MULTI-VARIATE CORRELATION AND MIXTURES OF PRODUCT MEASURES

Tim Austin

Total correlation ('TC') and dual total correlation ('DTC') are two classical ways to quantify the correlation among an n-tuple of random variables. They both reduce to mutual information when $n = 2$.

The first part of this paper sets up the theory of TC and DTC for general random variables, not necessarily finite-valued. This generality has not been exposed in the literature before.

The second part considers the structural implications when a joint distribution μ has small TC or DTC. If $\mathrm{TC}(\mu) = o(n)$, then μ is close to a product measure according to a suitable transportation metric: this follows directly from Marton's classical transportation-entropy inequality. If $\mathrm{DTC}(\mu) = o(n)$, then the structural consequence is more complicated: μ is a mixture of a controlled number of terms, most of them close to product measures in the transportation metric. This is the main new result of the paper.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be an n-tuple of random variables. When $n = 2$, the mutual information $\mathrm{I}(X_1 \, ; \, X_2)$ is a canonical way to quantify the dependence between them. Once $n \geq 3$, mutual information can be generalized in several different ways, suitable for different purposes.

This paper focuses on two of these. If $X_1, \ldots, X_n$ are finite-valued, then their 'total correlation' ('TC') is

$$\mathrm{TC}(X_1 \, ; \, \ldots \, ; \, X_n) = \Big(\sum_{i=1}^{n} \mathrm{H}(X_i) \Big) - \mathrm{H}(X_1, \ldots, X_n),$$

and their 'dual total correlation' ('DTC') is

$$\mathrm{DTC}(X_1 \, ; \, \ldots \, ; \, X_n) := \mathrm{H}(X_1, \ldots, X_n) - \sum_{i=1}^{n} \mathrm{H}(X_i \,|\, X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

These definitions are discussed more carefully in Subsection 4.1, and extended to general random variables as suprema over quantizations.

Part I of the present paper builds up the basic theory of TC and DTC. The main results in this part are some identities and inequalities relating TC, DTC, Shannon entropy, and mutual information. These can mostly be seen as generalizations of the chain rule and monotonicity properties of mutual information. We take care to define and study TC and DTC for general random variables, not just finite-valued ones, and this introduces various additional technicalities. In handling these technicalities we extend foundational work of Kolmogorov, Dobrushin, Gelfand, Yaglom and Perez on mutual information.

The proofs in Part I are quite routine: a large fraction of the work boils down to applications of the chain rule. Many of these calculations have been done before. I include the references I know, but suspect that more lie buried in the literature.

Both TC and DTC are non-negative, and zero only in the case of independent random variables. One can therefore look for stability results for these quantities: does a small value of TC or DTC imply some 'structure' which is close to independence? This is the topic of Part II of the paper.

In our main results of this kind, 'closeness' of distributions is in the sense of transportation metrics. Suppose that each X_i takes values in a complete and separable metric space (K_i, d_{K_i}), and assume that each d_{K_i} has diameter at most 1. We endow the product $\prod_i K_i$ with the normalized Hamming average of these metrics:

$$d_n(x, y) := \frac{1}{n} \sum_{i=1}^{n} d_{K_i}(x_i, y_i) \qquad \left(x, y \in \prod_i K_i\right). \tag{1}$$

If each K_i is just a finite set endowed with the discrete metric, then d_n is the normalized Hamming metric on $\prod_i K_i$. This case already involves most of the ideas that we need. The case of other metrics on finite sets is easily reduced to this one, since then each d_{K_i} is bounded above by the discrete metric on K_i, and the results for the latter imply those for the former. However, if one of the spaces K_i is uncountable, then switching from d_{K_i} to the discrete metric sacrifices separability and introduces new Borel sets on which our measures may not be defined. For the sake of this case we formulate and prove our results for general metrics of diameter at most 1.

Using d_n, we endow the set $\mathrm{Prob}(\prod_i K_i)$ of all probability distributions with the transportation metric:

$$\overline{d_n}(\mu, \nu) := \inf_{\lambda} \int d_n(x, y) \, \lambda(\mathrm{d}x, \mathrm{d}y) \qquad \left(\mu, \nu \in \mathrm{Prob}\left(\prod_i K_i\right)\right), \tag{2}$$

where λ ranges over all couplings of μ and ν.

For TC, existing results can easily be re-cast as a structural conclusion of the desired kind: if TC $= o(n)$, then the joint distribution μ of $X_1, \ldots, X_n$ is close in $\overline{d_n}$ to a product measure. This is a corollary of Marton's classical transportation-entropy inequality [34, 35]: see Proposition 7.1 and Corollary 7.2 below.

For DTC the situation is more complicated. We always have

$$\mathrm{TC} \leq (n-1) \cdot \mathrm{DTC}$$

(see Lemma 4.13), so the result mentioned above for TC can be applied if DTC $= o(1)$. Outside this range, μ need not be close to a product: indeed, any mixture of k product measures always has DTC at most $\log k$ (see Proposition 8.1).

However, it turns out that this is roughly the only possibility over a much larger range of DTC-values: if DTC $= o(n)$, then μ is close in $\overline{d_n}$ to a 'low-complexity' mixture of product measures. This is the main result of the paper.

Theorem A. Fix a parameter $\delta > 0$ and let $\mu \in \text{Prob}(\prod_i K_i)$. If $\text{DTC}(\mu) \leq \delta^3 n$, then μ may be written as a mixture

$$\mu = \int_L \mu_y \, \nu(\mathrm{d}y)$$

so that

  (a) the mutual information in the mixture satisfies $\text{I}(\nu, \mu_\bullet) \leq \text{DTC}(\mu)$, and

  (b) there is a measurable family $(\xi_y : \ y \in L)$ of product measures on $\prod_i K_i$ such that

$$\int_L \overline{d_n}(\mu_y, \xi_y) \, \nu(\mathrm{d}y) < 2\delta.$$

In this statement, the 'mutual information in the mixture' $\text{I}(\nu, \mu_\bullet)$ is the mutual information between $(X_1, \ldots, X_n) \sim \mu$ and the mixing parameter $y \sim \nu$ when they are coupled by the kernel $(\mu_y : \ y \in L)$: see Subsection 3.4.

In the proof of Theorem A, the mixture representing μ is obtained by conditioning on a small subset S of the coordinates in $\{1, 2, \ldots, n\}$, resulting in the explicit choice $L := \prod_{i \in S} K_i$.

The bounds provided by Theorem A do not depend on the alphabets K_i at all. However, Theorem A has variants in which such a dependence does appear. For instance, Theorem 8.7 below gives a nontrivial conclusion in the range $\text{DTC}(\mu) \ll \delta^2 n$, so it is slightly less restrictive than Theorem A, but the alphabets K_i must be finite. It gives an alternative to part (a) of Theorem A which depends on the cardinalities $|K_i|$. See Subsection 8.3 for more discussion.

Because our main structural results give a description only up to approximations in $\overline{d_n}$, they are not completely satisfactory. For TC, the results described above give:

$$\text{product measure}$$
$$\implies \quad \text{TC} = 0 \quad \text{(standard calculation)}$$
$$\implies \quad \text{small TC} \quad \text{(trivially)}$$
$$\implies \quad \textit{near}\text{-product measure} \quad \text{(Corollary 7.2)},$$

where 'nearness' refers to an approximation in $\overline{d_n}$. This leaves open the possibility that some near-product measures have small TC while others have large TC. In fact we can say a little more about this gap, because if each (K_i, d_{K_i}) is a finite set with its discrete metric then TC enjoys some 'continuity' in the metric $\overline{d_n}$. However, the quality of this continuity deteriorates as the cardinalities $|K_i|$ increase: see Lemma 7.3. So, if each K_i is finite, then any measure sufficiently close to a product must have small TC, but the

necessary closeness here also depends on those cardinalities. This still leaves a gap in our understanding when some of the values $|K_i|$ are large or infinite, and indeed there are measures which occupy that gap and have large TC: see Example 7.5.

Similarly, for DTC we have

low-complexity mixture of products

$\implies$ small DTC   (Proposition 8.1)

$\implies$ low-complexity mixture of *near*-products   (Theorem A).

But it remains unclear when a mixture of near-products has small or large DTC. Unlike for TC, there seems to be no useful 'continuity' for DTC at all: Example 4.11 has transportation distance less than $1/n$ from a product measure, but has large DTC.

Relation to previous work

The study of measures of multi-variate correlation began with McGill's notion of 'interaction information' [38]. Since then, numerous other proposals have been studied in the information theory literature, and numerous identities relating them have been uncovered. TC was first studied by Watanabe in [47], and DTC by Han in [28]. Confusingly, both of these quantities and several of the others seem to have been rediscovered multiple times, and given a new name each time. Here we use Watanabe's and Han's original names.

Basic theoretical work in this area has been driven by a search for ways to identify different kinds of dependence structure among several random variables: for instance, disjoint subcollections of them exhibiting some conditional independence. See [39, 41] for some early analyses along these lines, and [45] for a more recent overview.

Much of the literature on notions of multi-variate correlation concerns their application in other branches of science. The paper [46] recalls several of these notions and discusses the practical matter of choosing one for the sake of interpreting different kinds of experimental data. The note [13] gives a quicker survey of the many options. Both of these references contain a more complete guide to the literature. A concrete example of an application of TC can be found in [15, Example 4], where it provides an upper bound on the secret key capacity in a certain network communication model of secrecy generation.

TC and DTC retain one basic property of mutual information which is especially relevant to our work: if X_i determines Y_i for each i, then

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) \geq \mathrm{TC}(Y_1 \,;\, \ldots \,;\, Y_n),$$

and similarly with DTC in place of TC. In particular, TC and DTC are both non-negative. This non-negativity is included in a more general family of inequalities discovered by Han [29]; see [12, Theorem 17.6.1] for a textbook treatment. Some other generalizations of mutual information, including McGill's interaction information and several others listed in [13, 46], lose this monotonicity property. They can vanish for quite highly correlated random variables, depending on the nature of the correlation.

However, TC and DTC are not the only non-negative measures of multi-variate correlation. On the contrary, they can be seen as the two simplest members of a large

family, all obtained as the gaps in different entropy inequalities, and all non-negative as a result. These include the other inequalities from [29], which in turn are all special cases of an inequality due to Shearer (also from about 1978, but first published in [9]). Fujishige interpreted Han's inequalities using polymatroids in [26], and explicitly investigated possible values for the gaps. Further refinements of Shearer's inequality have recently been investigated by Madiman and Tetali [32] and Balister and Bollobás [7]. We restrict attention to TC and DTC in most of this paper, but return briefly to those other inequalities in the final section.

Beyond monotonicity, our specific interest in TC and DTC stems from previous structural results about those quantities. The TC of a joint distribution μ is equal to the Kullback–Leibler divergence between μ and the product of its marginals (see equation (26) below). In this guise, it appears in Csiszár's paper [14] on generalizing Sanov's theorem and conditional limit theorems, and is implicitly at work in Marton's concentration inequality for product measures (see Section 7 below).

DTC has received less theoretical attention than TC. However, in the recent paper [5] it plays a crucial role in a new decomposition theorem for measures on product spaces, of a similar flavour to Theorem A but with a different conclusion.

In [5], a measure on a product space is decomposed into a mixture whose terms mostly satisfy a kind of concentration inequality called a T-inequality. Following the terminology in [5], a measure $\mu \in \text{Prob}(\prod_i K_i)$ satisfies $\text{T}(a, r)$ for some parameters $a, r > 0$ if we have

$$\int e^{af} \, d\mu \le \exp\left(a \int f \, d\mu + ar\right) \tag{3}$$

whenever $f : \prod_i K_i \longrightarrow \mathbb{R}$ is 1-Lipschitz for the metric d_n. (To be precise, the notation $\text{T}(a, r)$ is used in [5] for a certain transportation-entropy inequality, which is then proven equivalent to (3) via a version of the Bobkov–Götze equivalence: see [5, Subsection 5.3].)

The decomposition result in [5] is initially formulated for TC (see [5, Theorem B]) in order to meet the needs of an application to ergodic theory. But that version is derived from the analogous result for DTC [5, Theorem 7.1]. Here are those results:

Theorem 1.1. For any $\varepsilon, r > 0$ there exist $c, \kappa > 0$ for which the following holds. Any $\mu \in \text{Prob}(\prod_i K_i)$ can be written as a mixture

$$\mu = p_1 \mu_1 + \cdots + p_m \mu_m$$

so that

(a) $m \le c \cdot \exp(c \cdot \text{TC}(\mu))$,

(b) $p_1 < \varepsilon$, and

(c) the measure μ_j satisfies $\text{T}(\kappa n, r)$ for each $j = 2, 3, \ldots, m$.

The same conclusion holds with DTC in place of TC, except the dependence of c and κ on ε and r may be different.

The switch from TC to DTC is an essential idea in the proof of Theorem 1.1 in [5]. DTC exhibits a 'decrement' under a certain splitting operation on measures, and this

phenomenon is central to the proof of Theorem 1.1. It seems to have no analog for TC. See [5, Sections 6 and 7].

This ability of DTC to find the structure in Theorem 1.1 was one of the main discoveries in [5]. The present paper is largely motivated by that discovery, although the proof of Theorem A above is quite different from the arguments in [5], and much shorter. Aside from Theorem 1.1, the literature contains few stability or structural results based on DTC (another possible example is the main result of [24]). The present paper fills in some more of this picture.

Any product measure on $\prod_i K_i$ satisfies $\mathrm{T}(8rn, r)$ for all $r > 0$. This is essentially McDiarmid's inequality [37], and is equivalent to Marton's transportation-entropy inequality (Proposition 7.1 below) via the Bobkov–Götze equivalence. In addition, perturbations in $\overline{d_n}$ preserve T-inequalities up to some fiddly trimming and a slight deterioration in the constants (see [5, Proposition 5.9]), so near-product measures still satisfy fairly good T-inequalities up to that trimming. But many measures that are far from any product measure in $\overline{d_n}$ also satisfy good T-inequalities: see [5, Subsection 5.2] for examples and more discussion. For this reason, the T-inequalities promised by part (c) of Theorem 1.1 are strictly weaker than the 'near-product' structure obtained in part (b) of Theorem A.

On the other hand, Theorem 1.1 applies for any value of TC or DTC, and part (a) of that theorem gives a straightforward exponential bound in this parameter. These features are crucial for the application of Theorem 1.1 in [5]. By contrast, the results of the present paper give nothing at all unless TC or DTC is sufficiently small compared to n. When TC and DTC are too large for the results of the present paper to apply, Theorem 1.1 remains applicable and possibly valuable. A relevant example is given in [5, Example 5.4]. It is a measure on A^n for a large finite alphabet A which (i) has both TC and DTC of order n, but small compared to $n \log |A|$, (ii) already satisfies a good T-inequality, but (iii) cannot be written as a mixture of near-product measures with any meaningful control on the complexity of the mixture. Clearly Theorem A of the present paper cannot be extended to cover that example.

Let us describe the choice between these different structural results more quantitatively in case each K_i is finite, say of size at most k. For TC, Theorem 1.1 is non-trivial when $\mathrm{TC} = o(n \log k)$ — if TC is larger than this, then we might as well partition $\prod_i K_i$ into singletons. But the simpler result of Corollary 7.2 below takes over when $\mathrm{TC} = o(n)$. As a result, the TC-part of Theorem 1.1 is really valuable only when k is large. This was already remarked in [5] following the statement of [5, Theorem B].

