# A measure of statistical complexity based on predictive information with application to finite spin systems

Samer A. Abdallah[a,*], Mark D. Plumbley[a]

[a]*School of Electronic Engineering and Computer Science*
*Queen Mary University of London*
*London E1 4NS, UK*

## Abstract

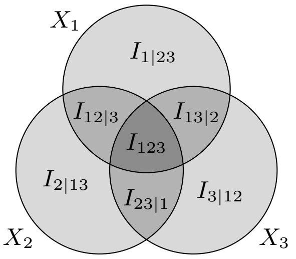
We propose the *binding information* as an information theoretic measure of complexity between multiple random variables, such as those found in the Ising or Potts models of interacting spins, and compare it with several previously proposed measures of statistical complexity, including excess entropy, Bialek *et al*'s predictive information, and the multi-information. We discuss and prove some of the properties of binding information, particularly in relation to multi-information and entropy, and show that, in the case of binary random variables interactions, the processes which maximise binding information are the 'parity' processes. The computation of binding information is demonstrated on Ising models of finite spin systems, showing that various upper and lower bounds are respected and also that there is a strong relationship between the introduction of high-order interactions and an increase of binding-information. Finally we discuss some of the implications this has for the use of the binding information as a measure of complexity.

*Keywords:* Information theory; Entropy; Statistical complexity; Ising model; Spin glasses.

## 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. 2–8] and is the one we adopt here. It is based on a consideration of the entropies and conditional entropies between the random variables in a probabilistic model of the system under investigation. A visualisation of some of the relevant quantities can be seen in Fig. 1; similar diagrams will be used to illustrate the various measures that will be examined in later sections.

In previous work, we defined the *predictive information rate* (PIR) [9] of a sequentially observed random process as the average information in one observation about future observations yet to be 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 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 letter, 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.



$$
\begin{aligned}
I_{1|23} &= H(X_1|X_2,X_3) \\
I_{13|2} &= I(X_1;X_3|X_2) \\
I_{1|23}+I_{13|2} &= H(X_1|X_2) \\
I_{12|3}+I_{123} &= I(X_1;X_2)
\end{aligned}
$$

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 [10]. Some other information measures are indicated in the legend.

We begin by reviewing a number of earlier proposals for measures of structure and complexity as well as the PIR. Then, in § 3, we define the *binding information* as the extensive counterpart of the PIR applicable to arbitrary countable sets of random variables. In § 4 we derive some upper and lower bounds on the binding information in relation to the entropy and multi-information, and in § 5, investigate which processes maximise binding information. Finally, in § 6, we illustrate how binding information

[*] Corresponding author.
*Email addresses:* `samer.abdallah@eecs.qmul.ac.uk` (Samer A. Abdallah), `mark.plumbley@eecs.qmul.ac.uk` (Mark D. Plumbley)

*Preprint submitted to Physics Letters A* *October 7, 2011*

behaves in an extended Ising model of a spin glass with high-order interactions, and conclude with some observations about binding information as a measure of complexity and its relationship to the alternatives.

In the following, if $X$ is a random process 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 $X_\alpha$ for each $\alpha \in \mathcal{B}$. $|\mathcal{B}|$ 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. Entropic measures of statistical structure

Suppose that $(\ldots, X_{-1}, X_0, X_1, \ldots)$ is a bi-infinite stationary sequence of random variables, and that $\forall t \in \mathbb{Z}$, the random variable $X_t$ takes values in a discrete set $\mathcal{X}$. Let $\mu$ be the associated shift-invariant probability measure over all configurations of the system. Stationarity implies that the probability distribution associated with any contiguous block of $N$ variables $(X_{t+1}, \ldots, X_{t+N})$ is independent of $t$, and therefore we can define a shift-invariant block entropy function:

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

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

A number of quantities can be expressed in terms of the block entropy $H(N)$. Firstly, the entropy rate $h_\mu$ can be written in two equivalent ways [1]:

$$h_\mu = \lim_{N\to\infty} \frac{H(N)}{N} = \lim_{N\to\infty} H(N) - H(N-1). \quad (2)$$

The entropy rate gives a measure of the overall randomness or unpredictability of the process.

*Excess entropy.* The block entropy function $H(\cdot)$ can be used to express the mutual information between two contiguous blocks of length $N$ and $M$ respectively:

$$I(X_{-N..-1}; X_{0..M-1}) = H(N) + H(M) - H(N+M). \quad (3)$$

If we let both block lengths $N$ and $M$ tend to infinity, we obtain what Crutchfield and Packard [11] 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:

$$E = \lim_{N\to\infty} 2H(N) - H(2N). \quad (4)$$

It can also be expressed in terms of the $h_\mu(N)$ defined by Crutchfield [12] 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 $p_\mu^N$. Crutchfield and Young [3] define the excess entropy in terms of $h_\mu(\cdot)$ as follows

$$E \triangleq \sum_{M=1}^{\infty} (h_\mu(M) - h_\mu), \quad (5)$$

but the result is equivalent to the mutual information between semi-infinite halves of the process (4).

