

PREDICTIVE INFORMATION,
MULTI-INFORMATION, AND BINDING
INFORMATION

Samer Abdallah and Mark Plumbley

Centre for Digital Music,
Queen Mary, University of London
Technical Report C4DM-TR10-10
Version 1.1 – December 9, 2010

Abstract

We introduce an information theoretic measure of dependency between multiple random variables, called ‘binding information’ and compare it with several previously proposed measures of statistical complexity, including excess entropy, Crutchfield *et al*’s entropy of causal states, Bialek *et al*’s predictive information, Dubnov’s information rate, and the multi-information. We discuss and prove some of the properties of the binding information, particularly in relation to the multi-information, and show that, in the case of sets of binary random variables, the processes which maximises binding information are the ‘parity’ processes. Finally we discuss some of the implications this has for the use of the binding information as a measure of complexity.

Predictive information, multi-information, and binding information

Samer Abdallah (samer.abdallah@elec.qmul.ac.uk)
Mark Plumbley (mark.plumbley@elec.qmul.ac.uk)

December 9, 2010

Abstract

We introduce an information theoretic measure of dependency between multiple random variables, called ‘binding information’ and compare it with several previously proposed measures of statistical complexity, including excess entropy, Crutchfield *et al*’s entropy of causal states, Bialek *et al*’s predictive information, Dubnov’s information rate, and the multi-information. We discuss and prove some of the properties of the binding information, particularly in relation to the multi-information, and show that, in the case of sets of binary random variables, the processes which maximises binding information are the ‘parity’ processes. Finally we discuss some of the implications this has for the use of the binding information as a measure of complexity.

1 Introduction

The concepts of ‘structure’, ‘pattern’ and ‘complexity’ are relevant in many fields of inquiry: physics, biology, cognitive sciences, machine learning, the arts and so on; but are vague enough to resist being quantified in a single definitive manner. One approach is to attempt to characterise them in statistical terms, for *distributions* over configurations of some system (that is, for a statistical ensemble rather than particular members of the ensemble) using the tools of information theory [1]. This approach has been taken by several researchers (e.g. [2, 3, 4, 5, 6, 7, 8]) and is the one we adopt here.

By way of a brief reminder of some of the relevant quantities of information theory and how they are related, the Venn diagram visualisation (formalised in [9]) of entropies, conditional entropies, mutual informations, and conditional mutual informations is illustrated for three variables in Fig. 1. Similar diagrams will be used to illustrate the various information measures that will be reviewed or defined in later sections.

In previous work, we defined the *predictive information rate* (PIR) [10] of a sequential random process as the average information in one observation about future observations yet to me made *given* the observations made so far; thus, it quantifies the *new* information in observations made as part of a sequence. The PIR captures a dimension of temporal dependency structure that is not

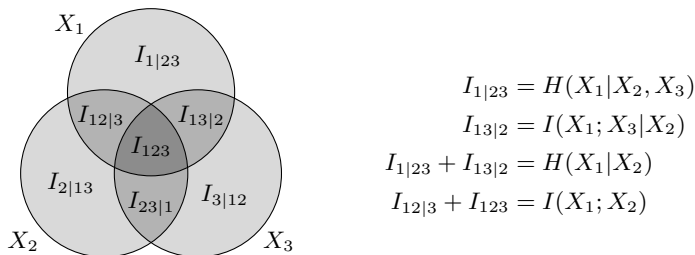


Figure 1: Venn diagram visualisation of entropies and mutual informations for three random variables X_1 , X_2 and X_3 . The areas of the three circles represent $H(X_1)$, $H(X_2)$ and $H(X_3)$ respectively. The total shaded area is the joint entropy $H(X_1, X_2, X_3)$. The central area I_{123} is the co-information [11]. Some other information measures are indicated in the legend.

accounted for by previously proposed measures that, loosely speaking, all focus on *redundancy*, or the extent to which different parts of a system all convey the same information. In this report, we propose a measure of statistical structure that is based on the PIR but is applicable to arbitrary countable sets of random variables and not just stationary sequences.

Overview We begin by reviewing a number of earlier proposals for measures of structure and complexity, and the PIR, in §2. Then, in §3, we define the *binding information* as the extensive counterpart of the PIR applicable to arbitrary countable sets of random variables. Some exploratory data illustrating binding information in relation to other information measures is presented (§4). These data suggest various constraints on the binding information, some which are proved in §5. We investigate some of the properties of processes that maximise binding information in §6, showing that for finite sets of binary random variables, the ‘parity’ processes (to be defined) are the unique processes that maximise binding information. Finally, in §7, we conclude with some observations about binding information as a measure of statistical complexity and its relationship to the alternatives.

Notational conventions In the following, if we have a random process X indexed by a set \mathcal{A} , and $\mathcal{B} \subseteq \mathcal{A}$, then $X_{\mathcal{B}}$ denotes the compound random variable (random ‘vector’) formed by taking X_{α} for each $\alpha \in \mathcal{B}$. (In order to be able to write the components of the vector in a definite order, the set \mathcal{A} must be equipped with a total order, but this need not have any further significance.) $|\mathcal{B}|$ denotes the cardinality of \mathcal{B} . The set of integers from M to N inclusive will be written $M..N$, and \setminus will denote the set difference operator, so, for example, $X_{1..4 \setminus \{2\}} \equiv (X_1, X_3, X_4)$.

2 Background

Much of the previous work on the quantification of statistical structure has been done with reference to infinite stationary sequences of random variables. Let $(\dots, X_{-1}, X_0, X_1, \dots)$ be such a sequence, infinite in both directions. Suppose that for each $t \in \mathbb{Z}$, the random variable X_t takes values in a discrete set \mathcal{X} , and let μ be the associated shift-invariant probability measure. Stationarity implies that the probability distribution associated with any finite contiguous block of N variables $(X_{t+1}, \dots, X_{t+N})$ is independent of t and therefore the joint entropy of these variables depends only on N . Hence we can define a shift-invariant block entropy function $H : \mathbb{N} \rightarrow [0, \infty)$

$$H(N) \triangleq H(X_1, \dots, X_N) = \sum_{\mathbf{x} \in \mathcal{X}^N} -p_\mu^N(\mathbf{x}) \log p_\mu^N(\mathbf{x}), \quad (1)$$

where $p_\mu^N : \mathcal{X}^N \rightarrow [0, 1]$ is the unique probability mass function for any N consecutive variables in the sequence, $p_\mu^N(\mathbf{x}) \triangleq \Pr(X_1 = x_1 \wedge \dots \wedge X_N = x_N)$.

A number of quantities can be defined in terms of the block entropy $H(N)$. Firstly, the entropy rate h_μ is defined as

$$h_\mu \triangleq \lim_{N \rightarrow \infty} \frac{H(N)}{N}. \quad (2)$$

As is well known [1], the entropy rate can also be defined equivalently as

$$h_\mu = \lim_{N \rightarrow \infty} H(N) - H(N-1). \quad (3)$$

Both expressions yield the same value, but the latter often converges to its limit faster than the former. The entropy rate gives a measure of the overall randomness of the process.

Excess entropy The block entropy function $H(\cdot)$ can also be used to express the mutual information between two contiguous segments of the sequence, i.e., two blocks with no gap between them. If the first is of length N and the second of M , then their mutual information is

$$I(X_{-N:-1}; X_{0:M-1}) = H(N) + H(M) - H(N+M), \quad (4)$$

where $X_{t:t'}$ is an abbreviation for $(X_t, \dots, X_{t'})$. If we let both block lengths N and M tend to infinity, we obtain what Crutchfield and Packard [12] called the *excess entropy*, and Grassberger [2] termed the *effective measure complexity* (EMC): it is the amount of information about the infinite future that can be obtained, on average, by observing the infinite past:

$$E = \lim_{N \rightarrow \infty} 2H(N) - H(2N). \quad (5)$$

It can also be expressed in terms of the $h_\mu(N)$ defined by Crutchfield [13] as $h_\mu(N) \triangleq H(X_N | X_{1:N-1}) = H(N) - H(N-1)$, which can be thought of as an estimate of the entropy rate obtained from the finite dimensional marginal

distribution p_μ^N . Clearly, $h_\mu = \lim_{N \rightarrow \infty} h_\mu(N)$. Crutchfield *et al.* [3] define the excess entropy in terms of $h_\mu(\cdot)$ as follows

$$E \triangleq \sum_{M=1}^{\infty} (h_\mu(M) - h_\mu), \quad (6)$$

but the result is equivalent to the mutual information between semi-infinite halves of the process (5).

Predictive information Grassberger [2] and others [4, 14] have commented on the manner in which $h_\mu(N)$ approaches its limit h_μ , noting that in certain types of random process with long-range correlations, the convergence can be so slow that the excess entropy is infinite, and that this is indicative of a certain kind of complexity. This phenomenon was examined in more detail by Bialek *et al.* [8], who defined the *predictive information* $\mathcal{I}_{\text{pred}}(N)$ as the mutual information between a block of length N and the infinite future following it:

$$\mathcal{I}_{\text{pred}}(N) \triangleq \lim_{M \rightarrow \infty} H(N) + H(M) - H(N + M). \quad (7)$$

Bialek *et al.* showed that even if $\mathcal{I}_{\text{pred}}(N)$ diverges as N tends to infinity, the *manner* of its divergence reveals something about the learnability of the underlying random process. Bialek *et al.* also emphasise that $\mathcal{I}_{\text{pred}}(N)$ is the *sub-extensive* component of the entropy: if the entropy rate h_μ is the intensive counterpart of the asymptotically extensive entropy $H(N)$, and Nh_μ is thought of as the purely extensive component of the entropy (i.e. the part that grows linearly with N), then $\mathcal{I}_{\text{pred}}(N)$ is the correction that gives

$$H(N) = Nh_\mu + \mathcal{I}_{\text{pred}}(N). \quad (8)$$

From this we can see that the sum of the first N terms of (6), which comes to $H(N) - Nh_\mu$, is equal to $\mathcal{I}_{\text{pred}}(N)$.

