Chain Independence and Common Information

Konstantin Makarychev and Yury Makarychev

***Abstract*—We present a new proof of a celebrated result of Gács and Körner that the common information is far less than the mutual information. Consider two sequences $\alpha_1, \dots \alpha_n$ and $\beta_1, \dots, \beta_n$ of random variables, where pairs $(\alpha_1, \beta_1), \dots, (\alpha_n, \beta_n)$ are independent and identically distributed. Gács and Körner proved that it is not possible to extract "common information" from these two sequences unless the joint distribution matrix of random variables (α_i, β_i) is a block matrix.**

In 2000, Romashchenko introduced a notion of *chain independent* random variables and gave a simple proof of the result of Gács and Körner for chain independent random variables. Furthermore, Romashchenko showed that *boolean* random variables α and β are chain independent unless $\alpha = \beta$ a.s. or $\alpha = 1 - \beta$ a.s. In this paper, we generalize this result to arbitrary (finite) distributions of α and β and thus give a simple proof of the result of Gács and Körner.

***Index Terms*—Chain independent random variables, common information, rate region.**

I. Introduction

Ahlswede, Gács, Körner, Witsenhausen and Wyner [1], [2], [4], [7], [8] studied the problem of extraction of "common information" from a pair of random variables. The simplest form of this problem is the following: Fix some distribution for a pair of random variables α and β. Consider n independent pairs $(\alpha_1, \beta_1), \dots, (\alpha_n, \beta_n)$; each has the same distribution as (α, β). We want to extract "common information" from the sequences $\alpha_1, \dots \alpha_n$ and $\beta_1, \dots, \beta_n$, i.e. find a random variable γ such that $H(\gamma|(\alpha_1, \dots, \alpha_n))$ and $H(\gamma|(\beta_1, \dots, \beta_n))$ are small. We say that "extraction of common information is impossible" if the entropy of any such variable γ is small.

Let us show that this is the case if α and β are independent. In this case $\alpha^n = (\alpha_1, \dots, \alpha_n)$ and $\beta^n = (\beta_1, \dots, \beta_n)$ are independent. Recall the well-known inequality

$$H(\gamma) \le H(\gamma|\alpha^n) + H(\gamma|\beta^n) + I(\alpha^n : \beta^n).$$

Here $I(\alpha^n : \beta^n) = 0$ (because α^n and β^n are independent); two other summands on the right hand side are small by our assumption.

It turns out that a similar statement holds for dependent random variables. However, there is one exception. If the joint probability matrix of (α, β) can be divided into blocks, there is a random variable τ that is a function of α and a function of β ("block number"). Then $\gamma = (\tau_1, \dots, \tau_n)$ is common information of α^n and β^n.

It was shown by Gács and Körner [4] that this is the only case when there exists common information. Their original proof is quite technical. Several years ago another approach was proposed by Romashchenko [5] using "chain independent" random variables. Romashchenko introduced the notion of chain independent random variables (which he called conditionally independent random variables) and showed that extraction of common information from chain independent random variables is impossible. We prove that if the joint probability matrix of a pair of random variables (α, β) is not a block matrix, then α and β are chain independent. We also show several new information inequalities for chain independent random variables.

Konstantin Makarychev is with Microsoft Research.

Yury Makarychev is with Toyota Technological Institute at Chicago. He is supported in part by the National Science Foundation Career Award CCF-1150062.

II. Chain independent random variables

Consider four random variables $\alpha, \beta, \alpha^*, \beta^*$. Suppose that α^* and β^* are independent, α and β are independent given α^*, and also independent given β^*, i.e., $I(\alpha^* : \beta^*) = 0$, $I(\alpha : \beta|\alpha^*) = 0$ and $I(\alpha : \beta|\beta^*) = 0$. Then we say that α and β are *chain independent of order* 1. (Chain independent random variables of order 0 are independent random variables.)

We consider chain independence of random variables as a property of their joint distributions. If a pair of random variables α and β has the same joint distribution as a pair of chain independent random variables α_0 and β_0 (on another probability space), we say that α and β are chain independent.

Replacing the requirement of independence of α^* and β^* by the requirement of chain independence of order 1, we get the definition of chain independent random variables (α and β) of order 2 and so on.

Definition 1: We say that α and β are *conditionally independent with respect to α^* and β^** if α and β are independent given α^*, and they are also independent given β^*, i.e., $I(\alpha : \beta|\alpha^*) = I(\alpha : \beta|\beta^*) = 0$.

Definition 2 (Romashchenko [5]): Two random variables α and β are called chain independent random variables of order k ($k \ge 0$) if there exists a probability space Ω and a sequence of pairs of random variables

$$(\alpha_0, \beta_0), (\alpha_1, \beta_1), \dots, (\alpha_k, \beta_k)$$

on it such that

(a) The pair (α_0, β_0) has the same distribution as (α, β).

(b) α_i and β_i are conditionally independent with respect to α_{i+1} and β_{i+1} when $0 \le i < k$.

(c) α_k and β_k are independent random variables.

The sequence

$$(\alpha_0, \beta_0), (\alpha_1, \beta_1), \dots, (\alpha_k, \beta_k)$$

is called a derivation for (α, β).

We say that random variables α and β are chain independent if they are chain independent of some order k.

The notion of chain independence can be applied for analysis of common information using the following observations (see below for proofs).

Lemma 1: Consider chain independent random variables α and β of order k. Let $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)$ be a sequence of random variables such that all (α_i, β_i) are independent, and each (α_i, β_i) is distributed as (α, β). Then the random variables $\alpha^n = (\alpha_1, \ldots, \alpha_n)$ and $\beta^n = (\beta_1, \ldots, \beta_n)$ are chain independent of order k.

The proof is given in Section IV (see statement (e)).

