < \varepsilon < 1$.

In order to fix ideas let us first take a new look at a one–decoder scheme for $\{(X_i, Y_i)\}_{i=1}^{\infty}$ and derive some consequences of the Slepian–Wolf theorem [6]. We shall say that a triple of positive reals (R_x, R_{xy}, R_y) is an element of the rate region $\mathcal{R}_0$ iff for every

[1] His result was again significantly improved in [12].

$\varepsilon > 0, \delta > 0$ and sufficiently large n there exists an ε–reproduction $(\tilde{X}^n, \tilde{Y}^n)$ of (X^n, Y^n) $(X^n = X_1 \dots X_n, Y^n = Y_1 \dots Y_n)$ such that for some deterministic functions f_n of X^n , g_n of Y^n , t_n of (X^n, Y^n) and a "decoding function" V_n we have

$$\text{(i)} \quad (\tilde{X}^n, \tilde{Y}^n) = V_n\big(f_n(X^n), t_n(X^n, Y^n) g_n(Y^n)\big)$$

$$\begin{aligned} \text{(ii)} \quad & \|f_n(X^n)\| \le \exp\{n(R_x + \delta)\} \\ & \|t_n(X^n, Y^n)\| \le \exp\{n(R_{xy} + \delta)\} \\ & \|g_n(Y^n)\| \le \exp\{n(R_y + \delta)\} \end{aligned}$$

Consider the quantities

$$\begin{aligned} (1) \quad & A_1(X,Y) = \sup R_{xy} \\ & R_{xy} + R_x \le H(X) \\ & R_{xy} + R_y \le H(Y) \\ & (R_x, R_{xy}, R_y) \in \mathcal{R}_0 \end{aligned}$$

and

$$\begin{aligned} (2) \quad & B_1(X,Y) = \inf R_{xy} \\ & R_x + R_{xy} + R_y \le H(X,Y) \\ & (R_x, R_{xy}, R_y) \in \mathcal{R}_0 \ . \end{aligned}$$

It is an immediate consequence of the Slepian–Wolf theorem that $A_1(X,Y) = I(X \wedge Y)$, the mutual information, and that $B_1(X,Y) = 0$. $A_1(X,Y)$ somehow measures how much knowledge about (X,Y) is simultaneously of interest for decoding X and Y in a lossles manner. Thus we arrived at a coding theoretic interpretation of mutual information, which allows us to view this quantity as a kind of "common information" for a one–decoder network.

The fact that $B_1(X,Y) = 0$ allows a simple and convincing interpretation. It means that the total entropy $H(X,Y)$ can be fully decomposed into two rates on the "sidelines", and it therefore makes sense to call $B_1(X,Y)$ the *indecomposable* entropy for a one decoder network. The two notions $A_1(X,Y)$ and $B_1(X,Y)$ are mathematically not very sophisticated; however, they help us in build up the right heuristic for two–decoder networks. Passing from the one–decoder to any two–decoder network (discussed below) the rate region *decreases* and therefore quantities defined with a "sup" decrease and those defined with an "inf" increase. It is therefore also clear that any possible reasonable notion of "common information" should lead to values *not* exceeding $A_1(X,Y) = I(X \wedge Y)$.

Let us now begin with a short description of the two–decoder networks we shall deal with. Consider a DMCSS $\{(X_i, Y_i)\}_{i=1}^{\infty}$.

Fig. 1

In our first network (Fig. 1) the sources $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are to be reproduced by two separate decoders, one for each of the sources. Similarly, there is one separate encoder for each of the sources, e.g. the encoder E_x can observe only $\{X_i\}_{i=1}^{\infty}$ and the block code he produces is available for the decoder D_x alone. However, there is a third encoder which allows us to exploit the correlation, since E_{xy} can observe both sources and its code is available for both individual decoders D_x and D_y . This is a modified version of a coding scheme of Gray and Wyner [3]. In their model all the three encoders can observe both sources (see Fig. 2).

Fig. 2

Finally, we introduce a coding scheme with four encoders (Fig. 3). The only difference between this and the coding scheme mentioned first (Fig. 1) is that the code exploiting the correlation is now supplied by two separate encoders, one for each of the sources. These codes are available for both individual decoders.