Multi-information Dubnov's [15] *information rate* was defined as the mutual information between the 'past' and the 'present' of a stationary random process. Dubnov considered a semi-infinite sequence of random variables (X_1, X_2, \dots) , and so defined the information rate at time $t = N$ as

$$\rho(X_{1:N}) \triangleq I(X_{1:N-1}; X_N) = H(N-1) + H(1) - H(N). \quad (9)$$

Dubnov pointed out that the information rate is equal to the rate of growth of the *multi-information* [16], which is defined for any collection of N random variables (X_1, \dots, X_N) as

$$I(X_{1:N}) \triangleq -H(X_{1:N}) + \sum_{i=1}^N H(X_i). \quad (10)$$

For $N = 2$, the multi-information reduces to the mutual information $I(X_1; X_2)$, while for $N > 2$, $I(X_{1:N})$ continues to be a measure of dependence, being zero if and only if the variables are statistically independent, that is, have a

fully factorisable probability distribution. In terms of the multi-information, Dubnov’s information rate is

$$\rho(X_{1:N}) = I(X_{1:N}) - I(X_{1:N-1}). \quad (11)$$

Letting N tend to infinity, we can define what would, in analogy with (3), more properly be called the (asymptotic) *multi-information rate* as

$$\rho_\mu \triangleq \lim_{N \rightarrow \infty} I(X_{1:N}) - I(X_{1:N-1}), \quad (12)$$

which, as we can see by consulting (9) and (7), is just $\mathcal{I}_{\text{pred}}(1)$, which, in turn, is equal to $H(1) - h_\mu$. This is illustrated in fig. 2(d).

Erb and Ay [17] also studied the behaviour of the multi-information in the thermodynamic limit (that is, for an infinitely large, shift-invariant system) and their thermodynamic $I \triangleq \lim_{N \rightarrow \infty} I(X_{1:N})/N$, is the same as ρ_μ in the same way that (2) and (3) define the same quantity. In addition, Erb and Ay found a complementary relationship between the multi-information and the ‘finite volume’ approximation to the excess entropy found by summing the first N terms of (6), which as we have already noted is equal to $\mathcal{I}_{\text{pred}}(N)$; in the present terminology,

$$I(X_{1:N}) + \mathcal{I}_{\text{pred}}(N) = N\rho_\mu. \quad (13)$$

Comparing this with (8), we see that, as well as being the sub-extensive component of the entropy, $\mathcal{I}_{\text{pred}}(N)$ is also the sub-extensive component of the multi-information. Thus, all of the measures considered so far, being linearly dependent in various ways, are closely related.

State machine based measures Moving on from measures that can be defined entirely in terms of the block entropy function $H(\cdot)$, Grassberger [2] defined the *true measure complexity* (TMC) as the average amount of information that some agent, having observed the infinite past (\dots, X_{-2}, X_{-1}) , must store in order to make optimal predictions about the next variable in the sequence, X_0 . Grassberger considered various stochastic automata or state machines that could provide sufficiently concrete models for the hypothetical agent to compute the TMC for particular random processes, but these were not applicable to arbitrary random processes.

A general method for constructing the required automata was provided by Crutchfield and Young’s ϵ -machine reconstruction algorithm [3], and later Shalizi *et al.*’s CSSR (Causal State Splitting Reconstruction) algorithm [18]). These are based on the concept of *causal states* [3, 13], which are the equivalence classes of sequences of observations that lead to the same conditional probability distribution over possible futures. Briefly, if \mathbf{x}' and \mathbf{x}'' are two *observation histories*, that is, semi-infinite sequences stretching back from time t into the infinite past, and both histories \mathbf{x}' and \mathbf{x}'' lead to the same conditional probabilities over the infinite future:

$$\forall \mathbf{x} \in \mathcal{X}^*. \Pr(X_{t+1:\infty} = \mathbf{x} | X_{-\infty:t} = \mathbf{x}') = \Pr(X_{t+1:\infty} = \mathbf{x} | X_{-\infty:t} = \mathbf{x}''), \quad (14)$$

then \mathbf{x}' and \mathbf{x}'' belong to the same causal state. (\mathcal{X}^* denotes the set of infinite sequences of symbols from \mathcal{X} .) This means that, in order to make optimal predictions, an agent needs only to keep track of the current causal state and not

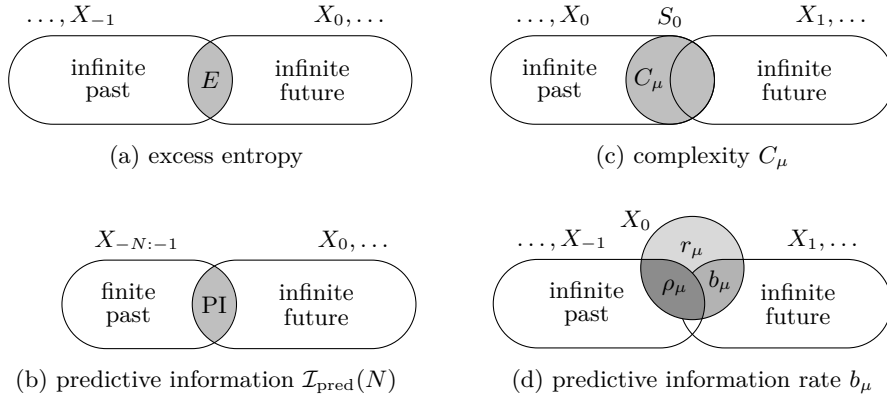


Figure 2: Venn diagram representation of several information measures for stationary random processes. Each circle or oval represents a random variable or sequence of random variables relative to time $t = 0$. Overlapped areas correspond to various mutual information as in Fig. 1. In (c), Crutchfield *et al*'s complexity C_μ is the entropy of the *causal state* S_0 , represented by the shaded circle, which is a many-to-one function of the infinite past, but retains all information relevant to predicting the future. In (d), the circle represents the ‘present’. Its total area is $H(X_0) = H(1) = \rho_\mu + r_\mu + b_\mu$, where ρ_μ is Dubnov’s (multi-)information rate, r_μ is the residual entropy rate, and b_μ is the predictive information rate. The entropy rate is $h_\mu = r_\mu + b_\mu$.

the entire observation history. An ϵ -machine, then, is a state machine whose states are the causal states and which emits an observable symbol from the alphabet \mathcal{X} for each transitions from one state to another. If the set of causal states is \mathcal{S} and the sequence of states occupied by the automaton is represented by a sequence of random variables $(\dots, S_{-1}, S_0, S_1, \dots)$, each taking values in \mathcal{S} , then this sequence forms a first-order Markov chain [19]. Crutchfield and Young [3] define their *statistical complexity* C_μ as the entropy of the stationary distribution over the causal states, which, for countable \mathcal{S} and any t , is

$$C_\mu = H(S_t) = \sum_{s \in \mathcal{S}} -\Pr(S_t = s) \log \Pr(S_t = s). \quad (15)$$

By the assumption of stationarity, $\Pr(S_t = s)$ is independent of t and can be computed from the transition probabilities of the Markov chain.

The causal state is a function of the observation history and so $H(S_t | X_{-\infty:t}) = 0$. Crutchfield and Feldman [20] show that $C_\mu \geq E$ (see fig. 2), and also that if X is a first order Markov chain, then the observable states *are* the causal states and $C_\mu = H(1)$.

C_μ can be interpreted as the average amount of information required, given a state of complete ignorance about the observable process X , to set the internal state of the ϵ -machine to pick-up the sequence from any given point in time and make optimal predictions about its future.

Predictive information rate In our previous work [10], we introduced the *instantaneous predictive information* (IPI), which, for a given sequence of observations assumed to come from a known random process, is the information in one observation about the infinite future given the infinite past. Suppose the random process X is observed up to time t as (\dots, x_{t-1}, x_t) . Let \overleftarrow{X}_t stand for the variables before time t , i.e., $(\dots, X_{t-2}, X_{t-1})$ and \overleftarrow{x}_t for their observed values. Similarly, let $\overrightarrow{X}_t = (X_{t+1}, X_{t+2}, \dots)$ represent the unknown future and let \overrightarrow{x} range over the set of possible futures \mathcal{X}^* (the set of infinite sequences of values from \mathcal{X}). Then the IPI at time t is

$$\mathcal{I}_t \triangleq \sum_{\overrightarrow{x} \in \mathcal{X}^*} P(\overrightarrow{x} | x_t, \overleftarrow{x}_t) \log \frac{P(\overrightarrow{x} | x_t, \overleftarrow{x}_t)}{P(\overrightarrow{x} | \overleftarrow{x}_t)}, \quad (16)$$

which should be read as the Kullback-Leibler divergence between the predictive distributions over possible futures before and after the latest observation $X_t = x_t$, both given \overleftarrow{x}_t , the sequence of observations before t ; hence, \mathcal{I}_t quantifies the *new* information gained about the future from the observation at time t . Note that \mathcal{I}_t is a function of both the current observation x_t and the observation history \overleftarrow{x}_t , and so, unlike the other measures considered so far, gives a dynamic (and causal) characterisation of a particular realisation of the random process, that is, the actually observed sequence as opposed to the statistical ensemble. When averaged over all possible realisations however, we obtain the *predictive information rate* (PIR), which *is* a property of the ensemble, and can be expressed as a conditional mutual information [1, Ch. 2]:

$$\underline{\mathcal{I}}_t \triangleq I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = H(\overrightarrow{X}_t | \overleftarrow{X}_t) - H(\overrightarrow{X}_t | X_t, \overleftarrow{X}_t). \quad (17)$$

The underline/overline notation follows that of [10, §3] and is intended to signify a two-fold expectation of the IPI considered as a function x_t and \overleftarrow{x}_t —first with respect to distribution of \overrightarrow{X}_t given $\overleftarrow{X}_t = \overleftarrow{x}_t$ and secondly with respect to distribution of possible histories \overleftarrow{X}_t . Equation (17) can be read as the average reduction in uncertainty about the future on learning X_t , given the past. However, due to the symmetry of the mutual information, it can also be written as $\underline{\mathcal{I}}_t = H(X_t | \overleftarrow{X}_t) - H(X_t | \overrightarrow{X}_t, \overleftarrow{X}_t)$. If, as we have been assuming so far, X is stationary, the first term is the entropy rate h_μ , but the second term is a quantity that does not appear to have been considered by other authors yet. It is the conditional entropy of one variable given *all* the others in the sequence, in the future as well as in the past. We call this the *residual entropy rate* r_μ , and define it as a limit:

$$r_\mu \triangleq \lim_{N \rightarrow \infty} H(X_{-N:N}) - H(X_{-N:-1}, X_{1:N}). \quad (18)$$

The second term, $H(X_{1:N}, X_{-N:-1})$, is the entropy of two non-adjacent blocks each of length N with a gap between them, and cannot be expressed as a function of block entropies alone. We will let b_μ denote the shift-invariant PIR. Thus, we have $b_\mu = h_\mu - r_\mu$. These relationships are illustrated in Fig. 2, along with several of the information measures we have discussed so far.

