On essentially conditional information inequalities

Tarik Kaced* Andrei Romashchenko†

June 1, 2011

Abstract

In 1997, Z. Zhang and R.W. Yeung found the first example of a conditional information inequality in four variables that is not "Shannon-type". This linear inequality for entropies is called conditional (or constraint) since it holds only under condition that some linear equations are satisfied for the involved entropies. Later, the same authors and other researchers discovered several unconditional information inequalities that do not follow from Shannon's inequalities for entropy.

In this paper we show that some non Shannon-type conditional inequalities are "essentially" conditional, i.e., they cannot be extended to any unconditional inequality. We prove one new essentially conditional information inequality for Shannon's entropy and discuss conditional information inequalities for Kolmogorov complexity.

1 Introduction

Let $(X_1, \ldots, X_n)$ be jointly distributed random variables on a finite domain. For this collection of random variables there are $2^n - 1$ non-empty subsets and for each subset we have a value of Shannon's entropy. We call this family of entropies the *entropy profile* of the distribution $(X_1, \ldots, X_n)$. Thus, to every n-tuple of jointly distributed random variables there corresponds its entropy profile which is a vector of values in $\mathbb{R}^{2^n-1}$. We say that a point in $\mathbb{R}^{2^n-1}$ is *constructible* if it is a vector of entropies for some distribution.

All constructible points satisfy different *information inequalities* that characterize the range of all entropies for X_i. The most known and understood are so-called *Shannon-type* inequalities, i.e., linear combinations of basic inequalities of type $I(U : V|W) \geq 0$, where U, V, W are any (possibly empty) subsets of the given family of random variables.

In 1998 Z. Zhang and R.W. Yeung proved the first example of an unconditional *non Shannon-type* information inequality, which was a linear inequality for entropies of (X_1, X_2, X_3, X_4) that cannot be represented as a combination of basic inequalities [5]. Since this seminal paper of Zhang and Yeung was

*LIF de Marseille, Univ. Aix–Marseille

†LIF de Marseille, CNRS & Univ. Aix–Marseille on leave from IITP of RAS, Moscow

published, many (in fact, *infinitely many*) non Shannon-type linear information inequalities were proven, see, e.g., [7, 8, 9, 12, 13]. Another very curious *piecewise linear* information inequality was proven in [15]. These new inequalities were applied in problems of network coding [14], secret sharing [16], etc. However, these inequalities and their 'physical meaning' are still not very well understood.

In this paper we discuss *conditional* (constraint) information inequalities. That is, we are interested in linear information inequalities that are true only given some linear constraint for entropies. Trivial examples of conditional inequalities can be easily derived from (unconditional) basic inequalities, e.g., if $H(X_1) = 0$ then $H(X_1, X_2) \leq H(X_2)$. However, some conditional inequalities cannot be obtained as a corollary of Shannon-type inequalities. The first example of a nontrivial conditional inequality was proven in [4] (even before the first example of an unconditional non Shannon-type inequality):

$$\begin{aligned} &\text{if } I(A:B) = I(A:B|C) = 0, \text{ then} \\ &\quad I(C:D) \leq I(C:D|A) + I(C:D|B) \end{aligned} \tag{1}$$

Another conditional inequality

$$\begin{aligned} &\text{if } I(A:B|C) = I(B:D|C) = 0, \text{ then} \\ &\quad I(C:D) \leq I(C:D|A) + I(C:D|B) + I(A:B) \end{aligned} \tag{2}$$

was proven by F. Matúš in [6].

In [7] it was conjectured that (1) can be extended to some unconditional inequality

$$\begin{aligned} I(C:D) \leq I(C:D|A) + I(C:D|B) + \\ + \kappa(I(A:B) + I(A:B|C)) \end{aligned} \tag{3}$$

(for some constant $\kappa > 0$). In this paper we prove that this conjecture is *wrong*: for any coefficient κ, inequality (3) is not true for some distributions. So, inequality (1) is "essentially conditional"; it cannot be extended to an unconditional information inequality. A similar statement can be proven for (2).

In this paper we also prove one new conditional linear inequality that cannot be extended to any unconditional inequality. So, now we have three examples of essentially conditional information inequality.

It is known that the class of unconditional linear information inequalities are the same for Shannon's entropy and for Kolmogorov complexity. The situation with conditional inequalities is more complicated: the known technique used to prove constraint information inequalities for Shannon's entropy cannot be directly adapted for Kolmogorov complexity. In fact, it is not even clear how to formulate Kolmogorov's version of constraint inequalities. However, we prove for Kolmogorov complexities some counterpart of inequality (1); this inequality holds only for some special tuples of words.

The paper is organized as follows. In Section 2 we use the technique from [4] and prove one new conditional information inequality. In Section 3 we prove that

this new inequality as well as (1) and (2) cannot be extended to any unconditional inequalities. In Section 4 we prove some version of conditional inequality for Kolmogorov complexities.

