Problems of Control and Information Theory, Vol 2 (2), pp 149–162 (1973)

COMMON INFORMATION IS FAR LESS THAN MUTUAL INFORMATION

P. GÁCS and J KÖRNER

(Budapest)

(Received February 5, 1972)

1. Introduction

Up to now mutual information has no immediate interpretation. It appears nowhere as the result of a coding problem. Imre Csiszár raised the problem whether there is any connection between the mutual information $E\left(\log_2 \frac{P(\xi = x, \eta = y)}{P(\xi = x)\cdot P(\eta = y)}\right)$ of two random variables and such a coding of them which allows to discern all their common information. More precisely: given the random variables ξ and η we are looking for a code consisting of three parts: A, B and C where ξ can be decoded from A and C, while η can be decoded if B and C are known. We shall call such a code a Y-scheme in the sequel

Our results are of negative character. Prescribing for the total length of the codes A, B and C that it should not exceed too much the optimum length of a common code of the pair (ξ, η), the optimum taken without any structural condition on the encoding, we shall prove in the case of memoryless sources that common information always corresponds to some deterministic interdependence of the two sources. Thus our aim is to show that common information has nothing to do with mutual information.

One can formulate the same question also in terms of Kolmogorov's "complexity" and "mutual information" of individual sequences. Our results remain here valid (See § 6 of the present paper).

2. Formulation of results

First of all let us introduce some notations and definitions.

Definition 1. A discrete memoryless stationary information source (DMSS) with finite alphabet X is a sequence $\{\xi_n\}_{n=1}^{\infty}$ of independent, identically distributed random variables ξ_n with values in the finite set X. We denote the source by $\mathfrak{X}$.

We shall often use the following notations:

$P(\xi = x)$:	the probability of the event $[\xi = x]$
$H(\xi)$:	the entropy of the random variable ξ.
$\xi^n = \xi_1\xi_1 \dots \xi_n$.	
X^n:	the n-th Cartesian power of the set X.
$I = \{0, 1\}$:	the set consisting of 0 and 1.
$I^* = \bigcup\limits_{k=0}^{\infty} I^k$	
$l(f)$:	the length of the binary word f.

Definition 2. A mapping $\varphi\colon X^n \to I^k$ is called a (binary) block-code. φ is an ε-code $(0 < \varepsilon < 1)$ of the DMSS $\mathcal{X} = \{\xi_n\}_{n=1}^{\infty}$ if there exists a mapping $\varphi'\colon I^k \to X^n$ such that the inequality $P[\varphi'(\varphi(x)) \neq x]$ holds. $P[\varphi'(\varphi(x)) \neq x]$ is the probability of error of the code described above. $< \varepsilon$

All the log's in this paper are to the base 2.

After these generally adopted definitions let us turn to the concepts of Y-scheme and common information.

Definition 3. Let $\mathcal{X} = \{\xi_n\}_{n=1}^{\infty}$ and $\mathcal{Y} = \{\eta_n\}_{n=1}^{\infty}$ be two DMSS's with finite alphabets X and Y. The joint distribution of the pair (ξ_1, ξ_1) — where the pairs (ξ_i, η_i) are supposed to be independent for different i's, — and $\varepsilon > 0$, $\delta > 0$ are given. We call a $Y_{\varepsilon,\delta}^{(n)}$-scheme (or simply Y-scheme) the following coding scheme:

(i) f and f' are binary block codes of the set X^n, g and g' are binary block codes of Y^n with

$$P[f(\xi^n) = g(\eta^n)] > 1 - \varepsilon. \tag{1}$$

(ii) the juxtaposition ff' of f and f' is an ε-code of $\mathcal{X}$ and similarly gg' is an ε-code of $\mathcal{Y}$.

(iii)
$$l(ff') = l(f) + l(f') < (1 + \delta)\, H(\eta^n)$$
$$l(gg') = l(g) + l(g') < (1 + \delta)\, H(\eta^n). \tag{2}$$

Definition 4. Let us write $J_n(\varepsilon, \delta) = \max l(f)$ where the maximum is taken over all the $Y_{\varepsilon,\delta}$-schemes coding the sources $\mathcal{X}$ and $\mathcal{Y}$. $J_n(\varepsilon, \delta)$ is the *common information* of ξ^n and η^n.

One could think that the quantity $J_n(\varepsilon, \delta)$ and the mutual information $I(\xi^n \wedge \eta^n)$ are closely related but this is not true. We shall show that in general cases of dependence $J_n(\varepsilon, \delta)$ is of smaller order of magnitude.