Measuring complexity As the reader may have guessed by the names chosen by their various proposers, many of the measures reviewed in the previous section were intended as measures of ‘complexity’, where ‘complexity’ is a quality that is somewhat open to interpretation [21, 6]; hence the variety of proposals. What is generally agreed, however, is that complexity should be low for systems that are deterministic or easy to compute or predict—‘ordered’—and also low for systems that are completely random and unpredictable—‘disordered’. A number of proposals we have not reviewed here (e.g. [5, 7]) attempt to achieve this by starting from some measure of ‘disorder’ D (usually an entropy or entropy rate), ascertaining a suitable maximum value D_{\max} , and constructing an expression which is zero for $D = 0$ and for $D = D_{\max}$, e.g., $D(D_{\max} - D)$. Clearly, this satisfies the required boundary conditions, but as Crutchfield and Feldman have argued persuasively [6, 22], measures that are simply functions of entropy or entropy rate are ‘over-universal’ in the sense that they fail to distinguish between the different strengths of temporal dependency—essentially, ‘structure’—that can be exhibited by systems at a given level of entropy rate.

The PIR is not simply a function of entropy rate; it satisfies the boundary conditions discussed above, but it does so in a different way from the other measures reviewed here. In our analysis of Markov chains [10], we found that processes which maximise the PIR are not those that maximise the multi-information rate ρ_μ , or, by extension, the excess entropy, which, in the case of first order Markov chains, is the same. Rather, the processes with maximal PIR have a certain kind of partial predictability that requires the observer continually to pay attention to the most recently observed values in order to make optimal predictions. And so, while Crutchfield *et al.* make a compelling case for the excess entropy E and the causal state entropy C_μ as measures of what would seem to correspond quite well with the concept of complexity, there is still room to suggest that the PIR captures a different and non-trivial aspect of temporal dependency structure that has not previously been examined.

3 Binding information

In this section, we address two questions. Firstly, if the predictive information rate (PIR) is a ‘rate’, what is it the rate of? Or more succinctly (not to mention grammatically), what is the extensive counterpart to the intensive PIR? Secondly, can the concept be extended to collections of random variables which are not organised sequentially?

When the PIR is accumulated over successive observations, one obtains a quantity which we call the *binding information*. To obtain an initial expression for it, we first reformulate the PIR so that it is applicable to a *finite* sequence of random variables (X_1, \dots, X_N) :

$$\bar{\mathcal{I}}_t(X_{1..N}) = I(X_t; X_{(t+1)..N} | X_{1..(t-1)}), \quad (19)$$

which is to be compared with the PIR for infinite sequences (17). Note that this is no longer shift-invariant and may depend on t . The binding information

$B(X_{1..N})$, then, is the sum

$$B(X_{1..N}) = \sum_{t=1}^N \bar{I}_t(X_{1..N}). \quad (20)$$

Expanding this sum in terms of conditional entropies yields

$$\begin{aligned} B(X_{1..N}) &= \sum_{t \in 1..N} I(X_t; X_{(t+1)..N} | X_{1..(t-1)}) \\ &= \sum_{t \in 1..N} H(X_{1..t}) - H(X_{1..(t-1)}) - H(X_t | X_{(t+1)..N}, X_{1..(t-1)}). \end{aligned} \quad (21)$$

Cancelling out entropies from successive terms of the sum and collecting the conditioning variables in the conditional entropy gives

$$B(X_{1..N}) = H(X_{1..N}) - \sum_{t=1}^N H(X_t | X_{1..N \setminus \{t\}}). \quad (22)$$

That is, the binding information is the joint entropy of all the variables minus the conditional entropy of each variable given the others. Like the multi-information, whose definition (10) is structurally similar in some ways, it measures the information that is common between a set of random variables, but in a different way. Referring to the Venn diagram visualisation of entropies (see Fig. 3), the binding information is the total area minus all the areas which are not overlapped with anything, and therefore measures the total area of overlap *without* multiply counting the areas that are overlapped. The multi-information, on the other hand, measures the overlap with repeated counting of areas that are multiply overlapped.

Though the binding information was derived by accumulating the PIR sequentially, the result is permutation invariant. This suggests that the concept might not be confined to sets of random variables configured one-dimensionally as sequence, but might also apply to arbitrary sets of random variables regardless of their topology. Accordingly, we formally define the binding information as follows:

Definition 1. *If $\{X_\alpha | \alpha \in \mathcal{A}\}$ is set of random variables indexed by a countable set \mathcal{A} , then their binding information is*

$$B(X_{\mathcal{A}}) \triangleq H(X_{\mathcal{A}}) - \sum_{\alpha \in \mathcal{A}} H(X_\alpha | X_{\mathcal{A} \setminus \{\alpha\}}). \quad (23)$$

This definition is quite general: it is applicable to any countable set of random variables, regardless of any topological structure that may be present in the indexing set \mathcal{A} and without requiring the process to be infinite and/or stationary. Since it can be expressed as a sum of (conditional) mutual informations, (21), it inherits a number of properties from the mutual information: it is (a) non-negative; (b) applicable to continuous-valued random variables as well as discrete-valued ones; and (c) invariant to invertible point-wise transformations of the variables; that is, if $Y_{\mathcal{A}}$ is a set of random variables taking values in \mathcal{Y} , and for all $\alpha \in \mathcal{A}$, there exists some invertible function $f_\alpha : \mathcal{X} \rightarrow \mathcal{Y}$ such that $Y_\alpha = f_\alpha(X_\alpha)$, then $B(Y_{\mathcal{A}}) = B(X_{\mathcal{A}})$.

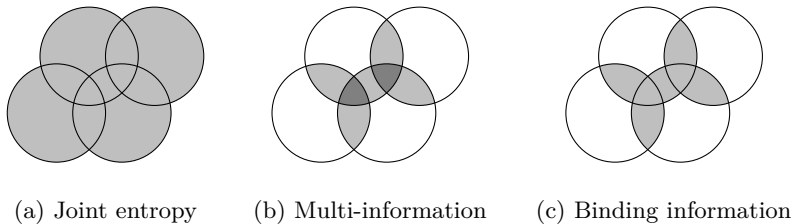


Figure 3: Illustration of binding information as compared with multi-information for a set of four random variables. In each case, the quantity is represented by the total amount of black ink, as it were, in the shaded parts of the diagram. Whereas the multi-information counts the triple-overlapped areas twice each, the binding information counts each overlapped areas just once.

Conditions for minimisation The binding information is zero for sets of independent random variables—the case of complete ‘disorder’—since in this case all the mutual informations in (21) are zero. The binding information is also zero when all variables have zero entropy, taking known, constant values and representing a certain kind of ‘order’. However, it is also possible to obtain low binding information for *random* systems which are nonetheless very ordered in a particular way. If for each pair of indices $\alpha, \alpha' \in \mathcal{A}$, there exists an invertible function $f_{\alpha'}^\alpha : \mathcal{X} \rightarrow \mathcal{X}$ such that $X_{\alpha'} = f_{\alpha'}^\alpha(X_\alpha)$, then there is, in effect, only one random variable: the state of the entire system can be read off from any one of its component variables. In this case, and the entropy of the whole system both equal the entropy of any one of its members: $\forall \alpha \in \mathcal{A}. B(X_{\mathcal{A}}) = H(X_{\mathcal{A}}) = H(X_\alpha)$, which is limited by the size of the alphabet \mathcal{X} . As we will see in §6, this is relatively low compared with what is possible for a system of N variables, since it is possible for the binding information to grow linearly with N . Thus, binding information is low for both highly ‘ordered’ and highly ‘disordered’ systems, but in this case, ‘highly ordered’ does *not* simply mean deterministic or known *a priori*: it means the whole is predictable from the smallest of its parts.

Relationship with disorder Like the predictive information rate in terms of which it was derived, the binding information is not simply a function of ‘disorder’ (entropy) and is not ‘over-universal’ in the sense of Crutchfield *et al.* [6, 22]. In fact, our analysis and experiments below suggest that the joint entropy, the binding information and the multi-information provide three distinct and functionally independent (within bounds) characterisations of randomness and interdependency, each of which may be relevant when studying a given random process.