2 Nontrivial conditional information inequalities

The very first example of an inequality that does not follow from basic (Shannon type) inequalities was the following result of Z. Zhang and R. W. Yeung:

Theorem 1 (Zhang–Yeung, [4])**.** *For all random variables A, B, C, D, if $I(A : B|C) = I(A : B) = 0$ then*

$$I(C : D) \leq I(C : D|A) + I(C : D|B).$$

With the same technique F. Matúš proved another conditional inequality (2), see [6]. Using a similar method, we prove one new conditional inequality[1]:

Theorem 2. *For all random variables A, B, C, D if*

$$H(C|A, B) = I(A : B|C) = 0,$$

then $I(C : D) \leq I(C : D|A) + I(C : D|B) + I(A : B)$.

Proof. The argument consists of two steps: *enforcing conditional independence* and *elimination of conditional entropy*. Let us have a joint distribution of random variables A, B, C, D. The first trick of the argument is a special transformation of this distribution: we keep the same distribution of the triples (A, C, D) and (B, C, D) but make A and B independent conditional on (C, D). Intuitively it means that we first choose at random (using the old distribution) values of C and D; then given fixed values of C, D we independently choose at random A and B (the conditional distributions of A given (C, D) and B given (C, D) are the same as in the original distribution).

More formally, we construct a new distribution $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$. If $\mathrm{Prob}[A = a, B = b, C = c, D = d]$ is the original distribution, then the new distribution is defined as follows:

$$\mathrm{Prob}[\tilde{A} = a, \tilde{B} = b, \tilde{C} = c, \tilde{D} = d] = \frac{\mathrm{Prob}[A = a, C = c, D = d] \cdot \mathrm{Prob}[B = b, C = c, D = d]}{\mathrm{Prob}[C = c, D = d]}$$

(with the convention $\frac{0}{0} = 0$ for all values a, b, c, d of the four random variables). From the construction ($\tilde{A}$ and $\tilde{B}$ are independent given $\tilde{C}, \tilde{D}$) it follows that

$$H(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = H(\tilde{C}, \tilde{D}) + H(\tilde{A}|\tilde{C}, \tilde{D}) + H(\tilde{B}|\tilde{C}, \tilde{D})$$

[1] The statement of this theorem (theorem 4(b)) in the first version of this paper was wrong.

Since $(\tilde{A}, \tilde{C}, \tilde{D})$ and $(\tilde{B}, \tilde{C}, \tilde{D})$ have exactly the same distributions as the original (A, C, D) and (B, C, D) respectively, we have

$$H(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = H(C, D) + H(A|C, D) + H(B|C, D)$$

The same entropy can be bounded in another way:

$$H(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) \leq H(\tilde{D}) + H(\tilde{A}|\tilde{D}) + H(\tilde{B}|\tilde{D}) + H(\tilde{C}|\tilde{A}, \tilde{B})$$

Notice that the entropies $H(\tilde{D})$, $H(\tilde{A}|\tilde{D})$ and $H(\tilde{B}|\tilde{D})$ are equal to $H(D)$, $H(A|D)$ and $H(B|D)$ respectively (we again use the fact that $\tilde{A}, \tilde{D}$ and $\tilde{B}, \tilde{D}$ have the same distributions as A, D and B, D respectively in the original distribution). Thus, we get

$$\begin{aligned} H(C, D) + H(A|C, D) + H(B|C, D) \leq \\ H(D) + H(A|D) + H(B|D) + H(\tilde{C}|\tilde{A}, \tilde{B}) \end{aligned}$$

It remains to estimate the value $H(\tilde{C}|\tilde{A}, \tilde{B})$. We will show that it is zero (and this is the second trick used in the argument).

Here we will use the two conditions of the theorem. We say that some values a, c (respectively, b, c or a, b) are *compatible* if in the original distribution these values can appear together, i.e., $\mathrm{Prob}[A = a, C = c] > 0$ (respectively, $\mathrm{Prob}[B = b, C = c] > 0$ or $\mathrm{Prob}[A = a, B = b] > 0$). Since A and B are independent given C, if some values a and b (of A and B) are compatible with the same value c of C, then these a and b are compatible with each other.

In the new distribution $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ values of $\tilde{A}$ and $\tilde{B}$ are compatible with each other *only if* they are compatible with one and the same value of $\tilde{C}$; hence, these values must be also compatible with each other in the original distribution (A, B). Further, since $H(C|A, B) = 0$, for each pair of compatible values of A, B there exists only one value of C. Thus, for each pair of values $(\tilde{A}, \tilde{B})$ with probability 1 there exists only one value of $\tilde{C}$. In a word, in the new distribution $H(\tilde{C}|\tilde{A}, \tilde{B}) = 0$.