Theorem 1 (Romashchenko [5]): If random variables α and β are chain independent of order k, and γ is an arbitrary random variable (on the same probability space), then

$$H(\gamma) \leq 2^k H(\gamma|\alpha) + 2^k H(\gamma|\beta).$$

The proof is given in Section III.

Definition 3: An $m \times n$ matrix is called a *block matrix* if (after some permutation of its rows and columns) it consists of four blocks; the blocks on the diagonal are not equal to zero; the blocks outside the diagonal are equal to zero.

Formally, A is a block matrix if the set of its rows $\{1, \ldots, m\}$ can be divided into two disjoint nonempty sets I_1 and I_2 ($I_1 \sqcup I_2 = \{1, \ldots, m\}$) and the set of its columns $\{1, \ldots, n\}$ can be divided into two sets J_1 and J_2 ($J_1 \sqcup J_2 = \{1, \ldots, n\}$) in such a way that each of the blocks $\{a_{ij} : i \in I_1, j \in J_1\}$ and $\{a_{ij} : i \in I_2, j \in J_2\}$ contains at least one nonzero element, and all the elements outside these two blocks are equal to 0, i.e. $a_{ij} = 0$ when $(i, j) \in (I_1 \times J_2) \cup (I_2 \times J_1)$.

Theorem 2: Random variables are chain independent if and only if their joint probability matrix is not a block matrix.

Using these statements, we conclude that if the joint probability matrix of a pair of random variables (α, β) is not a block matrix, then no information can be extracted from a sequence of n independent random variables each with the same distribution as (α, β):

$$H(\gamma) \leq 2^k H(\gamma|\alpha^n) + 2^k H(\gamma|\beta^n)$$

for some k (that does not depend on n) and for any random variable γ.

III. Proof of Theorem 1

Theorem 1 (Romashchenko [5]): If random variables α and β are chain independent of order k, and γ is an arbitrary random variable (on the same probability space), then

$$H(\gamma) \leq 2^k H(\gamma|\alpha) + 2^k H(\gamma|\beta).$$

Proof : The proof is by induction on k. The statement is already proved for independent random variables α and β ($k = 0$).

Suppose α and β are conditionally independent with respect to chain independent random variables α^* and β^* of order $k-1$. From the conditional form of the inequality

$$H(\gamma) \leq H(\gamma|\alpha) + H(\gamma|\beta) + I(\alpha : \beta)$$

(α^* is added everywhere as a condition) it follows that

$$H(\gamma|\alpha^*) \leq H(\gamma|\alpha\alpha^*) + H(\gamma|\beta\alpha^*) + I(\alpha : \beta|\alpha^*) = H(\gamma|\alpha\alpha^*) + H(\gamma|\beta\alpha^*) \leq H(\gamma|\alpha) + H(\gamma|\beta).$$

Similarly, $H(\gamma|\beta^*) \leq H(\gamma|\alpha) + H(\gamma|\beta)$. By the induction hypothesis $H(\gamma) \leq 2^{n-1}H(\gamma|\alpha^*) + 2^{n-1}H(\gamma|\beta^*)$. Replacing $H(\gamma|\alpha^*)$ and $H(\gamma|\beta^*)$ by their upper bounds, we get $H(\gamma) \leq 2^n H(\gamma|\alpha) + 2^n H(\gamma|\beta)$. ■

Corollary 1.1: If the joint probability matrix A of a pair of random variables is a block matrix, then these random variables are not chain independent.

Proof: Suppose that the joint probability matrix A of random variables (α, β) is a block matrix and these random variables are chain independent of order k.

Let us divide the matrix A into blocks $I_1 \times J_1$ and $I_2 \times J_2$ as in Definition 3. Observe, that $\alpha \in I_1$ if and only if $\beta \in J_1$. Define a new random variable γ. Let $\gamma = 1$, if $\alpha \in I_1$ and $\beta \in J_1$, and $\gamma = 2$, if $\alpha \in I_2$ and $\beta \in J_2$. Then, the random variable γ is a function of α and at the same time a function of β. Therefore, $H(\gamma|\alpha) = 0$ and $H(\gamma|\beta) = 0$. However, γ takes two different values with positive probability. Hence $H(\gamma) > 0$, which contradicts Theorem 1. ■

A similar argument shows that the order of chain independence should be large if the matrix is close to a block matrix.

IV. Proof of Theorem 2

For brevity, we call joint probability matrices of chain independent random variables *good matrices*.

The proof of Theorem 2 consists of three main steps. First, we prove, that the set of good matrices is dense in the set of all joint probability matrices. Then we prove that any matrix without zero elements is good. Finally, we consider the general case and prove that any matrix that is not a block matrix is good.

The following statements are used in the sequel.

(a) The joint probability matrix of independent random variables is a matrix of rank 1 and vice versa. In particular, all (joint probability) matrices of rank 1 are good.

(b) If α and β are chain independent, α' is a function of α, and β' is a function of β, then α' and β' are chain independent. (Indeed, if α and β are conditionally independent with respect to some α^* and β^*, then α' and β' are also conditionally independent with respect to α^* and β^*.)

(c) If two random variables are chain independent of order k, then they are chain independent of order l for any $l > k$. (We can add some constant random variables to the end of the derivation.)

(d) Assume that chain independent random variables α_1 and β_1 are defined on a probability space Ω_1 and chain independent random variables α_2 and β_2 are defined on a probability space Ω_2. Consider random variables (α_1, α_2) and (β_1, β_2) that are defined in a natural way on the product space $\Omega_1 \times \Omega_2$. Then (α_1, α_2) and (β_1, β_2) are chain independent. Indeed, for each pair (α_i, β_i) consider its derivation

$$(\alpha_i^0, \beta_i^0), (\alpha_i^1, \beta_i^1), \ldots, (\alpha_i^l, \beta_i^l)$$

(using (c), we may assume that both derivations have the same length l).