In the following sections, we will often compare properties of the binding information with those of the multi-information, and so, we include here the definition of the multi-information written in the same terms, for arbitrary indexing sets:

$$I(X_{\mathcal{A}}) \triangleq -H(X_{\mathcal{A}}) + \sum_{\alpha \in \mathcal{A}} H(X_\alpha). \quad (24)$$

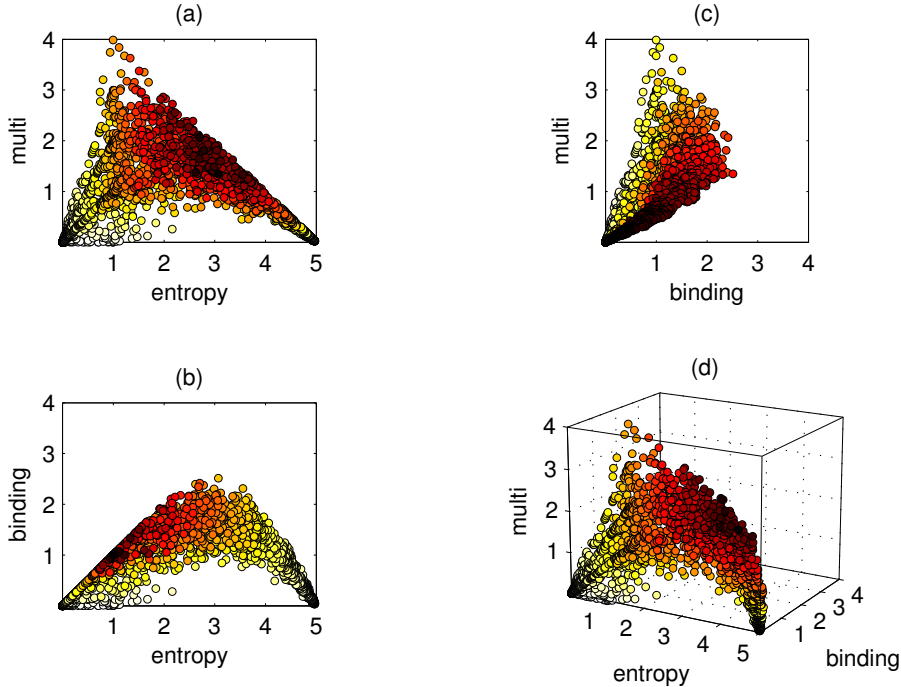


Figure 4: Joint entropy, binding information and multi-information plotted for 3200 systems of $N = 5$ binary random variables. The distribution for each system was sampled from a symmetric Dirichlet distribution with concentration parameters in the set $\{0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$. The shading of the points indicates one of the three measures; in (a) the binding information, (b) the entropy, (c) the multi-information and (d) the binding information.

4 Binding information in random energy models

Consider a system consisting of N binary random variables X_i for $i \in 1..N$. The system has 2^N distinct configurations and so the set of probability distributions over these states is the simplex in $2^N - 1$ dimensions. For small N , we can represent such probability distributions explicitly, and compute the total entropy, the multi-information, and the binding information numerically.

One way of generating distributions over the 2^N configurations is to sample them from a Dirichlet distribution with 2^N parameters $(\alpha_{\mathbf{x}})_{\mathbf{x} \in \{0,1\}^N}$. This is essentially a *random energy model* [23] where the effective energy of each configuration is sampled independently from a non-Gaussian distribution: if the energy of the configuration $\mathbf{x} \in \{0,1\}^N$ is $E_{\mathbf{x}}$, then $E_{\mathbf{x}}$ is distributed as the negative logarithm of a standard Gamma random variable with shape parameter $\alpha_{\mathbf{x}}$ and unit scale parameter:

$$E_{\mathbf{x}} \sim -\log \mathcal{G}(\alpha_{\mathbf{x}}, 1). \quad (25)$$

If the probabilities $P_{\mathbf{x}}$ are then set proportional to $e^{-E_{\mathbf{x}}}$, we recover the standard method for sampling from a Dirichlet distribution [24].

To generate the plots in Fig. 4, N was set at 5 and distributions were sampled from several Dirichlet distributions with symmetric parameters, $\alpha_{\mathbf{x}} = \alpha_0$ for all $\mathbf{x} \in \{0, 1\}^5$, with α_0 taking a range of positive values. Large values of α_0 tend to yield more uniform distributions over the configurations of the system, while small values tend to yield distributions which favour a few particular configurations. The joint entropy $H(X_{1..N})$, multi-information $I(X_{1..N})$ and binding information $B(X_{1..N})$ were computed for each sampled distribution and plotted in a 3-dimensional scatter plot. Observation of this and other similar plots suggested that various constraints exist between the information quantities, some of which were confirmed, as we will see in §5. The entropy ranges from 0 to $N = 5$ bits. The multi-information reaches a maximum of approximately 4 bits, and the binding information peaks at around 2.5 bits.

Viewed in the H - I plane, the points appear to be scattered over a triangular region with a peak near $(H, I) = (1, 4)$. This is consistent with provable upper bounds illustrated in red in Fig. 5 (see Theorem 1 in §5). Turning to the H - B plane, there appears to be an inverted-‘U’ shaped upper bound. The left-hand, rising segment roughly follows the line $B = H$, and does indeed reflect a provable upper bound (see Theorem 2 in §5). The right-hand, falling segment, however, does not represent an upper bound on the binding information, which we prove by counterexample in §6. It appears that sampling from a Dirichlet distribution in this way is extremely unlikely to result in those distributions which inhabit the upper-right region of the H - B plane. Finally, viewing the scatter plot in the B - I plane, we see what again seems to be a triangular region delineated by three constraints. All three of these turn out to be provable bounds, at least for all the values of N we were able to check (see Proposition 1).

5 Bounds on binding and multi-information

In this section we confine our attention to sets of discrete random variables taking values in a common alphabet containing K symbols, such as the system of binary variables described in the previous section. In this case, it is quite straightforward to derive upper bounds, as functions of the joint entropy, on both the multi-information and the binding information, and also upper bounds on multi-information and binding information as functions of each other.

Theorem 1. *If $\{X_\alpha | \alpha \in \mathcal{A}\}$ is a set of $|\mathcal{A}| = N$ random variables all taking values in a discrete set of cardinality K , then*

$$I(X_{\mathcal{A}}) \leq N \log K - H(X_{\mathcal{A}}) \quad (26)$$

$$\text{and } I(X_{\mathcal{A}}) \leq (N - 1)H(X_{\mathcal{A}}). \quad (27)$$

Proof. The multi-information is $I(X_{\mathcal{A}}) = \sum_{\alpha \in \mathcal{A}} H(X_\alpha) - H(X_{\mathcal{A}})$. Since each variable X_α can take one of only K values, $H(X_\alpha) \leq \log K$ for all $\alpha \in \mathcal{A}$. Therefore $\sum_{\alpha \in \mathcal{A}} H(X_\alpha) \leq N \log K$ and (26) follows directly. We also have, for all $\alpha \in \mathcal{A}$, $H(X_\alpha) \leq H(X_{\mathcal{A}})$, and so

$$I(X_{\mathcal{A}}) \leq NH(X_{\mathcal{A}}) - H(X_{\mathcal{A}}) = (N - 1)H(X_{\mathcal{A}}).$$

□

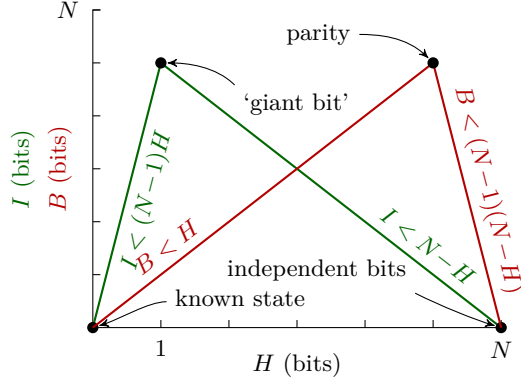


Figure 5: Upper bounds for multi-information $I(X_{1..N})$ (in green) and binding information $B(X_{1..N})$ (in red) for a system of $N = 6$ binary random variables. The labelled extremal points indicate identifiable distributions over the 2^N states that this system can occupy. ‘Known state’ means that the system is locked to just one configuration; ‘giant bit’ means that there are just two equiprobable configurations and one is the inverse of the other; ‘parity’ is one of the two parity processes to be described in Theorem 4; ‘independent’ is the system of independent unbiased random bits.

Theorem 2. *If $\{X_\alpha | \alpha \in \mathcal{A}\}$ is a set of $|\mathcal{A}| = N$ random variables all taking values in a discrete set of cardinality K , then*

$$B(X_{\mathcal{A}}) \leq H(X_{\mathcal{A}}) \quad (28)$$

$$\text{and } B(X_{\mathcal{A}}) \leq (N - 1)(N \log K - H(X_{\mathcal{A}})). \quad (29)$$

Proof. The first inequality comes directly from the definition of the binding information (23) or (22), since $H(X_\alpha | X_{\mathcal{A} \setminus \{\alpha\}}) \geq 0$ for any discrete random variable X_α . To obtain the second inequality, we expand the conditional entropies of (22):

$$\begin{aligned} B(X_{\mathcal{A}}) &= H(X_{\mathcal{A}}) - \sum_{\alpha \in \mathcal{A}} H(X_\alpha | X_{\mathcal{A} \setminus \{\alpha\}}) \\ &= H(X_{\mathcal{A}}) - \sum_{\alpha \in \mathcal{A}} [H(X_{\mathcal{A}}) - H(X_{\mathcal{A} \setminus \{\alpha\}})] \\ &= \sum_{\alpha \in \mathcal{A}} H(X_{\mathcal{A} \setminus \{\alpha\}}) - (N - 1)H(X_{\mathcal{A}}). \end{aligned}$$

But, for all α , $H(X_{\mathcal{A} \setminus \{\alpha\}}) \leq (N - 1) \log K$ bits, so

$$\begin{aligned} B(X_{\mathcal{A}}) &\leq N(N - 1) \log K - (N - 1)H(X_{\mathcal{A}}) \\ &= (N - 1)(N \log K - H(X_{\mathcal{A}})). \end{aligned}$$

□

These bounds restrict $I(X_{\mathcal{A}})$ and $B(X_{\mathcal{A}})$ to two triangular regions of the plane when plotted against the joint entropy $H(X_{\mathcal{A}})$, illustrated for $N = 2$, $K = 2$ in Fig. 5.

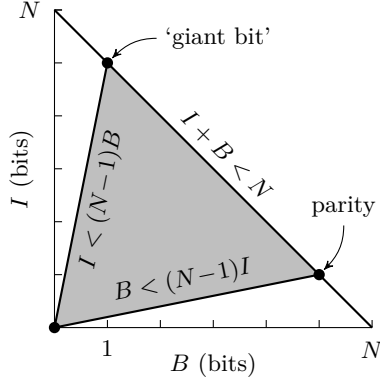


Figure 6: Constraints on the multi-information $I(X_{1..N})$ and the binding information $B(X_{1..N})$ for a system of $N = 6$ binary random variables. The labelled extremal points indicate identifiable distributions over the states that this system can occupy. ‘Giant bit’ means that there are just two equiprobable configurations and one is the inverse of the other; ‘parity’ is one of the two parity processes described in Theorem 4.