Summarizing our arguments, we get

$$\begin{aligned} H(C, D) + H(A|C, D) + H(B|C, D) \leq \\ H(D) + H(A|D) + H(B|D), \end{aligned}$$

which is equivalent to

$$I(C : D) \leq I(C : D|A) + I(C : D|B) + I(A : B).$$

□

The proof of Theorem 2 presented above is based implicitly on non-negativity of the Kullbak–Leibler divergence. The same idea can be presented in a slightly different form, with an explicit reference to the Kullbak–Leibler inequality. The argument is almost the same as the proof of the second part of Proposition 2.1 in[6]:

Second version of the proof of Theorem 2: Let $p[a,b,c,d]$ be a distribution of (A,B,C,D) such that $H(C|AB) = I(A:B|C) = 0$. With some abuse of notations for we denote projections of this distribution as

$$p[a,c,d] = \mathrm{Prob}[A=a, C=c, D=d],\ p[a,d] = \mathrm{Prob}[A=a, D=d],\ \text{etc.}$$

We construct two new distributions, $\tilde{p}[a,b,c,d] = \mathrm{Prob}[\tilde{A}=a, \tilde{B}=b, \tilde{C}=c, \tilde{D}=d]$, and $\hat{p}[a,b,c,d] = \mathrm{Prob}[\hat{A}=a, \hat{B}=b, \hat{C}=c, \hat{D}=d]$. We define them as follows:

$$\tilde{p}[a,b,c,d] = \frac{p[a,c,d]\cdot p[b,c,d]}{p[c,d]}$$

and

$$\hat{p}[a,b,c,d] = \begin{cases} \frac{p[a,d]\cdot p[b,d]}{p_D[d]}, & \text{if } p[a,b,c]>0, \\ 0, & \text{otherwise.} \end{cases}$$

Since $I(A:B|C)=0$, the condition $p[a,b,c]>0$ is true if and only if $p[a,c]>0$ and $p[b,c]>0$.

Then we use non-negativity of the Kullback–Leibler divergence:

$$0 \le D(\hat{p}||\tilde{p}) = \sum \frac{p[a,c,d]\cdot p[b,c,d]}{p[c,d]} \cdot \log \frac{p[a,c,d]\cdot p[b,c,d]\cdot p[d]}{p[c,d]\cdot p[a,d]\cdot p[b,d]}$$

(the sum over all values a,b,c,d such that $p[a,b,c]>0$). It follows immediately that

$$0 \le H(A,D) + H(B,D) + H(C,D) - H(A,C,D) - H(B,C,D) - H(D).$$

Now we add the values $I(B:C|A) = H(A,C) + H(B,C) - H(A,B,C) - H(C)$ and $H(C|A,B) = H(A,B,C) - H(A,B)$ to the right-hand side of the inequality (both these values are equal to 0 for our distribution). We obtain

$$0 \le I(C:D|A) + I(C:D|B) + I(A:B) - I(C:D),$$

and we are done. □

3 Conditional inequalities that cannot be extended to any unconditional inequalities

In [7] it was conjectured that the conditional inequality from Theorem 1 is a corollary of some *unconditional* information inequality (which was not discovered yet):

Conjecture 1 ([7])**.** *For some constant $\kappa > 0$ inequality (3) is true for all random variables A,B,C,D.*

Obviously, if such an inequality could be proven, it would imply the statement of Theorem 1. Similar conjectures could be formulated for (2) and the conditional inequality from Theorem 2. We prove that these conjectures are false, i.e., these three conditional inequalities cannot be converted into unconditional inequalities:

Theorem 3. *(a) For any κ the inequality (3) is not true for some distributions (A, B, C, D).*

(b) For any κ the inequality

$$I(C:D) \le I(C:D|A) + I(C:D|B) + I(A:B) + \\ + \kappa(I(A:B|C) + H(C|A,B)) \tag{4}$$

is not true for some distributions (A, B, C, D).

(c) For any κ the inequality

$$I(C:D) \le I(C:D|A) + I(C:D|B) + I(A:B) + \\ + \kappa(I(A:B|C) + H(B:D|C)) \tag{5}$$

is not true for some distributions (A, B, C, D). Thus, (2) cannot be extended to an unconditional inequality.

Proof. (a) For all $\varepsilon \in [0, 1]$ we us consider the following joint distribution of binary variables (A, B, C, D):

$$\begin{array}{rcl} \mathrm{Prob}[A=0,\ B=0,\ C=0,\ D=1] & = & (1-\varepsilon)/4, \\ \mathrm{Prob}[A=0,\ B=1,\ C=0,\ D=0] & = & (1-\varepsilon)/4, \\ \mathrm{Prob}[A=1,\ B=0,\ C=0,\ D=1] & = & (1-\varepsilon)/4, \\ \mathrm{Prob}[A=1,\ B=1,\ C=0,\ D=1] & = & (1-\varepsilon)/4, \\ \mathrm{Prob}[A=1,\ B=0,\ C=1,\ D=1] & = & \varepsilon. \end{array}$$

For each value of A and for each values of B, the value of at least one of variables C, D is uniquely determined: if $A = 0$ then $C = 0$; if $A = 1$ then $D = 1$; if $B = 0$ then $D = 1$; and if $B = 1$ then $C = 0$. Hence, $I(C : D|A) = I(C : D|B) = 0$. Also it is easy to see that $I(A : B|C) = 0$. Thus, if (3) is true, then $I(C : D) \le \kappa I(A : B)$.