The analogous discussion for DTC was left incomplete in [5]. Theorem A from the present paper completes it, up to approximations in $\overline{d_n}$. As above, Theorem 1.1 is non-trivial when $\mathrm{DTC} = o(n \log k)$, but now Theorem A takes over when $\mathrm{DTC} = o(n)$. As for TC, this means that the DTC-part of Theorem 1.1 is really valuable only when k is large.

These different structural results from [5] and the present paper are summarized in this table:

Range where nontrivial	Conclusion
$\mathrm{TC} = o(n)$	$\overline{d}_n$-close to a product
$\quad= o(n \log k)$	mixture with mut.inf. $= O(\mathrm{TC})$, most terms concentrated
$\mathrm{DTC} = o(1)$	$\overline{d}_n$-close to a product
$\quad= o(n)$	mixture with mut.inf. $= O(\mathrm{DTC})$, most terms $\overline{d}_n$-close to products
$\quad= o(n \log k)$	mixture with mut.inf. $= O(\mathrm{DTC})$, most terms concentrated

Relation to nonlinear large deviations

Several recent papers in probability have begun to develop a new theory of 'nonlinear large deviations', starting with Chatterjee and Dembo's work in [8]. This theory includes examples of measures for which Theorem A is nontrivial.

Let $(X_1, \ldots, X_n)$ be a tuple of independent random variables, say taking values in $K_1 \times \cdots \times K_n$ and having joint distribution $\lambda_1 \times \cdots \times \lambda_n$. This new theory considers functions $f(X_1, \ldots, X_n)$ which are more complicated than sums of functions of the individual coordinates (this is the meaning of 'nonlinear'), and seeks conditions on f under which one can still obtain good estimates on probabilities of large deviations of $f(X_1, \ldots, X_n)$ from its mean. Problems of this kind occur naturally in the study of triangle- or other subgraph-counts in Erdős–Rényi random graphs, and in questions of a similar flavour about arithmetic progressions in random arithmetic sets.

After suitably modifying the function f, this large deviations problem can be translated into that of estimating the partition function $Z = \int \mathrm{e}^f \, \mathrm{d}(\lambda_1 \times \cdots \times \lambda_n)$ that normalizes the Gibbs measure

$$\mu(\mathrm{d}x) := \frac{\mathrm{e}^{f(x)}}{Z} \lambda_1(\mathrm{d}x_1) \cdots \lambda_n(\mathrm{d}x_n).$$

In [8], Chatterjee and Dembo give a good approximation to Z when $K_i = \{0, 1\}$, λ_i is uniform for each i, and f satisfies a 'low complexity' assumption on its 'discrete gradients'. When $K_i = \{0, 1\}$ for each i, the discrete gradient $\nabla_x f$ at a point $x \in \{0, 1\}^n$ is the vector $(\partial_i f(x))_{i=1}^n$, where

$$\partial_i f(x) = f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n).$$

The function f is linear if and only if $\nabla_x f$ is the same vector for all x. Beyond this case, Chatterjee and Dembo consider functions f for which the set of discrete gradients $\{\nabla_x f : x \in \{0, 1\}^n\}$ can be covered by relatively few (intuitively, $\exp(o(n))$) small-radius balls in $\mathbb{R}^n$ — this is the meaning 'low complexity'. Chatterjee and Dembo's results are extended to more general product spaces in [48].

Following the insight that this 'low complexity' assumption enables good approximations to Z, Eldan showed that it also implies a relatively simple structure for the Gibbs measure μ. In [21] he considers the product space $\{-1, 1\}^n$ (technically more convenient than $\{0, 1\}^n$ for his proofs), assumes that $\{\nabla_x f : x \in \{-1, 1\}^n\}$ is small in a slightly stronger sense given by the notion of 'Gaussian width', and deduces that μ is a mixture

of near-product measures with relatively small mutual information in the mixture. His later papers [22, 23] with Gross give a more precise description of *which* products measures appear in the mixture, with even finer results in the setting of exponential random graph models.

So Eldan's structural conclusion is similar to Theorem A, and it is natural to wonder whether these 'low complexity' assumptions on f imply a bound on the DTC of the Gibbs measure μ. It turns out that they do, even under Chatterjee and Dembo's original and weaker covering-number assumption, and in a way that can easily be formulated over arbitrary product spaces. This is proved in the preprint [4],

The resulting bound on DTC implies a structural conclusion similar to Eldan's. But this implication is somewhat misleading. In fact, the bound on DTC obtained in the main result of [4] is a by-product of another, more direct proof that μ is a mixture of near-products with low mutual information. That more direct proof uses basic information theoretic principles such as Gibbs' variational principle and Marton's transportation-entropy inequality. It applies over arbitrary product spaces once one has the right general definition of discrete gradients. It gives better estimates on the structure of μ than one would obtain by taking the resulting bound on DTC and then applying Theorem A. So [4] actually takes very little from the present paper: in addition to the non-negativity of DTC (Han's inequality), it just needs the alternative formula for DTC that we prove below in part (b) of Proposition 5.1. But it does treat a large class of probabilistic models that illustrate the structure appearing in Theorem A.

Part I: Some basic theory of multi-variate correlation

2. BACKGROUND FROM PROBABILITY

2.1. Basic conventions

We assume standard results and notation from measure-theoretic probability. In Part II we restrict attention to standard Borel probability spaces, in order to use disintegrations freely. We often denote a measurable space by a single letter such as K. If it is standard Borel, then we denote its sigma-algebra by $\mathcal{B}_K$, and we write $\mathrm{Prob}(K)$ for the convex set of Borel probability measures on it. If K is a complete and separable metric space, then it is endowed with its Borel sigma-algebra by default.

If X is a random variable taking values in a measurable space K, and $\mathcal{P}$ is a finite measurable partition of K, then we write $[X]_\mathcal{P}$ for the **quantization** of X by $\mathcal{P}$: the $\mathcal{P}$-valued random variable defined by

$$[X]_\mathcal{P} = P \quad \Longleftrightarrow \quad X \in P.$$

Similarly, if μ is a probability measure on X, then $[\mu]_\mathcal{P}$ is the $\mathcal{P}$-indexed stochastic vector $(\mu(P): \ P \in \mathcal{P})$.

Whenever several random variables are under discussion simultaneously, they are defined on the same underlying probability space.

2.2. Kernels, mixtures and randomizations

Let $(\Omega, \mathcal{F})$ be a measurable space and K a standard Borel space. We assume familiarity with the notion of a kernel from Ω to K. We usually denote such a kernel by $\omega \mapsto \mu_\omega$ or something similar. Given such a kernel and also a probability measure P on Ω, the resulting **mixture** is the measure μ on K defined by

$$\mu(A) := \int_\Omega \mu_\omega(A) \, P(\mathrm{d}\omega) \quad \forall A \in \mathcal{B}_K. \tag{4}$$

If $\mu_\bullet$ and P are as above, then they also define a measure on the product space $(\Omega \times K, \mathcal{F} \otimes \mathcal{B}_K)$: the measure of $E \in \mathcal{F} \otimes \mathcal{B}_K$ is

$$\int_\Omega \mu_\omega \{x : (\omega, x) \in E\} \, P(\mathrm{d}\omega).$$

The correctness of this definition is a standard extension of the proof of Fubini's theorem: see, for instance, [19, Theorem 10.2.1(II)]. We denote this new measure by $P \ltimes \mu_\bullet$. It may also be described as the mixture obtained from the measure P and the $(\Omega \times K)$-valued kernel $\omega \mapsto \delta_\omega \times \mu_\omega$. The marginal of $P \ltimes \mu_\bullet$ on the K-coordinate is precisely the mixture (4).

Since K is standard Borel, any probability measure λ on $(\Omega \times K, \mathcal{F} \otimes \mathcal{B}_K)$ can be written as $P \ltimes \mu_\bullet$ in an essentially unique way. In this representation, P is simply the marginal of λ on Ω, and then μ_ω is unique up to agreement for P-almost every ω. These assertions are the existence and uniqueness parts of the measure disintegration theorem: see, for instance, [19, Theorem 10.2.2].

When $\mu \in \mathrm{Prob}(K)$ is represented by the mixture (4), a **randomization** of that mixture is any pair of random variables (Y, X) on some background probability space whose joint distribution is $P \ltimes \mu_\bullet$ (so Y takes values in Ω and X takes values in K). We invoke randomizations a few times below, because some information theoretic quantities are more easily described in terms of random variables, even though they depend only on the distributions of those random variables.

3. BACKGROUND FROM INFORMATION THEORY

3.1. Finite-valued random variables

Let X, Y and Z be finite-valued random variables on a common probability space. We assume familiarity with the Shannon entropy $\mathrm{H}(X)$ and mutual information

$$\mathrm{I}(X \,;\, Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y); \tag{5}$$

with their conditional versions $\mathrm{H}(X \,|\, Z)$ and $\mathrm{I}(X \,;\, Y \,|\, Z)$; and with the **chain rules** that these quantities satisfy:

$$\mathrm{H}(X, Z) := \mathrm{H}(Z) + \mathrm{H}(X \,|\, Z) \quad \text{and} \quad \mathrm{I}(X \,;\, Y, Z) = \mathrm{I}(X \,;\, Z) + \mathrm{I}(X \,;\, Y \,|\, Z).$$

See, for instance, [12, Chapter 2].

3.2. Extension to general random variables

The definitions of $H(X \mid Z)$ and $I(X \,;\, Y \mid Z)$ are easily extended to the case of an arbitrary random variable Z, provided X and Y are still finite-valued. For Shannon entropy, we have

$$H(X \mid Z) := \inf_{\mathcal{P}} H(X \mid [Z]_{\mathcal{P}}), \qquad (6)$$

where the infimum runs over all finite quantizations of Z. Then we define

$$I(X \,;\, Y \mid Z) := H(X \mid Z) - H(X \mid Y, Z) \qquad (7)$$

as before.

Mutual information can be defined for a general pair of random variables X and Y by quantizing:

$$I(X \,;\, Y) = \sup_{\mathcal{P}, \mathcal{Q}} I([X]_{\mathcal{P}} \,;\, [Y]_{\mathcal{Q}}), \qquad (8)$$

where the supremum runs over all finite quantizations of X and Y. This mutual information may still be finite for non-discrete random variables: indeed, it is zero whenever X and Y are independent.

The analysis of mutual information for general random variables began with Kolmogorov [30] and Dobrushin [17]. It is recounted carefully in Pinsker's classic book [42]. Some of their proofs require that the random variables take values in standard Borel spaces (see the translator's remarks to [42, Chapter 3]). One can avoid this assumption by working entirely through quantizations, following Gelfand, Kolmogorov, Yaglom and Perez [27, 40]. We include a quick account here for reference.

The infimum in (6) and supremum in (8) are monotone under refinement of the partitions $\mathcal{P}$ and $\mathcal{Q}$: non-increasing in (6) and non-decreasing in (8). This fact is crucial to the understanding of these quantities. It allows us to think of these infima and suprema as limits under refinement of partitions, and this point of view simplifies several proofs. We meet similar situations later when we study TC and DTC.

It is also useful to know that these infima and suprema can be restricted to special families of partitions.

The following terminology helps us to formulate these properties precisely.

Definition 3.1. For any measurable space K, a family $\mathfrak{P}$ of finite measurable partitions of K is

- **directed** if any two members of $\mathfrak{P}$ have a common refinement in $\mathfrak{P}$;

- **generating** if together the members of $\mathfrak{P}$ generate the whole sigma-algebra of K.

Examples 3.2.   1. The family of all finite measurable partitions of K is clearly directed and generating.

2. Suppose that K is a product $K_1 \times K_2$ of two other measurable spaces with the product sigma-algebra, and that $\mathfrak{P}_i$ is a family of finite partitions of K_i for $i = 1, 2$. Whenever $\mathcal{P}_i \in \mathfrak{P}_i$ for $i = 1, 2$, we can identify $\mathcal{P}_1 \times \mathcal{P}_2$ with the family

$$\{A \times B : \ A \in \mathcal{P}_1, \ B \in \mathcal{P}_2\},$$

which is a partition of K into measurable rectangles. If each $\mathfrak{P}_i$ is directed (respectively, generating), then the collection $\{\mathcal{P}_1 \times \mathcal{P}_2 : \mathcal{P}_i \in \mathfrak{P}_i\}$ is directed (respectively, generating) in K.

This example generalizes directly to larger finite products.

3. If the sigma-algebra of K is countably generated, then it has a generating filtration

$$\{\emptyset, K\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$$

consisting of finite algebras of subsets. Let $\mathcal{P}_i$ be the partition of K into the atoms of $\mathcal{F}_i$ for each i. Then the collection $\mathfrak{P} := \{\mathcal{P}_i : i \geq 1\}$ is directed and generating.

We can now formulate the notion of 'limit' that we need.

Definition 3.3. Let $\mathfrak{P}_1$, $\ldots$, $\mathfrak{P}_n$ be directed families of partitions of the measurable spaces K_1, $\ldots$, K_n, respectively, and let φ be a function from $\mathfrak{P}_1 \times \cdots \times \mathfrak{P}_n$ to the extended real line $[-\infty, \infty]$. Let $c \in [-\infty, \infty]$. We write

$$\lim_{\mathcal{P}_1 \in \mathfrak{P}_1, \mathcal{P}_2 \in \mathfrak{P}_2, \ldots, \mathcal{P}_n \in \mathfrak{P}_n} \varphi(\mathcal{P}_1, \ldots, \mathcal{P}_n) = c$$

to mean that, for every neighbourhood U of c, there exist $\mathcal{Q}_1 \in \mathfrak{P}_1$, $\ldots$, $\mathcal{Q}_n \in \mathfrak{P}_n$ such that

$$\varphi(\mathcal{P}_1, \ldots, \mathcal{P}_n) \in U$$

whenever $\mathcal{P}_i \in \mathfrak{P}_i$ and $\mathcal{P}_i$ refines $\mathcal{Q}_i$ for each i.

Since our families of partitions are directed under refinement, the function φ in Definition 3.3 may be regarded as an extended-real-valued net indexed by a directed family. Then Definition 3.3 is simply an instance of convergence of a net in the sense of Moore and Smith: see, for instance, [19, pp28–30]. We assume some basic properties of limits of nets in the sequel; the properties we need are proved in the same way as for sequences.

In the sequel, we sometimes abbreviate $\lim_{\mathcal{P} \in \mathfrak{P}}$ to $\lim_{\mathcal{P}}$ when the choice of directed family is clear from the context.

We may rewrite (6) and (8) as limits according to the following classical results of Dobrushin: see [42, Theorems 2.4.1 and 3.5.1].