Fig. 3

Let us denote by $\mathcal{R}_i$ the rate region of the coding problems described in figure i $(i = 1, 2, 3)$. Replacing in definition (1) and (2) $\mathcal{R}_0$ by $\mathcal{R}_1$ the situation changes dramatically. Denoting an arbitrary element of $\mathcal{R}_1$ by (R_x, R_{xy}, R_y) where R_x is the rate of the code produced by E_x ; R_{xy} that of E_{xy} and R_y the rate of encoder E_y , we define the quantities

$$(3) \quad A_2(X,Y) = \sup R_{xy}$$

$$(R_x, R_{xy}, R_y) \in \mathcal{R}_1$$

$$R_x + R_{xy} \leq H(X)$$

$$R_y + R_{xy} \leq H(Y)$$

and

$$(4) \quad B_2(X,Y) = \inf R_{xy}$$

$$(R_x, R_{xy}, R_y) \in \mathcal{R}_1$$

$$R_x + R_y + R_{xy} \leq H(X,Y) .$$

Again we refer to the first quantity defined as "common information", because it measures how much knowledge about (X, Y) is simultaneously of interest for decoding X and Y is a lossless manner. Since X and Y are decoded separately now, this quantity seems to be a natural measure. However, we prove (Corollary 1, Section 2) that $A_2(X, Y)$, which is by definition not smaller than the common information of [1], is actually equal to that quantity.

The quantity $B_2(X, Y)$ is in some sense a dual to $A_2(X, Y)$. $B_2(X, Y)$ is that minimal portion of the joint entropy $H(X, Y)$ of the DMCSS $\big\{(X_i, Y_i)\big\}_{i=1}^{\infty}$ which one has to encode by a *joint encoder* observing both sources; otherwise the coding scheme of Fig. 1 would not be optimal. In other words this entropy can not be encoded by separate encoders without a loss in the total rate, and therefore it is *indecomposable*.

Wyner [4] has earlier introduced the quantity

$$\begin{aligned} C(X,Y) = \inf\, & R_{xy} \\ & (R_x, R_{xy}, R_y) \in \mathcal{R}_2 \\ & R_x + R_y + R_{xy} \le H(X,Y)\ . \end{aligned}$$

He has independently [10] also introduced the quantity $B_2(X,Y)$ and observed that $C(X,Y) = B_2(X,Y)$.

He calls $C(X,Y)$ *the* common information. However we believe that this would be a misleading name not only because of the large variety of analogous notions which can be obtained using different coding schemes but also and more importantly because it suggests a wrong heuristic. We have explained earlier that a quantity called common information should not exceed the mutual information $I(X \wedge Y)$. However, one easily sees that $I(X \wedge Y) \le B_2(X,Y) \le \min\{H(X), H(Y)\}$.

A single letter characterisation of the region $\mathcal{R}_2$ is known [3], [4]. We give such a characterisation for $\mathcal{R}_1$ (Theorem 2, Section 2) and therefore also for the quantities $A_2(X,Y)$ and $B_2(X,Y)$. Our method is that of [5], which proves to be quite general and easily adaptable to various source coding problems. The identity $\mathcal{R}_1 = \mathcal{R}_2$ follows as a byproduct. During the preparation of this manuscript we learnt that in an independent paper and by a different method Wyner [10] also obtained Theorem 2.

In Section 3, Corollary 2, we prove the somewhat surprising fact that

$$B_2(X,Y) = I(X \wedge Y) \text{ iff } I(X \wedge Y) = A_2(X,Y).$$

The determination of the rate region $\mathcal{R}_3$ corresponding to the coding scheme of Fig. 3 is still unsolved. Stating the problem here serves three purposes:

1.) It shows the relativity of any notion of common information.

2.) The two basic coding theorems for correlated sources, that is, the Slepian–Wolf theorem and the source coding theorem in case of side information [5], [10] do not provide all the tools to deal successfully with somewhat more complex networks.

    Probably the "most canonical" network of this kind, which is intimately related to the one above, is obtained by considering a correlated source $\{(X_i, Y_i, Z_i)\}_{i=1}^{\infty}$ with three separate encoders for each source and one decoder, who wants to reproduce $\{X_i\}$ and gets side information from $\{X_i\}$ as well as from $\{Z_i\}$.

3.) Similarly to $B_2(X,Y)$ we shall introduce the quantity

$$\begin{aligned} B_2^*(X,Y) = \inf\, & R_x^* + R_y^* \\ & R_x^* + R_x + R_y^* + R_y \le H(X,Y) \\ & (R_x, R_y, R_x^*, R_y^*) \in \mathcal{R}_3 \end{aligned}$$

    and call it the *strong indecomposable entropy* of the DMCSS $\{(X_i, Y_i)\}_{i=1}^{\infty}$.