Then the sequence

$$((\alpha_1^0,\alpha_2^0),(\beta_1^0,\beta_2^0)),\dots,((\alpha_1^l,\alpha_2^l),(\beta_1^l,\beta_2^l))$$

is a derivation for the pair of random variables $((\alpha_1,\alpha_2),(\beta_1,\beta_2))$. For example, random variables $(\alpha_1,\alpha_2)=(\alpha_1^0,\alpha_2^0)$ and $(\beta_1,\beta_2)=(\beta_1^0,\beta_2^0)$ are independent given the value of (α_1^1,α_2^1), because α_1 and β_1 are independent given α_1^1, variables α_2 and β_2 are independent given α_2^1, and the measure on $\Omega_1\times\Omega_2$ is equal to the product of the measures on Ω_1 and Ω_2.

Applying statement (d) $(n-1)$ times, we get Lemma 1. Combining Lemma 1 and statement (b), we get the following statement:

(e) Let $(\alpha_1,\beta_1),\dots,(\alpha_n,\beta_n)$ be independent and identically distributed random variables. Assume that the variables in each pair (α_i,β_i) are chain independent. Then any random variables α' and β', where α' depends only on $\alpha_1,\dots,\alpha_n$ and β' depends only on $\beta_1,\dots,\beta_n$, are chain independent.

Definition 4: Denote

$$D_\varepsilon=\begin{pmatrix}1/2-\varepsilon & \varepsilon\\ \varepsilon & 1/2-\varepsilon\end{pmatrix}$$

(where $0\le\varepsilon\le 1/2$).

The matrix $D_{1/4}$ corresponds to a pair of independent random bits; as ε tends to 0 these bits become more dependent (though each is still uniformly distributed over $\{0,1\}$).

Lemma 2 (Special case of Lemma 4 in [5]): (i) $D_{1/4}$ is a good matrix.

(ii) If D_ε is a good matrix then $D_{\varepsilon(1-\varepsilon)}$ is good.

(iii) There exists an arbitrary small ε such that D_ε is good.

Proof:

(i) The matrix $D_{1/4}$ is of rank 1, hence it is good (independent random bits).

(ii) Consider a pair of random variables α and β distributed according to D_ε.

Define new random variables α' and β' as follows:

- if $(\alpha,\beta)=(0,0)$ then $(\alpha',\beta')=(0,0)$;
- if $(\alpha,\beta)=(1,1)$ then $(\alpha',\beta')=(1,1)$;
- if $(\alpha,\beta)=(0,1)$ or $(\alpha,\beta)=(1,0)$ then

$$(\alpha',\beta')=\begin{cases}(0,0) & \text{with probability } \varepsilon/2;\\ (0,1) & \text{with probability } (1-\varepsilon)/2;\\ (1,0) & \text{with probability } (1-\varepsilon)/2;\\ (1,1) & \text{with probability } \varepsilon/2.\end{cases}$$

The joint probability matrix of α' and β' given $\alpha=0$ is equal to

$$\begin{pmatrix}(1-\varepsilon)^2 & \varepsilon(1-\varepsilon)\\ \varepsilon(1-\varepsilon) & \varepsilon^2\end{pmatrix}$$

and its rank equals 1. Therefore, α' and β' are independent given $\alpha=0$.

Similarly, the joint probability matrix of α' and β' given $\alpha=1$, $\beta=0$ or $\beta=1$ has rank 1. This yields that α' and β' are conditionally independent with respect to α and β, hence α' and β' are chain independent.

The joint distribution of α' and β' is

$$\begin{pmatrix}1/2-\varepsilon(1-\varepsilon) & \varepsilon(1-\varepsilon)\\ \varepsilon(1-\varepsilon) & 1/2-\varepsilon(1-\varepsilon)\end{pmatrix},$$

hence $D_{\varepsilon(1-\varepsilon)}$ is a good matrix.

(iii) Consider the sequence ε_n defined by $\varepsilon_0=1/4$ and $\varepsilon_{n+1}=\varepsilon_n(1-\varepsilon_n)$. The sequence ε_n tends to zero (its limit is a root of the equation $x=x(1-x)$). It follows from statements (i) and (ii) that all matrices D_{ε_n} are good. ■

Note: The order of chain independence of D_ε tends to infinity as $\varepsilon\to 0$. Indeed, applying Theorem 1 to random variables α and β with joint distribution D_ε and to $\gamma=\alpha$, we obtain

$$H(\alpha)\le 2^k(H(\alpha|\alpha)+H(\alpha|\beta))=2^kH(\alpha|\beta).$$

Here $H(\alpha)=1$; for any fixed value of β the random variable α takes two values with probabilities 2ε and $1-2\varepsilon$, therefore

$$H(\alpha|\beta)=-(1-2\varepsilon)\log_2(1-2\varepsilon)-2\varepsilon\log_2(2\varepsilon)=O(-\varepsilon\log_2\varepsilon)$$

and, if D_ε corresponds to chain independent variables of order k, then

$$2^k\ge H(\alpha)/H(\alpha|\beta)=1/O(-\varepsilon\log_2\varepsilon)\to\infty$$

as $\varepsilon\to 0$.

We showed that for every positive ε, there exists two chain independent random variables α and β each distributed uniformly in $\{0,1\}$ such that $\Pr(\alpha\ne\beta)\le\varepsilon$. We generalize this statement to distributions on a larger domain.

Corollary 2.1: For every positive ε, for every natural T, there exist two chain independent random variables α and β uniformly distributed on the boolean cube $\{0,1\}^T$ such that $\Pr(\alpha\ne\beta)\le\varepsilon$ and for every $u,v\in\{0,1\}^T$, $\Pr(\alpha=u,\beta=v)=\Pr(\alpha=v,\beta=u)$.