Next, we examine the bounds in the B - I plane suggested by the scatter plot of Fig. 4(c), proving the existence of one and finding strong evidence for the other two.

Theorem 3. *If $\{X_\alpha | \alpha \in \mathcal{A}\}$ is a set of $|\mathcal{A}| = N$ random variables all taking values in a discrete set of cardinality K , then*

$$B(X_{\mathcal{A}}) + I(X_{\mathcal{A}}) \leq N \log K. \quad (30)$$

Proof. Expanding the definitions of binding and multi-informations yields

$$\begin{aligned} B(X_{\mathcal{A}}) + I(X_{\mathcal{A}}) &= H(X_{\mathcal{A}}) - \sum_{\alpha \in \mathcal{A}} H(X_\alpha | X_{\mathcal{A} \setminus \{\alpha\}}) + \sum_{\alpha \in \mathcal{A}} H(X_\alpha) - H(X_{\mathcal{A}}) \\ &= \sum_{\alpha \in \mathcal{A}} H(X_\alpha) + H(X_{\mathcal{A} \setminus \{\alpha\}}) - H(X_{\mathcal{A}}) \\ &= \sum_{\alpha \in \mathcal{A}} I(X_\alpha; X_{\mathcal{A} \setminus \{\alpha\}}). \end{aligned}$$

Since the mutual information of any pair of variables is no more than either of their entropies, and $H(X_\alpha) \leq \log K$ for discrete random variables with K possible values, we obtain a chain of two inequalities:

$$\sum_{\alpha \in \mathcal{A}} I(X_\alpha; X_{\mathcal{A} \setminus \{\alpha\}}) \leq \sum_{\alpha \in \mathcal{A}} H(X_\alpha) \leq N \log K.$$

and the theorem is proved. \square

Proposition 1. *If $\{X_\alpha | \alpha \in \mathcal{A}\}$ is a set of $N = |\mathcal{A}|$ discrete random variables, then the following two inequalities hold:*

$$I(X_{\mathcal{A}}) \leq (N - 1)B(X_{\mathcal{A}}) \quad (31)$$

$$\text{and } B(X_{\mathcal{A}}) \leq (N - 1)I(X_{\mathcal{A}}). \quad (32)$$

We will not prove Proposition 1 here but we will sketch out a potential proof suggested by a result that can be obtained for $\mathcal{A} = \{1, 2, 3\}$. In this case, it is relatively easy to find that

$$\begin{aligned} 2B(X_{1..3}) - I(X_{1..3}) &= I(X_1; X_2|X_3) + I(X_1; X_3|X_2) + I(X_2; X_3|X_1), \\ 2I(X_{1..3}) - B(X_{1..3}) &= I(X_1; X_2) + I(X_1; X_3) + I(X_2; X_3). \end{aligned}$$

Since both quantities are sums of non-negative mutual informations or conditional mutual informations, the proposition is proved for $N = 3$. This suggests that it may be possible to achieve a similar decomposition for $N > 3$, but as the number of terms involved increases rapidly with N , it is much harder to identify by intuition alone which mutual information terms are required and a more systematic approach is needed. We constructed a numerical algorithm to express the two quantities $\eta_1 = (N - 1)B(X_{1..N}) - I(X_{1..N})$ and $\eta_2 = (N - 1)I(X_{1..N}) - B(X_{1..N})$ as weighted sums of the ‘level-specific’ dependency measures $\Delta(r)$ defined by Studený and Vejnarová [16]. These are defined as sums of mutual informations and conditional mutual informations and so are guaranteed to be non-negative. Thus, if η_1 and η_2 can be expressed as weighted sums of the $\Delta(r)$ such that the weights are all non-negative, then they too will be proved non-negative. We found that our algorithm produced rational non-negative weights for all values of N up to 37, at which point insufficient numerical precision became the limiting factor.

6 Maximising binding information

Now that we have established some constraints on the binding information in relation to the entropy and the multi-information, it is instructive to examine what kind of processes maximise the binding information and in particular, whether the absolute maximum of $(N - 1) \log K$ implied by Theorem 2 is attainable. The answer (for finite sets of discrete variables over a common alphabet) is surprisingly simple.

Theorem 4. *If $\{X_1, \dots, X_N\}$ is a set of binary random variables each taking values in $\{0, 1\}$, then the binding information $B(X_{1..N})$ is maximised by the two ‘parity processes’ $P_{2,0}^N$ and $P_{2,1}^N$. For $m \in \{0, 1\}$, the probability of observing $\mathbf{x} \in \{0, 1\}^N$ under each process is*

$$P_{2,m}^N(\mathbf{x}) = \begin{cases} 2^{1-N} & : \text{if } \left(\sum_{i=1}^N x_i\right) \bmod 2 = m, \\ 0 & : \text{otherwise.} \end{cases} \quad (33)$$

The binding information of these processes is $N - 1$ bits.

$P_{2,0}^N$ is the ‘even’ process, which assigns equal probability to all configurations with an even number of 1s and zero to all others. $P_{2,1}^N$ is the ‘odd’ process, which assigns uniform probabilities over the complementary set. When observed sequentially, the parity processes yield the maximum possible 1 bit of instantaneous predictive information (16) for each observation except the last, which cannot provide any predictive information as there is nothing left to predict. This is true regardless of the order in which the values are observed.

Consider now the multi-information of the parity processes. Since the joint entropy of either of them is $N - 1$ bits and the marginal entropy of each variable is 1 bit, the multi-information, consulting (10), is 1 bit. By contrast, if we look for binary processes which maximise the multi-information, we find that they have low binding information. From Theorem 1, we know that the maximal multi-information is $(N - 1)$ bits, which can only be achieved at a joint entropy of 1 bit. At this entropy, Theorem 2 tells us that the binding information can be at most 1 bit. We can easily find such processes: consider a system in which the indices $1..N$ are partitioned into two non-intersecting sets \mathcal{B} and its complement $\bar{\mathcal{B}} = 1..N \setminus \mathcal{B}$, and the probabilities assigned to configurations $\mathbf{x} \in \{0, 1\}^N$ as follows:

$$P_{\mathcal{B}}^N(\mathbf{x}) = \begin{cases} \frac{1}{2} & : \text{if } \forall i \in 1..N . x_i = \mathbb{I}(i \in \mathcal{B}), \\ \frac{1}{2} & : \text{if } \forall i \in 1..N . x_i = \mathbb{I}(i \in \bar{\mathcal{B}}), \\ 0 & : \text{otherwise,} \end{cases} \quad (34)$$

where $\mathbb{I}(\cdot)$ is 1 if the proposition it contains is true and 0 otherwise. For all i , the marginal entropy $H(X_i) = 1$ bit, the conditional entropy $H(X_i | X_{1..N \setminus \{i\}}) = 0$, and the joint entropy $H(X_{1..N}) = 1$ bit. These 'giant bit' processes are marked on Figures 5 and 6: they have $I(X_{1..N}) = N - 1$ bits and $B(X_{1..N}) = 1$ bit. Thus we see that binary process that maximise the multi-information and the binding information are very different in character.

Rather than proving Theorem 4 directly, we prove the following, more general result for sets of discrete random variables over an alphabet of K symbols.

Theorem 5. *If $\{X_1, \dots, X_N\}$ is a set of discrete random variables each taking values in $0..(K-1)$, then the binding information $B(X_{1..N})$ is maximised by the K 'modulo- K processes' $P_{K,m}^N$ for $m \in 0..(K-1)$, under which the probability of a configuration $\mathbf{x} \in (0..K-1)^N$ is*

$$P_{K,m}^N(\mathbf{x}) = \begin{cases} K^{1-N} & \text{if } \left(\sum_{i=1}^N x_i \right) \bmod K = m, \\ 0 & \text{otherwise.} \end{cases} \quad (35)$$

The binding information of these processes is $(N-1) \log_2 K$ bits.

Proof. Firstly, the intersection of the two upper bounds on the binding information in Theorem 2 is at $H(X_{1..N}) = B(X_{1..N}) = (N - 1) \log K$. Hence, the binding information can be at most $(N - 1) \log_2 K$ bits. To prove that the modulo- K processes yield this maximal value, let $\sigma_K^N : (0..K-1)^N \rightarrow 0..(K-1)$ be the sum modulo- K function:

$$\sigma_K^N(\mathbf{x}) = \left(\sum_{i=1}^N x_i \right) \bmod K. \quad (36)$$

In terms of this, we can define $\Sigma_{K,m}^N$ as the support of m^{th} modulo- K process:

$$\Sigma_{K,m}^N = \{ \mathbf{x} \in (0..K-1)^N | \sigma_K^N(\mathbf{x}) = m \}. \quad (37)$$

Observe that, for every $m \in 0..(K-1)$, there are K^{N-1} distinct sequences of N values that receive non-zero probability from a modulo- K process distribution

$P_{K,m}^N$, since for each of the K^{N-1} sequences that could comprise the first $N-1$ values $x_{1:N-1}$, there is exactly one value $x_N = [m - \sigma_K^{N-1}(x_{1:N-1})] \bmod K$ that can occur in the final position to ensure that $\sigma_K^N(x_{1:N}) = m$. By the definition of the modulo- K processes, these sequences are equiprobable, so the joint entropy is $H(X_{1..N}) = \log_2 K^{N-1} = (N-1) \log_2 K$.

By the same logic, the value of any single variable X_i is completely determined by the values of the other $N-1$ variables, and so the conditional entropy $H(X_i | X_{1..N \setminus \{i\}})$ must be zero for all $i \in 1..N$. Thus, $B(X_{1..N}) = H(X_{1..N}) = (N-1) \log_2 K$ bits. This value reaches the upper bound on the binding information and therefore the modulo processes maximise the binding information at $(N-1) \log_2 K$ bits. \square

In fact, *any* realisation of a modulo- K process observed in any order yields the maximum possible $\log_2 K$ bits of IPI at every observation except the last. At every position i except $i = N$, there are K equiprobable possibilities for the observation x_i . This in turn determines which of the K modulo- K processes the subsequent $N-i$ values must belong to, that is, it determines the sum modulo K of the remaining values. The supports of these processes are mutually exclusive and of the same cardinality and total probability, so the observation at position i yields the full $\log_2 K$ bits of information about the rest of the sequence. Next, we examine the uniqueness of the parity processes in maximising binding information, for which the following definitions and lemmas will be useful.