    Whereas $B_2(X,Y)$ equals $C(X,Y)$, $B_2^*(X,Y)$ seems to be a new correlation measure.

1. Auxiliary results

This section is analogous to Section 1, Part I of [5] as far as we shall prove some convexity properties of the functions we have to deal with in the sequel. The ideas are those of Ahlswede–Körner [7], Section 4, where entropy inequalities for multiple–access channels (see [8]) were derived. Our aim is to generalize Lemmas 1 and 2 of [5].

We introduce the notation $X_1 \to X_2 \to X_3 \to X_4$ for saying that the r.v.'s X_1, X_2, X_3 and X_4 form a Markov chain in this order. For an arbitrary sequence $\{Z_i\}_{i\in N}$ of r.v.'s we put

$$Z^n = Z_1 Z_2 \dots Z_n .$$

Let us be given a sequence of independent and identically distributed triples $\big\{(S_i, X_i, Y_i)\big\}_{i\in N}$. For any positive real c we put:

Definition 1.1:

$\tau_n(c) = \big\{(R_x, R_y) : R_x \geq \frac{1}{n} H(X^n|U), R_y \geq \frac{1}{n} H(Y^n|U)$;
$U \to S^n \to (X^n, Y^n); H(S^n|U) \geq c\big\}$

We shall write $\tau(c) = \tau_1(c)$.

This is a natural generalization of the functions $T_n(c)$ defined in [5]. We shall write $(b_1, b_2) \leq (b_1', b_2')$ iff $b_1 \leq b_1'$ and $b_2 \leq b_2'$.

Lemma 1.

a) $\tau(c') \subset \tau(c)$ *for* $c \leq c'$ *(monotonicity)*

b) *For any* $0 \leq \alpha \leq 1$ *and* $c = \alpha c_1 + (1-\alpha) c_2$

$\alpha\tau(c_1) \oplus (1-\alpha)\tau(c_2) \subset \tau(x)$,

where

$\alpha\tau(c_1) \oplus (1-\alpha)\tau(c_2) = \big\{\alpha\vec{b_1} + (1-\alpha)\vec{b_2} : \vec{b_1} \in \tau(c_1); \vec{b_2} \in \tau(c_2)\big\}$ *(convexity).*

Proof:

a) is an immediate consequence of Definition (1.1). In order to prove b) we assume that $(R_x^1, R_y^1) \in \tau(c_1)$ and $(R_x^2, R_y^2) \in \tau(c_2)$, i.e. for suitable $U^{(i)}$ $(i = 1, 2)$ we have

$$H(S|U^{(i)}) \geq c_i \tag{1.1}$$

and $\big(H(X|U^{(i)}), H(Y|U^{(i)})\big) \leq (R_x^{(i)}, R_y^{(i)})$ where $U^{(i)} \to S \to (X, Y)$.

We introduce now the new quadruple of r.v.'s $\tilde{U}, \tilde{S}, \tilde{X}, \tilde{Y}$ such that $\Pr(\tilde{U}, \tilde{S}, \tilde{X}, \tilde{Y} = U^{(1)}, S^{(1)}, X^{(1)}, Y^{(1)}) = \alpha$ and $\Pr(\tilde{U}, \tilde{S}, \tilde{X}, \tilde{Y} = U^{(2)}, S^{(2)}, X^{(2)}, Y^{(2)}) = 1 - \alpha$ and furthermore, a r.v. I ranging over the set $\{1, 2\}$ with $\Pr(I = 1) = \alpha$ and such that $(I, \tilde{U}) \to \tilde{S} \to (\tilde{X}, \tilde{Y})$.

We have $H(\tilde{S}|\tilde{U}, I) = \alpha c_1 + (1-\alpha)c_2 = c$. Hence

$$\big(H(\tilde{H}|\tilde{U}, I), H(\tilde{Y}|\tilde{U}, I)\big) \in \tau(c).$$

On the other hand

$$\big(H(\tilde{X}|\tilde{U}, I), H(\tilde{Y}|\tilde{U}, I)\big) = \alpha\big(H(X|U^{(1)}), H(Y|U^{(1)})\big) + (1-\alpha)\cdot\big(H(X|U^{(2)}), H(Y|U^{(2)})\big)$$

and the statement of b) follows.

Remark:

It follows by a usual argument (see e.g. Lemma 3 of [5]) that the set $\tau(c)$ remains the same if in Definition 1.1 we limit ourselves to r.v.'s U satisfying the bound

$$\|U\| \le \|S\| + 2.$$

Lemma 2.