Proof: Consider T independent pairs of random variables $(\alpha_1,\beta_1),\dots,(\alpha_T,\beta_T)$; each variable (α_i,β_i) is distributed according to D_ε (for sufficiently small ε). The random variables $\alpha=(\alpha_1,\dots,\alpha_T)$ and $\beta=(\beta_1,\dots,\beta_T)$ are uniformly distribute on the boolean cube $\{0,1\}^T$. They are chain independent by Lemma 1. We have

$$\Pr(\alpha\ne\beta)\le\sum_{i=1}^T\Pr(\alpha_i\ne\beta_i)\le\varepsilon'\equiv T\varepsilon.$$

Since the joint probability matrix of α_i and β_i is symmetric for each i, and all random variables (α_i,β_i) are independent, we have $\Pr(\alpha=u,\beta=v)=\Pr(\alpha=v,\beta=u)$. ■

Lemma 3: For every sequence $p_1,\dots,p_n\in(0,1)$ with $p_1+\dots+p_n=1$, and for every positive $\varepsilon>0$, there exist two chain independent random variables α and β such that $\Pr(\alpha=i)=\Pr(\beta=i)=p_i$ (for $i\in\{1,\dots,n\}$), $\Pr(\alpha\ne\beta)\le\varepsilon$ and the joint probability matrix of α and β is symmetric i.e., $\Pr(\alpha=i,\beta=j)=\Pr(\alpha=j,\beta=i)$ (for $i,j\in\{1,\dots,n\}$).

Proof: Fix an integer $T\in\mathbb{N}$ such that $2^{-T}n\le\varepsilon/2$ and let α and β be two chain independent random variables uniformly distributed on $\{0,1\}^T$ as in Corollary 2.1 such that $\Pr(\alpha\ne\beta)\le\varepsilon/2$. Loosely speaking, our goal is to represent the desired random variables α' and β' as $\alpha'=f(\alpha)$ and $\beta'=f(\beta)$ for some function f. Then, α' and β' are chain

independent (by statement (b)), and $\Pr(\alpha' \neq \beta') \leq \Pr(\alpha \neq \beta) \leq \varepsilon$. To do so, we express each p_i as $p_i = q_i/2^T + r_i/2^T$, where $q_i = \lfloor p_i 2^T \rfloor$ is an integer, and $r_i = (p_i 2^T - q_i) \in (0,1)$. Then, we arbitrary partition 2^T vertices of the boolean cube $\{0,1\}^T$ in n sets $Q_1, \ldots, Q_n$ of size $|Q_i| = q_i$ and a set R of remaining vertices. This is possible because $\sum_{i=1}^n |Q_i| = \sum_{i=1}^n \lfloor 2^T p_i \rfloor \leq 2^T$. If all r_i were equal to 0, we would let $f(u) = i$ for $u \in Q_i$, and get $\Pr(f(\alpha) = i) = \Pr(\alpha \in Q_i) = p_i$. In the general case, we define two independent random variables ξ and η,

$$\Pr(\xi = i) = \Pr(\eta = i) = \frac{r_i}{|R|}.$$

Observe, that $\sum_i r_i = |R|$, hence ξ and η are well defined. We define $f : \{0,1\}^T \times \{1, \ldots, n\} \to \{1, \ldots, n\}$ as follows:

$$f(u,j) = \begin{cases} i, & \text{if } u \in Q_i \text{ for some } i; \\ j, & \text{if } u \in R; \end{cases}$$

and let $\alpha' = f(\alpha, \xi)$ and $\beta' = (\beta, \eta)$. The random variables α and β are chain independent, the random variables ξ and η are independent, thus the pairs (α, ξ) and (β, η) are chain independent. By statement (b), $f(\alpha, \xi)$ and $f(\beta, \eta)$ are chain independent. Verify that α' and β' satisfy the desired properties.

$$\begin{aligned} \Pr(f(\alpha,\xi) = i) &= \Pr(\alpha \in Q_i) + \Pr(\alpha \in R) \cdot \Pr(\xi = i) \\ &= \frac{q_i}{2^T} + \frac{|R|}{2^T} \cdot \frac{r_i}{|R|} = p_i. \end{aligned}$$

Then,

$$\begin{aligned} \Pr(f(\alpha,\xi) \neq f(\beta,\eta)) &\leq \Pr(\alpha \neq \beta) + \Pr(\alpha \in R) \\ &= \frac{\varepsilon}{2} + \frac{n}{2^T} \leq \varepsilon. \end{aligned}$$

Finally, our construction is completely symmetric with respect to α' and β', thus $\Pr(\alpha' = i, \beta' = j) = \Pr(\alpha' = j, \beta' = i)$ for all i and j. ∎

We now prove a lemma that defines chain independence in terms of matrices. Denote by $S(M)$ the sum of all elements of a matrix M.

Lemma 4: Consider a matrix N whose elements are matrices N_{ij} of the same size. If

(a) all N_{ij} contain only nonnegative elements;

(b) the sum of matrices in each row and in each column of the matrix N is a matrix of rank 1;

(c) the matrix P with elements $p_{ij} = S(N_{ij})$ is a good joint probability matrix;

then the sum of all the matrices N_{ij} is a good matrix.

Proof: This lemma is a reformulation of the definition of chain independent random variables. Consider random variables α^*, β^* such that the probability of the event $(\alpha^*, \beta^*) = (i,j)$ is equal to p_{ij}, and the probability of the event

$$\alpha = k,\ \beta = l,\ \alpha^* = i,\ \beta^* = j$$

is equal to the (k,l)-th element of the matrix N_{ij}.