In order to formulate our result let us introduce a notion of accessibility with respect to the joint distribution of ξ_1 and η_1, which is closely related to the one introduced in the theory of Markov-chains (see [5]). We can assume without loss of generality that for every $x \in X$ and $y \in Y$ $P[\xi = x] > 0$ and $P[\eta = y] > 0$. We shall say that $x \in X$ and $x' \in X$ are *communicating* (write $x \sim x'$) if there exists some sequence $x_0 y_1 x_1 \dots x_N$ with $x_i \in X$, $y_i \in Y$ and $P[\xi = x_{i-1}, \eta = y_i] > 0$, $P[\xi = x_i, \eta = y_i] > 0$, where $x_1 = x$ and $x_N = x'$. It is obvious that $x' \sim x$ is an equivalence relation on X. Let the corresponding equivalence classes be X_i for $i = 1, 2, \dots, r$. Similarly we define a relation $y \sim y'$ on Y equivalence classes Y_i, $i = 1, 2, \dots, s$. It is easy to see that one can pass from a given X_i only to a single Y_j with positive probability and vice versa. Hence one has $r = s$. If we give the same index to these "communicating" classes, we obtain

$$P(\eta \in Y_j \mid \xi \in X_i) = P(\xi \in X_i \mid \eta \in Y_i) = \delta_{ij} \tag{3}$$

where δ_{ij} denotes the Kronecker symbol.

Definition 5. We shall call the above unique partition of X and Y an *ergodic decomposition.*

This terminology will be motivated later on. The introduction of a random variable ζ taking its values from the set of the indices of the equivalence classes and defined by

$$\zeta = i \quad \text{if} \quad \xi \in X_i$$

proves to be useful. One has obviously:

$$P[\zeta = i] = P[\xi \in X_i, \qquad \eta \in Y_i]. \tag{4}$$

Passing to the product space the sequence $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ defines a sequence of random variables $\zeta_1 \dots \zeta_n$ in a similar way. It is evident that the ζ_i's are independent and identically distributed. The different classes of the ergodic decomposition of X^n and Y^n correspond to the values of ζ^n in a one-to-one way.

It is clear that ζ^n represents some "common information" of ξ^n and η^n. The amount of this information is $H(\zeta^n) = nH(\zeta_1)$. The following theorem states that there is no more common information in ξ^n and η^n.

Theorem 1. Two DMSS's $\mathfrak{X}$ and $\mathfrak{Y}$ and ε $(0 < \varepsilon < 1)$ are given. If δ_n tends to 0, we have

$$\limsup_{n \to \infty} \frac{1}{n} J_n(\varepsilon, \delta_n) \leq H(\zeta_1)\,, \tag{5a}$$

Furthermore there exists a sequence $\delta_n \to 0$ such that the equality

$$\lim_{n\to\infty} \frac{1}{n} J_n(\varepsilon, \delta_n) = H(\zeta_1) \tag{5b}$$

holds We remark that the above limit does not depend on ε.

Originally we were interested to know in which cases is equality between the limit of (5b) and the mutual information $\lim\limits_{n\to\infty} \frac{1}{n} I(\xi^n \wedge \eta^n) = I(\xi_1 \wedge \eta_1)$. The inequality $H(\zeta_1) \leq I(\xi_1 \leq \eta_1)$ is obvious.

Corollary 1. The limit $H(\zeta_1)$ of the common information equals $I(\xi_1 \wedge \eta_1)$ only in the trivial case when ξ_1 and η_1 are "conditionally independent", i.e. if and only if ξ_1 and η_1 are independent for any fixed value of ζ_1.

Proof of the corollary. The statement is an immediate consequence of the following trivial identity:

$$I(\xi \wedge \eta) = H(\zeta) + \sum_z P(\zeta = z) \cdot I(\xi \wedge \eta \mid \zeta = z).$$

Hence follows that $I(\xi \wedge \eta) = H(\zeta)$ holds if and only if $I(\xi \wedge \eta \mid \zeta = z) =$ $= 0$ for any fixed z

Proving Theorem 1 we shall answer also the following question which we believe to be interesting in itself.

Definition 6. Given $\lambda (0 < \lambda \leq 1)$ we call a λ-block a pair of sets (A, B) with $A \subset X^n$ and $B \subset Y^n$ when

$$P(\xi^n \in A \mid \eta^n \in B) \geq \lambda \qquad P(\eta^n \in B \mid \xi^n \in A) \geq \lambda \tag{6}$$

holds.

Comparing (3) and (6) it is evident that our ergodic classes are λ-blocks for every λ. The question is whether a λ-block may be essentially smaller than the smallest ergodic class. According to Theorem 2 the answer is negative.