For all $n \in N$ *and* $c \ge 0$

$$\tau_n(c) = \tau(c) \quad \textit{(stationarity).}$$

Proof:

Let (U, S^n, X^n, Y^n) be a quadruple of r.v.'s satisfying $U \to S^n \to (X^n, Y^n)$.

We can write

$$\begin{aligned} H(X^n|U) &= \sum_{i=1}^n H(X_i|U, X^{i-1}) \ge \sum_{i=1}^n H(X_i|U, X^{i-1}, S^{i-1}) \\ &= \sum_{i=1}^n H(X_i|U, S^{i-1}) \end{aligned} \tag{1.2}$$

where the last identity follows by the fact that $U \to S^n \to (X^n, Y^n)$ and the triples (S_i, X_i, Y_i) are independent.

Similarly, one deduces that

$$H(Y^n|U) \ge \sum_{i=1}^n H(Y_i|U, S^{i-1}). \tag{1.3}$$

By the definition of $\tau(c)$ we have $\big(H(X_i|U, S^{i-1} = s^{i-1}), H(Y_i|U, S^{i-1} = s^{i-1})\big) \in \tau(c)$ for $c = H(S_i|U, S^{i-1} = s^{i-1})$ and hence by the convexity of $\tau(c)$ averaging over all the possible values of S^{i-1} , yields for the corresponding expected values

$$\big(H(X_i|U, S^{i-1}), H(Y_i|U, S^{i-1})\big) \in \tau(c_i) \tag{1.4}$$

where $c_i = H(S_i|U, S^{i-1})$.

This, (1.2), (1.3), and the monotonicity of $\tau(\cdot)$ yield

$$\big(H(X^n|U), H(Y^n|U)\big) \in \sum_{i=1}^n \tau(c_i), \tag{1.5}$$

where $\sum_{i=1}^n \tau(c_i) = \left\{ \vec{b} : \vec{b} = \sum_{i=1}^n \vec{b}_i,\ \vec{b}_i \in \tau(c_i) \right\}$.

From (1.5) and the convexity of $\tau(\cdot)$ it follows that

$$\left(\frac{1}{n}H(X^n|U), \frac{1}{n}H(Y^n|U)\right) \in \tau\left(\frac{1}{n}\sum_{i=1}^n c_i\right) = \tau(c^*)$$

where $c^* = \frac{1}{n}H(S^n|U)$.

Hence $\tau_n(c) \subset \tau(c)$, whereas $\tau_n(c) \supset \tau(c)$ is trivial. This completes the proof.

2. Common information

We begin with two definitions.

Definition 2.1:

A triple of positive reals (R_x, R_{xy}, R_y) is an element of the rate region $\mathcal{R}_1$ iff for every $\varepsilon > 0; \delta > 0$ and sufficiently large $n\big(n > n_0(\varepsilon,\delta)\big)$ there exists an ε–reproduction $(\tilde{X}^n, \tilde{Y}^n)$ of (X^n, Y^n) satisfying the following conditions:

There exist some deterministic functions f_n of X^n , g_n of Y^n , t_n of (X^n, Y^n) , and two decoding functions V_n and W_n with

$$\text{(i)}\quad \tilde{X}^n = V_n\big(f_n(X^n), t_n(X^n, Y^n)\big)$$
$$\tilde{Y}^n = W_n\big(g_n(Y^n), t_n(X^n, Y^n)\big)$$
$$\text{(ii)}\quad \|f_n(X^n)\| \le \exp\big\{n(R_x + \delta)\big\}$$
$$\|t_n(X^n, Y^n)\| \le \exp\big\{n(R_{xy} + \delta)\big\}$$
$$\|g_n(Y^n)\| \le \exp\big\{n(R_y + \delta)\big\} .$$

Definition 2.2:

$$A_2(X,Y) = \sup\ R_{xy}$$
$$R_{xy} + R_x \le H(X)$$
$$R_{xy} + R_y \le H(Y)$$
$$(R_x, R_{xy}, R_y) \in \mathcal{R}_1$$

is called the "common information" of the DMCSS $\big\{(X_i,Y_i)\big\}_{i=1}^{\infty}$.

After deriving from Theorems 1 in [5] and Lemmas 1 and 2 a single–letter description of $\mathcal{R}_1$, we shall prove that $A_2(X,Y)$ equals the common information in the sense of Gács and Körner [1]. Especially, for an X and Y having an indecomposable joint distribution (e.g.: $\forall\ x \in \mathcal{X}, y \in \mathcal{Y}\quad \Pr(X=x, Y=y) > 0$) it will follow that $A_2(X,Y)=0$.