Denote the right-hand and left-hand sides of this inequality by $L(\varepsilon) = I(C : D)$ and $R(\varepsilon) = \kappa I(A : B)$. Both functions $L(\varepsilon)$ and $R(\varepsilon)$ are continuous, and $L(0) = R(0) = 0$ (for $\varepsilon = 0$ both sides of the inequality are equal to 0). However the asymptotics of $L(\varepsilon)$ and $R(\varepsilon)$ as $\varepsilon \to 0$ are different: it is not hard to check that $L(\varepsilon) = \Theta(\varepsilon)$, but $R(\varepsilon) = O(\varepsilon^2)$. From (3) we have $\Theta(\varepsilon) \le O(\varepsilon^2)$, which is a contradiction.

(b) For every value of $\varepsilon \in [0, 1]$ we consider the following joint distribution of binary variables (A, B, C, D):

$$\begin{array}{rcl} \mathrm{Prob}[A=1,\ B=1,\ C=0,\ D=0] & = & 1/2-\varepsilon, \\ \mathrm{Prob}[A=0,\ B=1,\ C=1,\ D=0] & = & \varepsilon, \\ \mathrm{Prob}[A=1,\ B=0,\ C=1,\ D=0] & = & \varepsilon, \\ \mathrm{Prob}[A=0,\ B=0,\ C=1,\ D=1] & = & 1/2-\varepsilon. \end{array}$$

The argument is similar to the proof if (a). First, it is not hard to check that $I(C : D|A) = I(C : D|B) = H(C|AB) = 0$ for every ε. Second,

$$\begin{aligned} I(A:B) &= 1+(2-2/\ln 2)\varepsilon+2\varepsilon\log\varepsilon+O(\varepsilon^2),\\ I(C:D) &= 1+(4-2/\ln 2)\varepsilon+2\varepsilon\log\varepsilon+O(\varepsilon^2), \end{aligned}$$

so $I(A:B)$ and $I(C:D)$ both tend to 1 as $\varepsilon\to 0$, but their asymptotics are different. Similarly,

$$I(A:B|C)=O(\varepsilon^2).$$

It follows from (4) that

$$2\varepsilon+O(\varepsilon^2)\le O(\varepsilon^2)+O(\kappa\varepsilon^2),$$

and with any κ we get a contradiction for small enough ε.

(c) For the sake of contradiction we consider the following joint distribution of binary variables (A,B,C,D) for every value of $\varepsilon\in[0,1]$:

$$\begin{aligned} \mathrm{Prob}[A=0,\ B=0,\ C=0,\ D=0] &= 3\varepsilon,\\ \mathrm{Prob}[A=1,\ B=1,\ C=0,\ D=0] &= 1/3-\varepsilon,\\ \mathrm{Prob}[A=1,\ B=0,\ C=1,\ D=0] &= 1/3-\varepsilon,\\ \mathrm{Prob}[A=0,\ B=1,\ C=0,\ D=1] &= 1/3-\varepsilon. \end{aligned}$$

We substitute this distribution in (5) and obtain

$$I_0+O(\varepsilon)\le I_0+3\varepsilon\log\varepsilon+O(\varepsilon)+O(\kappa\varepsilon),$$

where I_0 is the mutual information between C and D for $\varepsilon=0$ (which is equal to the mutual information between A and B for $\varepsilon=0$). We get a contradiction as $\varepsilon\to 0$. □

The theorem above implies that the set of all linear information inequalities for 4-tuples must have rather complicated structure. Let us remind that a point $\vec{a}\in\mathbb{R}^{15}$ is called *constructible* is there exists a joint distribution (A,B,C,D) such that

$$\vec{a}=(H(A),H(B),\dots,H(A,B,C,D))$$

($\vec{a}$ consists of entropies of all non-empty tuples of random variables A,B,C,D). Further, a point $\vec{a}$ is called *asymptotically constructible* if for every $\varepsilon>0$ in ε-neighborhood of $\vec{a}$ there exists an constructible point $\vec{a}'$. In a similar way the set of (asymptotically) constructible points is defined for any number of random variables (in $\mathbb{R}^{2^n-1}$ for n-tuples of random variables). It is known (see, e.g., [5, 18]) that for every n the set of asymptotically constructible points representable by n-tuples of random variables make a closed convex cone in $\mathbb{R}^{2^n-1}$. The dual representation of this cone is the set of all linear information inequalities. We will show that for $n\ge 4$ the structure of this cone is not trivial.