The sum of matrices N_{ij} in a row i corresponds to the distribution of the pair (α, β) given $\alpha^* = i$; the sum of matrices N_{ij} in a column j corresponds to the distribution of the pair (α, β) given $\beta^* = j$; the sum of all the matrices N_{ij} corresponds to the distribution of the pair (α, β). ∎

Example: Using Lemma 4, we can prove Lemma 2 as follows. Consider the matrix

$$N = \frac{1}{2} \left(\begin{array}{cc|cc} 1-2\varepsilon & 0 & \varepsilon^2 & \varepsilon(1-\varepsilon) \\ 0 & 0 & \varepsilon(1-\varepsilon) & \varepsilon^2 \\ \hline \varepsilon^2 & \varepsilon(1-\varepsilon) & 0 & 0 \\ \varepsilon(1-\varepsilon) & \varepsilon^2 & 0 & 1-2\varepsilon \end{array} \right).$$

The sum of all elements N_{ij} equals $D_{\varepsilon(1-\varepsilon)}$. The sum of matrices in each row and each column has rank 1. The matrix P (with $p_{ij} = S(N_{ij})$, see Lemma 4 item (c)) equals D_ε. Hence, if D_ε is good, then $D_{\varepsilon(1-\varepsilon)}$ is good.

We will use the following definition.

Definition 5: An *r-matrix* is a matrix with nonnegative elements and with a "rectangular" support i.e., a matrix A is an r-matrix if for some set of rows I and some set of columns J, $a_{ij} > 0$, if $(i,j) \in I \times J$ and $a_{ij} = 0$, otherwise.

Lemma 5: Every r-matrix M is the sum of some r-matrices of rank 1 with the same support as M.

Proof: Let the rectangle $I \times J$ be the support of M. Consider the basis E_{ij} in the vector space of matrices whose support is a subset of $I \times J$. (Here E_{ij} is the matrix that has 1 in the (i,j)-position and 0 elsewhere.)

The matrix M has positive coordinates in the basis E_{ij}. Let us approximate each matrix E_{ij} by a slightly different matrix E'_{ij} of rank 1 with support $I \times J$:

$$E'_{ij} = \left(\bar{e}_i + \varepsilon \sum_{k \in I} \bar{e}_k \right) \cdot \left(\bar{e}_j + \varepsilon \sum_{l \in J} \bar{e}_l \right)^T,$$

where $\bar{e}_1, \ldots, \bar{e}_n$ is the standard basis in $\mathbb{R}^n$.

The coordinates c_{ij} of M in the new basis E'_{ij} continuously depend on ε. Thus they remain positive if ε is sufficiently small. So taking a sufficiently small ε we get the required representation of M as the sum of matrices of rank 1 with support $I \times J$:

$$M = \sum_{(i,j) \in I \times J} c_{ij} E'_{ij}.$$

∎

Lemma 6: Every r-matrix is good.

Proof: Let M be an r-matrix with support $I \times J$. Using Lemma 5, we represent it as a sum of n rank 1 matrices: $M = A^{(1)} + \cdots + A^{(T)}$. To illustrate the idea of the proof, consider a matrix $\tilde{N}$ as in Lemma 4:

$$\begin{pmatrix} A^{(1)} & 0 & 0 & 0 \\ 0 & A^{(2)} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & A^{(T)} \end{pmatrix}.$$

The sum of the matrices in each row and in each column is a matrix of rank 1. The sum of all entries equals M. The only problem is that the matrix $\tilde{p}_{ij} = S(\tilde{N}_{ij})$ is diagonal and hence it is not good. To overcome this obstacle, we replace the matrix $\tilde{P}$ with a good "almost diagonal" matrix P using Lemma 3, and then modify matrices N_{ij} to satisfy $p_{ij} = S(N_{ij})$.

Pick an arbitrary $i^* \in I$ an $j^* \in J$ (then $a^{(t)}_{i^*j^*} > 0$ for every t), fix a positive $\varepsilon < \min_t a^{(t)}_{i^*j^*}$, and let $\mathcal{E} = E_{i^*j^*}$ (as in Lemma 5, $E_{i^*j^*}$ has 1 in the (i^*, j^*)-position, and 0 elsewhere else). By Lemma 3, there exist two chain independent random variables α and β such that $\Pr(\alpha = t) = \Pr(\beta = t) = S(A^{(t)})$, $\Pr(\alpha \neq \beta) \leq \varepsilon$ and the joint probability matrix of α and β is symmetric. Denote this matrix by P.

Now define matrices N_{ij}: $N_{ii} = A^{(i)} - \sum_{j\neq i} p_{ij} \cdot \mathcal{E}$ and $N_{ij} = p_{ij} \cdot \mathcal{E}$ (for $i \neq j$). Observe that $S(N_{ii}) = S(A^{(i)}) - \sum_{j\neq i} p_{ij} = \Pr(\alpha = i) - \Pr(\alpha = i, \beta \neq i) = \Pr(\alpha = i, \beta = i) \equiv p_{ii}$. Verify that matrix N satisfies all conditions of Lemma 4. All elements of each matrix N_{ij} are nonnegative, because elements of $A^{(i)}$ are nonnegative, elements of $\mathcal{E}$ are nonnegative, and i^*j^*-th element of N_{ii} equals

$$a^{(i)}_{i^*j^*} - \sum_{j:j\neq i^*} \Pr(\alpha = i^*,\ \beta = j) > \varepsilon - \Pr(\alpha \neq \beta) > 0.$$

The sum of matrices N_{ij} in row i equals $\left(A^{(i)} - \sum_{j\neq i} p_{ij}\mathcal{E}\right) + \sum_{j\neq i} p_{ij}\mathcal{E} = A^{(i)}$, which is a rank 1 matrix. Similarly, the sum all matrices in each column is a rank 1 matrix. The sum of all matrices N_{ij} equals M, and therefore M is good. ■

Our proof relies on the existence of a position i^*j^* such that all $a^{(t)}_{i^*j^*}$ are positive. Such i^*, j^* exist if M is an r-matrix, but not necessarily in the general case. To deal with the general case we need to define an *r-decomposition* of M.