Theorem 1.

Let $\big\{(X_i,Y_i)\big\}_{i\in N}$ *be a discrete memoryless correlated source with finite alphabets. The* rate region $\mathcal{R}_1$ *(as defined by Definition 2.1) satisfies*

$$\mathcal{R}_1 = \bigg\{\bigg(\frac{1}{n}H\big(X^n|t_n(X^n,Y^n)\big), \frac{1}{n}H\big(t_n(X^n,Y^n)\big)\ ,$$
$$\frac{1}{n}H\big(Y^n|t_n(X^n,Y^n)\big)\bigg)\ n\in N; t_n : \mathcal{X}^n \times \mathcal{Y}^n + N\bigg\}.$$

The proof is based on the simple observation that the coding scheme of Fig. 1 can be considered as a simultaneous "source coding with side information" for the DMCSS's $\big\{(X_i^*,Y_i^*)\big\}_{i\in N}$ and $\big\{(X_i^{**},Y_i^{**})\big\}_{i\in N}$ where (using the notation of Theorem 1 and 2 of [5])

$$X_i^* = X_i^{**} = (X_i,Y_i);\ Y_i^* = X_i;\ Y_i^{**} = Y_i$$

and where the same code has to be used for $\{X_i^*\} = \{X_i^{**}\} = \big\{(X_i,Y_i)\big\}$, serving in both cases as side information.

Now the proof of Theorem 1 in [5] literally applies and gives the assertion of the theorem.

As in [5] we shall give a single–letter description of $\mathcal{R}_1$ by rewriting our former description by means of the convexity arguments of Section 1.

Theorem 2.

$$\mathcal{R}_1 = \big\{(R_x,R_{xy},R_y) : R_x \ge H(X|Z), R_{xy} \ge I\big((X,Y)\wedge Z\big), R_y \ge H(Y|Z);$$
$$\|Z\| \le \|X\|\cdot\|Y\| + 2\big\}. \tag{2.1}$$

Proof:

We denote by $\mathcal{R}_1^*$ the set defined by the right–hand side of (2.1). We show first that

$$\mathcal{R}_1 \subset \mathcal{R}_1^*.$$

Suppose that for $K = t_n(X^n,Y^n)$ we have

$$R_x = \frac{1}{n}H(X^n|K), R_{xy} = \frac{1}{n}H(K)\ \text{ and }\ R_y = \frac{1}{n}H(Y^n|K).$$

We have to show that there exists a triple (X, Y, Z) such that the joint pr.d. of (X, Y) is that of the (X_i, Y_i)'s , $\|Z\| \le \|X\| \cdot \|Y\| + 2$ and $R_x \ge H(X|Z)$, $R_{xy} \ge I\big((X, Y) \wedge Z\big)$ and $R_y \ge H(Y|Z)$.

It is clear that

$$n \cdot R_{xy} = H(K) \ge I\big(K \wedge (X^n, Y^n)\big) = H(X^n, Y^n) - H(X^n, Y^n|K). \tag{2.2}$$

The independence of the (X_i, Y_i)'s and (2.2) yield

$$\frac{1}{n} H(X^n, Y^n|K) \ge H(X, Y) - R_{xy}. \tag{2.3}$$

We shall apply the Lemmas of Section 1 in the following set–up: $S_i = (X_i, Y_i)$.

By the definition of $\tau_n(c)$ we know that

$$\left(\frac{1}{n} H(X^n|K), \frac{1}{n} H(Y^n|K)\right) \in \tau_n \left(\frac{1}{n} H(X^n, Y^n|K)\right).$$

By Lemma 2 this gives

$$\left(\frac{1}{n} H(X^n|K), \frac{1}{n} H(Y^n|K)\right) \in \tau \left(\frac{1}{n} H(X^n, Y^n|K)\right). \tag{2.4}$$

Because of the monotonicity of the regions $\tau(\cdot)$ (see Lemma 1) the inequalities (2.3) and (2.4) yield

$$\left(\frac{1}{n} H(X^n|K), \frac{1}{n} H(Y^n|K)\right) \in \tau\big(H(X, Y) - R_{xy}\big). \tag{2.5}$$

By the definition of the region $\tau\big(H(X, Y) - R_{xy}\big)$ the last relation means that there exists a triple (Z, X, Y) such that

$$\begin{aligned} &R_x = \frac{1}{n} H(X^n|K) \ge H(X|Z), R_y = \frac{1}{n} H(Y^n|K) \ge H(Y|Z), \quad \text{and} \\ &\|Z\| \le \|X\| \cdot \|Y\| + 2, \end{aligned} \tag{2.6}$$

whereas $H(X, Y|Z) \ge H(X, Y) - R_{xy}$.

