

Conditionally independent random variables

Konstantin Makarychev and Yuri Makarychev

Abstract—In this paper we investigate the notion of conditional independence and prove several information inequalities for conditionally independent random variables.

Keywords—Conditionally 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 β_1, \dots, β_n , *i.e.*, to 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

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

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 Ahlswede, Gács and Körner [1], [2], [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 “conditionally independent” random variables. Romashchenko introduced the notion of conditionally independent random variables and showed that extraction of common information from conditionally independent random variables is impossible. We prove that if the joint probability matrix of a pair of random variables (α, β) is

Princeton University

E-mail: {kmakaryc, ymakaryc}@princeton.edu

This work was done while the authors were at Moscow State University.

Supported by Russian Foundation for Basic Research grant 01-01-01028.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

not a block matrix, then α and β are conditionally independent. We also show several new information inequalities for conditionally independent random variables.

II. CONDITIONALLY 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 *conditionally independent of order 1*. (Conditionally independent random variables of order 0 are independent random variables.)

We consider conditional 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 conditionally independent random variables α_0 and β_0 (on another probability space), we say that α and β are conditionally independent.

Replacing the requirement of independence of α^* and β^* by the requirement of conditional independence of order 1, we get the definition of conditionally independent random variables (α and β) of order 2 and so on. (Conditionally independent variables of order k are also called k -conditionally independent in the sequel.)

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 conditionally independent random variables of order k ($k \geq 0$) 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 $0 \leq i < k$.
- (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 conditionally independent if they are conditionally independent of some order k .

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

Lemma 1: Consider conditionally independent random variables α and β of order k . Let $\alpha^n [\beta^n]$ be a sequence

of independent random variables each with the same distribution as α [β]. Then the variables α^n and β^n are conditionally independent of order k .

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

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

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 first indices $\{1, \dots, m\}$ can be divided into two disjoint nonempty sets I_1 and I_2 ($I_1 \sqcup I_2 = \{1, \dots, m\}$) and the set of its second indices $\{1, \dots, n\}$ can be divided into two sets J_1 and J_2 ($J_1 \sqcup J_2 = \{1, \dots, n\}$) 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 conditionally independent *iff* 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: If random variables α and β are conditionally 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 conditionally 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

$$\begin{aligned} H(\gamma|\alpha^*) &\leq H(\gamma|\alpha\alpha^*) + H(\gamma|\beta\alpha^*) + I(\alpha : \beta|\alpha^*) = \\ &H(\gamma|\alpha\alpha^*) + H(\gamma|\beta\alpha^*) \leq H(\gamma|\alpha) + H(\gamma|\beta). \end{aligned}$$

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 conditionally independent.

Proof: Suppose that the joint probability matrix A of random variables (α, β) is a block matrix and these random variables are conditionally 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. Consider a random variable γ with two values that is equal to the block number that contains (α, β) :

$$\begin{aligned} \gamma = 1 &\Leftrightarrow \alpha \in I_1 \Leftrightarrow \beta \in J_1; \\ \gamma = 2 &\Leftrightarrow \alpha \in I_2 \Leftrightarrow \beta \in J_2. \end{aligned}$$

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 conditional 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 conditionally 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 matrices of rank 1 are good.

(b) If α and β are conditionally independent, α' is a function of α and β' is a function of β , then α' and β' are conditionally 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 k -conditionally independent, then they are l -conditionally independent for any $l > k$. (We can add some constant random variables to the end of the derivation.)

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

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

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

Then the sequence

$$((\alpha_1^0, \alpha_2^0), (\beta_1^0, \beta_2^0)), \dots, ((\alpha_1^1, \alpha_2^1), (\beta_1^1, \beta_2^1))$$

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 (d) several times, we get Lemma 1.

Combining Lemma 1 and (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 conditionally independent. Then any random variables α' and β' , where α' depends only on $\alpha_1, \dots, \alpha_n$ and β' depends only on β_1, \dots, β_n , are conditionally independent.

Definition 4: Let us introduce the following notation:

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

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

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: (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 conditionally 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 conditional independence of D_ε tends to infinity as $\varepsilon \rightarrow 0$. Indeed, applying Theorem 1 to random variables α and β with joint distribution D_ε and to $\gamma = \alpha$, we obtain

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

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 conditionally independent variables of order k)

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

as $\varepsilon \rightarrow 0$.