Proposition 3.4. Suppose that X and Y take values in the measurable spaces K and L respectively, and let $\mathfrak{P}$ and $\mathfrak{Q}$ be any directed and generating families of partitions of K and L. In case K is a finite set, we have

$$\mathrm{H}(X \,|\, Y) = \lim_{\mathcal{Q} \in \mathfrak{Q}} \mathrm{H}(X \,|\, [Y]_\mathcal{Q}). \tag{9}$$

For any choice of K and L, we have

$$\mathrm{I}(X \,;\, Y) = \lim_{\mathcal{P} \in \mathfrak{P}, \mathcal{Q} \in \mathfrak{Q}} \mathrm{I}([X]_\mathcal{P} \,;\, [Y]_\mathcal{Q}). \tag{10}$$

To complete the generalization of basic information theory, we must define conditional mutual information in general. This is done by combining (7) and (8): if X, Y and Z are general random variables, then we set

$$I(X\,;\,Y\mid Z) := \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}}\mid Z), \qquad (11)$$

where $\mathcal{P}$ and $\mathcal{Q}$ are as in (8). For each fixed $\mathcal{P}$ and $\mathcal{Q}$ here, the right-hand side is defined by (7), which in turn is a difference of quantities defined by infima as in (6). In light of Proposition 3.4, we now recognize (11) as an iterated limit. Suppose that X, Y and Z take values in K, L and M respectively, and let $\mathfrak{P}$, $\mathfrak{Q}$ and $\mathfrak{R}$ be any directed and generating families for those respective spaces. Then Proposition 3.4 gives

$$I(X\,;\,Y\mid Z) = \lim_{\mathcal{P}\in\mathfrak{P},\mathcal{Q}\in\mathfrak{Q}}\lim_{\mathcal{R}\in\mathfrak{R}} I([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}}\mid[Z]_{\mathcal{R}}). \qquad (12)$$

In general, the order of the iterated limit in (12) is important, and one cannot write it simply as the joint limit over $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{R}$.

Example 3.5. Let $K := L := (\mathbb{Z}/2\mathbb{Z})^{\mathbb{N}}$ and let $M := K \times L$. Let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ take values in K and L respectively, and assume that all the X_is and Y_is are uniform and independent. Finally, let $Z := (X, Y)$.

Since Z determines both X and Y, the definition (11) gives that $I(X\,;\,Y\mid Z) = 0$. (Once we know Z, there is no information left for X and Y to share!)

However, now let $\mathcal{P}_i$ and $\mathcal{Q}_i$ be the partitions of K and L generated by $(X_j)_{j=1}^{2i}$ and $(Y_j)_{j=1}^{2i}$, respectively, and let $\mathcal{R}_i$ be the partition of M generated by the random variables X_j for $1 \leq j \leq i$ and $X_j + Y_j(\mathrm{mod}\ 2)$ for $1 \leq j \leq 2i$. Clearly the partitions $\mathcal{P}_i$ (respectively, $\mathcal{Q}_i$, $\mathcal{R}_i$) become finer as $i \longrightarrow \infty$ and generate $\mathcal{B}_K$ (respectively, $\mathcal{B}_L$, $\mathcal{B}_M$). But for each fixed i we have

$$I([X]_{\mathcal{P}_i}\,;\,[Y]_{\mathcal{Q}_i}\mid[Z]_{\mathcal{R}_i}) = \sum_{j=1}^{i} I(X_i\,;\,Y_i\mid X_i, X_i+Y_i) + \sum_{j=i+1}^{2i} I(X_i\,;\,Y_i\mid X_i+Y_i)$$
$$= 0 + i = i,$$

by the independence among all the X_is and Y_is. As $i \longrightarrow \infty$ this tends to ∞, even though $I(X\,;\,Y\mid Z) = 0$. $\triangleleft$

Mutual information still satisfies the chain rule for general random variables X, Y and Z. We give a quick proof below which illustrates the use of limits along families of partitions. A proof using conditional distributions (and hence requiring standard Borel spaces) goes back to [17]; see [42, equation (3.6.3)], and also the translator's notes to [42, Chapter 3].

Lemma 3.6. Any random variables X, Y and Z satisfy

$$I(X\,;\,Y, Z) = I(X\,;\,Z) + I(X\,;\,Y\mid Z).$$

P r o o f. This is a textbook result if X, Y and Z are all finite-valued, already quoted above.

In the general case, let X, Y and Z take values in K, L and M respectively. Let $\mathfrak{P}$, $\mathfrak{Q}$ and $\mathfrak{R}$ be the families of all finite measurable partitions of those respective spaces (see Examples 3.2, Item 1). Then two appeals to Proposition 3.4 give

$$\mathrm{I}(X\,;\,Z) = \lim_{\mathcal{P},\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Z]_{\mathcal{R}})$$

and

$$\mathrm{I}(X\,;\,Y,Z) = \lim_{\mathcal{P},\mathcal{Q},\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y,Z]_{\mathcal{Q}\times\mathcal{R}}) = \lim_{\mathcal{P},\mathcal{Q},\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}},[Z]_{\mathcal{R}}).$$

The second of these formulae uses the construction from Examples 3.2, Item 2: in quantizing (Y,Z), we may restrict attention to partitions of the form $\mathcal{Q}\times\mathcal{R}$, since these form a directed and generating family for $L\times M$.

Moreover, since mutual information is non-decreasing under refinement of quantizations, the joint limits above agree with any corresponding iterated limits. Therefore we also have

$$\mathrm{I}(X\,;\,Z) = \lim_{\mathcal{P}}\lim_{\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Z]_{\mathcal{R}}) \quad \text{and} \quad \mathrm{I}(X\,;\,Y,Z) = \lim_{\mathcal{P}}\lim_{\mathcal{Q}}\lim_{\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}},[Z]_{\mathcal{R}}).$$

Finally, formula (12) gives

$$\mathrm{I}(X\,;\,Y\,|\,Z) = \lim_{\mathcal{P},\mathcal{Q}}\lim_{\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}}\,|\,[Z]_{\mathcal{R}}) = \lim_{\mathcal{P}}\lim_{\mathcal{Q}}\lim_{\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}}\,|\,[Z]_{\mathcal{R}}).$$

The order of limits is important in this equality, but not in the previous ones. Once again, the second equality holds here because $\lim_{\mathcal{R}} \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}}\,|\,[Z]_{\mathcal{R}})$ is non-decreasing in $\mathcal{P}$ and $\mathcal{Q}$.

With these limit formulae in hand, we can apply the particular iterated limit $\lim_{\mathcal{P}}\lim_{\mathcal{Q}}\lim_{\mathcal{R}}$ to the finite-valued chain rule

$$\mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}},[Z]_{\mathcal{R}}) = \mathrm{I}([X]_{\mathcal{P}}\,;\,[Z]_{\mathcal{R}}) + \mathrm{I}([X]_{\mathcal{P}}\,;\,[Y]_{\mathcal{Q}}\,|\,[Z]_{\mathcal{R}}),$$

and we arrive at the chain rule in general. $\qquad\square$

3.3. Kullback–Leibler divergence

Let μ,ν be two probability measures on the same measurable space K. The Kullback–Leibler ('KL') divergence between ν and μ is defined to be $+\infty$ unless $\nu \ll \mu$, and in that case it is

$$\int \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\nu$$

(which may still be $+\infty$). This value is denoted by $\mathrm{D}(\nu\,\|\,\mu)$. It is non-negative, and zero if and only if $\nu = \mu$.

KL divergence arises naturally in information theory and large deviations as a comparison between two distributions: see, for instance, [12, Chapters 2 and 11] for the case of distributions on finite sets, and [42, Section 2.4] or [16, Appendix D.3] for the general case.

KL divergence is related to entropy and mutual information by various identities. In particular, suppose that X and Y are arbitrary random variables, say taking values in K and L, let $\lambda \in \mathrm{Prob}(K \times L)$ be the joint distribution of X and Y, and let μ and ν be the marginals of λ. Then

$$\mathrm{I}(X\,;Y) = \mathrm{D}(\lambda \,\|\, \mu \times \nu). \tag{13}$$

A classical result of Gelfand, Kolmogorov, Yaglom and Perez equates this with the definition in (8): see [42, Theorem 2.4.2] and the translator's second note to that chapter of [42]. Equation (13) itself appears as [42, equation (2.4.4)].

KL divergence plays a role in this paper through formulae for TC and DTC that generalize (13): see Section 5. The formula for TC obtained there is important during the proof of Theorem A.

Like entropy and mutual information, KL divergence satisfies a chain rule. We need the following special case of this in the sequel. Assume that K and L are standard Borel. Let λ be a probability measure on $K \times L$, let μ and ν be its marginals on K and L respectively, and let $\nu_\bullet$ be a kernel from K to L such that $\lambda = \mu \ltimes \nu_\bullet$. Also, let μ' and ν' be two other probability measures on K and L respectively. Then

$$\mathrm{D}(\lambda \,\|\, \mu' \times \nu') = \mathrm{D}(\mu \,\|\, \mu') + \int \mathrm{D}(\nu_x \,\|\, \nu')\,\mu(\mathrm{d}x). \tag{14}$$

A more general version of (14) decomposes $\mathrm{D}(\lambda \,\|\, \lambda')$ for any pair of probability measures λ, λ' on $K \times L$, but we do not need this below. For distributions on finite sets, the chain rule for KL divergence can be found in standard references such as [12, Theorem 2.5.3]. The proof of that special case already shows the essential calculation. Dobrushin's paper [17] seems to have been the first to treat the general case, which involves some extra analytic considerations. That generality can also be found in [42, equation (3.11.5)], where it is attributed to Kolmogorov, and in [16, Theorem D.13].

Later we make use of (14) through various special cases. If we let $\mu' = \mu$ in (14), then the first right-hand term vanishes, leaving

$$\mathrm{D}(\lambda \,\|\, \mu \times \nu') = \int \mathrm{D}(\nu_x \,\|\, \nu')\,\mu(\mathrm{d}x). \tag{15}$$

Now let us insert (15) back into (14), then do the same with the roles of the two coordinates reversed, and then apply (13):

$$\begin{aligned}
\mathrm{D}(\lambda \,\|\, \mu' \times \nu') &= \mathrm{D}(\mu \,\|\, \mu') + \mathrm{D}(\lambda \,\|\, \mu \times \nu') \\
&= \mathrm{D}(\mu \,\|\, \mu') + \mathrm{D}(\nu \,\|\, \nu') + \mathrm{D}(\lambda \,\|\, \mu \times \nu) \\
&= \mathrm{D}(\mu \,\|\, \mu') + \mathrm{D}(\nu \,\|\, \nu') + \mathrm{I}(X\,;Y).
\end{aligned} \tag{16}$$

In case $\mu = \mu'$ and $\mathrm{D}(\nu \,\|\, \nu') < \infty$, we may re-arrange (16) and make another substitution from (15):

$$\mathrm{I}(X\,;Y) = \mathrm{D}(\lambda \,\|\, \mu \times \nu') - \mathrm{D}(\nu \,\|\, \nu') = \int \mathrm{D}(\nu_x \,\|\, \nu')\,\mu(\mathrm{d}x) - \mathrm{D}(\nu \,\|\, \nu'). \tag{17}$$

Finally, if we also let $\nu' = \nu$, then this simplifies to

$$\mathrm{I}(X\,;Y) = \int \mathrm{D}(\nu_x \,\|\, \nu)\,\mu(\mathrm{d}x). \tag{18}$$

3.4. Mutual information and mixtures

If $P \ltimes \mu_\bullet$ is a probability measure on $\Omega \times K$, and if Z, Y have joint distribution $P \ltimes \mu_\bullet$, then we define the **mutual information** in P and $\mu_\bullet$ to be $\mathrm{I}(Z\,;\,Y)$. Clearly this quantity depends only on the joint distribution $P \ltimes \mu_\bullet$. We may also refer to the 'mutual information in the mixture' $\int \mu_\bullet \, \mathrm{d}P$, and denote it by $\mathrm{I}(P, \mu_\bullet)$. In the case of a finite mixture

$$\mu = p_1 \mu_1 + \cdots + p_m \mu_m$$

this mutual information is bounded by $\mathrm{H}(p_1, \dots, p_m)$, and hence by $\log m$.

4. TOTAL CORRELATION AND DUAL TOTAL CORRELATION

4.1. Definitions

Consider random variables $X_1, \dots, X_n$ defined on the same probability space and taking values in measurable spaces $K_1, \dots, K_n$. Like mutual information, TC and DTC are defined by an explicit formula in case the range spaces K_i are finite, and by a supremum over quantizations in the general case.

Definition 4.1. If $K_1, \dots, K_n$ are finite sets, then the **total correlation** ('**TC**') of $X_1, \dots, X_n$ is

$$\mathrm{TC}(X_1\,;\,\dots\,;\,X_n) := \left(\sum_{i=1}^{n} \mathrm{H}(X_i) \right) - \mathrm{H}(X_1, \dots, X_n).$$

In the general case, their total correlation is

$$\mathrm{TC}(X_1\,;\,\dots\,;\,X_n) := \sup_{\mathcal{P}_1, \dots, \mathcal{P}_n} \mathrm{TC}([X_1]_{\mathcal{P}_1}\,;\,\dots\,;\,[X_n]_{\mathcal{P}_n}), \tag{19}$$

where this supremum runs over all tuples of finite measurable partitions $\mathcal{P}_i$ of the spaces K_i. This supremum may equal $+\infty$.

For discrete random variables, total correlation was introduced by Watanabe in [47], and this is his terminology. It goes by various other names, such as 'multi-information' in [1, 6, 44].

In the sequel we frequently write $[n] := \{1, 2, \dots, n\}$ when $n \in \mathbb{N}$, and we generally abbreviate $[n] \setminus \{i\} =: [n] \setminus i$ when $i \in [n]$. We set $[0] := \emptyset$ to allow for some degenerate cases. Given any tuple $x = (x_1, \dots, x_n)$ and nonempty $S \subseteq [n]$, we write $x_S := (x_i)_{i \in S}$. We also write $X_S := (X_i)_{i \in S}$ for the random variables, and we let $X_\emptyset$ be an arbitrary choice of deterministic (hence 'trivial') random variable. This is so that expressions such as $\mathrm{I}(Y\,;\,X_\emptyset)$ still make sense: for example, this one simply equals zero.

Definition 4.2. If $K_1, \dots, K_n$ are finite sets, then the **dual total correlation** ('**DTC**') of $X_1, \dots, X_n$ is

$$\mathrm{DTC}(X_1\,;\,\dots\,;\,X_n) := \mathrm{H}(X_1, \dots, X_n) - \sum_{i=1}^{n} \mathrm{H}(X_i \,|\, X_{[n] \setminus i}).$$

In the general case, their dual total correlation is

$$\mathrm{DTC}(X_1, \ldots, X_n) := \sup_{\mathcal{P}_1, \ldots, \mathcal{P}_n} \mathrm{DTC}([X_1]_{\mathcal{P}_1} \,;\, \ldots \,;\, [X_n]_{\mathcal{P}_n}), \qquad (20)$$

where this supremum runs over all tuples of finite measurable partitions $\mathcal{P}_i$ of the spaces K_i. Once again, this may be $+\infty$ in general.

For discrete random variables, dual total correlation was first studied by Han in [28]. It is sometimes called 'excess entropy', as in [6], or 'binding information', as in [1].

The use of suprema to define TC and DTC in the general case is justified by the monotonicity properties proved in Subsection 4.3 below.

It is clear that both TC and DTC depend only on the joint distribution μ of X_1, $\ldots$, X_n, so we sometimes write $\mathrm{TC}(\mu)$ or $\mathrm{DTC}(\mu)$. This leaves the relevant product structure to the reader's understanding.

4.2. Basic identities

The rest of this section is given to some basic identities and inequalities that relate TC, DTC, Shannon entropy, and mutual information. They are used frequently during other proofs in later sections.