Theorem 2. Putting

$$\begin{gathered} Q(n, \lambda) = \min P(\xi^n \in A,\ \eta^n \in B) \\ P(\xi^n \in A \mid \eta^n \in B) \geq \lambda \\ P(\eta^n \in B \mid \xi^n \in A) \geq \lambda \end{gathered}$$

for any sequence $\lambda_n = 2^{-n\varepsilon_n}$ with $\varepsilon_n \to 0$ we have

$$\lim_{n\to\infty} \frac{1}{n} \cdot \log Q(n, \lambda_n) = \min_z P(\zeta = z)\,.$$

In the following paragraphs we shall reduce the coding problem to a problem concerning blocks and prove Theorems 1 and 2 by means of a pivotal lemma. We shall often use a very elementary Markov-type inequality for bounded random variables. Throughout the paper it is referred to as the reverse Markov inequality. A proof – if needed   is to be found in Loeve [5] For the terminology of Markov-chains see [1] and [5].

3. Blocks

The consequences of inequality (1) are presented.

Proposition 1. The binary block codes f and g of ξ^n and η^n respectively are supposed to have the same length $l = l(f(\xi^n))$ and satisfy (1). Denoting by $C(\varepsilon)$ the set of those c's in I^l for which

$$P(f(\xi^n) = g(\eta^n) \mid f(\xi^n) = c) \geq 1 - \sqrt{\varepsilon} \qquad \text{and}$$
$$P(f(\xi^n) = g(\eta^n) \mid g(\eta^n) = c) \geq 1 - \sqrt{} \tag{7}$$

we have $P[f(\xi^n) = g(\eta^n) \in C(\varepsilon)] \geq 1 - 4\sqrt{\varepsilon}$.

Proof. Let us introduce the sets

$$A(\varepsilon') = \{c \in I^l;\ \ P(g(\eta^n) = c \mid f(\xi^n) = c) > 1 - \varepsilon'\} \qquad \text{and}$$
$$B(\varepsilon') = \{c \in I^l;\ \ P(f(\xi^n) = c \mid g(\eta^n) = c) > 1 - \varepsilon'\}.$$

A reverse Markov inequality combined with (1) implies that $P[f(\xi) \in A(\varepsilon')] \geq \geq 1 - \frac{\varepsilon}{\varepsilon'}$ and $P[g(\eta) \in B(\varepsilon') \geq 1 - \frac{\varepsilon}{\varepsilon'}$. Choosing $\varepsilon' = \sqrt{\varepsilon}$ we write $C(\varepsilon) = = A(\sqrt{\varepsilon}) \cap B(\sqrt{\varepsilon})$. Obviously for $c \subset C(\varepsilon)$ one has (7).

On the other hand we obtain

$$\sum_{c \in A(\sqrt{\varepsilon})} P[f(\xi^n) = g(\eta^n) = c] \geq (1 - \sqrt{\varepsilon}) \cdot P[f(\xi^n) \in A(\sqrt{\varepsilon})] \geq (1 - \sqrt{\varepsilon})^2 \geq 1 - 2\sqrt{\varepsilon}$$

and similarly: $P[f(\xi^n) = g(\eta^n),\ g(\eta^n) \in B(\sqrt{\varepsilon})] \geq 1 - 2\sqrt{\varepsilon}$.

For the intersection of $A(\sqrt{\varepsilon})$ and $B(\sqrt{\varepsilon})$ thus follows: $P[f(\xi^n) = g(\eta^n) \in \in C(\varepsilon)] \geq 1 \quad 4\sqrt{\varepsilon}$. Proposition 1 says that a large part of (X^n, Y^n) consists of $(1 \quad \sqrt{\varepsilon})$-block. In this way one can see that the number $l(f)$ – i.e. the length of the "stem" of the Y-scheme is related to the "number of blocks in (X^n, Y^n)". We shall show that every block is essentially an ergodic class or the union of several ones. Before doing this we need one more proposition. We use the notations of Definition 3. Let us be given a sequence of Y-schemes $Y^{(n)}_{\varepsilon,\delta_n}$ with codes

f_n, g_n, f'_n, g'_n, and δ_n tending to 0. Write $l_n = l(f_n)$. We want to obtain an upper bound for $\limsup_{n\to\infty} \frac{1}{n} l_n$. Doing this it will be useful to replace f_n by a fuitable code of it.

Proposition 2. If $\tilde{f}_n = U_n(f_n)$ is any code of f_n having error probability less than λ (where $\lambda < 1 - \varepsilon$), of some fixed length, the inequality

$$\limsup_{n\to\infty} \frac{1}{n} l(\tilde{f}_n) \geq \limsup_{n\to\infty} \frac{1}{n} l_n \qquad \text{follows.}$$

Proof. It is obvious that juxtaposing $\tilde{f}_n$ and f_n we obtain a code $\tilde{f}_n f'_n$ of ξ^n with error probability less than $\varepsilon + \lambda$. Thus Shannon's classical source coding theorem implies that $\limsup \frac{1}{n} l(\tilde{f}_n f'_n) \geq H(\xi_1)$. Since $\delta_n \to 0$ implies that $\lim_{n\to\infty} \frac{1}{n} l(\bar{f}_n f'_n) = H(\xi_1)$, we have $\limsup_{n\to\infty} \frac{1}{n} l(\tilde{f}_n) f'_n) \geq \lim \frac{1}{n} l(f_n f_n)$. Hence our assertion becomes trivial. Put

$$\varphi_n(\xi^n) = \big(f_n(\xi^n), \zeta^n\big) \quad \Psi_n(\eta^n) = \big(g_n(\eta^n), \zeta^n\big) \tag{8}$$