Rewriting the last inequality we get

$$R_{xy} \ge I\big((X, Y) \wedge Z\big). \tag{2.7}$$

Now we show that $\mathcal{R}_1^* \subset \mathcal{R}_1$ by the approximation argument of [5], Section 4. We have to prove that for every triple (Z, X, Y) with $\|Z\| \le \|X\| \cdot \|Y\| + 2$ there exists an n and a function t_n of (X^n, Y^n) such that

$$\frac{1}{n} H\big(X^n | t_n(X^n, Y^n)\big) \le H(X|Z), \frac{1}{n} H\big(Y^n | t_n(X^n, Y^n)\big) \le H(Y|Z) \text{ and}$$
$$\frac{1}{n} H\big(t_n(X^n, Y^n)\big) \le I\big((X, Y) \wedge Z\big).$$

It suffices to show that

$$\inf_n \inf_{\left(\frac{1}{n} H\left(X^n | t_n(X^n, Y^n)\right), \frac{1}{n} H\left(Y^n | t_n(X^n, Y^n)\right)\right) \le (x_1, x_2)} \frac{1}{n} H\big(t_n(X^n, Y^n)\big) \le$$
$$\inf_{\left(H(X|Z), H(Y|Z)\right) \le (x_1, x_2)} I\big((X, Y) \wedge Z\big). \tag{2.8}$$

From the independence of the (X_i, Y_i)'s and the fact that

$$I\big((X^n, Y^n) \wedge t_n(X^n, Y^n)\big) = H\big(t_n(X^n, Y^n)\big)$$

it follows that it is enough to show for $t_n = t_n(X^n, Y^n)$

$$\sup_n \sup_{\left(\frac{1}{n} H(X^n | t_n), \frac{1}{n} H(Y^n | t_n)\right) \le (x_1, x_2)} \frac{1}{n} H\big(X^n, Y^n | t_n(X^n, Y^n)\big) \ge$$
$$\sup_{\left(H(X|Z), H(Y|Z)\right) \le (x_1, x_2)} H(X, Y|Z). \tag{2.9}$$

Now we apply the construction of [5; Section 4] to the DMCSS's $\{X_i^*, Y_i^*\}_{i \in N}$ and $\{X_i^{**}, Y_i^{**}\}_{i \in N}$ and the r.v.'s U^* and U^{**} where as in the proof of Theorem 1

$$X_i^* = X_i^{**} = (X_i, Y_i), Y_i^* = X_i; Y_i^{**} = Y_i \text{ and } U^* = U^{**} = Z.$$

Observing that the construction of [5] depends only on the joint pr. d. of (U^*, X^*, Y^*), it becomes clear that — using the notation of [5] — the choice $t_n(X^n, Y^n) \triangleq f_n(X^{*n}) = f_n(X^{**n})$ actually establishes (2.9).

In what follows we shall use Theorem 1 to prove a generalization of Theorem 1, p. 151 of [1]. Actually, we prove that the common information $A_2(X, Y)$ of Definition 2.2, which is clearly not smaller than that of [1], is equal to it. We recall from [1] the following

Definition 2.3:

We suppose without loss of generality that for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ $\Pr(X_1 = x) > 0$ and $\Pr(Y_1 = y) > 0$. We consider the stochastic matrix of the conditional probabilities $\{\Pr(X = x | Y = y)\}$ and its ergodic decomposition. Clearly, the ergodic decompositions of the matrices $\{\Pr(X=x|Y=y)\}$ and $\{\Pr(Y=y|X=x)\}$ coincide and form a partition

$$\mathcal{X}\times\mathcal{Y}=\bigcup_j \mathcal{X}_j\times\mathcal{Y}_j$$

of $\mathcal{X}\times\mathcal{Y}$ where the $\mathcal{X}_j$'s and $\mathcal{Y}_j$'s having different subscripts are disjoint. We introduce the r.v. J such that

$$J=j \Leftrightarrow X\in\mathcal{X}_j \Leftrightarrow Y\in\mathcal{Y}_j.$$

It is clear that J is a function of both X and Y. We shall prove that the common information $A_2(X,Y)$ equals the entropy of this common function of X and Y.