Lemma 3: The set of good matrices is dense in the set of all joint probability matrices (*i.e.*, the set of $m \times n$ matrices with non-negative elements, whose sum is 1).

Proof: Any joint probability matrix A can be approximated as closely as desired by matrices with elements of the form $l/2^N$ for some N (where N is the same for all matrix elements).

Therefore, it suffices to prove that any joint probability matrix B with elements of the form $l/2^N$ can be approximated (as closely as desired) by good matrices. Take a pair of random variables (α, β) distributed according to D . The pair (α, β) can be represented as a function of N independent Bernoulli trials. The joint distribution matrix of each of these trials is D_0 and, by Lemma 2, can be approximated by a good matrix. Using statement (e), we get that (α, β) can also be approximated by a good matrix. Hence D can be approximated as closely as desired by good matrices. ■

Lemma 4: If $A = (a)_{ij}$ and $B = (b)_{ij}$ are stochastic matrices and M is a good matrix, then $A^T M B$ is a good matrix.

Proof: Consider a pair of random variables (α, β) distributed according to M . This pair of random variables is conditionally independent.

Roughly speaking, we define random variable α' [β'] as a transition from α [β] with transition matrix A [B]. The joint probability matrix of (α', β') is equal to $A^T M B$. But since the transitions are independent from α and β , the new random variables are conditionally independent.

More formally, let us randomly (independently from α and β) choose vectors \vec{c} and \vec{d} as follows

$$\Pr(\text{proj}_i(\vec{c}) = j) = a_{ij},$$

$$\Pr(\text{proj}_i(\vec{d}) = j) = b_{ij},$$

where proj_i is the projection onto the i -th component.

Define $\alpha' = \text{proj}_\alpha(\vec{c})$ and $\beta' = \text{proj}_\beta(\vec{d})$. Then

(i) the joint probability matrix of (α', β') is equal to $A^T M B$;

(ii) the pair (α, \vec{c}) is conditionally independent from the pair (β, \vec{d}) . Hence by statement (b), α' and β' are conditionally independent. ■

Now let us prove the following technical lemma.

Lemma 5: For any nonsingular $n \times n$ matrix M and a matrix $R = (r)_{ij}$ with the sum of its elements equal to 0, there exist matrices P and Q such that

1. $R = P^T M + M Q$;
2. the sum of all elements in each row of P is equal to 0;
3. the sum of all elements in each row of Q is equal to 0.

Proof: First, we assume that $M = I$ (here I is the identity matrix of the proper size), and find matrices P' and Q' such that

$$R = P'^T + Q'.$$

Let us define $P' = (p')_{ij}$ and $Q' = (q')_{ij}$ as follows:

$$q'_{ij} = \frac{1}{n} \sum_{k=1}^n r_{kj}.$$

Note that all rows of Q' are the same and equal to the average of rows of R .

$$P' = (R - Q')^T$$

It is easy to see that condition (1) holds. Condition (3) holds because the sum of all elements in any row of Q is equal to the sum of all elements of R divided by n , which is 0 by the condition. Condition (2) holds because

$$\sum_{j=1}^n p'_{ij} = \sum_{j=1}^n \left(r_{ji} - \frac{1}{n} \sum_{k=1}^n r_{ki} \right) = 0.$$

Now we consider the general case. Put $P = (M^{-1})^T P'$ and $Q = M^{-1} Q'$. Clearly (1) holds. Conditions (2) and (3) can be rewritten as $P\vec{u} = 0$ and $Q\vec{u} = 0$, where \vec{u} is the vector consisting of ones. But $P\vec{u} = (M^{-1})^T (P'\vec{u}) = 0$ and $Q\vec{u} = M^{-1} (Q'\vec{u}) = 0$. Hence (2) and (3) hold. ■

By altering the signs of P and Q we get Corollary 5.1.

Corollary 5.1: For any nonsingular matrix M and a matrix R with the sum of its elements equal to 0, there exist matrices P and Q such that

1. $R = -P^T M - M Q$;
2. the sum of all elements in each row of P is equal to 0;
3. the sum of all elements in each row of Q is equal to 0.

Lemma 6: Any nonsingular matrix M without zero elements is good.

Proof: Let M be a nonsingular $n \times n$ matrix without zero elements. By Lemma 4, it suffices to show that M can be represented as

$$M = A^T G B,$$

where G is a good matrix; A and B are stochastic matrices. In other words, we need to find invertible stochastic matrices A, B such that $(A^T)^{-1} M B^{-1}$ is a good matrix.