Definition 6: An *r-decomposition* of a matrix is its expression as a (finite) sum of r-matrices $M = M_1 + M_2 + \ldots$ of the same size such that the supports of M_i and M_{i+1} intersect (for any i). The *length* of the decomposition is the number of the summands; the *r-complexity* of a matrix is the length of its shortest decomposition (or $+\infty$, if there is no such decomposition).

Lemma 7: Every non-block matrix M with nonnegative elements has an r-decomposition.

Proof: Consider a graph whose vertices are nonzero entries of M. Two vertices are connected by an edge if and only if they are in the same row or column. By assumption, the matrix is a non-block matrix, hence the graph is connected and there exists a (possibly non-simple) path $(i_1, j_1) \ldots (i_m, j_m)$ that visits each vertex of the graph at least once.

Express M as the sum of matrices corresponding to the edges of the path: each edge corresponds to a matrix whose support consists of the endpoints of the edge; each positive element of M is distributed among matrices corresponding to the adjacent edges. Each of these matrices is of rank 1. So the expression of M as the sum of these matrices is an r-decomposition.

■

Corollary 7.1: The r-complexity of any non-block matrix is finite.

Lemma 8: Any non-block matrix M is good.

Proof: The proof is by induction on the r-complexity of M. For matrices of r-complexity 1, we apply Lemma 6.

Suppose that M has r-complexity 2. In this case M is equal to the sum of some r-matrices A and B such that their supports are intersecting rectangles. We use the same idea as in the proof of Lemma 6. By Lemma 5, each of the matrices A and B is the sum of matrices of rank 1 with the same support. Suppose, for example, that $A = A^{(1)} + A^{(2)}$ and $B = B^{(1)} + B^{(2)}$. Take a matrix $\mathcal{E}$ with only one nonzero element that is located in the intersection of the supports of A and B. If this nonzero element is sufficiently small, then all the elements of the matrix N:

$$\begin{pmatrix} A^{(1)} - 3\mathcal{E} & \mathcal{E} & \mathcal{E} & \mathcal{E} \\ \mathcal{E} & A^{(2)} - 3\mathcal{E} & \mathcal{E} & \mathcal{E} \\ \mathcal{E} & \mathcal{E} & B^{(1)} - 3\mathcal{E} & \mathcal{E} \\ \mathcal{E} & \mathcal{E} & \mathcal{E} & B^{(2)} - 3\mathcal{E} \end{pmatrix}$$

are nonnegative matrices. The sum of the elements of each of the matrices that form the matrix N is positive. Hence the matrix $p_{ij} = S(N_{ij})$ is an r-matrix and is good. The sum of the elements in each row and in each column is of rank 1 since it is either $A^{(t)}$ or $B^{(t)}$. Using Lemma 4 we conclude that the matrix M is good.

Suppose now that the r-complexity of matrix M is n. Represent M as a sum of r-matrices: $M = A(1) + \cdots + A(n)$, such that the supports of $A(i)$ and $A(i+1)$ are intersecting rectangles. Then, using Lemma 5, write each $A(i)$ as a sum of rank 1 matrices: $A(i) = A(i,1) + \cdots + A(i,T)$. We may assume that T is the same for every i, since we can always increase T by replacing $A(i,t)$ with two matrices A' and A'': $A'(i,t) = A''(i,t) = A(i,t)/2$. For each i pick a matrix $\mathcal{E}_i$ with only one nonzero element that is located in the intersection of the supports of $A(i)$ and $A(i+1)$; we let this element be small enough (we set $\mathcal{E}_0 = \mathcal{E}_n = 0$). We now define matrix N. The indices of rows and columns of N are pairs (i,t). Let $N_{(i,t),(i,t)} = A(i,t) - (2T-1)\mathcal{E}_{i-1} - (2T-1)\mathcal{E}_i$, $N_{(i,s),(i,t)} = \mathcal{E}_{i-1} + \mathcal{E}_i$ (for $s \neq t$), $N_{(i,s),(i+1,t)} = N_{(i+1,t),(i,s)} = \mathcal{E}_i$ and $N_{(i,s),(j,t)} = 0$ if $|i-j| > 1$. For example, if $n = 4$ and $T = 1$, then matrix N is defined as follows:

$$\begin{pmatrix} A(1)-\mathcal{E}_1 & \mathcal{E}_1 & 0 & 0 \\ \mathcal{E}_1 & A(2)-\mathcal{E}_1-\mathcal{E}_2 & \mathcal{E}_2 & 0 \\ 0 & \mathcal{E}_2 & A(3)-\mathcal{E}_2-\mathcal{E}_3 & \mathcal{E}_3 \\ 0 & 0 & \mathcal{E}_3 & A(4)-\mathcal{E}_3 \end{pmatrix}.$$

The sum of all matrices in row (i,t) (or column (i,t)) equals $A(i,t)$ and hence is of rank 1. The sum of all matrices equals M. Finally, the matrix $p_{(i,s)(j,t)} = S(N_{(i,s)(j,t)})$ has r-complexity $n-1$, because the support of M is the union of $n-1$ rectangles of the form $(\{i, i+1\}, *) \times (\{i, i+1\}, *)$. By the induction hypothesis, P is a good matrix. Therefore, M is a good matrix (by Lemma 4). ■

This concludes the proof of Theorem 2: Random variables are chain independent if and only if their joint probability matrix is a non-block matrix.

Note that this proof is "constructive" in the following sense. Assume that the joint probability matrix for α, β is given and this matrix is not a block matrix. (For simplicity we assume that matrix elements are rational numbers, though this is not an important restriction.) Then we can effectively find k such that α and β are k-independent, and find the joint distribution of all random variables that appear in the definition of chain independence. (Probabilities for that distribution are not necessarily rational numbers, but we can provide algorithms that compute approximations with arbitrary precision.)

