Do essentially conditional information inequalities have a physical meaning?

***Abstract*—We show that two essentially conditional linear information inequalities (including the Zhang–Yeung'97 conditional inequality) do not hold for asymptotically entropic points. This result raises the question of the "physical" meaning of these inequalities and the validity of their use in practice-oriented applications.**

I. Introduction

Following Pippenger [13] we can say that the most basic and general "laws of information theory" can be expressed in the language of information inequalities (inequalities which hold for the Shannon entropies of jointly distributed tuples of random variables for every distribution). The very first examples of information inequalities were proven (and used) in Shannon's seminal papers in the 1940s. Some of these inequalities have a clear intuitive meaning. For instance, the entropy of a pair of jointly distributed random variables a, b is not greater than the sum of the entropies of the marginal distributions, i.e., $H(a,b) \leq H(a) + H(b)$. In standard notations, this inequality means that the mutual information between a and b is non-negative, $I(a{:}b) \geq 0$; this inequality becomes an equality if and only if a and b are independent in the usual sense of probability theory. These properties have a very natural "physical" meaning: a pair cannot contain more "uncertainty" than the sum of "uncertainties" in both components. This basic statement can be easily explained, e.g., in term of standard coding theorems: the average length of an optimal code for a distribution (a, b) is not greater than the sum of the average lengths for two separate codes for a and b. Another classic information inequality $I(a{:}b|c) \geq 0$ is slightly more complicated from the mathematical point of view, but is also very natural and intuitive.

We believe that the success of Shannon's information theory in a myriad of applications (in engineering and natural sciences as well as in mathematics and computer science) is due to the intuitive simplicity and natural "physical" interpretations of the very basic properties of Shannon's entropy.

Formally, information inequalities are just a dual description of the set of all entropy profiles. That is, for every joint distribution of an n-tuple of random variables we have a vector of $2^n - 1$ ordered entropies (entropies of all random variables involved, entropies of all triples, of quadruples, etc). A vector in $\mathbb{R}^{2^n-1}$ is called *entropic* if it represents entropy values of some distribution. The fundamental (and probably very difficult) problem is to describe the set of entropic vectors for all n. It is known, [16], that for every n the closure of the set of all entropic vectors is a convex cone in $\mathbb{R}^{2^n-1}$. The points that belong to this closure are called *asymptotically entropic* or *asymptotically constructible* vectors, [11], say a.e. vectors for short. The class of all linear information inequalities is exactly the dual cone to the set of asymptotically entropic vectors. In [13] and [4] a natural question was raised: *What is the class of all universal information inequalities?* (Equivalently, how to describe the cone of a.e. vectors?) More specifically, does there exist any linear information inequality that cannot be represented as a combination of Shannon's basic inequality?

In 1998 Z. Zhang and R.W. Yeung came up with the first example of a *non-Shannon-type* information inequality [17]:

$$I(c{:}d) \leq 2I(c{:}d|a) + I(c{:}d|b) + I(a{:}b) + I(a{:}c|d) + I(a{:}d|c).$$

This unexpected result raised other challenging questions: *What does this inequality mean? How to understand it intuitively?* Although we still do not know a complete and comprehensive answer to the last questions, several interesting and meaningful interpretations of the inequality from [17] were found, see [15], [18]. In fact, this inequality is closely connected with *Ingleton's inequality* for ranks of linear spaces, [2], [5], [7], [10], [11].

Though the inequality from [17] was (partially) explained, there are a few related results which still have no satisfactory intuitive explanation. We mean other "universal laws of information theory", those that can be expressed as *conditional linear information inequalities* (linear inequalities for entropies which are true for distributions whose entropies satisfy some linear constraints). We do not give a general definition of a "conditional linear information inequality" since the entire list of all known nontrivial inequalities in this class is very short. In fact, we know only three nontrivial examples of such inequalities:

(1) [16]: if $I(a{:}b|c) = I(a{:}b) = 0$, then
$$I(c{:}d) \leq I(c{:}d|a) + I(c{:}d|b),$$
(2) [8]: if $I(a{:}b|c) = I(b{:}d|c) = 0$, then
$$I(c{:}d) \leq I(c{:}d|a) + I(c{:}d|b) + I(a{:}b),$$
(3) [6]: if $I(a{:}b|c) = H(c|a,b) = 0$, then
$$I(c{:}d) \leq I(c{:}d|a) + I(c{:}d|b) + I(a{:}b).$$