Let V be the affine space of all $n \times n$ matrices in which the sum of all the elements is equal to 1:

$$V = \{X : \sum_{i=1}^n \sum_{j=1}^n x_{ij} = 1\}.$$

(This space contains the set of all joint probability matrices.)

Let U be the affine space of all $n \times n$ matrices in which the sum of all elements in each row is equal to 1:

$$U = \{X : \sum_{j=1}^n x_{ij} = 1 \text{ for all } i\}.$$

(This space contains the set of stochastic matrices.)

Let \tilde{U} be a neighborhood of I in U such that all matrices from this neighborhood are invertible. Define a mapping $\psi : \tilde{U} \times \tilde{U} \rightarrow V$ as follows:

$$\psi(A, B) = (A^T)^{-1} M B^{-1}.$$

Let us show that the differential of this mapping at the point $A = B = I$ is a surjective mapping from $T_{(I,I)} \tilde{U} \times \tilde{U}$ (the tangent space of $\tilde{U} \times \tilde{U}$ at the point (I, I)) to $T_M V$ (the tangent space of V at the point M). Differentiate ψ at (I, I) :

$$d\psi|_{A=I, B=I} = d((A^T)^{-1} M B^{-1}) = -(dA)^T M - M dB.$$

We need to show that for any matrix $R \in T_M V$, there exist matrices $(P, Q) \in T_{(I,I)} \tilde{U} \times \tilde{U}$ such that

$$R = -P^T M - M Q.$$

But this is guaranteed by Corollary 5.1.

Since the mapping φ has a surjective differential at (I, I) , it has a surjective differential in some neighborhood N_1 of (I, I) in $\tilde{U} \times \tilde{U}$. Take a pair of stochastic matrices (A_0, B_0) from this neighborhood such that these matrices are interior points of the set of stochastic matrices.

Now take a small neighborhood N_2 of (A_0, B_0) from the intersection of N_1 and the set of stochastic matrices. Since the differential of φ at (A_0, B_0) is surjective, the image of N_2 has an interior point. Hence it contains a good matrix (recall that the set of good matrices is dense in the set of all joint probability matrices). In other words, $\psi(A_1, B_1) = (A_1^T)^{-1} M B_1^{-1}$ is a good matrix for some pair of stochastic matrices $(A_1, B_1) \in N_2$. This finishes the proof. ■

Lemma 7: Any joint probability matrix without zero elements is a good matrix.

Proof: Suppose that $X = (\vec{v}_1, \dots, \vec{v}_n)$ is an $m \times n$ ($m > n$) matrix of rank n . It is equal to the product of a

nonsingular matrix and stochastic matrix:

$$X = (\vec{v}_1 - \vec{u}_1 - \dots - \vec{u}_{m-n}, \vec{v}_2, \dots, \vec{v}_n, \vec{u}_1, \dots, \vec{u}_{m-n}) \times \\ \times \begin{pmatrix} & I & & \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix}$$

where $\vec{u}_1, \dots, \vec{u}_{m-n}$ are sufficiently small vectors with positive components that form a basis in \mathbb{R}^m together with $\vec{v}_1, \dots, \vec{v}_n$ (it is easy to see that such vectors do exist); vectors $\vec{u}_1, \dots, \vec{u}_{m-n}$ should be small enough to ensure that the vector $\vec{v}_1 - \vec{u}_1 - \dots - \vec{u}_{m-n}$ has positive elements.

The first factor is a nonsingular matrix with positive elements and hence is good. The second factor is a stochastic matrix, so the product is a good matrix.

Therefore, any matrix of full rank without zero elements is good. If a $m \times n$ matrix with positive elements does not have full rank, we can add (in a similar way) m linearly independent columns to get a matrix of full rank and then represent the given matrix as a product of a matrix of full rank and stochastic matrix. ■

We denote by $S(M)$ the sum of all elements of a matrix M .