V. Improved version of Theorem 1

The inequality

$$H(\gamma) \le 2^k H(\gamma|\alpha) + 2^k H(\gamma|\beta)$$

from Theorem 1 can be improved. In this section we prove a stronger theorem.

Theorem 3: If random variables α and β are chain independent of order k, and γ is an arbitrary random variable, then

$$H(\gamma) \le 2^k H(\gamma|\alpha) + 2^k H(\gamma|\beta) - (2^{k+1}-1)H(\gamma|\alpha\beta),$$

or, in another form,

$$I(\gamma : \alpha\beta) \le 2^k I(\gamma : \alpha|\beta) + 2^k I(\gamma : \beta|\alpha).$$

Proof: The proof is by induction on k.

We use the following inequality:

$$\begin{aligned} H(\gamma) = H(\gamma|\alpha) + H(\gamma|\beta) + \\ I(\alpha : \beta) - I(\alpha : \beta|\gamma) - H(\gamma|\alpha\beta) \le \\ H(\gamma|\alpha) + H(\gamma|\beta) + I(\alpha : \beta) - H(\gamma|\alpha\beta). \end{aligned}$$

If α and β are independent then $I(\alpha : \beta) = 0$, we get the required inequality.

Assume that α and β are conditionally independent with respect to α' and β'; α' and β' are chain independent of order $k-1$.

We can assume without loss of generality that two random variables, the pair (α', β'), and γ are independent given (α, β). Indeed, consider random variables (α^*, β^*) defined by the following formula

$$\begin{aligned} \Pr(\alpha^* = c, \beta^* = d|\alpha = a, \beta = b, \gamma = g) = \\ \Pr(\alpha' = c, \beta' = d|\alpha = a, \beta = b). \end{aligned}$$

The distribution of $(\alpha, \beta, \alpha^*, \beta^*)$ is the same as the distribution of $(\alpha, \beta, \alpha', \beta')$, and (α^*, β^*) is independent from γ given (α, β).

From the "relativized" form of the inequality

$$H(\gamma) \le H(\gamma|\alpha) + H(\gamma|\beta) + I(\alpha : \beta) - H(\gamma|\alpha\beta)$$

(α' is added as a condition everywhere) it follows that

$$\begin{aligned} H(\gamma|\alpha') \le \\ H(\gamma|\alpha\alpha') + H(\gamma|\beta\alpha') + I(\alpha : \beta|\alpha') - H(\gamma|\alpha'\alpha\beta) \le \\ H(\gamma|\alpha) + H(\gamma|\beta) - H(\gamma|\alpha'\alpha\beta). \end{aligned}$$

Note that according to our assumption α' and γ are independent given α and β, so $H(\gamma|\alpha'\alpha\beta) = H(\gamma|\alpha\beta)$.

Using the upper bound for $H(\gamma|\alpha')$, the similar bound for $H(\gamma|\beta')$ and the induction assumption, we conclude that

$$\begin{aligned} H(\gamma) \le 2^k H(\gamma|\alpha) + 2^k H(\gamma|\beta) \\ - 2^k H(\gamma|\alpha\beta) - (2^k - 1)H(\gamma|\alpha'\beta'). \end{aligned}$$

Applying the inequality

$$H(\gamma|\alpha'\beta') \ge H(\gamma|\alpha'\beta'\alpha\beta) = H(\gamma|\alpha\beta),$$

we get the statement of the theorem. ∎

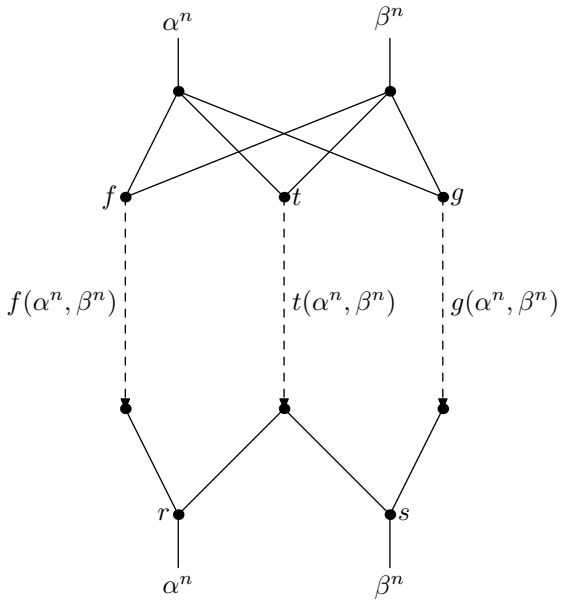

Fig. 1. Values of α^n and β^n are encoded by functions f, t and g and then transmitted via channels of limited capacity (dashed lines); decoder functions r and s have to reconstruct values α^n and β^n with high probability having access only to a part of transmitted information.

VI. Rate Regions

Definition 7: The rate region of a pair of random variables α, β is the set of triples of real numbers (u, v, w) such that for all $\varepsilon > 0$, $\delta > 0$ and sufficiently large n there exist

- "coding" functions t, f and g; their arguments are pairs (α^n, β^n); their values are binary strings of length $\lfloor (u + \delta)n \rfloor$, $\lfloor (v + \delta)n \rfloor$ and $\lfloor (w + \delta)n \rfloor$ (respectively).
- "decoding" functions r and s such that

$$r(t(\alpha^n, \beta^n), f(\alpha^n, \beta^n)) = \alpha^n$$

and

$$s(t(\alpha^n, \beta^n), g(\alpha^n, \beta^n)) = \beta^n$$

with probability more then $1 - \varepsilon$.

This definition (standard for multisource coding theory, see [3]) corresponds to the scheme of information transmission presented on Figure 1.

The following theorem was discovered by Vereshchagin. It gives a new constraint on the rate region when α and β are chain independent.

Theorem 4: Let α and β be chain independent random variables of order k. Then,

$$H(\alpha) + H(\beta) \le v + w + (2 - 2^{-k})u$$

for any triple (u, v, w) in the rate region.