These three inequalities are much less understood than the (unconditional) non-Shannon-type information inequality from [17]. It is known that (1-3) are "essentially conditional", i.e., they cannot be extended to any unconditional inequalities, [6], e.g., for (1) this means that for any values of "Lagrange multipliers" λ_1, λ_2 the corresponding unconditional extension

$$I(c{:}d) \leq I(c{:}d|a) + I(c{:}d|b) + \lambda_1 I(a{:}b) + \lambda_2 I(a{:}b|c)$$

does not hold for some distributions (a, b, c, d). In other words, (1-3) make some very special kind of "information laws": they cannot be represented as "shades" of any unconditional inequalities on the subspace corresponding to their linear constraints.

Inequalities (1-3) remain quite mysterious, and we do not know any intuitive explanation of their meaning. In this paper we reveal some evidence of why these inequalities (at least (1) and (3)) *must* be very different from "unconditional" inequalities. We prove that (1) and (3) do not hold for a.e. points. So, these inequalities are, in some sense, similar to the nonlinear (piecewise linear) conditional information inequality from [9].

Together with [6], where (1–3) are proven to be essentially conditional, our result indicates that (1) and (3) are very fragile and non-robust properties. We cannot hope that similar inequalities hold when the constraints become soft. For instance, assuming that $I(a{:}b)$ and $I(a{:}b|c)$ are "very small" we cannot say that

$$I(c{:}d) \le I(c{:}d|a) + I(c{:}d|b)$$

holds also with only "a small error"; even a negligible deviation from the conditions in (1) can result in a dramatic effect $I(c{:}d) \gg I(c{:}d|a) + I(c{:}d|b)$.

Conditional information inequalities (in particular, inequality (2)) were used in [8] to describe conditional independences among several jointly distributed random variables. Conditional independence is known to have wide applications in statistical theory (including methods of parameter identification, causal inference, data selection mechanisms, etc.), see, e.g., surveys in [1], [14]. We are not aware of any direct or implicit *practical* applications of (1-3), but it would not be surprising to see such applications in the future. However our results indicate that these inequalities are non-robust and therefore might be misleading in practice-oriented applications.

So, (1) and (3) can be used only with the assumption that the corresponding independence conditions hold exactly, without any error. Can we assume that some a and b are *absolutely independent* (respectively, absolutely independent conditional on c) when we deal with objects in the real world? We do not try to answer this question. Our knowledge is not enough to say whether "essentially conditional" information inequalities are just an artifact of the definition of Shannon's entropy for discrete distributions, or they still make some "physical" meaning. But certainly these inequalities must be handled and applied with great caution.

The rest of the paper is organized as follows. We provide a different proof of why two conditional inequalities are essentially conditional. This new proof uses a simple algebraic example of random variables. We also show that (1) and (3) are not valid for a.e. vectors, leaving the question for (2) open.

II. Why "essentially conditional" : an algebraic counterexample

Consider the quadruple $(a, b, c, d)_q$ of geometric objects, resp. $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$, on the affine plane over the finite field $\mathbb{F}_q$ defined as follows :

- First choose a random non-vertical line $\mathcal{C}$ defined by the equation $y = c_0 + c_1 x$ (the coefficients c_0 and c_1 are independent random elements of the field);
- pick points $\mathcal{A}$ and $\mathcal{B}$ on $\mathcal{C}$ independently and uniformly at random (these points coincide with probability $1/q$);
- then pick a parabola $\mathcal{D}$ uniformly at random in the set of all non-degenerate parabolas $y = d_0 + d_1 x + d_2 x^2$ (where $d_0, d_1, d_2 \in \mathbb{F}_q, d_2 \neq 0$) that intersect $\mathcal{C}$ at $\mathcal{A}$ and $\mathcal{B}$; (if $\mathcal{A} = \mathcal{B}$ we require that $\mathcal{C}$ is a tangent line to $\mathcal{D}$). When $\mathcal{C}$ and $\mathcal{A}, \mathcal{B}$ are chosen, there exist $(q-1)$ different parabolas $\mathcal{D}$ meeting these conditions.