Lemma 8: 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 conditionally 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) . ■

From Lemma 8 it follows that any 2×2 matrix of the form $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$ is good.¹ Indeed, let us apply Lemma 8 to the following matrix:

$$N = \left(\begin{array}{cc|cc} a & 0 & 0 & b/2 \\ 0 & 0 & 0 & 0 \\ \hline 0 & b/2 & 0 & 0 \\ 0 & 0 & 0 & c \end{array} \right).$$

The sum of matrices in each row and in each column is of rank 1. The sum of elements of each matrix N_{ij} is positive,

¹ a, b and c are positive numbers whose sum equals 1.

so (by Lemma 7) the matrix $p_{ij} = S(N_{ij})$ is a good matrix. Hence the sum of matrices N_{ij} is good.

Recalling that a, b and c stand for any positive numbers whose sum is 1, we conclude that any 2×2 -matrix with 0 in the left bottom corner and positive elements elsewhere is a good matrix. Combining this result with the result of Lemma 7, we get that any non-block 2×2 matrix is good.

In the general case (we have to prove that any non-block matrix is good) the proof is more complicated.

We will use the following definitions:

Definition 5: The *support* of a matrix is the set of positions of its nonzero elements. An *r-matrix* is a matrix with nonnegative elements and with a “rectangular” support (i.e., with support $A \times B$ where $A[B]$ is some set of rows[columns]).

Lemma 9: Any r-matrix M is the sum of some r-matrices of rank 1 with the same support as M .

Proof: Denote the support of M by $N = A \times B$. Consider the basis E_{ij} in the vector space of matrices whose support is a subset of N . (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 N :

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

where $\vec{e}_1, \dots, \vec{e}_n$ 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 N :

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

Definition 6: An *r-decomposition* of a matrix is its expression as a (finite) sum of r-matrices $M = M_1 + M_2 + \dots$ 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 10: Any 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 iff 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 $(i_1, j_1) \dots (i_m, j_m)$ 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. \blacksquare

Corollary 10.1: The r -complexity of any non-block matrix is finite.

Lemma 11: Any non-block matrix M is good.

Proof: The proof uses induction on r -complexity of M . For matrices of r -complexity 1, we apply Lemma 7.

Now 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. By Lemma 9, each of the matrices A and B is the sum of matrices of rank 1 with the same support.

Suppose, for example, that $A = A_1 + A_2 + A_3$ and $B = B_1 + B_2$. Consider the block matrix

$$\begin{pmatrix} A_1 & 0 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 & 0 \\ 0 & 0 & A_3 & 0 & 0 \\ 0 & 0 & 0 & B_1 & 0 \\ 0 & 0 & 0 & 0 & B_2 \end{pmatrix}.$$

The sum of the matrices in each row and in each column is a matrix of rank 1. The sum of all the entries is equal to $A + B$. All the conditions of Lemma 8 but one hold. The only problem is that the matrix p_{ij} is diagonal and hence is not good, where p_{ij} is the sum of the elements of the matrix in the (i, j) -th entry (see Lemma 8). To overcome this obstacle take a matrix 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

$$N = \begin{pmatrix} A_1 - 4e & e & e & e & e \\ e & A_2 - 4e & e & e & e \\ e & e & A_3 - 4e & e & e \\ e & e & e & B_1 - 4e & e \\ e & e & e & e & B_2 - 4e \end{pmatrix}$$

are nonnegative matrices. The sum of the elements of each of the matrices that form the matrix N is positive. And the sum of the elements in any row and in any column is not changed, so it is of rank 1. Using Lemma 8 we conclude that the matrix M is good.

The proof for matrices of r -complexity 3 is similar. For simplicity, consider the case where a matrix of complexity 3 has an r -decomposition $M = A + B + C$, where A, B, C are r -matrices of rank 1. Let e_1 be a matrix with one positive element that belongs to the intersection of the supports of A and B (all other matrix elements are zeros), and e_2 be a matrix with a positive element in the intersection of the supports of B and C .

Now consider the block matrix

$$N = \begin{pmatrix} A - e_1 & e_1 & 0 \\ e_1 & B - e_1 - e_2 & e_2 \\ 0 & e_2 & C - e_2 \end{pmatrix}.$$

Clearly, the sums of the matrices in each row and in each column are of rank 1. The support of the matrix $(p)_{ij}$ is of the form

$$\begin{pmatrix} * & * & 0 \\ * & * & * \\ 0 & * & * \end{pmatrix};$$

and $(p)_{ij}$ has r -complexity 2.² By the inductive assumption any matrix of r -complexity 2 is good. Therefore, M is a good matrix (Lemma 8).