In the case of finite-valued random variables, these identities follow from easy applications of the chain rule. However, the extension to the general case is tricky if we use the definitions as suprema in (19) and (20). It is made much easier by the fact that these suprema can instead be understood as limits in the sense of Definition 3.3, which follows from the finite-valued case of the monotonicity properties in Lemmas 4.8 and 4.9 below. Those, in turn, are proved using the finite-valued cases of some of our other identities. Therefore, although we formulate all of our results directly in their full generality, they are proved in two rounds: we first give all the proofs for finite-valued random variables, and then return to the general setting in Subsection 4.4.

The next lemma provides some alternative formulae for TC and DTC. These formulae do not require any limits or suprema, even for general random variables.

Lemma 4.3. TC and DTC satisfy these identities:

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) = \sum_{i=1}^{n} \mathrm{I}(X_i \,;\, X_{[i-1]}); \qquad (a)$$

$$\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n) = \sum_{i=1}^{n} \mathrm{I}(X_i \,;\, X_{[n]\setminus i} \mid X_{[i-1]}) = \sum_{i=1}^{n} \mathrm{I}(X_i \,;\, X_{\{i+1,\ldots,n\}} \mid X_{[i-1]}); \quad (b)$$

and

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) + \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n) = \sum_{i=1}^{n} \mathrm{I}(X_i \,;\, X_{[n]\setminus i}). \qquad (c)$$

In particular, in each of these equations, one side equals $+\infty$ only if they both do.

Remarks 4.4.

1. The right-hand sides in (a) and (b) depend on the order in which we label our random variables, as in the chain rule itself. The symmetry of TC and DTC can be recovered by averaging over all possible orderings. We use this trick later in the proof of Theorem A.

2. Parts (a) and (c) appear as [46, equations (14) and (18)]. In [1], part (b) appears as equation (20) and part (c) appears inside the proof of Theorem 3.

P r o o f. [Finite-valued case.] The middle and right-hand side in (b) are equal because, after conditioning on the variables $X_1, \ldots, X_{i-1}$, they cannot make any contribution to the mutual information $\mathrm{I}(X_i\,;\,X_{[n]\setminus i}\,|\,X_{[i-1]})$.

Using the chain rule to expand $\mathrm{H}(X_1,\ldots,X_n)$ into n terms, and inserting the result into Definitions 4.1 and 4.2, we have

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n) = \sum_{i=1}^{n} \big[\mathrm{H}(X_i) - \mathrm{H}(X_i\,|\,X_{[i-1]})\big]$$

and

$$\mathrm{DTC}(X_1,\ldots,X_n) = \sum_{i=1}^{n} \big[\mathrm{H}(X_i\,|\,X_{[i-1]}) - \mathrm{H}(X_i\,|\,X_{[n]\setminus i})\big].$$

Recalling formulae (5) and (7), we arrive immediately at parts (a) and (b).

On the other hand, if we add the definitions of TC and DTC for finite-valued random variables, then the term $\mathrm{H}(X_1,\ldots,X_n)$ cancels to leave

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n) + \mathrm{DTC}(X_1\,;\,\ldots\,;\,X_n) = \sum_{i=1}^{n} \big[\mathrm{H}(X_i) - \mathrm{H}(X_i\,|\,X_{[n]\setminus i})\big].$$

Now another appeal to formula (5) gives part (c). □

The next lemma gives recursive formulae for TC and DTC as the number of variables increases.

Lemma 4.5. If $n \geq 3$, then TC and DTC satisfy

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n) = \mathrm{TC}(X_1\,;\,\ldots\,;\,X_{n-1}) + \mathrm{I}(X_n\,;\,X_{[n-1]}) \qquad (a)$$

and

$$\mathrm{DTC}(X_1\,;\,\ldots\,;\,X_n) = \mathrm{DTC}(X_1\,;\,\ldots\,;\,X_{n-1}) + \sum_{i=1}^{n-1} \mathrm{I}(X_i\,;\,X_n\,|\,X_{[n-1]\setminus i}). \qquad (b)$$

In particular, in both of these equations, one side equals $+\infty$ only if they both do.

Remark 4.6. Part (b) appears as [6, equation (13)].

P r o o f .   [Finite-valued case.]   To prove (a), let us substitute for $\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_{n-1})$ from the right-hand side of Lemma 4.3(a):

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_{n-1}) + \mathrm{I}(X_n \,;\, X_{[n-1]}) = \sum_{i=1}^{n-1} \mathrm{I}(X_i \,;\, X_{[i-1]}) + \mathrm{I}(X_n \,;\, X_{[n-1]}).$$

This equals $\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n)$ by another use of Lemma 4.3(a).

Similarly, to prove (b), we substitute for $\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_{n-1})$ from the right-hand side of Lemma 4.3(b):

$$\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_{n-1}) + \sum_{i=1}^{n-1} \mathrm{I}(X_i \,;\, X_n \,|\, X_{[n-1]\setminus i})$$

$$= \sum_{i=1}^{n-1} \mathrm{I}(X_i \,;\, X_{\{i+1,\ldots,n-1\}} \,|\, X_{[i-1]}) + \sum_{i=1}^{n-1} \mathrm{I}(X_i \,;\, X_n \,|\, X_{[n-1]\setminus i})$$

$$= \sum_{i=1}^{n-1} \big(\mathrm{I}(X_i \,;\, X_{\{i+1,\ldots,n-1\}} \,|\, X_{[i-1]}) + \mathrm{I}(X_i \,;\, X_n \,|\, X_{[i-1]}, X_{\{i+1,\ldots,n-1\}}) \big).$$

By the chain rule for mutual information (Lemma 3.6), this is equal to

$$\sum_{i=1}^{n-1} \mathrm{I}(X_i \,;\, X_{\{i+1,\ldots,n\}} \,|\, X_{[i-1]}).$$

Adding the dummy term $\mathrm{I}(X_n \,;\, X_\emptyset \,|\, X_{[n-1]}) = 0$, we see that this agrees with $\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n)$ by another use of Lemma 4.3(b).   □

Corollary 4.7. If $n \geq 3$ then

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) \geq \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_{n-1})$$

and

$$\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n) \geq \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_{n-1}).$$

4.3. Monotonicity properties

This subsection introduces simple but useful comparisons of the TC or DTC values of two different n-tuples of random variables.

Lemma 4.8. For any random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ we have

$$\mathrm{TC}(Y_1 \,;\, \ldots \,;\, Y_n) \leq \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) + \sum_{i=1}^{n} \mathrm{H}(Y_i \,|\, X_i).$$

In particular, if X_i almost surely determines Y_i for each i, then

$$\mathrm{TC}(Y_1 \,;\, \ldots \,;\, Y_n) \leq \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n). \tag{21}$$

P r o o f . [Finite-valued case.] By symmetry and induction it suffices to prove the first inequality when $Y_i = X_i$ for $i = 1, 2, \ldots, n-1$. In that case two appeals to Lemma 4.5(a) give

$$\begin{aligned}
&\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_{n-1} \,;\, Y_n) - \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_{n-1} \,;\, X_n) \\
&= \mathrm{I}(Y_n \,;\, X_{[n-1]}) - \mathrm{I}(X_n \,;\, X_{[n-1]}) \\
&\leq \mathrm{I}(Y_n, X_n \,;\, X_{[n-1]}) - \mathrm{I}(X_n \,;\, X_{[n-1]}) \\
&= \mathrm{I}(Y_n \,;\, X_{[n-1]} \,|\, X_n) \qquad\qquad \text{(chain rule for I)} \\
&\leq \mathrm{H}(Y_n \,|\, X_n).
\end{aligned}$$

This proves the first inequality. The second inequality is a special case. $\qquad\square$

Lemma 4.9. For any random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ we have

$$\mathrm{DTC}(Y_1 \,;\, \ldots \,;\, Y_n) \leq \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n) + (n-1) \sum_{i=1}^{n} \mathrm{H}(Y_i \,|\, X_i).$$

In particular, if X_i almost surely determines Y_i for each i, then

$$\mathrm{DTC}(Y_1 \,;\, \ldots \,;\, Y_n) \leq \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n). \tag{22}$$

P r o o f . [Finite-valued case.] As for TC, it suffices to prove the first inequality when $Y_i = X_i$ for $i = 1, 2, \ldots, n - 1$. In that case two appeals to Lemma 4.5(b) give

$$\begin{aligned}
&\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_{n-1} \,;\, Y_n) - \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_{n-1} \,;\, X_n) \\
&= \sum_{i=1}^{n-1} \left[\mathrm{I}(X_i \,;\, Y_n \,|\, X_{[n-1]\setminus i}) - \mathrm{I}(X_i \,;\, X_n \,|\, X_{[n-1]\setminus i}) \right].
\end{aligned}$$

Each of these $n - 1$ terms may be bounded as follows:

$$\begin{aligned}
&\mathrm{I}(X_i \,;\, Y_n \,|\, X_{[n-1]\setminus i}) - \mathrm{I}(X_i \,;\, X_n \,|\, X_{[n-1]\setminus i}) \\
&\leq \mathrm{I}(X_i \,;\, Y_n, X_n \,|\, X_{[n-1]\setminus i}) - \mathrm{I}(X_i \,;\, X_n \,|\, X_{[n-1]\setminus i}) \\
&= \mathrm{I}(X_i \,;\, Y_n \,|\, X_{[n]\setminus i}) \qquad\qquad \text{(chain rule for I)} \\
&\leq \mathrm{H}(Y_n \,|\, X_{[n]\setminus i}) \\
&\leq \mathrm{H}(Y_n \,|\, X_n) \qquad\qquad \text{(since } n \in [n] \setminus i\text{).}
\end{aligned}$$

Therefore our original difference of DTC-values is at most $(n-1)\mathrm{H}(Y_n \,|\, X_n)$, as required for the first inequality. The second inequality is a special case. $\qquad\square$

Remark 4.10. I would not be surprised if (21) and (22) were already in the literature somewhere, but I have not found a reference.

The factor of $(n - 1)$ in the statement of Lemma 4.9 cannot be improved in general. This is seen in the following example, which is essentially taken from [5, Example 3.6]

Example 4.11. In the product group $(\mathbb{Z}/p\mathbb{Z})^n$, consider the subset

$$Z := \big\{(a_1, \ldots, a_n) \in (\mathbb{Z}/p\mathbb{Z})^n : \ a_1 + \cdots + a_n = 0\big\}$$

Let μ be the uniform distribution on Z and let $Y_1, \ldots, Y_n$ be random variables with joint distribution μ. Let $X_i := Y_i$ for $i \leq n-1$, and let X_n be a uniform random element of $\mathbb{Z}/p\mathbb{Z}$ independent of all the other random variables. Then $\mathrm{DTC}(X_1 ; \ldots ; X_n) = 0$, since all the X_is are independent, while a simple calculation gives $\mathrm{DTC}(Y_1 ; \ldots ; Y_n) = (n-1)\log p$. Finally,

$$\mathrm{H}(Y_i \,|\, X_i) = \begin{cases} 0 & \text{if } i \leq n-1 \\ \log p & \text{if } i = n. \end{cases}$$

So the inequality of Lemma 4.9 becomes an equality in this example. $\triangleleft$

It may be possible to improve Lemma 4.9 by using some other information theoretic quantities in the bound, in addition to the conditional entropies $\mathrm{H}(Y_i \,|\, X_i)$, but we do not explore that possibility here.

4.4. Extension to general random variables

For random variables $X_1, \ldots, X_n$ taking values in general measurable spaces $K_1, \ldots, K_n$, the definitions of TC and DTC in (19) and (20) are tractable using the monoticity provided by Lemmas 4.8 and 4.9. The next result is the analog of Proposition 3.4.

Lemma 4.12. Let $\mathfrak{P}_i$ be a directed and generating collection of partitions of K_i for each i (recall Definition 3.1). Then

$$\mathrm{TC}(X_1, \ldots, X_n) = \lim_{\mathcal{P}_1 \in \mathfrak{P}_1, \ldots, \mathcal{P}_n \in \mathfrak{P}_n} \mathrm{TC}([X_1]_{\mathcal{P}_1} ; \ldots ; [X_n]_{\mathcal{P}_n}) \tag{23}$$

and similarly with DTC in place of TC.

P r o o f .   The inequality '$\geq$' in (23) is immediate from the definition (19). So now let $\mathcal{Q}_i$ be an arbitrary finite measurable partition of K_i for each i, and let $\varepsilon > 0$. Since each $\mathfrak{P}_i$ is directed and generating, there are members $\mathcal{P}_i \in \mathfrak{P}_i$ such that

$$\mathrm{H}([X_i]_{\mathcal{Q}_i} \,|\, [X_i]_{\mathcal{P}_i}) < \varepsilon/n \quad \text{for each } i \in [n].$$

Now Lemma 4.8 gives

$$\mathrm{TC}([X_1]_{\mathcal{P}_1'} ; \ldots ; [X_n]_{\mathcal{P}_n'}) > \mathrm{TC}([X_1]_{\mathcal{Q}_1} ; \ldots ; [X_n]_{\mathcal{Q}_n}) - \varepsilon$$

whenever $\mathcal{P}_i' \in \mathfrak{P}_i$ is a further refinement of $\mathcal{P}_i$ for each i. Since $\varepsilon > 0$ was arbitrary, this completes the proof of (23). The argument for DTC is exactly analogous. $\square$

TC and DTC give two different ways to quantify the dependence among a tuple of random variables. In general they can give very different values. However, they are not completely independent; in particular, one cannot be infinite unless they both are. This is a corollary of the following more precise inequalities.

Lemma 4.13. The TC and DTC of $X_1, \ldots, X_n$ both lie between

$$\max_i \mathrm{I}(X_i \,;\, X_{[n]\setminus i}) \quad \text{and} \quad (n-1) \cdot \max_i \mathrm{I}(X_i \,;\, X_{[n]\setminus i}). \tag{24}$$

In particular,

$$\mathrm{DTC} \le (n-1) \cdot \mathrm{TC} \quad \text{and} \quad \mathrm{TC} \le (n-1) \cdot \mathrm{DTC},$$

and the quantities TC, DTC and $\max_i \mathrm{I}(X_i \,;\, X_{[n]\setminus i})$ are either all finite or all infinite.

P r o o f . *Step 1.* Suppose first that each K_i is finite. Then the finite-valued cases of parts (a) and (b) of Lemma 4.3 give

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) \ge \mathrm{I}(X_n \,;\, X_{[n]\setminus n})$$

and

$$\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n) \ge \mathrm{I}(X_1 \,;\, X_{[n]\setminus 1})$$

respectively, since all the terms in those sums are non-negative. This argument may be applied for any re-ordering of the random variables, so it implies the lower bound in (24). Now we can insert that lower bound into the left-hand side of Lemma 4.3(c) and then cancel a common term, giving the required upper bound.