A typical quadruple is represented on figure II :

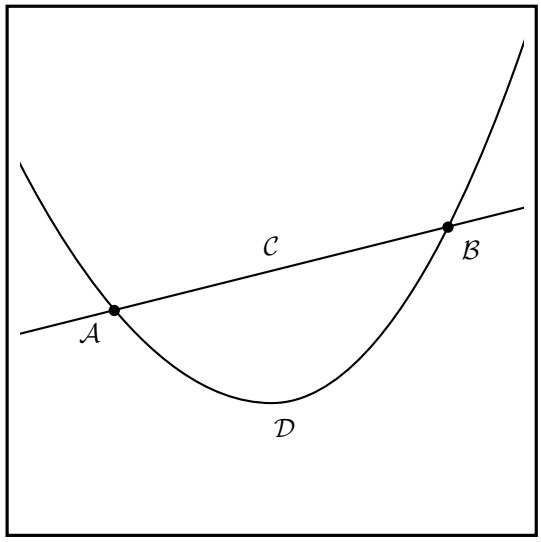

Fig. 1. An algebraic example

Remark: This picture is not strictly accurate, for the plane is discrete, but helps grasping the general idea since the relevant properties used are also valid in the continuous case.

Let us now describe the entropy vector of this quadruple.

- Every single random variable is uniform over its support.
- The line and the parabola share some mutual information, (the fact that they intersect) which is approximately one bit. Indeed, $\mathcal{C}$ and $\mathcal{D}$ intersect iff the corresponding equation discriminant is a quadratic residue, which happens almost half of the time.

$$I(c{:}d) = \frac{q-1}{q}$$

- When an intersection point is given, the line does not give more information about the parabola.

$$I(c{:}d|a) = I(c{:}d|b) = 0$$

- When the line is known, an intersection point does not help knowing the other (by construction).

$$I(a{:}b|c) = 0$$

- The probability that there is only one intersection point is $1/q$. In that case, the line can be any line going through

this point.

$$I(a{:}b) = H(c|a,b) = \frac{\log_2 q}{q}$$

Now we plug the computations into the following inequalities

$$I(c{:}d) \le I(c{:}d|a) + I(c{:}d|b) + \lambda_1 I(a{:}b) + \lambda_2 I(a{:}b|c)$$

or

$$I(c{:}d) \le I(c{:}d|a)+I(c{:}d|b)+I(a{:}b)+\lambda_1 I(a{:}b|c)+\lambda_2 I(c|a,b),$$

which are "unconditional" counterparts of (1) and (3) respectively. For any constants λ_1, λ_2 we get

$$1 - \frac{1}{q} \le O\left(\frac{\log_2 q}{q}\right)$$

and conclude they can not hold when q is large. Thus, we get the following theorem (originally proven in [6]):

Theorem 1. *Inequalities (1) and (3) are essentially conditional.*

III. Why (1) and (3) do not hold for a.e. vectors

We are going to use the previous example to show that conditional inequalities (1) and (3) are not valid for asymptotically entropic vectors. We will use the Slepian–Wolf coding theorem as our main tool.

Lemma 1 (Slepian–Wolf)**.** *Let (x, y) be joint random variables and (X, Y) be N independent copies of this distribution. Then there exists X' such that $H(X'|X) = 0$, $H(X') = H(X|Y) + o(N)$ and $H(X|X', Y) = o(N)$.*

This lemma constructs a hash of a random variable X which is almost independent of Y and has approximately the entropy of X given Y. We will say that X' is the Slepian–Wolf hash of X given Y and write $X' = SW(X|Y)$.

Theorem 2. *(1) and (3) are not valid for a.e. vectors.*

Proof: For each given inequality, we construct an asymptotically entropic vector which excludes it. The main step is to ensure, via Slepian–Wolf lemma, that the constraints are met.

a) Construction of a counterexample for inequality (1):

1. Start with the quadruple $(a, b, c, d)_q$ from the previous section for some fixed q to be defined later. Notice that it does not satisfy the constraints.
2. Serialize it: define a new quadruple (A, B, C, D) such that each entropy is N times greater. (A, B, C, D) is obtained by sampling N times independently (a_i, b_i, c_i, d_i) according to the distribution (a, b, c, d) and letting, e.g., $A = (a_1, a_2, \ldots, a_N)$.
3. Apply Slepian–Wolf lemma to get $A' = SW(A|B)$ such that $I(A'{:}B) = o(N)$, and replace A by A' in the quadruple. The entropy profile of (A', B, C, D) cannot vary much, at most by $I(A{:}B) + o(N) = O\left(\frac{\log_2 q}{q} N\right)$, from the entropy profile of the old quadruple. Notice that $I(A'{:}B|C) = 0$ since A' functionally depends on A and $I(a{:}b|c) = 0$.
4. Scale down the entropy vector of (A', B, C, D) by a factor of $1/N$. This standard operation can be done within a precision of, say, $o(N)$.
5. Tend N to infinity. This defines an a.e. vector for which inequality (1) does not hold when q is large. Indeed, for this a.e. vector, $I(A{:}B)$ and $I(A{:}B|C)$ both tend to zero as N approaches infinity. On the other hand, for the resulting limit of entropies (this limit is not an entropic but only asymptotically entropic point) inequality (1) turns into