From Theorem 3 we get a new proof of the result by F. Matúš [9]: the set of linear information inequalities for 4 random variables is not finitely generated.

Theorem 4 ([9])**.** *The cone of asymptotically constructible points for* 4 *random variables is not polyhedral (equivalently, the set of linear information inequalities for* 4*-tuples of random variables is not finitely generated).*

Proof. For the sake of contradiction we assume that the cone in $\mathbb{R}^{15}$ that consist of all asymptotically constructible points for 4 random variables (A, B, C, D) is polyhedral. The constraints $I(A : B) = I(A : B|C) = 0$ specify some face (of co-dimension 2) on the boundary of this polyhedron. The corresponding conditional inequality (from Theorem 1) specifies a non-degenerate linear functionals, which is non-negative on the corresponding faces. Technically, this functional is defined by the linear form $g = I(C : D|A) + I(C : D|B) - I(C : D)$, which is non-negative on this face of the cone. With the standard linear programming technique it can be proven that this functional g can be extended to a linear functional g' such that (a) g' is non-negative on the entire cone of asymptotically constructible points, and (b) g' coincides with g on the subspace of co-dimension 2 defined by the condition $I(A : B) = I(A : B|C) = 0$ (see Proposition 17 in [1]). In coordinates such a functional g' must have form

$$g' = I(C : D|A) + I(C : D|B) - I(C : D) \\ +d_1 I(A : B) + d_2 I(A : B|C).$$

(with some reals d_1 and d_2). It follows that $g' \geq 0$ for all constructible points, so we get (3) (where κ is equal to maximum of d_1 and d_2). This contradicts Theorem 3, and we are done. For the sake of being self-contained, we present a more detailed argument in Appendix. □

4 Constraint inequality for Kolmogorov complexity

Kolmogorov complexity of a finite binary string X is defined as the length of the shortest program that generates X; similarly, Kolmogorov complexity of a string X given another string Y is defined as the length of the shortest program that generates X given Y as an input. More formally, for any programming language L, Kolmogorov complexity $K_L(X|Y)$ is defined as

$$K_L(X|Y) = \min\{|p| \ : \ \text{program } p \text{ prints } X \text{ on input } Y\},$$

and unconditional complexity $K_L(X)$ is defined as complexity of X given the empty Y. The basic fact of Kolmogorov complexity theory is the invariance theorem: there exists a *universal* programming language U such that for any other language L we have $K_U(X|Y) \leq K_L(X|Y) + O(1)$ (the $O(1)$ depends on L but not on X and Y). We fix such a universal language U; in what follows we omit the subscript U and denote Kolmogorov complexity by $K(X)$, $K(X|Y)$. We refer the reader to an excellent book [10] for a survey of properties of Kolmogorov complexity.

Kolmogorov complexity was introduced in [2] as an algorithmic version of measure of information in an individual object. In some sense, properties of Kolmogorov complexity are quite similar to properties Shannon's entropy. For example, for the property of Shannon's entropy $H(A, B) = H(A) + H(B|A)$

there is a Kolmogorov's counterpart

$$K(A,B) = K(A) + K(B|A) + O(\log K(A,B)) \tag{6}$$

(the Kolmogorov–Levin theorem, [3]). This result justifies the definition of the mutual information, which is an algorithmic version of the standard Shannon's definition: the mutual information is defined as $I(A:B) := K(A) + K(B) - K(A,B)$, and the conditional mutual information is defined as

$$I(A:B|C) := K(A,C) + K(B,C) - K(A,B,C) - K(C).$$

From the Kolmogorov–Levin theorem it follows that $I(A:B)$ is equal to $K(A) - K(A|B)$, and the conditional mutual information $I(A:B|C)$ is equal to $K(A|C) - K(A|B,C)$ (all these equations hold only up to logarithmic terms).

In fact, we have a much more deep and general parallel between Shannon's and Kolmogorov's information theories; for every linear inequality for Shannon's entropy there exists a Kolmogorov's counterpart:

Theorem 5 ([11])**.** *For each family of coefficients $\{\lambda_W\}$ the inequality*

$$\sum_i \lambda_i H(\alpha_i) + \sum_{i<j} \lambda_{ij} H(\alpha_i, \alpha_j) + \ldots \geq 0$$

is true for every distribution $\{\alpha_i\}$ if and only if for some constant C the inequality

$$\sum_i \lambda_i K(a_i) + \sum_{i<j} \lambda_{ij} K(a_i, a_j) + \ldots C \log N \geq 0$$

is true for all tuples of strings $\{a_i\}$, $N = K(a_1, a_2, \ldots)$ (C does not depend on a_i).