*Predictive information.* Grassberger [2] and others [4, 13] have commented on the manner in which $h_\mu(N)$ approaches its limit $h_\mu$, 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:

$$\mathcal{I}_{\text{pred}}(N) \triangleq \lim_{M\to\infty} H(N) + H(M) - H(N+M). \quad (6)$$

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 in the following sense: if the entropy rate $h_\mu$ is the intensive counterpart of the asymptotically extensive entropy $H(N)$, and $Nh_\mu$ 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 difference, such that

$$H(N) = Nh_\mu + \mathcal{I}_{\text{pred}}(N). \quad (7)$$

From this we can see that the sum of the first $N$ terms of (5), which comes to $H(N) - Nh_\mu$, is equal to $\mathcal{I}_{\text{pred}}(N)$.

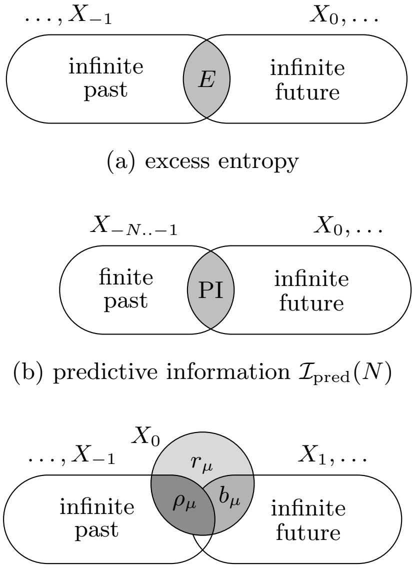
*Multi-information.* The *multi-information* [14] is defined for any collection of $N$ random variables $(X_1, \ldots, X_N)$ as

$$I(X_{1..N}) \triangleq -H(X_{1..N}) + \sum_{i\in 1..N} H(X_i). \quad (8)$$

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. In the thermodynamic limit, its intensive counterpart the *multi-information rate* can be defined as

$$\rho_\mu \triangleq \lim_{N\to\infty} I(X_{1..N}) - I(X_{1..N-1}). \quad (9)$$

It can be shown from (2), (8) and (6) that $\rho_\mu = \mathcal{I}_{\text{pred}}(1) = H(1) - h_\mu$ (see fig. 2). Erb and Ay [15] studied the behaviour of the multi-information in the thermodynamic limit and demonstrated a relationship between the multi-information rate $\rho_\mu$ (they call it $I$) and the 'finite volume'



(a) excess entropy

(b) predictive information $\mathcal{I}_{\text{pred}}(N)$

(c) predictive information rate $b_\mu$

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), the circle represents the 'present'. Its total area is $H(X_0) = H(1) = \rho_\mu + r_\mu + b_\mu$, where $\rho_\mu$ is the multi-information rate, $r_\mu$ is the residual entropy rate, and $b_\mu$ is the predictive information rate. The entropy rate is $h_\mu = r_\mu + b_\mu$.

approximation to the excess entropy found by summing the first $N$ terms of (5), which as we have already noted is equal to $\mathcal{I}_{\text{pred}}(N)$; in the present terminology,

$$I(X_{1..N}) + \mathcal{I}_{\text{pred}}(N) = N\rho_\mu. \qquad (10)$$

Comparing this with (7), 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.* Another class of measures, including Grassberger's *true measure complexity* [2] and Crutchfield *et al*'s *statistical complexity* $C_\mu$ [3, 16], is based on the properties of stochastic automata that model the process under consideration. These have some interesting properties that make them viable measures of complexity, but, due to space limitations, are beyond the scope of this letter.

*Predictive information rate.* In previous work [9], we introduced the *predictive information rate* (PIR), which is the average information in one observation about the infinite future given the infinite past. If $\overleftarrow{X}_t = (\ldots, X_{t-2}, X_{t-1})$ denotes the variables before time $t$, and $\overrightarrow{X}_t = (X_{t+1}, X_{t+2}, \ldots)$ denotes those after $t$, the PIR at time $t$ is defined as a conditional mutual information:

$$\underline{\overline{\mathcal{I}}}_t \triangleq I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = H(\overrightarrow{X}_t | \overleftarrow{X}_t) - H(\overrightarrow{X}_t | X_t, \overleftarrow{X}_t). \qquad (11)$$