Corollary 1.

$$A_2(X,Y)=H(J).$$

Proof:

It follows from our Theorem 2 that

$$\begin{aligned}
A_2(X,Y) &= \sup I\big((X,Y)\wedge Z\big)\\
&\quad I\big((X,Y)\wedge Z\big)+H(X|Z)\le H(X)\\
&\quad I\big((X,Y)\wedge Z\big)+H(Y|Z)\le H(Y)\\
&\quad \|Z\|\le\|X\|\cdot\|Y\|+2\ .
\end{aligned}$$

Looking at the constraint inequalities we find that from

$H(X)\ge I\big((X,Y)\wedge Z\big)+H(X|Z)=H(X,Z)-H(Z|X,Y)$

we get the inequality

$$H(Z|X,Y)\ge H(Z|X),$$

which gives $Z\to Y\to X$. Similarly, our other constraint gives $Z\to X\to Y$.

Now we shall analyze these conditions

$$Z\to Y\to X \quad\text{and}\quad Z\to X\to Y. \tag{2.10}$$

It follows from (2.10) that

$$\Pr(X{=}x,Y{=}y)>0\Rightarrow \Pr(Z{=}z|X{=}x,Y{=}y)=\Pr(Z{=}z|X{=}x)=\Pr(Z{=}z|Y{=}y). \tag{2.11}$$

Hence for any fixed value of Z and for every index j $\Pr(Z=z|X=\cdot,Y=\cdot)$ is constant over $\mathcal{X}_j\times\mathcal{Y}_j$ whenever it is defined. This means that

$$\Pr(Z{=}z|X{=}x,Y{=}y) = \Pr(Z{=}z|J{=}j) = \sum_{\mathcal{X}_j\times\mathcal{Y}_j} \Pr(Z{=}z|X{=}\hat{x},Y{=}\hat{y})\cdot\Pr(X{=}\hat{x}|Y{=}\hat{y}) \ .$$

The last relation means that given any value of J the r.v. Z is conditionally independent from (X,Y) , i.e. $I\big((X,Y)\wedge Z|J\big)=0$. However since J is a function of (X,Y) we have

$$I\big((X,Y)\wedge Z\big) = I\big((X,Y,J)\wedge Z\big) = I(J\wedge Z) + I\big((X,Y)\wedge Z|J\big) \tag{2.12}$$

where the last equality follows by a well–known identity (see e.g. Gallager [9], formula (2.2.29) pn p. 22). Comparing the two extremities of (2.12) we get

$$I\big((X,Y)\wedge Z\big) \le H(J) + I\big((X,Y)\wedge Z|J\big) = H(J). \tag{2.13}$$

Taking into account that J is a deterministic function of X and a deterministic function of Y and thus it satisfies the contraints of our second definition of $A_2(X,Y)$, we conclude that $A_2(X,Y)=H(J)$.

Remark:

The quantity

$$\begin{aligned}A(X,Y) &= \sup R_y\\ &\quad R_x+R_y \le H(X)\\ &\quad (R_x,0,R_y)\in\mathcal{R}_0\end{aligned}$$

is meaningful in a one–decoder situation. It says how much information about X we can extract from Y in a "lossless manner". It is easy to see that $H(J)\le A(X,Y)\le A_2(X,Y)$ and hence that also $A(X,Y)=H(J)$.

3. Indecomposable entropy

Definition 3.1:

$$\begin{aligned}B_2(X,Y) &= \inf R_{xy}\\ &\quad (R_x,R_{xy},R_y)\in\mathcal{R}_1\\ &\quad R_x+R_{xy}+R_y) \le H(X,Y)\end{aligned}$$

is called the "indecomposable entropy" of the DMCSS $\big\{(X_i,Y_i)\big\}_{i\in N}$. A justification for this terminology was given in the introduction. It is clear from the foregoing that

$$B_2(X,Y) = \inf_{\substack{H(X|Z)+H(Y|Z)+I\left((X,Y)\wedge Z\right)=H(X,Y)\\ \|Z\|\le\|X\|\cdot\|Y\|+2}} I\big((X,Y)\wedge Z\big) \tag{3.1}$$

and

$$B_2(X,Y) \geq I(X \wedge Y) \geq A_2(X,Y). \tag{3.2}$$

Looking into the constraint on the right hand side of (3.1) and taking into account that $H(X,Y|Z) + I\big((X,Y) \wedge Z\big) = (X,Y)$ always holds we conclude that the constraint is equivalent to $H(X,Y|Z) = H(X|Z) + H(Y|Z)$. This allows us to write

$$B_2(X,Y) = \min_{X \to Z \to Y} I\big((X,Y) \wedge Z\big)$$
$$\|Z\| \leq \|X\| \cdot \|Y\| + 2.$$