Thus, the class of unconditional inequalities valid for Shannon's entropy coincides with the class of (unconditional) inequalities valid for Kolmogorov complexity. What about conditional inequalities?

In the framework of Kolmogorov complexity we cannot say that some information quantity *exactly* equals zero. Indeed, even the definition of Kolmogorov complexity makes sense only up to an additive term that depends on the choice of the universal programming language. Moreover, such a natural basic statement as the Kolmogorov–Levin theorem (6) holds only up to a logarithmic term. So, if we want to prove a sensible conditional inequality for Kolmogorov complexity, the linear constraints must be formulated with some reasonable precision. A natural version of Theorem 1 is the following conjecture:

Conjecture 2. *There exist functions $f(n)$ and $g(n)$ such that $f(n) = o(n)$ and $g(n) = o(n)$, and for all strings A, B, C, D satisfying $I(A:B|C) \leq f(N)$, $I(A:B) \leq f(N)$ it holds $I(C:D) \leq I(C:D|A) + I(C:D|B) + g(N)$ (where $N = K(A,B,C,D)$).*

There is no hope to prove Conjecture 2 with $f(n)$ and $g(n)$ of order $\Theta(\log n)$. Indeed, using a counterexample from the proof of Theorem 3(a), we can construct binary strings A, B, C, D such that the quantities $I(A : B|C)$, $I(A : B)$, $I(C : D|A)$, and $I(C : D|B)$ are bounded by $O(\log N)$, but $I(C : D) = \Omega(\sqrt{N \log N})$. However, even if Conjecture 2 is false in general, similar conditional inequalities (even with logarithmic precision) can be true for some special tuples A, B, C, D. In what follows we show how to prove such an inequality for one natural example of strings A, B, C (and any D).

Let $\mathbb{F}_n$ be the finite field of 2^n elements. We consider the affine plane over $\mathbb{F}_n$. Let C be random line in this plane, and A and B be two points incident to this line. To specify the triple $\langle A, B, C\rangle$ we need at most $4n + O(1)$ bits of information: a line in a plane can be specified by two parameters in $\mathbb{F}_n$; to specify each point in a given line we need additional n bits of information.

We take a triple of strings $\langle A, B, C\rangle$ as specified above with maximal possible Kolmogorov complexity, i.e., such that $K(A, B, C) = 4n + O(1)$ (it follows from a simple counting argument that such a triple exists; moreover, there are about $2^{4n+O(1)}$ such triples). For these A, B and C we can easily estimate all their Kolmogorov complexities:

$$\begin{aligned} &K(A),\ K(B),\ \text{and } K(C) \text{ are equal to } 2n + O(1),\\ &K(A, C) = 3n + O(1),\ K(B, C) = 3n + O(1),\\ &H(A, B) = 4n + O(1). \end{aligned}$$

For this triple of strings the quantities $I(A : B)$ and $I(A : B|C)$ are negligible (logarithmic). This condition is very similar to the condition on random variables A, B, C in Theorem 1. So, it is not very surprising that Kolmogorov's counterpart of Theorem 1 holds for these strings:

Proposition 1. *For the strings A, B, C defined above and for* all *strings D we have*

$$I(C : D) \le I(C : D|A) + I(C : D|B) + O(\log N),$$

where $N = K(A, B, C, D)$.

This statement can be proven by an argument similar to the proof of Theorem 2. Let us explain this argument in full detail.

Proof. We may identify C with a linear function $c_1x + c_2$ over $\mathbb{F}_n$, where c_1 and c_2 are elements of the field (since Kolmogorov complexity of C is large, it cannot be a *vertical* line on the plane). Further, the points A and B in this line can be represented as pairs $\langle a_1, a_2\rangle$ and $\langle b_1, b_2\rangle$ such that

$$c_1 \cdot a_1 + c_2 = a_2 \text{ and } c_1 \cdot b_1 + c_2 = b_2$$

(here a_i and b_i are also elements of $\mathbb{F}_n$). By assumption, complexity of the pair (A, B) is close to $4n$. It means that $A \ne B$; hence, $a_1 \ne b_1$. Let i be one of indexes such that the ith bits of a_1 and b_1 are different. W.l.o.g. we assume that the ith bit in a_1 is equal to 0 and the ith bit in b_1 is equal to 1.

Now we split the affine plane over $\mathbb{F}_n$ into two halves: P_0 will consist of all points (x, y) such that the ith bit of x is 0, and P_1 will consist of the points (x, y) such that the ith bit of x is 1. So, point $A = (a_1, a_2)$ belongs to P_0, and $B = (b_1, b_2)$ belongs to P_1.