$$1 + O\left(\frac{\log_2 q}{q}\right) \le O\left(\frac{\log_2 q}{q}\right),$$

which can not hold if q is bigger than some constant.

b) Construction of a counterexample for inequality (3): We start with another lemma based on the Slepian–Wolf coding theorem.

Lemma 2. *For every distribution (a, b, c, d) and every integer N there exists a distribution (A', B', C', D') such that*

- $H(C'|A', B') = o(N)$,
- *Denote $\vec{h}$ the entropy profile of (a, b, c, d) and $\vec{H}$ the entropy profile of (A', B', C', D'); then the components of $\vec{H}$ differ from the corresponding components of $N \cdot \vec{h}$ by at most $N \cdot H(c|a, b) + o(N)$.*

Proof of lemma: First we serialize (a, b, c, d). This means that we take N i.i.d. copies of the initial distribution. The result is a distribution (A, B, C, D) whose entropy profile is the one of (a, b, c, d) multiplied by N. Then we apply Lemma 1 and obtain a $Z = SW(C|A, B)$ such that

- $H(Z) = H(C|A, B) + o(N)$,
- $H(C|A, B, Z) = o(N)$.

Then, it is not hard to show that for most values ζ of the new random variable Z the corresponding conditional distribution (A', B', C', D') (the distribution on (A, B, C, D) with the condition $Z = \zeta$) satisfies all the required conditions.

Remark: This time we are not interested in the Slepian–Wolf hash as a random variable. We use its information content as an oracle which allows to perform a "relativization".

c) Rest of the proof for (3):

1. Start with the distribution $(a, b, c, d)_q$ for some q, to be fixed later, from the previous section.
2. Apply the "relativization" lemma 2 and get (A', B', C', D') such that $H(C'|A', B') = o(N)$. Lemma 2 guarantees that other entropies are about N times larger then the corresponding entropies for (a, b, c, d), possibly with an overhead of size

$$O(N \cdot H(c|a, b)) = O\left(\frac{\log_2 q}{q} N\right).$$

   Moreover, since the quadruple (a, b, c, d) satisfies $I(a{:}b|c) = 0$, we also have $I(A'{:}B'|C') = 0$ by construction of the random variables in Lemma 2.
3. Scale down the entropy vector of (A', B', C', D') by a factor of $1/N$ within a $o(N)$ precision.

4. Tend N to infinity to get an a.e. vector. Indeed, all entropies from the previous profile converge when N goes to infinity. Conditions of inequality (3) are satisfied for $I(A'{:}B'|C')$ and $H(C'|A',B')$ both vanish at the limit. Inequality (3) eventually reduces to

$$1 + O\left(\frac{\log_2 q}{q}\right) \le O\left(\frac{\log_2 q}{q}\right)$$

which can not hold if q is bigger than some constant.

■

Notice that in both cases, even one fixed value of q suffices to prove the result. The choice of the value of q provides some freedom in controlling the gap between the lhs and rhs of both inequalities.

In fact, we may combine the two above constructions into one to get a single a.e. vector to prove the previous result.

Proposition 1. *There exists one a.e. vector which excludes both (1) and (3) simultaneously.*

Proof sketch:

1. Generate (A,B,C,D) from $(a,b,c,d)_q$ with entropies N times greater.
2. Construct $A'' = SW(A|B)$ and $C' = SW(C|A,B)$ simultaneously (with the same serialization (A,B,C,D)).
3. Since A'' is a Slepian–Wolf hash of A given B, we have
   - $H(C|A'',B) = H(C|A,B) + o(N)$ and
   - $H(C|A'',B,C') = H(C|A,B,C') + o(N) = o(N)$.