We shall prove that

Corollary 2.

$$B_2(X,Y) = I(X \wedge Y) \Leftrightarrow I(X \wedge Y) = A_2(X,Y).$$

Remark:

Since $A_2(X,Y) = H(J)$, the entropy of the ergodic class index which is a common function of X and Y , the statement of Corollary 2 means that $B_2(X,Y)$ equals the mutual information iff all the correlation between X and Y is of deterministic character. Expecially if X and Y have an indecomposable joint pr.d. the corollary says that $B_2(X,Y) = I(X \wedge Y)$ implies $B_2(X,Y) = 0$.

Proof:

We suppose that for a r.v. Z satisfying the constraint of minimization of (3.3) we have

$$I\big((X,Y) \wedge Z\big) = I(X \wedge Y). \tag{3.4}$$

Using the identity

$$H(X,Y) = I(X \wedge Y) + H(X|Y) + H(Y|X) \tag{3.4}$$

becomes equivalent to

$$H(X,Y|Z) = H(X|Y) + H(Y|X). \tag{3.5}$$

Since by our supposition $X \to Z \to Y$ we have

$$H(X,Y|Z) = H(X|Z) + H(Y|Z) = H(X|Z,Y) + H(Y|Z,X). \tag{3.6}$$

Comparing (3.5) and (3.6) we obtain that (3.4) is equivalent to the condition

$$H(X|Y) + H(Y|X) = H(X|Z,Y) + H(Y|Z,X).$$

Rewriting this we get

$$I(X \wedge Z|Y) + I(Y \wedge Z|X) = 0. \tag{3.7}$$

Since conditional mutual informations are non–negative, (3.7) is equivalent to

$$I(X \wedge Z|X) = 0 \ \text{ and } \ I(Y \wedge Z|X) = 0.$$

Hence we get that

$$X \rightarrow Y \rightarrow Z \ \text{ and } \ Z \rightarrow X \rightarrow Y.$$

Observing that this is just (2.10), the deduction consecutive to relation (2.10) in Section 2 applies and we get that $I\big((X,Y) \wedge Z\big) = H(J)$.

This completes the proof.

References

[1] P. Gács and J. Körner, "Common information is far less than mutual information", Problems of Contr. and Inf. Th., Vol. 2, 149–162, 1973.

[2] H.S. Witsenhausen, "On sequences of pairs of dependent random variables", SIAM J. of Appl. Math., Vol. 28, 100–113, 1975.

[3] R.M. Gray and A.D. Wyner, "Source coding for a simple network", Bell System. Techn. J., Dec. 1974.

[4] A.D. Wyner, "The common information of two dependent random variables", IEEE Trans. Inform. Theory, Vol. IT–21, 163–179, Mar. 1975.

[5] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels", IEEE Trans. on Inf. Th., Vol. IT–21, No. 6, 629–637, Nov., 1975.

[6] D. Slepian and J.K. Wolf, "Noiseless coding for correlated information sources", IEEE Trans. on Inf. Th., Vol. IT–19, 471–480, July 1973.

[7] R. Ahlswede and J. Körner, "On the connection between the entropies of input and output distributions of discrete memoryless channels", Proceedings of the 5th Conference on Probability Theory, Brasov 1974, Editura Academeiei Rep. Soc. Romania, Bucaresti, 13–23, 1977.

[8] R. Ahlswede, "Multi–way communication channels", Proc. 2nd Int. Symp. Inf. Th., Tsahkadsor, Armenian S.S.R., 1971, 23–52. Publishing House of the Hungarian Academy of Sciences, 1973.

[9] R.G. Gallager, "Information Theory and Reliable Communication", Wiley and Sons, New York 1968.

[10] A.D. Wyner, "On source coding sith side information at the decoder", IEEE Trans. on Inf. Th., Vol. IT–21, No. 3, May 1975.

[11] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi–user communication", Zeitschr. für Wahrscheinlichkeitstheorie und verw. Gebiete, Vol. 34, 157–177, 1976.

[12] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator", Ann. of Probability, Vol. 4, No. 6, 925–939, 1976.