Definition 2. *If X and Y are discrete random variables taking values in \mathcal{X} and \mathcal{Y} respectively, the information about Y in the event $X=x$ is:*

$$I(X=x; Y) \triangleq \sum_{y \in \mathcal{Y}} \Pr(Y=y|X=x) \log \frac{\Pr(Y=y|X=x)}{\Pr(Y=y)}. \quad (38)$$

Lemma 1. *If X and Y are random variables taking values in finite sets \mathcal{X} and \mathcal{Y} respectively, and $I(X; Y) = \log |\mathcal{X}|$, then all of the following are true: (a) X is uniformly distributed on \mathcal{X} , (b) X is a function of Y , and (c) every event $X=x$ yields $\log_2 |\mathcal{X}|$ bits of information about Y :*

$$\forall x \in \mathcal{X} . I(X=x; Y) = \log |\mathcal{X}|. \quad (39)$$

Proof. Since \mathcal{X} is finite, $H(X) \leq \log |\mathcal{X}|$. This yields a chain of inequalities:

$$\log |\mathcal{X}| = I(X; Y) = H(X) - H(X|Y) \leq H(X) \leq \log |\mathcal{X}|.$$

Therefore, $H(X) = \log |\mathcal{X}|$, which implies that X is uniformly distributed over \mathcal{X} , and $H(X|Y) = 0$, which implies that there exists a function $f : \mathcal{Y} \rightarrow \mathcal{X}$ such that $X = f(Y)$, or more precisely,

$$\forall (x, y) . \Pr(X=x \wedge Y=y) > 0 \Rightarrow x = f(y). \quad (40)$$

This in turn implies that for each $x \in \mathcal{X}$, there exists a set $\Sigma_x = \{y \in \mathcal{Y} | f(y) = x\}$ which is the pre-image of x in \mathcal{Y} under f , and that these sets are disjoint, i.e., $\forall x \neq x' . \Sigma_x \cap \Sigma_{x'} = \emptyset$. Definition 2 gives

$$I(X=x; Y) = \sum_{y \in \mathcal{Y}} \Pr(Y=y|X=x) \log \frac{\Pr(Y=y|X=x)}{\Pr(Y=y)},$$

but $\Pr(Y=y|X=x) > 0$ only if $y \in \Sigma_x$, which in turn implies that $\Pr(Y=y) = \Pr(Y=y \wedge X=x)$ (because the event $Y=y$ implies the event $X=x$ when $y \in \Sigma_x$), and therefore the information is

$$\begin{aligned} I(X=x; Y) &= \sum_{y \in \mathcal{Y}} \Pr(Y=y|X=x) \log \frac{\Pr(Y=y|X=x)}{\Pr(Y=y \wedge X=x)} \\ &= \sum_{y \in \mathcal{Y}} \Pr(Y=y|X=x) \log \frac{1}{\Pr(X=x)} = -\log \Pr(X=x). \end{aligned}$$

Since X is uniformly distributed on \mathcal{X} , $\Pr(X=x) = |\mathcal{X}|^{-1}$ for all $x \in \mathcal{X}$ and the final part of the theorem is proved. \square

Lemma 2. *If X, Y and Z are random variables taking values in finite sets \mathcal{X}, \mathcal{Y} and \mathcal{Z} respectively, and $I(X; Y|Z) = \log|\mathcal{X}|$, then for all $z \in \mathcal{Z}$,*

$$\Pr(Z=z) > 0 \Rightarrow I(X; Y|Z=z) = \log|\mathcal{X}|. \quad (41)$$

Proof. The conditional mutual information $I(X; Y|Z)$ is the expectation of the mutual information $I(X; Y|Z=z)$ conditioned on events $Z=z$ for all $z \in \mathcal{Z}$:

$$I(X; Y|Z) = \sum_{z \in \mathcal{Z}} I(X; Y|Z=z) \Pr(Z=z). \quad (42)$$

If $I(X; Y|Z) = \log|\mathcal{X}|$, then $\log|\mathcal{X}| - I(X; Y|Z) = 0$ and hence

$$\sum_{z \in \mathcal{Z}} [\log|\mathcal{X}| - I(X; Y|Z=z)] \Pr(Z=z) = 0. \quad (43)$$

Since \mathcal{X} is finite, $I(X; Y|Z=z) = H(X|Z=z) - H(X|Y, Z=z) \leq \log|\mathcal{X}|$, so every term in the above sum is non-negative. Therefore, every term must be zero, and for all $z \in \mathcal{Z}$, either $\Pr(Z=z) = 0$ or $I(X; Y|Z=z) = \log|\mathcal{X}|$. \square

Lemma 3. *If $\{X_\alpha | \alpha \in \mathcal{A}\}$ is a set of N random variables taking values in a discrete set of cardinality K , and $B(X_{\mathcal{A}}) = (N-1) \log K$, then for any permutation of the variables (X_1, \dots, X_N) ,*

$$\forall i \in 1..(N-1). I(X_i; X_{(i+1)..N} | X_{1..(i-1)}) = \log K. \quad (44)$$

Proof. From the sequential formulation of the binding information (21), $B(X_{\mathcal{A}})$ equals the sum of the predictive information rates:

$$B(X_{\mathcal{A}}) = B(X_{1..N}) = \sum_{i=1}^N I(X_i; X_{(i+1)..N} | X_{1..(i-1)}).$$

The last term is zero because $I(X_N; X_\emptyset | X_{1..(N-1)}) = 0$ (there is nothing left to predict). Thus, if $B(X_{1..N})$ is maximal at $(N-1) \log K$, then

$$\sum_{i=1}^{N-1} I(X_i; X_{(i+1)..N} | X_{1..(i-1)}) = (N-1) \log K.$$

For each i , $I(X_i; X_{(i+1)..N} | X_{1..(i-1)}) \leq H(X_i) \leq \log K$, and there are only $N-1$ terms in the above sum, therefore, every term must equal $\log K$. \square

Lemma 4. *If (X_1, \dots, X_N) is any permutation of set of random variables $\{X_\alpha | \alpha \in \mathcal{A}\}$ taking values in a discrete set of cardinality K and having the maximal binding information of $B(X_{\mathcal{A}}) = (N - 1) \log K$, then any realisation of the process, when observed sequentially as (x_1, \dots, x_N) , yields $\log_2 K$ bits of instantaneous predictive information at every observation except the last.*

Proof. Applying Lemma 2 to the results of Lemma 3 gives

$$\forall i \in 1..(N - 1). I(X_i; X_{i+1:N} | X_{1:i-1} = x_{1:i-1}) = \log K$$

for all $x_{1:i-1}$ such that $\Pr(X_{1..i-1} = x_{1:i-1}) > 0$. Lemma 1 then implies that

$$\forall i \in 1..(N - 1). I(X_i = x_i; X_{i+1:N} | X_{1:i-1} = x_{1:i-1}) = \log K,$$

and so the IPI is $\log K$ at every observation from $i = 1$ to $N - 1$. \square

Theorem 6. *If $\{X_1, \dots, X_N\}$ is a set of binary random variables each taking values in $\{0, 1\}$, then the parity processes are the only two random processes that yield the maximal $N - 1$ bits of binding information.*

Proof. We will argue by induction. First, we prove that *if* only the odd and even parity processes maximise the binding information of $N - 1$ binary variables, *then* only the odd and even parity process maximise the binding information of N binary variables. Then we prove the base case for two random variables. Assume below that all random variables are binary and that all informations are in bits.

Induction step The premise of the induction step is that given some N , then for all sets of $N - 1$ binary random variables $\{Y_1, \dots, Y_{N-1}\}$,

$$B(Y_{1..N-1}) = N - 2 \Rightarrow \exists m \in \{0, 1\}. Y_{1..N-1} \sim P_{2,m}^{N-1}. \quad (45)$$

Assume that $B(X_{1..N}) = N - 1$. Therefore, by Lemma 3, $I(X_1; X_{2..N}) = 1$, which by Lemma 1, implies that $\Pr(X_1=0) = \Pr(X_1=1) = \frac{1}{2}$ and that $X_{2..N}$ is a function of X_1 . It also implies that binding information of the remaining variables $X_{2..N}$, *given* the observation $X_1=k$, must be $N - 2$, since the binding information is the expectation of the sum of the IPI at each observation, which by Lemma 4, is 1 bit at every observation except the last:

$$\forall k \in \{0, 1\}. B(X_{2..N} | X_1=k) = N - 2. \quad (46)$$

By the premise of our induction, this implies that the $N - 1$ variables $X_{2..N}$, *given* the observation $X_1=k$, must be distributed as one of the parity processes:

$$\forall k \in \{0, 1\}. \exists m \in \{0, 1\}. (X_{2..N} | X_1=k) \sim P_{2,m}^{N-1}. \quad (47)$$

As X_1 is a function of $X_{2..N}$, the supports of the distributions of $(X_{2..N} | X_1=0)$ and $(X_{2..N} | X_1=1)$ must be disjoint, so there are only two possibilities for assigning m in (47):

$$[(X_{2..N} | X_1=k) \sim P_{2,k}^{N-1}] \vee [(X_{2..N} | X_1=k) \sim P_{2,1-k}^{N-1}]. \quad (48)$$

Both possibilities result in a definite parity for $X_{1..N}$; in conjunction with the requirement that $\Pr(X_1=0) = \Pr(X_1=1) = \frac{1}{2}$, this implies that both result in one of the parity processes:

$$\begin{aligned} (X_{2..N}|X_1=k) \sim P_{2,k}^{N-1} &\Rightarrow X_{1..N} \sim P_{2,0}^N \\ \wedge (X_{2..N}|X_1=k) \sim P_{2,1-k}^{N-1} &\Rightarrow X_{1..N} \sim P_{2,1}^N. \end{aligned} \quad (49)$$

Thus, the induction premise is extended to N variables.