Here ζ_n is the random variable defined in § 1. φ_n and ψ_n are such refinements of f_n and g_n which take different values in different ergodic classes. Obviously·

$$P\big(\varphi_n(\xi^n) = \Psi_n(n^n)\big) = P\big(f_n(\xi_n) = g_n(\eta^n)\big) > 1 - \varepsilon\,. \tag{9}$$

Thus the codes φ_n and ψ_n satisfy the conditions of Proposition 1 and so the assertion remains valid for them. From now on, $C(\varepsilon)$ corresponds to φ_n and ψ_n. Since φ_n is a refinement of f_n, from Proposition 2 we obtain for any code $\tilde{\varphi}_n = U_n(\varphi_n)$ having error probability less than λ that

$$\limsup_{n\to\infty} \frac{1}{n} \cdot l(\tilde{\varphi}_n) \geq \limsup_{n\to\infty} \frac{1}{n} \cdot l(f_n)\,.$$

In the following we shall give an upper bound of $\limsup_{n\to\infty} \frac{1}{n} \cdot l(\tilde{\varphi}_n)$ for a suitable series of codes U_n.

4. A lemma

First of all we prove that φ_n (defined by (8)) does not take too many different values in a single ergodic class. The proof is based upon our main

Lemma. Suppose for the sets $A_n \subset X^n$ and $B_n \subset Y^n$ that

$$P(\xi^n \in A_n \mid \eta^n \in B_n) \geq 2^{n-\varepsilon_n} \text{ and } P(\eta^n \in B_n \mid \xi^n \in A_n) \geq 2^{n-\varepsilon_n} \tag{10}$$

where $\varepsilon_n \to 0$. For the sake of simplicity let $\varepsilon_n > \frac{1}{n}$. If $A_n \subset \tilde{A}_n$ and $B_n \subset \tilde{B}_n$ where $\tilde{A}_n$ and $\tilde{B}_n$ are ergodic classes, it follows that

$$\frac{P(\xi^n \in A_n, \eta^n \in B_n)}{P(\xi^n \in \widetilde{A}_n, \eta^n \in \widetilde{B}_n)} \geq 2^{-nC \cdot \varepsilon_n \cdot \log^2 \varepsilon_n}$$

where C is a constant depending only on the distribution of the pair (ξ_1, η_1).

Proof. The simple main idea of the proof is that if we get from A_n to B_n with a large probability and vice versa then applying formally both transitions one after the other we get from A_n back to A_n with quite a large probability. On the other hand iterating this formal transition many times, the set A_n "spreads" over its whole ergodic class. Passing over to the demonstration, we write

$$W(x' \mid x) = \sum_y P(\xi_1 = x' \mid \eta_1 = y) \cdot P(\eta_1 = y \mid \xi_1 = x) \quad \text{if} \quad x, x' \in X.$$

Now we define Markov chains $\{\xi_{nk}\}_{k=1}^{\infty}$. Let $\xi_{n0} = \xi_n$ and $P(\xi_{n(k+1)} = x' \mid \xi_{nk} = x) = W(x' \mid x)$. For different n's we suppose the variables ξ_{nk} to be independent. Let us write $W^{(k)}(x' \mid x) = P(\xi_{1k} = x' \mid \xi_{10} = x)$. It is easy to see that the ergodic classes of the transition matrix $W(x' \mid x)$ just coincide with the X_i's of (3). That is why they were called ergodic classes. Especially the matrix $W(x' \mid x)$ has no transient element. This follows also from the fact that $P[\xi_1 = x]$ — evidently an invariant distribution of W — is positive for every $x \in X$. Moreover $W(x \mid x)$ is positive for every $x \in X$, hence no ergodic class is cyclic. Thus we can apply the well-known theorem on convergence of the transition probabilities to the case of $W(x' \mid x)$. It states that there exist probability distributions $\Pi_i(x)$ for which $\Pi_i(x) > 0$ if and only if $x \in X_i$, and also a δ $(0 \leq \delta < 1)$ depending only on the matrix W such that $\mid W^{(k)}(x' \mid x) - \Pi_i(x') \mid < \delta^k$ for every x and x' in X_i. From the last inequality we obtain

$$W^{(k)}(x' \mid x) \leq 2^{c_1 \delta^k} \Pi_i(x') \tag{11}$$

where C_1 is a constant depending only on the matrix W. Let us write $\xi^{nk} = \xi_{1k}\, \xi_{2k} \ldots \xi_{nk}$ and especially $\xi^{n0} = \xi^n$. We define