Step 2. Now consider arbitrary measurable spaces $K_1, \ldots, K_n$. For any finite measurable partitions $\mathcal{P}_1, \ldots, \mathcal{P}_n$ of those spaces, Step 1 above gives

$$\max_i \mathrm{I}([X_i]_{\mathcal{P}_i} \,;\, [X_{[n]\setminus i}]_{\mathcal{P}_{[n]\setminus i}})$$
$$\le \mathrm{TC}([X_1]_{\mathcal{P}_1} \,;\, \ldots \,;\, [X_n]_{\mathcal{P}_n}), \;\; \mathrm{DTC}([X_1]_{\mathcal{P}_1} \,;\, \ldots \,;\, [X_n]_{\mathcal{P}_n})$$
$$\le (n-1) \cdot \max_i \mathrm{I}([X_i]_{\mathcal{P}_i} \,;\, [X_{[n]\setminus i}]_{\mathcal{P}_{[n]\setminus i}}), \quad (25)$$

where $\mathcal{P}_{[n]\setminus i}$ denotes the partition of $\prod_{j \in [n]\setminus i} K_j$ into the sets $\prod_{j \in [n]\setminus i} C_j$ with $C_j \in \mathcal{P}_j$. Taking the limit along $\mathcal{P}_1, \ldots, \mathcal{P}_n$ in (25) proves the bounds (24) in general.

Step 3. The remaining assertions of Lemma 4.13 follow directly from the bounds (24). $\square$

Definition 4.14. The random variables $X_1, \ldots, X_n$ are **finitely correlated** if any (and hence all) of the quantities

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n), \quad \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n), \quad \max_i \mathrm{I}(X_i \,;\, X_{[n]\setminus i})$$

are finite. We also refer to a joint distribution of such random variables as **finitely correlated**. If this property does not hold, then the random variables and their joint distribution are **infinitely correlated**.

Clearly finite-valued random variables are always finitely correlated, and so are arbitrary independent random variables.

We can now extend our previous identities for TC and DTC to general random variables. In the following three proofs, $\mathfrak{P}_i$ is the family of all finite measurable partitions of K_i for each i.

P r o o f . [Lemma 4.3 in general case.]

Parts (a) and (c) follow by applying the finite-valued case to the quantizations $[X_1]_{\mathcal{P}_1}$, $\ldots$, $[X_n]_{\mathcal{P}_n}$ and then taking the limit along $\mathcal{P}_1 \in \mathfrak{P}_1$, $\ldots$, $\mathcal{P}_n \in \mathfrak{P}_n$ on both sides.

The proof of part (b) is complicated by the presence of conditional mutual information values on the right. Recall from Subsection 3.2 that, if these are to be handled as limits, then they require iterated limits taken in the right order. However, we can instead give an indirect proof of (b) using (a) and (c).

If X_1, $\ldots$, X_n are infinitely correlated, then both sides of (b) are infinite. If they are finitely correlated, then every quantity appearing in (a) or (c) is finite, and therefore so is every quantity appearing in (b) by the chain rule for mutual information (Lemma 3.6). As a result, we may subtract (c) from (c) to leave exactly (b), by another use of that chain rule.   □

P r o o f . [Lemma 4.5 and Corollary 4.7 in general case.]

Now that we have the general case of Lemma 4.3, these two results follow from it exactly as in the finite-valued case. All of the required re-arrangements are still correct if some of the values are equal to $+\infty$, because no cancellation or subtraction are involved.   □

P r o o f . [Lemmas 4.8 and 4.9 in general case.]

Suppose that Y_1, $\ldots$, Y_n take values in L_1, $\ldots$, L_n, and let $\mathfrak{Q}_i$ be the family of all finite measurable partitions of L_i for each i. The finite-valued case of Lemma 4.8 gives

$$\mathrm{TC}([Y_1]_{\mathcal{Q}_1} \,;\, \ldots \,;\, [Y_n]_{\mathcal{Q}_n}) \leq \mathrm{TC}([X_1]_{\mathcal{P}_1} \,;\, \ldots \,;\, [X_n]_{\mathcal{P}_n}) + \sum_{i=1}^{n} \mathrm{H}([Y_i]_{\mathcal{Q}_i} \,|\, [X_i]_{\mathcal{P}_i})$$

for any $\mathcal{P}_i \in \mathfrak{P}_i$ and $\mathcal{Q}_i \in \mathfrak{Q}_i$. Taking the limit along $\mathcal{P}_1 \in \mathfrak{P}_1$, $\ldots$, $\mathcal{P}_n \in \mathfrak{P}_n$, and using the monotonicity of Shannon entropy, this gives

$$\mathrm{TC}([Y_1]_{\mathcal{Q}_1} \,;\, \ldots \,;\, [Y_n]_{\mathcal{Q}_n}) \leq \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) + \sum_{i=1}^{n} \mathrm{H}([Y_i]_{\mathcal{Q}_i} \,|\, X_i)$$

$$\leq \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) + \sum_{i=1}^{n} \mathrm{H}(Y_i \,|\, X_i).$$

Now the limit along $\mathcal{Q}_1 \in \mathfrak{Q}_1$, $\ldots$, $\mathcal{Q}_n \in \mathfrak{Q}_n$ completes the proof for TC. The argument for DTC is analogous.   □

5. TC AND DTC IN TERMS OF REFERENCE MEASURES

Another approach to TC and DTC for general random variables starts with the formula (13) for mutual information in terms of KL divergence. In the case of TC the relevant formula is already well-used in the literature:

$$\mathrm{TC}(\mu) = \mathrm{D}(\mu \,\|\, \mu_1 \times \cdots \times \mu_n), \tag{26}$$

where μ_i is the ith marginal of μ. When $n = 2$, this is just (13). The case $n \geq 3$ follows easily from that special case by induction, using Lemma 4.5(a) for the left-hand side and the general chain rule for KL divergence for the right-hand side.

In the form given by (26), TC has already played an important role in some results of theoretical probability, such as those of Csiszár [14] and Marton [34, 35] mentioned in the Introduction.

If we assume that the measurable spaces K_i are standard, then DTC also has a formula in terms of KL divergences. Before we derive this, however, let us add another layer of generality. Rather than compare μ to its own marginals, as in (26), one can compare it to a prior choice of reference measures on each K_i.

Proposition 5.1. Assume that each K_i is standard, and let $\lambda_i \in \text{Prob}(K_i)$ be a reference measure for each i.

(a) If $D(\mu_i \,\|\, \lambda_i) < \infty$ for each i, then $TC(\mu)$ is equal to

$$D(\mu \,\|\, \lambda_1 \times \cdots \times \lambda_n) - \sum_{i=1}^{n} D(\mu_i \,\|\, \lambda_i).$$

(b) If μ is finitely correlated and $D(\mu_i \,\|\, \lambda_i) < \infty$ for each i, then $DTC(\mu)$ is equal to

$$\sum_{i=1}^{n} \int D(\mu_{i,z} \,\|\, \lambda_i) \, \mu_{[n] \setminus i}(\mathrm{d}z) - D(\mu \,\|\, \lambda_1 \times \cdots \times \lambda_n), \tag{27}$$

where $\mu_{[n] \setminus i}$ is the projection of μ to $\prod_{j \in [n] \setminus i} K_j$, and $(\mu_{i,z} \,:\, z \in \prod_{j \in [n] \setminus i} K_j)$ is a conditional distribution for the ith coordinate given the other coordinates according to μ.

P r o o f. *Part (a).* Given the formula (26), part (a) is equivalent to

$$D(\mu \,\|\, \lambda_1 \times \cdots \times \lambda_n) = D(\mu \,\|\, \mu_1 \times \cdots \times \mu_n) + \sum_{i=1}^{n} D(\mu_i \,\|\, \lambda_i). \tag{28}$$

This is a classical identity due to Csiszár: see [14, equation (2.11)]. It generalizes the earlier formula (16) relating mutual information and KL divergences. We may rearrange Csiszár's identity in the required way because we have assumed that each of the quantities $D(\mu_i \,\|\, \lambda_i)$ is finite.

Part (b). This can be derived quickly by combining part (a) and Lemma 4.3(c). Those ingredients give

$$DTC(\mu) = \sum_{i=1}^{n} I(X_i \,;\, X_{[n] \setminus i}) - \left[D(\mu \,\|\, \lambda_1 \times \cdots \times \lambda_n) - \sum_{i=1}^{n} D(\mu_i \,\|\, \lambda_i) \right], \tag{29}$$

where $X_1, \ldots, X_n$ are random variables with joint distribution μ. The bracketed expression on the right equals $TC(\mu)$, and we may subtract it like this because we have assumed that μ is finitely correlated.

For each $i = 1, 2, \ldots, n$, we now apply the special case (17) of the chain rule for KL divergence:

$$\mathrm{I}(X_i \, ; \, X_{[n]\setminus i}) = \int \mathrm{D}(\mu_{i,z} \, \| \, \lambda_i) \, \mu_{[n]\setminus i}(\mathrm{d}z) - \mathrm{D}(\mu_i \, \| \, \lambda_i).$$

Substituting this into the first sum in (29), the appearances of the quantities $\mathrm{D}(\mu_i \, \| \, \lambda_i)$ cancel, and we are left with (27). □

Remark 5.2. In view of Csiszár's identity (28), the assumptions for Proposition 5.1(b) are equivalent to the single assumption

$$\mathrm{D}(\mu \, \| \, \lambda_1 \times \cdots \times \lambda_n) < \infty.$$

Since DTC is non-negative, Proposition 5.1(b) implies that

$$\mathrm{D}(\mu \, \| \, \lambda_1 \times \cdots \times \lambda_n) \leq \sum_{i=1}^{n} \int \mathrm{D}(\mu_{i,z} \, \| \, \lambda_i) \, \mu_{[n]\setminus i}(\mathrm{d}z). \tag{30}$$

This inequality is widely known. It offers a simple explanation for the tensorization property of logarithmic Sobolev inequalities, as exposed in [31]: see [31, Proposition 4.1] for an elementary family of inequalities that includes (30), attributed to Bobkov in that paper. As far as I know, it is a new observation that the gap in (30) is actually the same quantity as DTC (which is the gap in the analogous Han inequality) after passing to a supremum over quantizations.

Before leaving this section, let us mention another setting that offers its own formulae for TC and DTC. If $X_1, \ldots, X_n$ are real-valued and jointly absolutely continuous, then

$$\mathrm{TC}(\mu) = \sum_{i=1}^{n} \mathrm{h}(X_i) - \mathrm{h}(X_1, \ldots, X_n)$$

and

$$\mathrm{DTC}(\mu) = \mathrm{h}(X_1, \ldots, X_n) - \sum_{i=1}^{n} \mathrm{h}(X_i \, | \, X_{[n]\setminus i}).$$

Here h stands for Shannon's differential entropy (see [12, Chapter 8]), and we assume enough smoothness of the joint PDF that the subtracted terms are all finite. These identities generalize a standard formula for the mutual information between jointly continuous random variables: see [12, Section 8.5]. The proof of these formulae is analogous to the proof of Proposition 5.1, and we omit it. Alternatively, they may be subsumed into Proposition 5.1 itself if one expands the definition of KL divergence to allow arbitrary reference measures.

6. CONDITIONAL TC AND DTC

6.1. Definitions and first properties

Consider a random n-tuple $X_1, \ldots, X_n$ and another random variable Y on the same probability space. Assume that the X_is are finite-valued, but let Y be arbitrary. Then

we define the **conditional total correlation** by

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n\,|\,Y) := \Big(\sum_{i=1}^{n}\mathrm{H}(X_i\,|\,Y)\Big) - \mathrm{H}(X_1,\ldots,X_n\,|\,Y) \qquad (31)$$

and the **conditional dual total correlation** by

$$\mathrm{DTC}(X_1\,;\,\ldots\,;\,X_n\,|\,Y) := \mathrm{H}(X_1,\ldots,X_n\,|\,Y) - \sum_{i=1}^{n}\mathrm{H}(X_i\,|\,X_{[n]\setminus i},Y). \qquad (32)$$

In case the X_is are also general random variables, we extend the definitions of conditional TC and DTC just as we did in the unconditional case:

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n\,|\,Y) := \sup_{\mathcal{P}_1,\ldots,\mathcal{P}_n}\ \mathrm{TC}\big([X_1]_{\mathcal{P}_1}\,;\,\ldots\,;\,[X_n]_{\mathcal{P}_n}\,\big|\,Y\big)$$

and similarly with DTC in place of TC. When $n = 2$, this generalizes the definition of conditional mutual information in (11). Just as in Lemma 4.12 for unconditional TC and DTC, we may replace the supremum above with a limit along any directed and generating families of partitions $\mathfrak{P}_i$. The proof is unchanged from the unconditional case.

6.2. Expression in terms of disintegrations

Let X_1, $\ldots$, X_n and Y be random variables with respective target spaces K_1, $\ldots$, K_n and L, let μ be the joint distribution of X_1, $\ldots$, X_n on $\prod_i K_i$, and let ν be the distribution of Y on L. In case the spaces K_i are standard Borel, μ has a disintegration $(\mu_y : y \in L)$ over Y. Then conditional TC and DTC satisfy the obvious formula in terms of this disintegration, generalizing a classical identity for mutual information.

Proposition 6.1. In the situation above, we have

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n\,|\,Y) = \int \mathrm{TC}(\mu_y)\,\nu(\mathrm{d}y)$$

and similarly with DTC in place of TC.

P r o o f.   We give the proof for TC; the proof for DTC is analogous.

Suppose first that each K_i is finite. In that case we have

$$\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n\,|\,Y) = \sum_{i}\mathrm{H}(X_i\,|\,Y) - \mathrm{H}(X_1,\ldots,X_n\,|\,Y).$$

Let $(\mu_{i,y} : y \in L)$ be a conditional distribution for X_i given Y. Then $\mu_{i,y}$ is the ith marginal of μ_y for almost every y, by the essential uniqueness of conditional distributions. Therefore, by the classical formula for conditional entropy in terms of conditional distributions, the above is equal to

$$\int \Big[\sum_{i}\mathrm{H}(\mu_{i,y}) - \mathrm{H}(\mu_y)\Big]\,\nu(\mathrm{d}y) = \int \mathrm{TC}(\mu_y)\,\nu(\mathrm{d}y).$$

For the general case, since each K_i is standard Borel, it has a refining and generating sequence of finite partitions $\mathcal{P}_{i,1}$, $\mathcal{P}_{i,2}$, ... For each finite t, the special case proved above gives

$$\mathrm{TC}([X_1]_{\mathcal{P}_{1,t}} \, ; \dots ; \, [X_n]_{\mathcal{P}_{n,t}} \,|\, Y) = \int \mathrm{TC}([\mu_y]_{\mathcal{P}_{1,t} \times \dots \times \mathcal{P}_{n,t}}) \, \nu(\mathrm{d}y), \qquad (33)$$

where $[\mu_y]_{\mathcal{P}_{1,t} \times \dots \times \mathcal{P}_{n,t}}$ is the discrete probability distribution induced on the product partition $\mathcal{P}_{1,t} \times \dots \times \mathcal{P}_{n,t}$ by the measure μ_y. As $t \longrightarrow \infty$:

- the left-hand side of (33) converges to $\mathrm{TC}(X_1 \, ; \dots ; \, X_n \,|\, Y)$, by the generalization of Lemma 4.12 to conditional random variables; and

- the right-hand side of (33) converges to the integral of $\mathrm{TC}(\mu_y)$ with respect to $\nu(\mathrm{d}y)$, by Lemma 4.12 and the monotone convergence theorem.