Base case With 2 variables, an exhaustive examination of the possibilities proves that there are only two distributions which have $I(X_1; X_2) = 1$:

$$\begin{aligned} \Pr(X_1=0 \wedge X_2=0) = \Pr(X_1=1 \wedge X_2=1) &= \frac{1}{2} \\ \vee \Pr(X_1=0 \wedge X_2=1) = \Pr(X_1=1 \wedge X_2=0) &= \frac{1}{2}, \end{aligned} \quad (50)$$

which are the even and odd parity process $P_{2,0}^2$ and $P_{2,1}^2$ respectively. \square

By analogy, it would be appealing to suggest that the only processes which maximise binding information in a set of K -ary random variables are the modulo- K processes. However, the above proof cannot be generalised in this way, because the logic applied at (47), (49) and (50) fails when $K > 2$, essentially because there are more than K permutations of K values when $K > 2$. To see this, a counter example at the base case will suffice. If $K = 3$ and we have two random variables X_1 and X_2 , then to obtain $I(X_1; X_2) = \log_2 3$ bits we must have $\Pr(X_1=k) = 1/3$ for all $k \in \{0, 1, 2\}$, and X_2 can be *any* one-to-one function of X_1 . This includes the 3 modulo-preserving functions

$$\begin{aligned} \{0 \mapsto 0, \quad 1 \mapsto 2, \quad 2 \mapsto 1\} &= \Sigma_{3,0}^2, \\ \{0 \mapsto 1, \quad 1 \mapsto 0, \quad 2 \mapsto 2\} &= \Sigma_{3,1}^2, \\ \{0 \mapsto 2, \quad 1 \mapsto 1, \quad 2 \mapsto 0\} &= \Sigma_{3,2}^2, \end{aligned} \quad (51)$$

but it also includes 3 non-modulo-preserving functions

$$\begin{aligned} \{0 \mapsto 0, \quad 1 \mapsto 1, \quad 2 \mapsto 2\}, \\ \{0 \mapsto 1, \quad 1 \mapsto 2, \quad 2 \mapsto 0\}, \\ \{0 \mapsto 2, \quad 1 \mapsto 0, \quad 2 \mapsto 1\}, \end{aligned} \quad (52)$$

which do not correspond to any modulo-3 process. Hence, when we try to apply the inductive formula to add another variable at $N = 3$, we are no longer confined to the 3 modulo-3 processes when choosing the 3 conditional distributions for $(X_{2..3}|X_1=k)$ for $k \in \{0, 1, 2\}$.

With hindsight, we should not be surprised by this. The binding information is invariant to invertible transformations of the variables, so if we have N invertible functions $f_i : 0..(K-1) \rightarrow 0..(K-1)$ for $i \in 1..N$ and $Y_i = f_i(X_i)$, then $B(Y_{1..N}) = B(X_{1..N})$. If $X_{1..N}$ is a modulo- K process with the maximal binding information, $Y_{1..N}$ will also have the maximal binding information without necessarily being a modulo- K process. However, looking at the example above with $N = 2$, we can see that permuting the value space at X_2 can map any of the non-modulo-3 processes in (52) on to one of the modulo-3 processes in (51). Thus, we propose (but do not prove here) the following:

Proposition 2. *If $\{X_1, \dots, X_N\}$ is a set of random variables each taking values in $0..(K-1)$, and $X_{1..N}$ has the maximal binding information $(N-1) \log_2 K$ bits, then their distribution must either be one of the K modulo- K process distributions $P_{m,K}^N$, or there exist N one-to-one functions $f_i : 0..(K-1) \rightarrow 0..(K-1)$ such that the set of variables $\{f_1(X_1), \dots, f_N(X_N)\}$ is distributed according to one of the K modulo- K distributions.*

To prove this, one would proceed inductively as before, showing that if a distribution over N variables with maximal binding information must be a modulo- K process or an invertible function of one, then a distribution over $N+1$ variables that maximises binding information must also be a modulo- K process or an invertible function of one.

7 Discussion

Binding information and noise One interpretation of the binding information is to say that the conditional entropy of each variable given *all* the others, what one might call the ‘residual entropy’ as a counterpart to the residual entropy rate (18), is essentially a kind of ‘noise’, since it represents the randomness in each variable which is not informative about any other variable. It is a sort of irreducible ‘non-predictive’ information.

Binding information and sub-extensive entropy As noted in §1, Bialek *et al* [8] point out that the predictive information $\mathcal{I}_{\text{pred}}(N)$ is the sub-extensive component of the entropy, that is, the part that grows slower than linearly with N , and that this is a desirable feature of a measure of complexity. However, we know that for stationary Markov chains at least [10], the binding information can have an extensive component, since the predictive information rate has a well defined time-invariant value, and the binding information grows linearly at this rate. Thus we have something of a contradiction between our proposal that the binding information is a useful measure of complexity and Bialek *et al*’s assertion that the non-extensive component of the entropy is the unique measure of complexity that satisfies certain reasonable desiderata, including transformation invariance for continuous valued variables [8, §5.3]. To begin to address this, we must examine their arguments rather carefully.

Firstly, transformation invariance does *not*, as they state [8, p. 2450], demand sub-extensivity: the binding information *is* transformation invariant since it is the accumulation of conditional mutual informations, and yet it *can* have an extensive component, as in a stationary Markov chain.

Secondly, Bialek *et al* imply that local correlations should not contribute to complexity [8, p. 2451–2452]. It is precisely these local correlations that give rise to the extensive component of the binding information, so we ask, is it reasonable to reduce all processes with only local correlations to the same class of ‘simple’ processes, no matter how large the finite correlation distance? Should a 5th order Markov chain be considered as simple as a sequence of independent variables? From the point of view of asymptotics, this may be appropriate, but when dealing with finite systems, it may not necessarily be so.

Complexity of the parity process Levels of stochastic dependence are discussed by Studený and Vejnarová, [16, §4], who formulate a level-specific measure of dependence which captures the dependency visible when fixed size subsets of variables are examined in isolation. Studený and Vejnarová [16, p. 277] use the parity process as an example of a random process in which the dependence is only visible at the highest level, that is, amongst all N variables. If fewer than N variables are examined, they appear to be independent. They note that such models were called ‘pseudo-independent’ by Xiang *et al* [25], who concluded that standard algorithms for Bayesian network construction fail on such processes. It is intriguing, then, that these are singled out as ‘most complex’ according to the binding information criterion.

Complexity and interestingness in music The fact that the parity/modulo process maximise the binding information raises some questions about the possible applications of binding information. In our previous work [10], we have begun to investigate the notion that instantaneous predictive information, computed from the point of view of an observer with some (subjective) probability model might be related to the perception of ‘interestingness’, or formal aesthetic value, during the sequential observation of some time-based art forms, specifically music. If this is indeed the case, we seem to have proved here that the most interesting sequences (of discrete variables over a common alphabet at least) are those drawn from one of the modulo processes. What are we to make of this? Though we defer a full discussion of this to a later article, we point out here that, to obtain the full $(N-1) \log_2 K$ bits of cumulative predictive information from a sequence drawn from one of the modulo- K processes $P_{K,m}^N$, the observer’s subjective probability model must be exactly this $P_{K,m}^N$ and the observer must know that values of N , K and m beforehand.

To make a musical analogy, suppose the X_t for $t \in 1..N$ correspond to a regular succession of beats and the binary value of X_t indicates whether or not a drum is struck at time t . Suppose further that distributions over $X_{1..N}$ correspond to ‘styles’ of music, in this case, drum patterns with N beats. To reach the maximal $N-1$ bits of total IPI, the listener must know that the piece is going to be N beats long and whether it is going to be in the ‘even style’ $P_{2,0}^N$ or the ‘odd style’ $P_{2,1}^N$. Furthermore, the listener’s enjoyment of the full $N-1$ bits of IPI is very brittle—they must correctly hear every beat to count the parity of the drum strikes so far; if a single beat is misheard then *all* the information is destroyed. There is no way for a listener to tell just by listening whether the piece is in the ‘odd’ style or the ‘even’ style until the very end. If we expand the alphabet of event types to, say, $K = 5$ pitched notes, then there are many more information maximising ‘styles’, all of which will be indistinguishable until the last note is heard!

All of this suggests that, even though the modulo processes might deliver the most IPI in theory, physical constraints and limitations on the cognitive capacities of observers would render this optimum unattainable in practice except in certain limited contexts. If binding information is relevant to perceived interestingness or formal aesthetic value, we must look for evidence of this within the limits of subjective probabilistic models that are accessible to human observers.

Summary To summarise, we have introduced the binding information as a measure of statistical structure that can be applied to any countable set of random variables, regardless of any topological structure inherent in the system. The binding information is minimised by systems of independent variables and by strongly ordered, but not necessarily deterministic, systems, and we showed that in a finite set of discrete random variables, it is maximised by the ‘modulo process’ and local transformations thereof. For finite sets of *binary* variables, the parity process are the *only* two processes that maximise the binding information.

Acknowledgements

This research was supported by EPSRC grant EP/H01294X/1.

References

- [1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
- [2] P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, 1986.
- [3] James P. Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63(2):105–108, Jul 1989.
- [4] K. Lindgren and M.G. Nordahl. Complexity measures and cellular automata. *Complex Systems*, 2(4):409–440, 1988.
- [5] Ricardo Lopez-Ruiz, Hector Mancini, and Xavier Calbet. A statistical measure of complexity. *Physics Letters A*, 209:321–326, 1995.
- [6] David P. Feldman and James P. Crutchfield. Measures of statistical complexity: Why? *Physics Letters A*, 238(4-5):244–252, 1998.
- [7] J. S. Shiner, Matt Davison, and P. T. Landsberg. Simple measure for complexity. *Physical Review E*, 59(2):1459–1464, Feb 1999.
- [8] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.
- [9] R.W. Yeung. A new outlook on Shannon’s information measures. *Information Theory, IEEE Transactions on*, 37(3):466–474, 1991.
- [10] Samer A. Abdallah and Mark D. Plumbley. Information dynamics: Patterns of expectation and surprise in the perception of music. *Connection Science*, 21(2):89–117, 2009.
- [11] W. McGill. Multivariate information transmission. *Information Theory, IRE Professional Group on*, 4(4):93–111, 1954.
- [12] JP Crutchfield and NH Packard. Symbolic dynamics of noisy chaos. *Physica D: Nonlinear Phenomena*, 7(1-3):201–223, 1983.