$$|W_n^{(k)}(\bar{x} \mid x^n) = P(\xi^{nk} = \bar{x}^n \mid \xi^{n0} = x^n)$$

and especially $W_1^{(1)}(\bar{x}|x) = W(\bar{x}|x)$. Similarly $W_n^{(k)}(A'|A) = P(\xi^{nk} \in A' | \xi^{n0} \in A)$. $\{\xi^{ni}\}_{i=0}^{\infty}$ is a Markov chain with k-step transition probabilities $W_n^{(k)}(\bar{x}^n \mid x^n)$ The identity

$$W_n^{(1)}(\bar{x}_n \mid x^n) = \sum_{y^n \in Y^n} P(\xi^n = \bar{x}^n \mid \eta^n = y^n) \cdot P(\eta^n = y^n \mid \xi^n = x^n)$$

remains valid, hence $W_n^{(1)}$ plays the role of the matrix $W(x' \mid x)$ on the product space. On the other hand

$$W_n^{(k)}(\bar{x}^n \mid x^n) = \prod_{j=1}^{n} W^{(k)}(\bar{x}_j \mid x_j) . \tag{12}$$

Every sequence $t = i_1 i_2 \dots i_n$ of the indices correspond to the ergodic class $X^{(t)} = X_{i_1} \times X_{i_2} \cdot \cdot \times X_{i_n} \subset X^n$. We write $\Pi_t(\bar{x}^n) = \prod_{j=1}^{n} \Pi_{i_j}(\bar{x}_j)$. The invariant distribution Π_t plays the role of Π_i in the product space. From (11) and (12) we obtain for every $t = i_1 i_2 \dots i_n$ and $\bar{x}^n$, $x^n \in X^{(t)}$ that

$$W_n^{(k)}(\bar{x}^n \mid x^n) \leq 2^{C_1 \cdot n\delta^k} \Pi_t(\bar{x}^n) .$$

Now $\left\{\frac{P(\xi^n = x^n)}{P(\xi^n \in X^{(t)})}\right\}_{x^n \in X^{(t)}}$ is an invariant distribution on $X^{(t)}$. However every ergodic class has a unique invariant distribution and therefore $\frac{P(\xi^n = x^n)}{P(\xi^n \in X^t)} = \Pi_t(x^n)$, i e.

$$W_n^{(k)}(\bar{x}_n \mid x^n) \leq 2^{C_1 \cdot n\, \delta^k} \frac{P(\xi^n = \bar{x}^n)}{P(\xi^n \in X^{(t)})} . \tag{13}$$

Let us now consider the set A_n of the lemma. $A_n \subset \tilde{A}_n$ with $\tilde{A}_n$ being an ergodic class. Replacing $X^{(t)}$ by $\tilde{A}_n$ (12) results in the inequality

$$W_n^{(k)}(A_n \mid A_n) = \frac{1}{P(\xi^n \in A_n)} \cdot \sum_{x^n \in A_n} W_n^{(k)}(A_n \mid x^n) \cdot P(\xi^n = x^n) \leq$$
$$\leq \frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \cdot 2^{C_1 \cdot n\, \delta^k} .$$

We want to obtain a lower bound for $W_n^{(k)}(A_n \mid A_n)$. Let us write $\lambda_n = 2^{-n\varepsilon_n}$. Starting with the estimstion of $W_n^{(1)}(\mathrm{A}_n \mid A_n)$ we have

$$W_n^{(1)}(A_n \mid A_n) = \sum_{y^n \in Y^n} P(\xi^n \in A_n \mid \eta^n = y^n) \cdot P(\eta^n = y^n \mid \xi^n \in A_n) =$$
$$= \frac{1}{P(\xi^n \in A_n)} \cdot \sum_{y^n \in Y^n} P^2(\xi^n \in A_n \mid \eta^n = y^n) \geq \tag{14}$$
$$\geq \frac{1}{P(\xi^n \in A_n)} \cdot \sum_{y^n \in B_n} P^2(\xi^n \in A_n \mid \eta^n = y^n) \cdot P(\eta^n = y^n) .$$

From the conditions of the lemma we have

$$\lambda_n \leq \sum_{y^n \in B_n} P(\xi^n \in A_n \mid \eta^n = y^n) \cdot \frac{P(\eta^n = y^n)}{P(\eta^n \in B^n)} .$$

Applying a reverse Markov inequality we obtain

$$\sum_{y^n,\, P(\xi^n \in A_n \mid \eta^n = y^n) > \frac{1}{2}\lambda_n} \frac{P(\eta^n = y^n)}{P(\eta^n \in B_n)} \geq \frac{1}{2}\lambda_n .$$