$\square$

6.3. Clumping rules

Let X_1, ..., X_n be random variables taking values in arbitrary measurable spaces K_1, ..., K_n. Let $[n] = S_1 \cup \dots \cup S_m$ be a partition into non-empty subsets, and enumerate $S_j = \{k_{j,1} < \dots < k_{j,s_j}\}$ for each j. Then the product spaces

$$\prod_{i=1}^{n} K_i \quad \text{and} \quad \prod_{j=1}^{m} \Big(\prod_{i \in S_j} K_i \Big)$$

have a canonical identification. Corresponding to this, we may define the new m-tuple of random variable $Y_j := X_{S_j}$ taking values in the product spaces $K_{S_j} := \prod_{i \in S_j} K_i$. We refer to this construction as **clumping**, and to Y_1, ..., Y_m as **clumped** random variables.

In the statements of the next two lemmas, we need the following convention: for a single random variable Z (that is, a '1-tuple') we always have

$$\mathrm{TC}(Z) = \mathrm{DTC}(Z) = 0.$$

Lemma 6.2. (TC clumping rules) The random variables above satisfy

$$\mathrm{TC}(X_1 \, ; , \dots ; \, X_n) = \mathrm{TC}(Y_1 \, ; \dots ; \, Y_m) + \sum_{j=1}^{m} \mathrm{TC}(X_{k_{j,1}} \, ; \dots ; \, X_{k_{j,s_j}}).$$

In particular, one side is $+\infty$ only if they both are.

Remark 6.3. The case $m = 2$ of this rule appears as [41, equation (4.1)], and [6, equation (5)] follows by applying a special case of this rule repeatedly.

P r o o f . By induction, we may assume that all but one of the sets S_j are singletons — applying that case to m separate 'clumpings' gives the general formula. Having done so, we may also assume that our partition of $[n]$ is

$$[n] = \{1, 2, \ldots, \ell\} \cup \{\ell + 1\} \cup \cdots \cup \{n\}$$

for some $\ell \in \{1, \ldots, n-1\}$, by symmetry among the random variables. In this case the desired formula is

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) = \mathrm{TC}(X_1 \,;\, \ldots \,;\, X_\ell) + \mathrm{TC}(Y \,;\, X_{\ell+1} \,;\, \ldots \,;\, X_n), \qquad (34)$$

where $Y := (X_1, \ldots, X_\ell)$. This formula can be deduced quickly from part (a) of Lemma 4.3:

$$\mathrm{TC}(X_1 \,;\, \ldots \,;\, X_n) = \sum_{i=1}^{n} \mathrm{I}(X_i \,;\, X_{[i-1]}) = \sum_{i=1}^{\ell} \mathrm{I}(X_i \,;\, X_{[i-1]}) + \sum_{i=\ell+1}^{n} \mathrm{I}(X_i \,;\, X_{[i-1]})$$

$$= \sum_{i=1}^{\ell} \mathrm{I}(X_i \,;\, X_{[i-1]}) + \sum_{i=\ell+1}^{n} \mathrm{I}(X_i \,;\, Y, X_{\ell+1}, \ldots, X_{i-1}).$$

By another appeal to Lemma 4.3(a), the two sums on the line above are equal to the two right-hand terms in (34), respectively. For the second sum, this becomes apparent after we insert the dummy term $\mathrm{I}(Y \,;\, X_\emptyset) = 0$. $\qquad\square$

Lemma 6.4. (DTC clumping rules) The random variables above satisfy

$$\mathrm{DTC}(X_1 \,;\, , \ldots \,;\, X_n) = \mathrm{DTC}(Y_1 \,;\, \ldots \,;\, Y_m) + \sum_{j=1}^{m} \mathrm{DTC}(X_{k_{j,1}} \,;\, \ldots \,;\, X_{k_{j,s_j}} \,|\, X_{[n]\setminus S_j}).$$

In particular, one side is $+\infty$ only if they both are.

Remark 6.5. A special case of this rule may be applied repeatedly to recover [6, equation (14)], where DTC is called 'excess entropy'.

P r o o f . Similarly to the proof of Lemma 6.2, we may reduce to the case in which all but one of the sets S_j are singletons, and then, by symmetry, to the case of the partition

$$[n] = \{1\} \cup \cdots \cup \{\ell - 1\} \cup \{\ell, \ell + 1, \ldots, n\}$$

for some $\ell \in \{1, \ldots, n-1\}$. Note that this time we put the non-singleton last. For this partition the desired formula is

$$\mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_n) = \mathrm{DTC}(X_1 \,;\, \ldots \,;\, X_{\ell-1} \,;\, Y) + \mathrm{DTC}(X_\ell \,;\, \ldots \,;\, X_n \,|\, X_{[\ell-1]}), \quad (35)$$

where $Y := (X_\ell, \ldots, X_n)$. Now we proceed using the expression given by Lemma 4.3(b):

$$\mathrm{DTC}(X_1 \, ; \, \dots \, ; \, X_n)$$

$$= \sum_{i=1}^{n} \mathrm{I}(X_i \, ; \, X_{\{i+1,\dots,n\}} \,|\, X_{[i-1]})$$

$$= \sum_{i=1}^{\ell-1} \mathrm{I}(X_i \, ; \, X_{\{i+1,\dots,\ell-1\}}, Y \,|\, X_{[i-1]}) + \sum_{i=\ell}^{n} \mathrm{I}(X_i \, ; \, X_{\{i+1,\dots,n\}} \,|\, X_{[\ell-1]}, X_\ell, \dots, X_{i-1}).$$

By two more appeals to Lemma 4.3(b), the two sums on the previous line are equal to the two terms on the right-hand side of (35), respectively. For the first sum, this becomes apparent after we insert the dummy term $\mathrm{I}(Y \, ; \, X_\emptyset \,|\, X_{[\ell-1]}) = 0$. □

The following special case of Lemma 6.4 is quite intuitive, and is needed by itself in the sequel.

Corollary 6.6. Any random variables $X_1, \dots, X_n$ and Y satisfy

$$\mathrm{DTC}(X_1 \, ; \, \dots \, ; \, X_n \, ; \, Y) = \mathrm{I}(X_1, \dots, X_n \, ; \, Y) + \mathrm{DTC}(X_1 \, ; \, \dots \, ; \, X_n \,|\, Y).$$

Corollary 6.7. In the setting of Lemmas 6.2 and 6.4, we have

$$\mathrm{TC}(Y_1 \, ; \, \dots \, ; \, Y_m) \leq \mathrm{TC}(X_1 \, ; \, \dots \, ; \, X_n),$$

and similarly with DTC in place of TC. In particular, if S, S^c is a binary partition of $[n]$, then

$$\mathrm{I}(X_S \, ; \, X_{S^c}) \leq \min \big\{ \mathrm{TC}(X_1 \, ; \, \dots \, ; \, X_n), \mathrm{DTC}(X_1 \, ; \, \dots \, ; \, X_n) \big\}.$$

The second inequality here generalizes the lower bound in (24).

P r o o f . The first pair of inequalities follows from Lemmas 6.2 and 6.4 because TC and DTC are non-negative. The second pair is a special case of the first, because both TC and DTC reduce to mutual information in the case of two random variables. □

Part II: Structure of measures with low correlation

7. TC AND PRODUCT MEASURES

The following concentration inequality lies behind much of Part II of this paper.

Proposition 7.1. (Marton's transportation inequality for product measures)
Let $(K_1, d_{K_1}), \dots, (K_n, d_{K_n})$ be complete and separable metric spaces of diameter at most 1. If $\nu \in \mathrm{Prob}(\prod_i K_i)$ is a product measure, and $\mu \in \mathrm{Prob}(\prod_i K_i)$ is arbitrary, then

$$\overline{d_n}(\mu, \nu) \leq \sqrt{\frac{1}{2n} \mathrm{D}(\mu \,\|\, \nu)}.$$

In Marton's original proofs in [34, 35], each K_i is equal to a fixed finite set and given its disrete metric, but those proofs generalize with only cosmetic changes to give the result stated above.

Combining Proposition 7.1 with the identity (26) gives the main structural result for measures with small TC.

Corollary 7.2. Any $\mu \in \mathrm{Prob}(\prod_i K_i)$ satisfies

$$\overline{d_n}(\mu, \mu_1 \times \cdots \times \mu_n) \leq \sqrt{\frac{1}{2n}\mathrm{TC}(\mu)},$$

where μ_i is the ith marginal of μ.

For finite alphabets with their discrete metrics, the relation between $\overline{d_n}$ and TC in Corollary 7.2 can be partially reversed using the next lemma, which gives a kind of 'continuity' for TC in the transportation metric. It strengthens [5, Lemma 4.4], but here we give a slightly more efficient proof based on Lemma 4.8.

Lemma 7.3. Let $\mu, \nu \in \mathrm{Prob}(\prod_i K_i)$, and suppose that each K_i has cardinality at most k and is given its discrete metric: $d_{K_i}(x,y) := 1_{\{x \neq y\}}$. Let $\delta := \overline{d_n}(\mu, \nu)$. Then

$$|\mathrm{TC}(\mu) - \mathrm{TC}(\nu)| \leq \big(\mathrm{H}(\delta, 1-\delta) + \delta \log(k-1)\big)n.$$

P r o o f. Let λ be a coupling of μ and ν that witnesses the distance $\overline{d_n}(\mu, \nu)$, and let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be random pairs that take values in $K_1 \times K_1, \ldots, K_n \times K_n$ and have joint distribution λ. Also, let

$$\delta_i := \lambda\{(x,y) : x_i \neq y_i\} \quad \text{for each } i = 1, 2, \ldots, n.$$

Then Lemma 4.8 and Fano's inequality give

$$\mathrm{TC}(\nu) \leq \mathrm{TC}(\mu) + \sum_{i=1}^{n} \mathrm{H}(Y_i \mid X_i) \leq \mathrm{TC}(\mu) + \sum_{i=1}^{n} \big(\mathrm{H}(\delta_i, 1-\delta_i) + \delta_i \log(k-1)\big).$$

Since we chose λ to be an optimal coupling, δ is the average of the δ_is. Therefore the concavity of the entropy function turns this into

$$\mathrm{TC}(\nu) \leq \mathrm{TC}(\mu) + \big(\mathrm{H}(\delta, 1-\delta) + \delta \log(k-1)\big)n.$$

By symmetry in μ and ν, this completes the proof. $\square$

Corollary 7.4. In the setting of Lemma 7.3, if ν is a product measure, then

$$\mathrm{TC}(\mu) \leq \big(\mathrm{H}(\delta, 1-\delta) + \delta \log(k-1)\big)n.$$

Unlike the inequality in Corollary 7.2, the inequality in Corollary 7.4 deteriorates for larger alphabets. The next example shows that this feature of Corollary 7.4 is essential. For large or infinite alphabets, I do not know a refined description of exactly which near-product measures have small TC.

Example 7.5. Let $\delta > 0$ and $k \geq 2$, and let $K_i := \{0, 1, \ldots, k-1\}$ for each $i = 1, 2, \ldots, n$. For each $j = 0, 1, \ldots, k-1$, let ν_j be the Dirac point mass on the n-tuple $(j, j, \ldots, j)$. Finally, let

$$\mu := (1 - \delta)\nu_0 + \frac{\delta}{k-1} \sum_{j=1}^{k-1} \nu_j.$$

The product $\mu \times \nu_0$ is the only coupling of μ and ν_0, and it gives

$$\overline{d_n}(\mu, \nu_0) = \delta.$$

However,

$$\mathrm{TC}(\mu) = n \cdot \mathrm{H}\Big(1 - \delta, \frac{\delta}{k-1}, \ldots, \frac{\delta}{k-1}\Big) - \mathrm{H}\Big(1 - \delta, \frac{\delta}{k-1}, \ldots, \frac{\delta}{k-1}\Big)$$
$$\geq (n-1)\delta \log(k-1).$$

If k is large enough, then this TC can be a large multiple of n, even if δ is very small. $\triangleleft$

In addition to giving Corollary 7.2, Proposition 7.1 is ultimately responsible for the approximation between μ_y and a product measure ξ_y in part (b) of Theorem A.

8. DTC AND MIXTURES OF PRODUCT MEASURES

8.1. The DTC of a mixture of products

Proposition 8.1. Let $\mu \in \mathrm{Prob}(\prod_i K_i)$, and let

$$\mu = \int_L \mu_y \, \nu(\mathrm{d}y) \tag{36}$$

be a representation of μ as a mixture. Then

$$\mathrm{DTC}(\mu) \leq \int_L \mathrm{DTC}(\mu_y) \, \nu(\mathrm{d}y) + \mathrm{I}(\nu, \mu_\bullet). \tag{37}$$

In particular, if every μ_y is a product measure, then $\mathrm{DTC}(\mu) \leq \mathrm{I}(\nu, \mu_\bullet)$.

P r o o f. Let $X_1, \ldots, X_n$ be random variables with joint distribution μ, and let Y be another random variable so that the pair $Y, (X_1, \ldots, X_n)$ is a randomization of the mixture (36). In terms of these random variables, the right-hand side of (37) is

$$\mathrm{DTC}(X_1; \ldots; X_n \mid Y) + \mathrm{I}\big((X_1, \ldots, X_n); Y\big).$$

This is equal to $\mathrm{DTC}(X_1; \ldots; X_n; Y)$ by Corollary 6.6, and this is greater than or equal to $\mathrm{DTC}(X_1; \ldots; X_n)$ by Lemma 4.5(b).

The final assertion follows because DTC is zero for any product measure. $\square$

DTC is even more sensitive to small perturbations in $\overline{d_n}$ than TC: near-products can have large DTC even for small alphabets. Because of this, a gap in our understanding remains between Proposition 8.1 and Theorem A.

Example 8.2. Let the measure $\mu \in \text{Prob}((\mathbb{Z}/p\mathbb{Z})^n)$ and random variables X_i and Y_i be as in Example 4.11. Let ν be the uniform distribution on $(\mathbb{Z}/p\mathbb{Z})^n$, so this is the joint distribution of $X_1, \ldots, X_n$. Then $\text{DTC}(\mu) = (n-1)\log p$. But the construction of the Y_is and X_is gives a coupling of μ and ν under which (i) the first $n-1$ coordinates always agree and (ii) the last coordinates agree with probability $1/p > 0$, so $\overline{d_n}(\mu, \nu) < 1/n$. ◁

8.2. Proof of Theorem A

Let $(X_1, \ldots, X_n)$ be a random n-tuple with joint distribution μ.

Lemma 8.3. There is a subset $S \subseteq [n]$ with $|S| \le \text{DTC}(\mu)/\delta^2$ such that

$$\text{TC}(X_1 ; \ldots ; X_n \,|\, X_S) + \text{DTC}(X_1 ; \ldots ; X_n \,|\, X_S) \le \delta^2 |[n] \setminus S|.$$

(In fact we need only the weaker conclusion $\text{TC}(X_1 ; \ldots ; X_n \,|\, X_S) \le \delta^2 |[n] \setminus S|$ to prove Theorem A.)

P r o o f. The formula from Lemma 4.3(b) is available for any ordering of the index set $[n]$. We may therefore consider its average over all possible orderings:

$$\text{DTC}(X_1 ; \ldots ; X_n) = \sum_{i=1}^{n} \mathbf{E}\big[\text{I}\big(X_{\sigma(i)} ; X_{[n]\setminus\sigma(i)} \,\big|\, X_{\{\sigma(1),\ldots,\sigma(i-1)\}}\big)\big], \qquad (38)$$

where $\mathbf{E} = \frac{1}{n!}\sum_\sigma$ denotes expectation over a uniformly random permutation σ of $[n]$.