- [13] James P. Crutchfield. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1-3):11–54, 1994.
- [14] Wentian Li. On the relationship between complexity and entropy for markov chains and regular languages. *Complex systems*, 5(4):381–399, 1991.
- [15] Shlomo Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.
- [16] M. Studený and J. Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 261–297. MIT Press, 1998.
- [17] Ionas Erb and Nihat Ay. Multi-information in the thermodynamic limit. *Journal of Statistical Physics*, 115:949–976, 2004.
- [18] C.R. Shalizi, K.L. Shalizi, and J.P. Crutchfield. Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence. *Journal of Machine Learning Research*, pages 02–10, 2002.
- [19] Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3):817–879, 2001.
- [20] James P. Crutchfield and David P. Feldman. Statistical complexity of simple 1D spin systems. *Physical Review E*, 55(2):1239R–1243R, 1997.
- [21] Charles H. Bennett. How to define complexity in physics, and why. In Wojciech H. Zurek, editor, *Complexity, Entropy, and the Physics of Information*, pages 137–148. Addison-Wesley, 1990.
- [22] James P. Crutchfield, David P. Feldman, and Cosma Rohilla Shalizi. Comment I on “Simple measure for complexity”. *Physical Review E*, 62(2):2996–2997, Aug 2000.
- [23] B. Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, 45(2):79–82, 1980.
- [24] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- [25] Y. Xiang, S.K.M. Wong, and N. Cercone. Critical remarks on single link search in learning belief networks. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, pages 564–571, 1996.
- [26] A.J. Bell. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003*, 2003.

A Code for proof of Proposition 1

Proposition 1 in § 5 states that for a set of N random variables $X_{1..N}$,

$$\eta_1 = (N - 1)B(X_{1..N}) - I(X_{1..N}) \geq 0 \quad (53)$$

$$\text{and } \eta_2 = (N - 1)I(X_{1..N}) - B(X_{1..N}) \geq 0. \quad (54)$$

We attempt to prove this by expressing each of η_1 and η_2 as weighted sums of Studený and Vejnarová's level-specific dependency measures $\Delta(r)$, which are defined as follows for $r \in 1..(N - 1)$ (see [16, p. 278]):

$$\Delta(r) \triangleq \sum_{\substack{\alpha \subseteq 1..N \\ |\alpha|=r+1}} \sum_{\{i,j\} \subseteq \alpha} I(X_i; X_j | X_{\alpha \setminus \{i,j\}}). \quad (55)$$

Since each $\Delta(r)$ is sum of non-negative conditional mutual informations, they are also non-negative, and so if η_1 and η_2 can be expressed as positively weighted sums of the $\Delta(r)$, the proposition will be proved.

To do this, we consider an N -dimensional vector space in which the random process $X_{1..N}$ is represented by a point whose coordinates are determined by the entropies of various subsets of the N component variables. One coordinate frame for this vector space is the *aggregate entropy* frame. If the coordinates of the process with respect to this frame are denoted by ξ_i^h for $i \in 1..N$, then

$$\xi_i^h = \sum_{\substack{\alpha \subseteq 1..N \\ |\alpha|=i}} H(X_\alpha). \quad (56)$$

That is, the i^{th} coordinate is the sum of all the joint entropies of subsets of i variables. For example, if $N = 3$, then

$$\begin{aligned} \xi_1^h &= H(X_1) + H(X_2) + H(X_3), \\ \xi_2^h &= H(X_{\{1,2\}}) + H(X_{\{1,3\}}) + H(X_{\{2,3\}}), \\ \xi_3^h &= H(X_{\{1,2,3\}}). \end{aligned}$$

An alternative coordinate frame is the *aggregate conditional co-information* frame, in which the coordinates ξ_i^c of the system X are the sums of conditional co-informations [26]:

$$\xi_i^c = \sum_{\substack{\alpha \subseteq 1..N \\ |\alpha|=i}} C(X_\alpha | X_{1..N \setminus \alpha}), \quad (57)$$

where $C(X_\alpha | X_\beta)$ denotes the conditional co-information of all the variables in X_α given all the variables in X_β . (Note that Bell [26] writes $I(\cdot)$ for the co-information, but we are already using $I(\cdot)$ to denote the multi-information.) Again, for $N = 3$,

$$\begin{aligned} \xi_1^c &= C(X_1 | X_{\{2,3\}}) + C(X_2 | X_{\{1,3\}}) + C(X_3 | X_{\{1,2\}}), \\ \xi_2^c &= C(X_{\{1,2\}} | X_3) + C(X_{\{1,3\}} | X_2) + C(X_{\{2,3\}} | X_1), \\ \xi_3^c &= C(X_{\{1,2,3\}}). \end{aligned}$$

A third coordinate frame can be defined by using the residual entropy as the first coordinate and $\Delta(r)$ for $r \in 1..(N-1)$ as the remaining $N-1$ coordinates:

$$\xi_i^\Delta = \begin{cases} \sum_{j \in 1..N} H(X_j | X_{1..N \setminus \{j\}}) & : \text{if } i = 1, \\ \Delta(i-1) & : \text{if } i \in 2..N. \end{cases} \quad (58)$$

Linear operators can be found for transforming between these three coordinate frames. In the MATLAB code below, the function `h2ci` computes the matrix to map the entropy coordinates ξ^h to the conditional co-information coordinates ξ^c . The function `delta_ci` returns the matrix to transform the coordinates ξ^c to the coordinates ξ^Δ .

If sets of random variables are represented as points a vector space, the corresponding *dual* space is a space of linear functionals, which when applied to a point in the original space, yield a scalar value which is *symmetric* in the N variables of $X_{1..N}$. For example, if $N = 3$, then $\xi_1^h - \xi_3^h = I(X_{1..3})$ and $\xi_3^c - \xi_1^c = B(X_{1..3})$.

Thus, we proceed by expressing the binding information and the multi-information as linear functionals in the dual space using whichever coordinate frame is most convenient: entropy coordinates for the multi-information (function `multi_h`) and conditional co-information coordinates for the binding information (`binding_ci`). Then we compute the the linear functionals for η_1 and η_2 in ξ^c coordinates (`bounds_ci`). Finally, we transform to ξ^Δ coordinates (`bounds_delta`) by inverting the transformation matrix returned by `delta_ci`. If the coefficients returned by `bounds_delta(N)` are all non-negative, the proposition is proved for that value of N .

We found that multiplying the result by the least common multiple of the integers $1..(N-1)$ (see `lcmv`) was sufficient to preserve exact integer arithmetic up to $N = 37$; the proposition therefore appears to be true for all $N \in 1..37$.

```

% bounds_delta - coefficients of  $\eta_1$  and  $\eta_2$  in delta basis
%
% bounds_delta :: N:natural  $\rightarrow$  [[2, N]].
function W=bounds_delta(N)
    % Coefficients in delta basis obtained from coefficients in aggregate
    % conditional co-information basis by using inverse of delta_ci(N).
    % Premultiplication by lowest common multiple of integers 1..N-1
    % keeps everything integer valued.
    W=(lcmv(1:N-1)*bounds_ci(N))/delta_ci(N);
end

% -----
% bounds_ci :: N:natural  $\rightarrow$  [[2, N]].
% Express  $\eta_1$  and  $\eta_2$  as weighted sums of aggregate conditional
% co-information. Use inverse of h2ci to transform multi-information
% in aggregate entropy frame to aggregate cond. co-inf frame.
function B=bounds_ci(N)
    B = [N-1, -1; -1, N-1]*[ binding_ci(N); multi_h(N)/h2ci(N)];
end

```

```

% binding_ci - Binding information in agg. cond. co-information frame
% binding_ci :: N:natural → [[1, N]].
function H=binding_ci(N),
    % counts each level of aggregate cond co-info once, except for 1st
    % level, which is the residual entropy  $\sum_{i \in 1..N} H(X_i | X_{1..N \setminus \{i\}})$ 
    H=[0,ones(1,N-1)];
end

% multi_h -Multi-information operator in aggregate entropy frame
% multi_h :: N:natural → [[1, N]].
function H=multi_h(N),
    H=[1,zeros(1,N-2),- 1]; % sum of entropies minus joint entropy of all.
end

% -----
% delta_ci - Level specific dependency operators in symmetric CI basis
% delta_ci :: N:natural → [[N,N]].
function D=delta_ci(N)
    % this formula comes from the combinatorial definition of  $\Delta(r)$ 
    % for  $r \in 2..N$ , with an extra component equal to the residual
    % entropy to make it into a complete basis.
    D=[ 1,          zeros(1,N-1); ...
        zeros(N-1,1),  fliplr(pasc(N-1)).*repmat(trinum(N-1),N-1,1)];
end

% trinum - First N triangular numbers
% trinum :: N:natural → [[1, N]→natural].
function t=trinum(N), i=1:N; t=i.*(1+i)/2; end

% -----
% h2ci - transform from agg. entropy frame to agg. cond. co-info frame
% h2ci : N:natural → [[N,N]].
function T=h2ci(N),
    T=fliplr(pasc(N+1));
    T=T(2:end,2:end).*fliplr(checker(N));
end

% checker - checkerboard pattern of +1 and -1
% checker :: N:natural → [[N,N]→{-1,1}].
function T=checker(N)
    M=(N+mod(N,2))/2;
    T=repmat([1,-1;-1,1],M,M);
    T=T(1:N,1:N);
end

```

```

% -----
% pasc - Pascal's triangle in upper right triangular matrix form.
% pasc :: N:natural → [[N,N]→natural].
function P=pasc(N)
    P=eye(N);
    for i=2:N
        P(1,i)=1;
        for j=2:i
            P(j,i)=P(j,i-1)+P(j-1,i-1);
        end
    end
end

% -----
% lcmv - Lowest common multiple of array of values
% lcmv :: [[N]→natural] → natural.
function y=lcmv(x),
    y=1; for i=1:length(x), y=lcm(y,x(i)); end
end

```