(It is easy to see that $H(\alpha) \le u + v$ since α^n can be reconstructed with high probability from strings of length approximately nu and nv. For similar reasons we have $H(\beta) \le u + w$. Therefore,

$$H(\alpha) + H(\beta) \le v + w + 2u$$

for any α and β. Theorem 4 gives a stronger bound for the case when α and β are k-independent.)

Proof: Consider random variables

$$\gamma = t(\alpha^n, \beta^n), \xi = f(\alpha^n, \beta^n), \eta = g(\alpha^n, \beta^n)$$

from the definition of the rate region (for some fixed $\varepsilon > 0$). By Theorem 1, we have

$$H(\gamma) \le 2^k (H(\gamma|\alpha^n) + H(\gamma|\beta^n)).$$

We can rewrite this inequality as

$$2^{-k} H(\gamma) \le H((\gamma, \alpha^n)) + H((\gamma, \beta^n)) - H(\alpha^n) - H(\beta^n)$$

or

$$H(\xi) + H(\eta) + (2 - 2^{-k}) H(\gamma) \ge H(\xi) + H(\eta) + 2H(\gamma) - H((\gamma, \alpha^n)) - H((\gamma, \beta^n)) + H(\alpha^n) + H(\beta^n).$$

We will prove the following inequality

$$H(\xi) + H(\gamma) - H((\gamma, \alpha^n)) \ge -c\varepsilon n$$

for some constant c that does not depend on ε and for sufficiently large n. Using this inequality and the symmetric inequality

$$H(\eta) + H(\gamma) - H((\gamma, \beta^n)) \ge -c\varepsilon n$$

we conclude that

$$H(\xi) + H(\eta) + (2 - 2^{-k}) H(\gamma) \ge \ge H(\alpha^n) + H(\beta^n) - 2c\varepsilon n.$$

Recall that values of ξ are $(v + \delta)n$-bit strings; therefore $H(\xi) \le (v + \delta)n$. Using similar arguments for η and γ and recalling that $H(\alpha^n) = nH(\alpha)$ and $H(\beta^n) = nH(\beta)$ (independence) we conclude that

$$(v + \delta)n + (w + \delta)n + (2 - 2^{-k})(u + \delta)n \ge \ge nH(\alpha) + nH(\beta) - 2c\varepsilon n.$$

Dividing over n and recalling that ε and δ may be chosen arbitrarily small (according to the definition of the rate region), we get the statement of Theorem 4.

It remains to prove that

$$H(\xi) + H(\gamma) - H((\gamma, \alpha^n)) \ge -c\varepsilon n$$

for some c that does not depend on ε and for sufficiently large n. For that we need the following simple bound:

Lemma 9: Let μ and μ' be two random variables that coincide with probability $(1 - \varepsilon)$ where $\varepsilon < 1/2$. Then

$$H(\mu') \le H(\mu) + 1 + \varepsilon \log m$$

where m is the number of possible values of μ'.

Proof: Consider a new random variable σ with $m + 1$ values that is equal to μ' if $\mu \ne \mu'$ and takes a special value if $\mu = \mu'$. We can use at most $1 + \varepsilon \log m$ bits on average to encode σ ($\log m$ bits with probability ε, if $\mu \ne \mu'$, and one additional bit to distinguish between the cases $\mu = \mu'$ and $\mu \ne \mu'$). Therefore, $H(\sigma) \le 1 + \varepsilon \log m$. If we know the values of μ and σ, we can determine the value of μ', therefore

$$H(\mu') \le H(\mu) + H(\sigma) \le H(\mu) + 1 + \varepsilon \log m.$$

■

The statement of Lemma 9 remains true if μ' can be reconstructed from μ with probability at least $(1 - \varepsilon)$ (just replace μ with a function of μ).

Now recall that the pair (γ, α^n) can be reconstructed from ξ and γ (using the decoding function r) with probability $(1-\varepsilon)$. Therefore, $H((\gamma, \alpha^n))$ does not exceed $H((\xi, \gamma)) + 1 + c\varepsilon n$ (for some c and large enough n) because both γ and α^n have range of cardinality $O(1)^n$. It remains to note that $H((\xi, \gamma)) \le H(\xi) + H(\gamma)$. ■

ACKNOWLEDGEMENTS

We thank participants of the Kolmogorov seminar, and especially Alexander Shen and Nikolai Vereshchagin for the formulation of the problem, helpful discussions and comments.

REFERENCES

[1] R. Ahlswede, J. Körner, On the connection between the entropies of input and output distributions of discrete memoryless channels, *Proceedings of the 5th Brasov Conference on Probability Theory*, Brasov, 1974; *Editura Academiei*, Bucuresti, pp. 13–23, 1977.

[2] R. Ahlswede, J. Körner. On common information and related characteristics of correlated information sources. [Online]. Available: www.mathematik.uni-bielefeld.de/ahlswede/homepage.

[3] I. Csiszár, J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Second Edition, *Akadémiai Kiadó*, 1997

[4] P. Gács, J. Körner, Common information is far less than mutual information, *Problems of Control and Information Theory*, vol. 2(2), pp. 149–162, 1973.

[5] A. E. Romashchenko, Pairs of Words with Nonmaterializable Mutual Information, *Problems of Information Transmission*, vol. 36, no. 1, pp. 3–20, 2000.

[6] C. E. Shannon, A mathematical theory of communication. *Bell System Tech. J.*, vol. 27, pp. 379–423, pp. 623–656.

[7] H. S. Witsenhausen, On sequences of pairs of dependent random variables, *SIAM J. Appl. Math*, vol. 28, pp. 100–113, 1975

[8] A. D. Wyner, The Common Information of two Dependent Random Variables, *IEEE Trans. on Information Theory, IT-21*, pp. 163–179, 1975.