By Markov's inequality, fewer than $\text{DTC}(\mu)/\delta^2$ of the summands in (38) can exceed δ^2, so some $i \le \text{DTC}(\mu)/\delta^2 + 1$ satisfies

$$\mathbf{E}\big[\text{I}\big(X_{\sigma(i)} ; X_{[n]\setminus\sigma(i)} \,\big|\, X_{\{\sigma(1),\ldots,\sigma(i-1)\}}\big)\big] \le \delta^2.$$

By the tower property of iterated conditional expectations, it follows that some subset $S \subseteq [n]$ of cardinality $i-1$ satisfies

$$\mathbf{E}\Big[\text{I}\big(X_{\sigma(i)} ; X_{[n]\setminus\sigma(i)} \,\big|\, X_{\{\sigma(1),\ldots,\sigma(i-1)\}}\big) \,\Big|\, \{\sigma(1),\ldots,\sigma(i-1)\} = S\Big] \le \delta^2.$$

However, conditionally on $\{\sigma(1),\ldots,\sigma(i-1)\} = S$, the random image $\sigma(i)$ is equally likely to be any element of $[n] \setminus S$. Therefore the conditional expectation above is equal to

$$\frac{1}{|[n] \setminus S|} \sum_{j \in [n] \setminus S} \text{I}(X_j ; X_{[n]\setminus j} \,|\, X_S).$$

(We used a random permutation σ in (38) in order to arrive at this average over $j \in [n] \setminus S$.) By Lemma 4.3(c), this average is equal to

$$\frac{1}{|[n] \setminus S|}\big(\text{TC}(X_1 ; \ldots ; X_n \,|\, X_S) + \text{DTC}(X_1 ; \ldots ; X_n \,|\, X_S)\big).$$

$\square$

Proof. [Theorem A.]

Let S be given by Lemma 8.3, so the hypotheses of Theorem A give $|S| \leq \mathrm{DTC}(\mu)/\delta^2 \leq \delta n$. Let μ_S be the joint distribution of X_S, and similarly for any other subfamily of the random variables X_i. Let $L := \prod_{i \in S} K_i$ and let $(\nu_y : y \in L)$ be a conditional distribution for $(X_1, \ldots, X_n)$ given X_S. Then μ is equal to the mixture

$$\mu = \int_L \nu_y \ \mu_S(\mathrm{d}y). \tag{39}$$

At this stage we have no control on the mutual information of this mixture. But we regain such control if we project to coordinates in $S^{\mathrm{c}} := [n] \setminus S$:

$$\mu_{S^{\mathrm{c}}} = \int_L (\nu_y)_{S^{\mathrm{c}}} \ \mu_S(\mathrm{d}y). \tag{40}$$

Indeed, the mutual information in the projected mixture (40) is precisely the mutual information between X_S and $X_{S^{\mathrm{c}}}$. By the last inequality of Corollary 6.7, this is at most $\mathrm{DTC}(\mu)$. In particular, it is finite, and so $(\nu_y)_{S^{\mathrm{c}}} \ll \mu_{S^{\mathrm{c}}}$ for μ_S-almost every y because of (18).

Let $d_{S^{\mathrm{c}}}$ be the normalized Hamming average of the metrics d_{K_i} for $i \in S^{\mathrm{c}}$: that is, the analog of (1) in which the sum extends only over $i \in S^{\mathrm{c}}$ and with normalizing constant $|S^{\mathrm{c}}|$. Let $\overline{d_{S^{\mathrm{c}}}}$ be the transportation metric associated to $d_{S^{\mathrm{c}}}$, defined as in (2) with $d_{S^{\mathrm{c}}}$ in place of d_n.

For each $y \in L$, let ξ'_y be the product measure on $\prod_{i \in S^{\mathrm{c}}} K_i$ with the same marginals as $(\nu_y)_{S^{\mathrm{c}}}$. These measures satisfy the following estimates:

$$
\int_L \overline{d_{S^{\mathrm{c}}}}\big((\nu_y)_{S^{\mathrm{c}}}, \xi'_y\big) \, \mu_S(\mathrm{d}y) \leq \int_L \sqrt{\frac{1}{2n} \mathrm{TC}\big((\nu_y)_{S^{\mathrm{c}}}\big)} \, \mu_S(\mathrm{d}y) \quad \text{(by Corollary 7.2)}
$$

$$
\leq \sqrt{\frac{1}{2n} \int_L \mathrm{TC}\big((\nu_y)_{S^{\mathrm{c}}}\big) \, \mu_S(\mathrm{d}y)} \quad \text{(by Hölder's ineq.)}
$$

$$
= \sqrt{\frac{1}{2n} \mathrm{TC}(X_1 \, ; \, \ldots \, ; \, X_n \,|\, X_S)} \quad \text{(by Proposition 6.1)}
$$

$$
\leq \delta/\sqrt{2} < \delta \qquad \text{(by Lemma 8.3)}.
$$

To finish the construction: let ξ_y be any lift of ξ'_y to a product measure on $\prod_i K_i$, chosen measurably in y; let ρ_y be the Radon–Nikodym derivative $\mathrm{d}(\nu_y)_{S^{\mathrm{c}}}/\mathrm{d}\mu_{S^{\mathrm{c}}}$; and define a new measure μ_y on $\prod_i K_i$ by

$$\mu_y(\mathrm{d}x) := \rho_y(x_{S^{\mathrm{c}}}) \cdot \mu(\mathrm{d}x) \quad \text{for each } y.$$

These new measures satisfy $(\mu_y)_{S^{\mathrm{c}}} = (\nu_y)_{S^{\mathrm{c}}}$ for every y.

Observe that the average $\int \rho_y \, \mu_S(\mathrm{d}y)$ is equal to the Radon–Nikodym derivative of $\int (\nu_y)_{S^{\mathrm{c}}} \, \mu_S(\mathrm{d}y)$ with respect to $\mu_{S^{\mathrm{c}}}$, which is identically equal to 1, by (40). This fact and Fubini's theorem give

$$\int_L \mu_y(A) \, \mu_S(\mathrm{d}y) = \int_A \Big(\int_L \rho_y(x_{S^{\mathrm{c}}}) \, \mu_S(\mathrm{d}y) \Big) \, \mu(\mathrm{d}x) = \int_A 1 \, \mu(dx) = \mu(A)$$

for all measurable sets $A \subseteq \prod_i K_i$, so we have expressed μ itself as a mixture:

$$\mu = \int_L \mu_y \, \mu_S(\mathrm{d}y). \tag{41}$$

This new mixture has the same mutual information as the mixture in (40), because by (18) that mutual information may be written as

$$\int \mathrm{D}(\mu_y \, \| \, \mu) \, \mu_S(\mathrm{d}y),$$

and the integrand appearing here is equal to

$$\mathrm{D}(\mu_y \, \| \, \mu) = \int \frac{\mathrm{d}\mu_y}{\mathrm{d}\mu} \log \frac{\mathrm{d}\mu_y}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int \rho_y(x_{S^c}) \log \rho_y(x_{S^c}) \, \mu(\mathrm{d}x)$$

$$= \int \rho_y(x') \log \rho_y(x') \, \mu_{S^c}(\mathrm{d}x') = \mathrm{D}((\nu_y)_{S^c} \, \| \, \mu_{S^c}).$$

So the mutual information in the mixture (41) is at most $\mathrm{DTC}(\mu)$. On the other hand, since $|S| \leq \delta n$, we have

$$\int_L \overline{d_n}(\mu_y, \xi_y) \, \mu_S(\mathrm{d}y) \leq \frac{|S|}{n} + \frac{|S^c|}{n} \int_L \overline{d_{S^c}}\big((\mu_y)_{S^c}, \xi_y'\big) \, \mu_S(\mathrm{d}y) < 2\delta, \tag{42}$$

where the first inequality holds by lifting couplings of $(\mu_y)_{S^c}$ and ξ_y' arbitrarily to couplings of μ_y and ξ_y and integrating the inequality

$$d_n(x, x') \leq \frac{|S|}{n} + \frac{|S^c|}{n} d_{S^c}(x_{S^c}, x'_{S^c}) \quad \left(x, x' \in \prod_i K_i \right).$$

$\square$

Remark 8.4. The proof of Theorem A obtains the required mixture of near-products by conditioning on a small, carefully-chosen set S of the coordinates. In this respect, it is a version of the 'wringing method' introduced by Dueck [20] for a proof of the strong converse to the coding theorem for the multiple access channel. See also [2, Section 4] and [3, Section III] for further development of 'wringing' in the context of the multiple access channel. A very similar method is introduced independently in the recent works [10, 11] to decompose certain Gibbs measures arising from factor-graph models into mixtures of approximate 'pure states'. In those works, as in ours, a large collection of random variables with modest dependence is brought close to independence by conditioning on a few of their number. However, all of those papers obtain only decorrelations among bounded-size subfamilies of the random variables (usually just pairs). By contrast, we obtain an approximation in $\overline{d_n}$, which is qualitatively stronger because it provides a coupling under which the joint behaviour of all coordinates is controlled at once. As a result the specific shape of our estimates is different from those predecessors.

8.3. Bounding number of terms instead of mutual information

Theorem A provides a mixture of near-product measures with a bound on the mutual information in the mixture. We could be slightly more demanding and ask for a bound on the number of terms.

Here we present two results of this kind. The first is obtained by combining Theorem A with the following lemma about 'sampling' from a mixture.

Lemma 8.5. (Sampling from a mixture with bounded mutual information) Let (K, μ) be a probability space, and let μ be represented as a mixture $\int_L \mu_\bullet \, d\nu$. Let $\varepsilon \in (0, 1/2)$, let $L' \subseteq L$ be measurable with $\nu(L') > 1 - \varepsilon/2$, and assume that the mutual information $I := \mathrm{I}(\nu, \mu_\bullet)$ is finite. Finally, let $m := \lceil 16\varepsilon^{-2} \mathrm{e}^{16(I+1)/\varepsilon} \rceil$. Then there exist $y_1, \ldots, y_m \in L'$ such that

$$\left\| \frac{1}{m} \sum_{j=1}^m \mu_{y_j} - \mu \right\| < 3\varepsilon, \tag{43}$$

where $\| \cdot \|$ denotes the total variation norm.

This lemma is taken directly from [5, Proposition 7.2]. To prove it, the elements $y_1, \ldots, y_m$ are chosen independently at random from ν and shown to have the desired property with high probability. The details are similar to elementary proofs of the law of large numbers: a truncation followed by a variance estimate. In [5], the value I in this lemma is written as the ν-integral of the KL divergences $\mathrm{D}(\mu_\bullet \,\|\, \mu)$, but that integral is equal to $\mathrm{I}(\nu, \mu_\bullet)$ by (18). Also, in [5, Proposition 7.2] it is assumed that K is standard Borel, but a quick check shows that this assumption plays no role in the proof.

By applying Lemma 8.5 to the mixture in Theorem A, we obtain the following variant of that theorem.

Theorem 8.6. In the setting of Theorem A, fix parameters $\varepsilon, \delta > 0$ and let μ be a probability measure on $\prod_i K_i$. If $\mathrm{DTC}(\mu) \leq \delta^3 n$, then μ may be written as a mixture

$$\mu = \frac{1}{m}\mu_1' + \cdots + \frac{1}{m}\mu_m'$$

so that

(a) $m = O(\varepsilon^{-2}) \exp(O(\mathrm{DTC}/\varepsilon))$, and

(b) there are product measures $\xi_1', \ldots, \xi_m'$ on $\prod_i K_i$ such that

$$\frac{1}{m} \sum_{j=1}^m \overline{d_n}(\mu_j', \xi_j') < 3\varepsilon + 4\delta/\varepsilon.$$

Proof. Let

$$\mu = \int_L \mu_y \, \nu(\mathrm{d}y)$$

be the representation of μ promised by Theorem A. By part (b) of Theorem A and Markov's inequality, the set

$$L' := \big\{ y \in L : \overline{d_n}(\mu_y, \xi_y) < 4\delta/\varepsilon \big\}$$

has $\nu(L') > 1 - \varepsilon/2$. Therefore Lemma 8.5 gives $y_1, \ldots, y_m \in L'$ such that (43) holds, where m satisfies part (a) of the present theorem.

We have not quite found a representation of μ as a finite mixture yet, but (43) says we are close. Let

$$\gamma := \frac{1}{m} \sum_{j=1}^{m} \mu_{y_j}. \tag{44}$$

Since $\overline{d_n}$ is always bounded by total variation (see, for instance, [5, Subsection 4.2]), it follows from (43) that some coupling λ of μ and γ satisfies

$$\int d_n \, \mathrm{d}\lambda < 3\varepsilon.$$

For each $j = 1, 2, \ldots, m$, define

$$\rho_j := \mathrm{d}\mu_{y_j}/\mathrm{d}\gamma \quad \text{and} \quad \lambda_j(\mathrm{d}x, \mathrm{d}y) := \rho_j(y) \cdot \lambda(\mathrm{d}x, \mathrm{d}y),$$

and let μ'_j be the first marginal of λ_j. Then the average

$$\frac{1}{m} \sum_{j=1}^{m} \mu'_j(\mathrm{d}x) \tag{45}$$

is the first marginal of

$$\frac{1}{m} \sum_{j=1}^{m} \lambda_j(\mathrm{d}x, \mathrm{d}y) = \Big(\frac{1}{m} \sum_{j=1}^{m} \rho_j(y) \Big) \cdot \lambda(\mathrm{d}x, \mathrm{d}y) = \lambda(\mathrm{d}x, \mathrm{d}y),$$

where the second equality holds because of (44). Therefore the average (45) equals μ exactly.

To finish, let us prove part (b) of the present theorem for these measures μ'_j. Let $\xi'_j := \xi_{y_j}$ for each j. By construction, λ_j is a coupling of μ'_j and μ_{y_j} for each j, and therefore

$$\frac{1}{m} \sum_{j=1}^{m} \overline{d_n}(\mu'_j, \xi'_j) \leq \frac{1}{m} \sum_{j=1}^{m} \Big[\int d_n \, \mathrm{d}\lambda_j + \overline{d_n}(\mu_{y_j}, \xi_{y_j}) \Big]$$

$$= \int d_n \, \mathrm{d}\lambda + \frac{1}{m} \sum_{j=1}^{m} \overline{d_n}(\mu_{y_j}, \xi_{y_j})$$

$$< 3\varepsilon + 4\delta/\varepsilon,$$

where the last line follows by our initial choice of λ and the definition of L'. $\qquad\square$

By taking $\varepsilon := \sqrt{\delta}$ in Theorem 8.6, we obtain $m = O(\delta^{-1}) \exp(\mathrm{DTC}(\mu)/\sqrt{\delta})$ and

$$\frac{1}{m} \sum_{j=1}^{m} \overline{d_n}(\mu'_j, \xi'_j) = O(\sqrt{\delta}).$$

I do not know whether this dependence on δ can be qualitatively improved.