Now we are going to variate the points A and B: we will substitute A and B by their 'clones' A' and B' so that the triples $\langle A', B', C\rangle$ remain "similar" to the initial one $\langle A, B, C\rangle$. More precisely, we say that A' is a *clone* of A if

- $A' = (a'_1, a'_2)$ is a point in line C, and $A' \in P_0$ (i.e., $c_1 \cdot a'_1 + c_2 = a'_2$, and the ith bit of a'_1 is equal to 0);
- complexities $K(A')$, $K(A', C)$, $K(A', D)$, and $K(A', C, D)$ are equal (up to an additive term $O(\log N)$) to the corresponding complexities $K(A)$, $K(A, C)$, $K(A, D)$, and $K(A, C, D)$.

Similarly, we say that B' is a *clone* of B if

- $B' = (b'_1, b'_2)$ is a point in line C, and $B' \in P_1$, and
- complexities $K(B')$, $K(B', C)$, $K(B', D)$, and $K(B', C, D)$ are equal (up to an additive term $O(\log N)$) to the corresponding complexities $K(B)$, $K(B, C)$, $K(B, D)$, and $K(B, C, D)$.

From a simple counting argument it follows that there exist $2^{K(A|C,D)-O(\log N)}$ different clones of A and $2^{K(B|C,D)-O(\log N)}$ clones of B (see, e.g., [11, Lemma 2] or [17, Lemmas 1–2]).

Let us take a pair of clones A' and B' with maximal complexity given (C, D). Then

$$\begin{aligned} K(A', B', C, D) = \\ K(C, D) + K(A'|CD) + K(B'|CD) + O(\log N) = \\ K(C, D) + K(A|C, D) + K(B|C, D) + O(\log N) \end{aligned}$$

On the other hand,

$$\begin{aligned} K(A', B', C, D) \le K(D) + \\ K(A'|D) + K(B'|D) + K(C|A', B') + O(\log N) \end{aligned}$$

By definition of clones, complexities $K(A'|D)$ and $K(B'|D)$ are equal (up to $O(\log N)$ term) to $K(A|D)$ and $K(B|D)$ respectively. Since A' and B' belong to P_0 and P_1 respectively, they cannot be equal to each other. Hence, A' and B' uniquely determine line C. So, we get

$$\begin{aligned} K(C, D) + K(A|CD) + K(B|CD) \le \\ K(D) + K(A|D) + K(B|D) + O(\log N), \end{aligned}$$

which is equivalent (by the Kolmogorov–Levin theorem) to

$$I(C : D) \le I(C : D|A) + I(C : D|B) + O(\log N).$$

□

5 Acknowledgement

This work was partially supported by EMC ANR-09-BLAN-0164-01 and NAFIT ANR-08-EMER-008-01 grants.

We thank anonymous referees of ISIT 2011 for useful comments that helped us to substantially rework the original manuscript.

References

[1] O. Hustad. Extension of Positive Linear Functionals. Mathematica Scandinavica, issue 11 (1962) pp. 63–78.

[2] A.N. Kolmogorov. Three approaches to the quantitative definition of information. Problems Inform. Transmission, 1(1):1–7, 1965.

[3] A.K. Zvonkin and L.A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. Russian Math. Surveys, 25(6):83–124, 1970.

[4] Z. Zhang and R. W. Yeung. A non-Shannon-type conditional information inequality. IEEE Transactions on Information Theory, 43(1997), pp. 1982–1986.

[5] Z. Zhang and R. W. Yeung. On Characterization of entropy function via information inequalities. IEEE Transactions on Information Theory, 44(1998), pp. 1440–1450.

[6] F. Matúš. Conditional independences among four random variables III: final conclusion. Combinatorics, Probability & Computing, 8 (1999), pp. 269–276.

[7] K. Makarychev, Yu. Makarychev, A. Romashchenko, N. Vereshchagin. A New Class of non-Shannon Type Inequalities for Entropies. Communications in Information and Systems. 2 (2002) No. 2, pp. 147–166.

[8] F. Matúš. Adhesivity of polymatroids. Discrete Mathematics 307, 2007, 2464–2477.

[9] F. Matúš. Infinitely many information inequalities. In IEEE ISIT 2007, pp. 41–44.

[10] M. Li and P. Vitányi. An introduction to Kolmogorov complexity and its applications, 3rd Edition, Springer-Verlag, 2007.

[11] D. Hammer, A. Romashchenko, A. Shen, and N. Vereshchagin. Inequalities for Shannon entropy and Kolmogorov complexity, Journal of Computer and Systems Sciences, 60(2000), pp. 442–464.

[12] R. Dougherty, C. Freiling, and K. Zeger. Six new non-Shannon information inequalities. In IEEE ISIT 2006, pp. 233–236, 2006.

[13] W. Xu, J. Wang, and J. Sun. A projection method for derivation of non-Shannon-type information inequalities. In IEEE ISIT 2008, pp. 2116–2120, 2008.

[14] T. Chan, A. Grant. Dualities Between Entropy Functions and Network Codes. IEEE transactions on information theory. 2008, vol. 54, no 10, pp. 4470–4487.