Multiplying both sides by $P(\eta^n \in B_n)$ this gives

$$\sum_{y^n,\, P(\xi^n \in A_n \mid \eta_n) > \frac{1}{2}\lambda_n} P(\eta^n = y^n) \geq \frac{1}{2}\lambda_n \cdot P(\eta^n \in B_n) .$$

Combining the last inequality with (14) and (10) we obtain

$$W_n^{(1)}(A_n \mid A_n) \geq \frac{P(\eta^n \in B_n)}{P(\xi^n \in A_n)} \cdot \frac{1}{8}\lambda_n^3 \geq \frac{P(\xi^n \in A_n, \eta^n \in B_n)}{P(\xi^n \in A_n)} \cdot \left(\frac{\lambda_n}{2}\right)^3 \geq 2 \cdot \left(\frac{\lambda_n}{2}\right)^4 .$$

Now we iterate the above estimates. By induction

$$W_n^{(2^r)}(A_n \mid A_n) \geq 2 \cdot \left(\frac{\lambda_n}{2}\right)^{4^{r+1}}$$

Putting $k = 2^r$ and $\lambda_n = 2^{-n\varepsilon_n}$

$$W_n^{(k)}(A_n \mid A_n) \geq 2^{-4k^2(n \cdot \varepsilon_n + 1)} .$$

From the assumption that $\varepsilon_n > \frac{1}{n}$ we thus obtain

$$W_n^{(k)}(A_n \mid A_n) \geq 2^{-8k^2\, n \cdot \varepsilon_n} .$$

Comparing this inequality with (13) we have

$$2^{-8k^3 n \varepsilon_n} \leq W_n^{(k)}(A_n \mid A_n) \leq \frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \cdot 2^{c_1 n \delta^k} ,$$

i.e.

$$\frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \geq 2^{-n[8k^2 \varepsilon_n + c_1 \cdot \delta^k]} .$$

Finally let us choose k in the exponent of the right-hand side in a suitable manner. Let $\delta^k = 2^{-\delta_1 k}$. We prescribe for k that $\frac{2}{\delta_1}\log\frac{1}{\varepsilon_n} \leq k \leq \frac{2}{\delta_1}\log\frac{1}{\varepsilon_n}$. It is possible to choose $k = 2^r$ within these bounds. Now

$$\frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \geq 2^{-n(C \cdot \varepsilon_n \log^2 \varepsilon_n)}$$

Using (10) this yields

$$\frac{P(\xi^n \in A_n . \eta^n \in B_n)}{P(\xi^n \in \tilde{A}_n, \eta^n \in \tilde{B}_n)} \geq 2^{-n\varepsilon_n} \cdot \frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \geq 2^{-n\, C\, \varepsilon_n\, \log^2 \varepsilon_n}$$

what we wanted to prove.

5. Proof of Theorems 1 and 2. Random common length

We have all the tools to finish the proofs of Theorems 1 and 2.

Proof of Theorem 2. This theorem is an immediate consequence of the lemma. Given any series $\lambda_n = 2^{-n\varepsilon_n}$ with ε_n tending to 0 one has the inequality:

$$\frac{1}{n} \cdot \log Q(n, \lambda_n) \geq - C \cdot \varepsilon_n \cdot \log^2 \varepsilon_n + \min_t \left(\frac{1}{n} \cdot \log P(\xi^n \in X^{(t)}, \eta^n \in Y^{(t)}) \right) .$$

Since $\varepsilon_n \cdot \log^2 \varepsilon_n$ tends to 0 when $\varepsilon_n \to 0$, evidently

$$\liminf_{n \to \infty} \frac{1}{n} \cdot \log Q(n_1 \lambda_n) \geq \min_z P(\zeta = z) .$$

On the other hand the choice $(A_n, B_n) = X_{z_0}, Y_{z_0})^n$ where z_0 is the index of the ergodic class having minimum probability, gives that $\lim_{n \to \infty} \frac{1}{n} \log Q(n_1 \lambda_n) = \min_z P(\zeta = z]$.

End of the proof of Theorem 1. Now we turn to the proof of the inequality (5a). Let us consider the function φ_n defined in (8). Apart from a set of probability $4 \sqrt{\bar{\varepsilon}}$ this function is defined on $(1 - \sqrt{\bar{\varepsilon}})$-blocks (see Proposition 1). Let us choose $\lambda > 4 \sqrt{\varepsilon}$. We shall prove that there exists a code $\tilde{\varphi}_n = U_n(\varphi_n)$ of $\varphi_n(\xi_n)$ with some fixed length and error probability less than λ, which performs the inequality $\limsup_{n \to \infty} \frac{1}{n} \cdot l(\varphi_n) \leq H(\zeta_1)$.