A different approach to obtaining a finite mixture is available if each K_i is finite, but it gives bounds that depend on their cardinalitites. Suppose for simplicity that these cardinalities are all at most k. Then we can make the following slight adjustment within the proof of Theorem A itself. Consider again the mixture (39). If we simply take this for the mixture produced by the theorem (rather than projecting to K_{S^c}, bounding the mutual information, and lifting back to $K_{[n]}$), then a different bound results. The number of terms in the mixture (39) is at most

$$|L| = \prod_{i \in S} |K_i| \leq k^{\mathrm{DTC}(\mu)/\delta^2}.$$

Now, for each $y \in L$, let ξ_y be the product measure on $\prod_i K_i$ with the same marginals as ν_y. As in the proof of Theorem A, we have the estimates

$$\int_L \overline{d_n}(\nu_y, \xi_y)\, \mu_S(\mathrm{d}y) \leq \int_L \sqrt{\frac{1}{2n}\mathrm{TC}(\nu_y)}\, \mu_S(\mathrm{d}y)$$

$$\leq \sqrt{\frac{1}{2n}\mathrm{TC}(X_1\,;\,\ldots\,;\,X_n\,|\,X_S)} \leq \delta/\sqrt{2} < \delta.$$

Thus we obtain the following alternative to Theorem A.

Theorem 8.7. Assume that each K_i has cardinality at most k. Fix a parameter $\delta > 0$. Then any $\mu \in \mathrm{Prob}(\prod_i K_i)$ may be written as a mixture

$$\mu = p_1\mu_1 + \cdots + p_m\mu_m \tag{46}$$

in which

(a) $m \leq k^{\mathrm{DTC}(\mu)/\delta^2}$,

(b) there are product measures ξ_j on $\prod_i K_i$, $j = 1, 2, \ldots, m$, such that

$$\sum_{j=1}^{m} p_j \cdot \overline{d_n}(\mu_j, \xi_j) < \delta.$$

Remark 8.8. Note that Theorem 8.7 can be applied for any value of $\mathrm{DTC}(\mu)$, and gives a nontrivial conclusion provided $\mathrm{DTC}(\mu) \ll \delta^2 n$. This is slightly less restrictive than Theorem A itself, which requires $\mathrm{DTC}(\mu) \leq \delta^3 n$. The extra power of δ is needed in the proof of Theorem A because the cardinality $|S|$ shows up in the estimate (42), and we need this to be comparable to the integral of $\overline{d_{S^c}}\big((\mu_y)_{S^c}, \xi'_y\big)$ in that estimate. Estimate (42) is not used in the proof of Theorem 8.7.

9. DIRECTIONS FOR FURTHER RESEARCH

9.1. 'Very small' alphabets

Theorems A and 8.6 do not depend on $\max_i |K_i|$ — indeed, they allow infinite alphabets. But Theorem 8.7 does depend on $\max_i |K_i|$.

One can also ask after 'very small' alphabets. This could mean, for instance, that $K_i = \{0,1\}$ for each i, and also that μ is very biased in every coordinate, say

$$\mu\{x_i = 1\} \le p \quad \forall i,$$

where $p \in (0,1)$ is small.

Question 9.1. Do our estimates in Part II have any natural refinements if one also allows them to depend on p?

9.2. Finer approximations than in transportation

As discussed in the Introduction, our stability results for TC and DTC are incomplete. They promise a certain structure only up to a perturbation in $\overline{d_n}$, but neither TC nor DTC is uniformly continuous for that kind of perturbation (TC does satisfy Lemma 7.3, but that estimate depends on the size of the alphabets).

Question 9.2. Is there a finer kind of approximation than $\overline{d_n}$ which enables a complete description of the structure behind small values of TC or DTC?

The simplest possibility to consider would be approximation in total variation, but I expect this is too strong.

9.3. Gaps in other entropy inequalities

The facts that TC and DTC are non-negative can both be subsumed into a much more general inequality of Shearer from [9]. If $\mathcal{S}$ is a family of subsets of $[n]$ with the property that every member of $[n]$ lies in at least k members of $\mathcal{S}$, then Shearer's inequality asserts that

$$H(X_1, \dots, X_n) \le \frac{1}{k} \sum_{S \in \mathcal{S}} H(X_i : \; i \in S). \tag{47}$$

Han [29] established this previously in case $\mathcal{S} = \binom{[n]}{\ell}$ for some ℓ. Among these special cases, TC is the gap in (47) when $\mathcal{S} = \binom{[n]}{1}$. After a simple application of the chain rule, DTC is the gap in (47) when $\mathcal{S} = \binom{[n]}{n-1}$ (up to a normalization by $n-1$). Several other relations among the gaps in Han's inequalities were explored by Fujishige [26].

By now, Shearer's inequality has opened a rich vein of applications in combinatorics: see, for instance, [43]. More recently it has been refined further by Madiman and Tetali [32] and Balister and Bollobás [7]. Madiman and Tetali's paper gives a more careful overview of other work in this direction. From any of these inequalities, one can define a new notion of multi-variate correlation by considering the gap between the two sides. Madiman and Tetali also explicitly introduce the gaps in their inequalities in [32, Section VII], and establish some identities between them, generalizing Fujishige's work.

As far as I know these investigations do not include most of our work from Part I above, let alone Part II. Even the fact of monotonicity, which generalizes non-negativity, does not seem to be in the literature.

Question 9.3. Do the inequalities and identities from Part I have natural generalizations to the gaps in some of the other inequalities mentioned above?

Question 9.4. Do any of those other gaps enjoy stability or structural results of a similar flavour to Theorem A?

It would be especially interesting if the right choice of gap could be used to capture other structural features than near-products or their low-information mixtures.

All of the entropy inequalities mentioned in the previous subsection are 'Shannon inequalities', meaning that they are corollaries of the strong subadditivity of Shannon entropy. More recently, other 'non-Shannon inequalities' have also been investigated: see [33, 36, 49] for an introduction to these. The paper [25] uses both Shannon and non-Shannon inequalities to derive necessary conditions for the existence of a joint probability distribution given some constraints on its marginals, and the paper [18] applies some non-Shannon inequalities to a network information theory model.

Question 9.5. Do any of these non-Shannon inequalities admit natural stability results?

REFERENCES

[1] S. A. Abdallah and M. D. Plumbley: Predictive Information, Multi-Information and Binding Information. Technical Report C4DM-TR-10-10, Queen Mary University of London, 2010.

[2] R. Ahlswede: An elementary proof of the strong converse theorem for the multiple-access channel. J. Combin. Inform. System Sci. *7* (1982), 3, 216–230.

[3] R. Ahlswede: The rate-distortion region for multiple descriptions without excess rate. IEEE Trans. Inform. Theory *31* (1985), 6, 721–726. DOI:10.1109/tit.1985.1057102

[4] T. Austin: The structure of low-complexity Gibbs measures on product spaces. Ann. Probab. *47* (2019), 6, 4002–4023. DOI:10.1214/19-aop1352

[5] T. Austin: Measure concentration and the weak Pinsker property. Publ. Math. Inst. Hautes Etudes Sci. *128* (2018), 1–119. DOI:10.1007/s10240-018-0098-3

[6] N. Ay, E. Olbrich, N. Bertschinger, and J. Jost: A unifying framework for complexity measures of finite systems. Working Paper 06-08-028, Santa Fe Institute, 2006.

[7] P. Balister and B. Bollobás: Projections, entropy and sumsets. Combinatorica *32* (2012), 2, 125–141. DOI:10.1007/s00493-012-2453-1

[8] S. Chatterjee and A. Dembo: Nonlinear large deviations. Adv. Math. *299* (2016), 396–450. DOI:10.1016/j.aim.2016.05.017

[9] F. R. K. Chung, R. L. Graham, P. Frankl, and J. B. Shearer: Some intersection theorems for ordered sets and graphs. J. Combin. Theory Ser. A *43* (1986), 1, 23–37. DOI:10.1016/0097-3165(86)90019-1

[10] A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborová: Information-theoretic thresholds from the cavity method. Adv. Math. *333* (2018), 694–795. DOI:10.1016/j.aim.2018.05.029

[11] A. Coja-Oghlan and W. Perkins: Bethe states of random factor graphs. Comm. Math. Phys. *366* (2019), 1, 173–201. DOI:10.1007/s00220-019-03387-7

[12] T. M. Cover and J. A. Thomas: Elements of Information Theory. Second edition. Wiley-Interscience, John Wiley and Sons, Hoboken, NJ 2006.

[13] G. Crooks: On Measures of Entropy and Information. Technical note.

[14] I. Csiszár: Sanov property, generalized *I*-projection and a conditional limit theorem. Ann. Probab. *12* (1984), 3, 768–793. DOI:10.1214/aop/1176993227

[15] I. Csiszár and P. Narayan: Secrecy capacities for multiple terminals. IEEE Trans. Inform. Theory *50* (2004), 12, 3047–3061. DOI:10.1109/tit.2004.838380

[16] A. Dembo and O. Zeitouni: Large deviations techniques and applications. Second edition. Springer-Verlag, Stochastic Modelling and Applied Probability *38*, Berlin 2010. DOI:10.1007/978-1-4612-5320-4

[17] R. L. Dobrušin: A general formulation of Shannon's fundamental theorem in the theory of information. Dokl. Akad. Nauk SSSR *126* (1959), 474–477.

[18] R. Dougherty, C. Freiling, and K. Zeger: Networks, matroids, and non-Shannon information inequalities. IEEE Trans. Inform. Theory *53* (2007), 6, 1949–1969. DOI:10.1109/tit.2007.896862

[19] R. M. Dudley: Real Analysis and Probability. Cambridge University Press, Cambridge Studies in Advanced Mathematics *74*, Cambridge 2002. DOI:10.1017/cbo9780511755347

[20] G. Dueck: The strong converse of the coding theorem for the multiple-access channel. J. Combin. Inform. System Sci. *6* (1981), 3, 187–196.

[21] R. Eldan: Gaussian-width gradient complexity, reverse log-sobolev inequalities and non-linear large deviations. Geometr. Funct. Anal. *28* (2018), 6, 1548–1596.

[22] R. Eldan and R. Gross: Exponential random graphs behave like mixtures of stochastic block models. Preprint, available online at arXiv.org: 1707.01227

[23] R. Eldan and R. Gross: Decomposition of mean-field Gibbs distributions into product measures. Electron. J. Probab. *23* (2018), 35, 24.

[24] D. Ellis, E. Friedgut, G. Kindler, and A. Yehudayoff: Geometric stability via information theory. Discrete Anal. *10* (2016), 28 pp. DOI:10.19086/da.784

[25] T. Fritz and R. Chaves: Entropic inequalities and marginal problems. IEEE Trans. Inform. Theory *59* (2013), 2, 803–817. DOI:10.1109/tit.2012.2222863

[26] S. Fujishige: Polymatroidal dependence structure of a set of random variables. Inform. Control *39* (1978), 1, 55–72. DOI:10.1016/s0019-9958(78)91063-x

[27] I. Gelfand, A. Kolmogorov, and I. Yaglom: On the general definition of the quantity of information. Doklady Akad. Nauk SSSR *111* (1956), 4, 745–748.

[28] T. S. Han: Linear dependence structure of the entropy space. Inform. Control *29* (1975), 4, 337–368. DOI:10.1016/s0019-9958(75)80004-0

[29] T. S. Han: Nonnegative entropy measures of multivariate symmetric correlations. Inform. Control *36* (1978), 2, 133–156. DOI:10.1016/s0019-9958(78)90275-9

[30] A. N. Kolmogorov: On the shannon theory of information transmission in the case of continuous signals. IEEE Trans. Inform. Theory *IT-2* (1956), 102–108. DOI:10.1016/s0019-9958(78)90275-9

[31] M. Ledoux: On Talagrand's deviation inequalities for product measures. ESAIM Probab. Statist. *1* (1995/97), 63–87. DOI:10.1051/ps:1997103

[32] M. Madiman and P. Tetali: Information inequalities for joint distributions, with interpretations and applications. IEEE Trans. Inform. Theory *56* (2010), 6, 2699–2713. DOI:10.1109/tit.2010.2046253

[33] K. Makarychev, Y. Makarychev, A. Romashchenko, and N. Vereshchagin: A new class of non-Shannon-type inequalities for entropies. Commun. Inf. Syst. *2* (2002), 2, 147–165. DOI:10.4310/cis.2002.v2.n2.a3

[34] K. Marton: A simple proof of the blowing-up lemma. IEEE Trans. Inform. Theory *32* (1986), 3, 445–446. DOI:10.1109/tit.1986.1057176

[35] K. Marton: Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. Ann. Probab. *24* (1996), 2, 857–866. DOI:10.1214/aop/1039639365

[36] F. Matúš: Two constructions on limits of entropy functions. IEEE Trans. Inform. Theory *53* (2007), 1, 320–330. DOI:10.1109/tit.2006.887090

[37] C. McDiarmid: On the method of bounded differences. In: Surveys in Combinatorics, Norwich 1989, London Math. Soc. Lecture Note Ser. *141*, Cambridge Univ. Press, Cambridge 1989, pp. 148–188. DOI:10.1017/cbo9781107359949.008

[38] W. J. McGill: Multivariate information transmission. Trans. I.R.E. *PGIT-4* (1954), 93–111. DOI:10.1109/tit.1954.1057469

[39] J. Pearl and A. Paz: Graphoids: a graph-based logic for reasoning about relevance relations. In: Advances in Artificial Intelligence – II (B. Du Boulay, D. Hogg, and L. Steels, eds.), North Holland, Amsterdam 1987, pp. 357–363.

[40] A. Perez: Information theory with an abstract alphabet. Generalized forms of McMillan's limit theorem for the case of discrete and continuous times. Theory Probab. Appl. *4* (1959), 99–102. DOI:10.1137/1104007

[41] A. Perez: ϵ-admissible simplifications of the dependence structure of a set of random variables. Kybernetika *13* (1977), 6, 439–449.

[42] M. S. Pinsker: Information and information Stability of Random Variables and Processes. Holden-Day, Inc., San Francisco 1964.

[43] J. Radhakrishnan: Entropy and counting. In: Computational Mathematics, Modelling and Applications (IIT Kharagpur, Golden Jubilee Volume) (J. Mishra, ed.), Narosa Publishers, 2001, pp. 146–168.

[44] E. Schneidman, S. Still, M. J. Berry, and W. Bialek: Network information and connected correlations. Phys. Rev. Lett. *91* (2003), 238701. DOI:10.1103/physrevlett.91.238701

[45] M. Studený and J. Vejnarová: The multiinformation function as a tool for measuring stochastic dependence. In: Proc. NATO Advanced Study Institute on Learning in Graphical Models, Kluwer Academic Publishers, Norwell 1998, pp. 261–297. DOI:10.1007/978-94-011-5014-9_10

[46] N. Timme, W. Alford, B. Flecker, and J. M. Beggs: Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. J. Comput. Neurosci. *36* (2014), 2, 119–140. DOI:10.1007/s10827-013-0458-4

[47] S. Watanabe: Information theoretical analysis of multivariate correlation. IBM J. Res. Develop. *4* (1960), 66–82. DOI:10.1147/rd.41.0066

[48] J. Yan: Nonlinear large deviations: beyond the hypercube. Preprint, available online at arXiv.org: 1703.08887

[49] Z. Zhang and R. W. Yeung: On characterization of entropy function via information inequalities. IEEE Trans. Inform. Theory *44* (1998), 4, 1440–1452. DOI:10.1109/18.681320

Tim Austin, UCLA Mathematics Department, Box 951555, Los Angeles CA 90095. U. S. A.

e-mail: tim@math.ucla.edu