[15] F. Matúš. Piecewise linear conditional information inequality. IEEE Transactions Information Theory, 52 (2008), pp. 236–238.

[16] A. Beimel and I. Orlov. Secret Sharing and Non-Shannon Information Inequalities. In Proc. of the 6th Theory of Cryptography Conference, volume 5444 of Lecture Notes in Computer Science, pp. 539–557, Springer-Verlag, 2009.

[17] An. Muchnik, A. Romashchenko. Stability of properties of Kolmogorov complexity under relativization. Problems of information transmission. vol. 46, no. 1, 2010.

[18] R.W. Yeung. A first course in information theory. Norwell, MA/New York: Kluwer/Plenum, 2002.

6 Appendix

Here we present a detailed proof of theorem 4.

Proof. For the sake of contradiction we assume that the cone in $\mathbb{R}^{15}$ that consist of all asymptotically constructible vectors for 4 random variables (A, B, C, D) is polyhedral. That is, the set of asymptotically constructible points is the set of solutions for some finite system of linear inequalities

$$f_1 \geq 0, \ldots, f_s \geq 0,$$

where each f_j is a linear functional on $\mathbb{R}^{15}$.

The constraints $I(A : B) = I(A : B|C) = 0$ specify a face (of co-dimension 2) on the boundary of the cone. The corresponding conditional inequality (from Theorem 1) specifies a non-degenerate linear functional which is non-negative on the corresponding face. Technically, this functional is defined by the linear form

$$g = I(C : D|A) + I(C : D|B) - I(C : D)$$

Let us show that this linear functional can be extended to the entire space $\mathbb{R}^{15}$ (as $g' = g + \kappa(I(A : B) + I(A : B|C))$ so that the resulting linear form g' is nonnegative on the polyhedral cone. This will imply a contradiction with Theorem 3.

Now change the coordinate system. Instead of coordinates $(x_1, \ldots, x_{15})$ corresponding to the entropy values $(H(A), H(B), \ldots, H(A, B, C, D))$ we introduce another coordinate systems $(y_1, \ldots, y_{15})$ such $y_1 = I(A : B)$ and

$y_2 = I(A : B|C)$. The choice of $y_3, \ldots, y_{15}$ is not important, we only require that the linear transformation

$$G : (x_1, \ldots, x_{15}) \mapsto (y_1, \ldots, y_{15})$$

is not degenerate. For example, we can set $y_1 = I(A : B)$ and $y_2 = I(A : B|C)$, and for coordinates $y_3, \ldots, y_{15}$ keep the values of entropies

$$H(B), H(D), H(A, B) \ldots, H(A, B, C, D)$$

(i.e., we take all entropies except for only $H(A)$ and $H(C)$).

The conditional inequality from Theorem 1 can reformulate as follows: if $y_1 = y_2 = 0$ then $g \geq 0$ (linear functional g can be represented as a linear form $g = a_3 y_3 + \ldots + a_{15} y_{15}$). On the other hand, in the new coordinate system we can represent each functional f_j as

$$f_j = a_{j,1} y_1 + a_{j,2} y_2 + \ldots + a_{j,15} y_{15}$$

Restrictions of f_j onto the subspace $y_1 = y_2 = 0$ can be specified by 13 coordinates (instead of 15). Denote

$$f'_j = a_{j,3} y_3 + a_{j,4} y_4 + \ldots + a_{j,15} y_{15}$$

Now from Theorem 1 we get the following fact: for all points $\bar{y} = (y_3, \ldots, y_1 5)$ such that $f'_j(\bar{y}) \geq 0$ for $j = 1, \ldots, s$, the inequality $g(\bar{y}) \geq 0$ holds. It follows from Farkas' lemma that for some reals $c_j \geq 0$

$$g(\bar{y}) = c_1 f'_1(\bar{y}) + \ldots + c_s f'_s(\bar{y}).$$

From the definition of f'_j we get

$$g(\bar{y}) = c_1(f_1 - a_{j,1} y_1 - a_{j,2} y_2) + \ldots + c_s(f_s - a_{s,1} y_1 - a_{s,2} y_2).$$

This is an identity for linear forms, so their coordinate representation must be equal to each other. Hence, the forms with these coordinate representations are equal to each other entire $\mathbb{R}^{15}$. Coming back to the original system of coordinates, we obtain $I(C : D|A) + I(C : D|B) - I(C : D) + d_1 I(A : B) + d_2 I(A : B|C) = \sum c_j f_j$ with some constants d_1 and d_2. The right-hand side of this equation is non-negative on the entire cone of asymptotically constructible points since all f_j by definition are non-negative on this cone. Thus, we get the inequality

$$I(C : D) \leq I(C : D|A) + I(C : D|B) + d_1 I(A : B) + d_2 I(A : B|C),$$

which must be true for all distributions (A, B, C, D). This contradicts Theorem 1, and we are done. □