   By inspecting the proof of the Slepian–Wolf theorem we conclude that A'' can be plugged into the argument of Lemma 2 instead of A.
4. The entropy profile of the constructed quadruple (A',B',C',D') is approximately N times the entropy profile of $(a,b,c,d)_q$ with a possible overhead of

$$O(I(A{:}B) + H(C|A,B)) + o(N) = O\left(\frac{\log_2 q}{q} N\right),$$

   and further :
   - $I(A'{:}B'|C') = 0$
   - $I(A'{:}B) = o(N)$
   - $H(C'|A',B') = o(N)$
5. Scale the corresponding entropy profile by a factor $1/N$ and tend N to infinity to define the desired a.e. vector.

IV. Conclusion

In this paper we discussed the known conditional information inequalities. We presented a simple algebraic example which provides a new proof that two conditional information inequalities are essentially conditional (they cannot be obtained as a direct corollary of any unconditional information inequality). Then, we prove a stronger result: two of the main three nontrivial linear conditional information inequalities are not valid for *asymptotically entropic* vectors.

This last result has a counterpart in the Kolmogorov complexity framework. It is known that unconditional linear information inequalities for Shannon's entropy can be directly translated into equivalent linear inequalities for Kolmogorov complexity, [3]. For conditional inequalities the things are more complicated. Inequalities (1) and (3) could be rephrased in the Kolmogorov complexity setting; but natural counterparts of these inequalities are not valid for Kolmogorov complexity. The proof of this fact is very similar to the argument in Theorem 2 (we need to use Muchik's theorem on conditional descriptions [12] instead of the Slepian–Wolf theorem employed in Shannon's framework). We skip details for the lack of space.

Open problems. Several natural questions remain open:

- Does (2) hold for a.e. vectors?
- Does there exist any other essentially conditional inequality that holds for a.e. vectors?
- Does there exist any essentially conditional linear inequality for Kolmogorov complexity?

Notice that any essentially conditional inequality for asymptotically entropic vectors would have an interesting geometric interpretation: it would means that the convex cone of a.e. vectors has smoothly rounded edges on some of its flat faces.

References

[1] A.P. Dawid, Conditional independence in statistical theory, Journal of the Royal Statistical Society. 41(1), 1979, pp. 1–31.

[2] R Dougherty, C. Freiling, and K. Zeger. Networks, Matroids, and Non-Shannon Information Inequalities. IEEE Transactions on Information Theory 53(6), June 2007, pp. 1949–1969.

[3] D. Hammer, A. Romashchenko, A. Shen, N. Vereshchagin. Inequalities for Shannon Entropy and Kolmogorov Complexity. Journal of Computer and System Sciences. 60 (2000) pp. 442–464.

[4] T.S. Han, A uniqueness of Shannon's information distance and related nonnegativity problems, J. Comb., Inform. Syst. Sci., vol. 6, pp. 320–321, 1981.

[5] A. W. Ingleton, Representation of Matroids in Combinatorial Matheatics and Its Applications, D. J. A. Welsh, Ed. London, U.K.: Academic, 1971, pp. 149–167.

[6] T. Kaced, A. Romashchenko. On essentially conditional information inequalities. IEEE ISIT 2011, pp. 1935–1939.

[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(2), 2002, pp. 147–166.

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

[9] F. Matúš, Piecewise linear conditional information inequality, IEEE Transactions on Information Theory, 2006, pp. 236–238.

[10] F. Matúš, Adhesivity of polymatroids, Discrete Math., 307, 2007, pp. 2464–2477.

[11] F. Matúš, Two constructions on limits of entropy functions. IEEE Transactions on Information Theory, 53(1), Jan. 2007, pp. 320–330.

[12] An. Muchnik, Conditional complexity and codes, Theoretical Computer Science, 271(1–2), 2002, pp. 97–109.

[13] N. Pippenger, What are the laws of information theory. 1986 Special Problems on Communication and Computation Conf., Palo Alto, CA.

[14] Judea Pearl, Causal inference in statistics: An overview. Statistics Surveys, 3, 2009, pp. 96–146.

[15] A. Romashchenko. Extracting the Mutual Information for a Triple of Binary Strings. Proc. 18th Annual IEEE Conference on Computational Complexity (2003). Aarhus, Denmark, July 2003, pp. 221–235.

[16] Z. Zhang and R. W. Yeung. A non-Shannon-type conditional information inequality. IEEE Trans. Inform. Theory, vol. 43, pp. 1982–1985, Nov. 1997.

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

[18] Z. Zhang. On a new non-Shannon-type information inequality, Communications in Information and Systems. 3(1), pp. 47–60, June 2003.