The construction of the code U_n is the following: let U_n take a constant value on the complementary set of $C(\varepsilon)$ defined in Proposition 1. On $C(\varepsilon)$ we code the values of $\varphi_n(\xi^n) = (f_n(\xi^n), \zeta^n)$ in two steps. First we construct *a ($\lambda - 4 \sqrt{\bar{\varepsilon}}$) — code (see Definition 2) of ζ^n with codewords' length less than* $n H(\zeta_1) + k_1 \sqrt{\bar{n}}$. According to the lemma φ_n takes less than $2^{K_1 \log^2 n}$ different values in an ergodic class, i.e. for any fixed value of ζ^n. This is so because the different values of the restriction $\varphi_n \mid C(\varepsilon)$ are taken of different $(1 - \sqrt{\bar{\varepsilon}})$-blocks. Thus the whole code U_n may have length less than $nH(\zeta_1) + K_1 \sqrt{n} + K_2 \log^2 n$.

So we obtain that $\limsup\limits_{n\to\infty} \frac{1}{n} l(\tilde{\varphi}_n) \leq H(\zeta_1)$. This proves (5a). The second part of the theorem is rather trivial. It is enough to prove at this point that any ε_1-code of the sequence ζ^n with $\varepsilon_1 < \varepsilon$ can be completed by addition of a complementary code of length $nH(\xi_1 \mid \zeta_1) + K\sqrt{n}$ to an $\varepsilon = \varepsilon_1 + \varepsilon_2$-code of the sequence ξ^n. Although this is very simple to show we prefer to give a reference where the proof is done (see [4]). Now our demonstration is complete. We have shown that it is not possible in general to code two DMSS's so that the resulting codes have some fixed common length of order n. All the same problem concerning the length of a possible random coincidence of codes has remained open. Which is the behaviour of the maximum expected length of random concidence? We shall show that this length is asymptotically equal to $nH(\zeta_1)$. Let us formulate first an exact form of the question.

Definition 6. Let $F_n(\xi^n)$ and $G_n(\eta^n)$ be ε-codes of the DMSS's $\mathfrak{X} = \{\xi_n\}$ and $\mathfrak{Y} = \{\eta^n\}$ respectively. Suppose that

$$l(F_n(\xi^n)) < (1 + \delta_n) \cdot H(\xi^n) \text{ and } l(G_n(\eta^n)) < (1 + \delta_n) \cdot H(\eta^n) .$$

Let the integer valued random variable ν_n be the length of the maximum common beginning segment of the binary sequences $F_n(\xi^n)$ and $G_n(\eta^n)$. The random variable ν_n is called the *random common length* of the codes F_n and G_n.

Theorem 3. Let $E(\nu_n)$ denote the expected value of ν_n. Supposing that $\delta_n \to 0$ we have

$$\limsup_{n\to\infty} E(\nu_n) \leq H(\zeta_1) .$$

Proof. Let F_n and G_n be given. From a reverse Markov inequality we obtain for $0 \leq x < 1$ that

$$P[\nu_n > (1-\alpha)\cdot E(\nu_n)] \geq \frac{\alpha \cdot E(\nu_n)}{[(1+\delta_n)\cdot H(\xi^n)] - (1-\alpha)\cdot E(\nu_n)} \geq$$

$$\geq \frac{\alpha}{2H(\xi_1)} \cdot \frac{E(\nu_n)}{n} . \tag{15}$$

Let us fix α independently of n. Put $l_n = (1-\alpha)\cdot E(\nu_n)$. If $f_n(\xi^n)$ and $g_n(\eta^n)$ denote the first l_n letters of the binary sequences $F_n(\xi^n)$ and $G_n(\eta^n)$ respectively, (15) yields

$$P[f_n(\xi^n) = g_n(\eta^n)] > \frac{\alpha}{2H(\xi_1)} \cdot \frac{E(\nu_n)}{n} .$$

Choosing for f_n' and g_n' the remaining segments of F_n and G_n we obtain Y-schemes having error probabilities equal to $\max\left(\varepsilon, 1 - \frac{\alpha}{2H(\xi_1)} \cdot \frac{E(\nu_n)}{n}\right)$. The common length is these Y-schemes is $l_n = (1 - \alpha) \cdot E(\nu_n)$. Thus Theorem 1 implies that

$$\limsup_{n \to \infty} \frac{E(\nu_n)}{n} \leq H(\zeta_1)$$

The complexity interpretation of the result

This section will be understandable only for readers acquainted with the notion of complexity of a finite word given by Kolmogorov [1, 2] (see also [3], [4]). Let X be a finite alphabet containing 0 and 1, $|X| = S$ Let T be a universal Turing machine the operation of which results in a partial function $T(p, x)$, $T : I^* \times X^* \to X^*$. We define