In the general case (any matrix of r -complexity 3) the reasoning is similar. Each of the matrices A, B, C is represented as the sum of some matrices of rank 1 (by Lemma 9). Then we need several entries e_1 (e_2) (as it was for matrices of r -complexity 2). In the same way, we prove the lemma for matrices of r -complexity 4 *etc.* \blacksquare

This concludes the proof of Theorem 2: Random variables are conditionally 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 k -conditional 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

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

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

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

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

or, in another form,

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

Proof: The proof is by induction on k .

We use the following inequality:

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

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

²Its support is the union of two intersecting rectangles, so the matrix is the sum of two r -matrices.

Assume that α and β are conditionally independent with respect to α' and β' ; α' and β' are conditionally 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

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

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

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

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

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

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

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

Applying the inequality

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

we get the statement of the theorem. \blacksquare

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 than $1 - \varepsilon$.

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 conditionally independent.

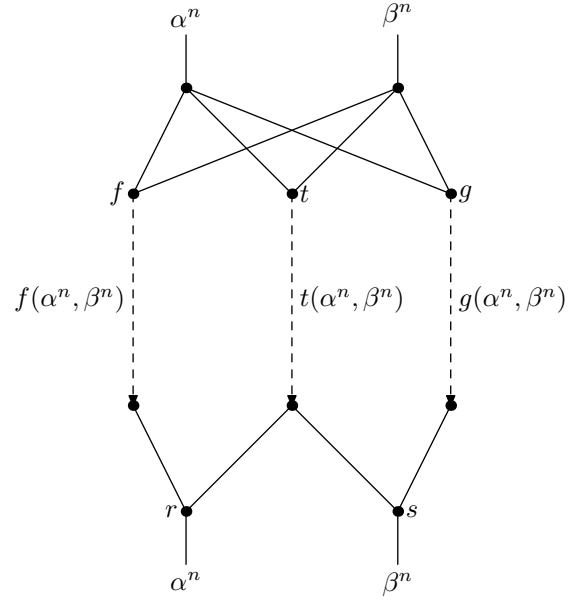


Fig. 1. Values of α^n and β^n are encoded by functions f, t and g and then transmitted via channels of limited capacity (dashed lines); decoder functions r and s have to reconstruct values α^n and β^n with high probability having access only to a part of transmitted information.

Theorem 4: Let α and β be k -conditionally independent random variables. Then,

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

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

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

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

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

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

We can rewrite this inequality as

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

or

$$\begin{aligned} H(\xi) + H(\eta) + (2 - 2^{-k})H(\gamma) &\geq H(\xi) + H(\eta) + \\ 2H(\gamma) - H((\gamma, \alpha^n)) - H((\gamma, \beta^n)) &+ H(\alpha^n) + H(\beta^n). \end{aligned}$$

We will prove the following inequality

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

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

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

we conclude that

$$\begin{aligned} H(\xi) + H(\eta) + (2 - 2^{-k})H(\gamma) &\geq \\ &\geq H(\alpha^n) + H(\beta^n) - 2c\varepsilon n. \end{aligned}$$

Recall that values of ξ are $(v + \delta)n$ -bit strings; therefore $H(\xi) \leq (v + \delta)n$. 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

$$\begin{aligned} (v + \delta)n + (w + \delta)n + (2 - 2^{-k})(u + \delta)n &\geq \\ &\geq nH(\alpha) + nH(\beta) - 2c\varepsilon n. \end{aligned}$$

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

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

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

Lemma 12: Let μ and μ' be two random variables that coincide with probability $(1 - \varepsilon)$ where $\varepsilon < 1/2$. Then

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

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 $\mu \neq \mu'$ 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 $\mu \neq \mu'$, and one additional bit to distinguish between the cases $\mu = \mu'$ and $\mu \neq \mu'$). Therefore, $H(\sigma) \leq 1 + \varepsilon \log m$. If we know the values of μ and σ , we can determine the value of μ' , therefore

$$H(\mu') \leq H(\mu) + H(\sigma) \leq H(\mu) + 1 + \varepsilon \log m. \quad \blacksquare$$

The statement of Lemma 12 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 $H((\xi, \gamma)) \leq H(\xi) + H(\gamma)$. \blacksquare

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.

We wish to thank Emily Cavalcanti, Daniel J. Webre and the referees for useful comments and suggestions.

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/ahlswe/de/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.