$$K_T(y \mid x) = \min_{T(p,x)=y} l(p) .$$

It is known that if T' is another universal machine then $|K_T - K_{T'}| \leq C_{TT'}$ where the constant $C_{TT'}$ depends only on the two machines Let us fix T once for ever and denote $K(y \mid x) = K_T(y \mid x)$ the "conditional compexity" of y with respect to x, $K(x) = K(x \mid \Lambda)$ the "complexity" of x, $K(x, y)$ the "common complexity" of x and y. The last can be defined for example as the length of the shortest program $p \in I^*$ for which $T(0p, \Lambda) = x$, $T(1p, \Lambda) = y$. We define further $I(x : y) = K(x) - K(x \mid y)$, the "mutual information" of x and y We cite from [3] the following facts among which only (16) is not very easy to see Let $\leq$ resp $\asymp$ denote the inequality resp equality up to an additive constant. By $\approx$ we mean equality up to an additive term $O(\log K(x, y))$. Then we have

$$K(x \mid y) \leq K(x) \leq l(x) \log s$$

$$K(z \mid y) + K(y \mid x) \geq K(z \mid x)$$

hence denoting $\Delta(x, y) = K(x \mid y) + K(y \mid x)$ Δ has approximately the properties of a metrics. Further

$$K(x, y) \approx K(x) + K(y \mid x) \tag{16}$$

hence $I(x : y) \approx I(y : x)$.

If $(\mathfrak{X}, \mathfrak{Y}) = \{\xi_n, \eta_n\}$ is a stationary source which is ergodic then for almost all $x = x_1x_2 \ldots, y = y_1y_2 \ldots$

$$\lim_{n\to\infty} \frac{1}{n} K(x^n) = \Pi(\mathfrak{X}) = \lim_{n\to\infty} \frac{1}{n} H(\xi^n), \ \lim_{n\to\infty} \frac{1}{n} K(x^n, y^n) = \lim_{n\to\infty} \frac{1}{n} H(\xi^n, \eta^n)$$

$$\lim_{n\to\infty} \frac{1}{n} K(x^n \mid y^n) = \lim_{n\to\infty} \frac{1}{n} H(\xi^n \mid \eta^n), \ \lim_{n\to\infty} \frac{1}{n} I(x^n : y^n) = \lim_{n\to\infty} \frac{1}{n} I(\xi^n, \eta^n)$$

The approximate symmetry of $I(x . y)$ suggested here the question: can we give a symmetric definition of it? For example is there a triple $a, b, c \in I^*$ of programs with

$$T(ca, A) = x, \ T(cb, A) = y$$

$$l(a) \approx K(x \mid y), \ l(b) \approx K(y \mid x)\ , \ l(c) = I(x : y).$$

More generally let $x, y \in X^n$, we shall say that z represents some common information of x and y if $K(z \mid x) \approx 0$, $K(z \mid y) \approx 0$. Let us define

$$J(x, y) = \max K(z)$$
$$z : K(z \mid x) \approx 0$$
$$K(z \mid y) \approx 0.$$

Here $\approx$ must be given in a bit more exact sense but we do not wish to be quite precise here. If $K(z) \approx J(x, y)$ then z is almost determined in the sense that it is in a given "sphere" of radius ≈ 0 in the metrics Δ. In this case we shall say that z *is* the common information of x and y. Our question was whether $J(x, y) \approx I(x \ y)$ The answer is negative in a rather general case. Let ξ^n and η^n denote the random variables defined in § 2. (Suppose that $X = Y$.) Then in the product space with a probability tending to 1 we have by Theorem 1 that ξ^n *is* the common information in ξ^n and η^n.

We are indebted to Imre Csiszár sr. for his stating of the problem and constant aid.

REFERENCES

1 *Doob, J L* Stochastic processes Wiley, New York, 1953, pp 170–185.
2 *Kolmogorov, A N* Logical basis for information theory and probability theory. IEEE Transactions IT–14 (1968) pp 662–664
3 *Kolmogorov, A N* Three approaches to the concept of "information quantity" (in Russian) Problemi peredachi informatsii 1 . 1 (1965) pp 3–7.
4. *Korner, J* On a property of conditional entropy Stud Sci Math Hung (1971) (to appear)
5 *Loeve, M* Probability Theory Van Nostrand New York, 1955 pp. 157 and 28–42.
6 *Martin Lof, P* The definition of random sequences. Inf and Control 9 (1966) pp. 602–619
7 *Zvonkin, A K –Levin, L A* Complexity of finite objects etc. (in Russian) Uspekhi Mat Nauk 25 6 (1970) pp 85–127.

Общая информация существенно больше, чем взаимная информация

П. ГААЧ–Й. КОРНЕР

(Будапешт)

С помощью трех представленных теорем уточняется понятие взаимной информации и доказывается, что термин «общая информация» является более общим, чем термин «взаимная информация».

P. Gács, J. Kórner
Mathematical Research Institute
Budapest, V. Reáltanoda u. 13–15. Hungary