Equation (11) can be read as the average reduction in uncertainty about the future on learning $X_t$, given the past. Due to the symmetry of the mutual information, it can also be written as

$$I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = H(X_t | \overleftarrow{X}_t) - H(X_t | \overrightarrow{X}_t, \overleftarrow{X}_t). \qquad (12)$$

Now, in the shift-invariant case, $H(X_t | \overleftarrow{X}_t)$ is the familiar entropy rate $h_\mu$, but $H(X_t | \overrightarrow{X}_t, \overleftarrow{X}_t)$, the conditional entropy of one variable given *all* the others in the sequence, future as well as past, is what we called the *residual entropy rate* $r_\mu$ in [17], but was previously identified by Verdú and Weissman [18] as the *erasure entropy rate*. It can be defined as the limit

$$r_\mu \triangleq \lim_{N \to \infty} H(X_{-N..N}) - H(X_{-N..-1}, X_{1..N}). \qquad (13)$$

The second term, $H(X_{1..N}, X_{-N..-1})$, is the joint 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. Thus, the shift-invariant PIR (which we will write as $b_\mu$) is the difference between the entropy rate and the erasure entropy rate: $b_\mu = h_\mu - r_\mu$. These relationships are illustrated in Fig. 2, along with several of the information measures we have discussed so far.

*Measuring complexity.* Many of the measures reviewed above were intended as measures of 'complexity', where 'complexity' is a quality that is somewhat open to interpretation [6, 19]; hence the variety of proposals. What is generally agreed [e.g. 5–7, 20], however, is that a plausible measure of complexity should be low for systems that are deterministic, or easy to compute or predict—'ordered'—and also low for systems that a completely random and unpredictable—'disordered'. The PIR satisfies these conditions without being 'over-universal' in the sense of Crutchfield *et al* [6, 20]: it is not simply a function of entropy or entropy rate that fails to distinguish between the different strengths of temporal dependency that can be exhibited by systems at a given level of entropy. In our analysis of Markov chains [9], we found that the processes which maximise the PIR are not those that maximise the multi-information rate $\rho_\mu$ (or the excess entropy, which is the same in this case), and do not have the kind of predictability associated with highly redundant processes. What they do have, however, is a sort of partial predictability that requires the observer continually to pay attention to the most recent observations in order to make optimal predictions. And so, while Crutchfield *et al* make a compelling case for the excess entropy $E$ and their statistical complexity $C_\mu$ as measures of complexity, there is still room to suggest that the PIR captures a different and non trivial aspect of temporal dependency structure not previously examined.

## 3. Binding information

In this section, we address two questions. Firstly, if the PIR is considered an intensive quantity like the entropy rate or the multi-information rate, what is its extensive counterpart? Secondly, can the concept be extended to collections of random variables which are not organised sequentially, such as spin systems in two or more dimensions?

When the PIR is accumulated over successive observations, one obtains a quantity which we call the *binding information.* To proceed, we first reformulate the PIR in a form that is applicable to a *finite* sequence of random variables $(X_1, \ldots, X_N)$:

$$\overline{\mathcal{I}}_t(X_{1..N}) = I(X_t; X_{(t+1)..N}|X_{1..(t-1)}), \tag{14}$$

which is to be compared with the PIR for infinite sequences (11). Note that this is no longer shift-invariant and may depend on $t$. The binding information $B(X_{1..N})$, then, is the sum

$$B(X_{1..N}) = \sum_{t=1}^{N} \overline{\mathcal{I}}_t(X_{1..N}). \tag{15}$$

Expanding this sum in terms of entropies and conditional entropies, cancelling terms and simplifying yields

$$B(X_{1..N}) = H(X_{1..N}) - \sum_{t=1}^{N} H(X_t|X_{1..N\setminus\{t\}}). \tag{16}$$

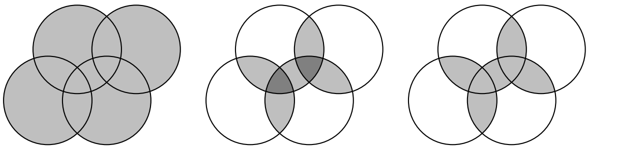
That is, the binding information is the difference between the joint entropy of all the variables and the *residual* or *erasure entropy* as defined by Verdú and Weissman [18]. Like the multi-information, it measures the information shared between a set of random variables, but in a different way (see Fig. 3).

Though the binding information was derived by accumulating the PIR sequentially, the result is permutation invariant. This suggests that the concept might 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 a countable indexed set of random variables, then its binding information is*

$$B(X_\mathcal{A}) \triangleq H(X_\mathcal{A}) - \sum_{\alpha\in\mathcal{A}} H(X_\alpha|X_{\mathcal{A}\setminus\{\alpha\}}). \tag{17}$$

Since it can be expressed as a sum of (conditional) mutual informations (15), 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 pointwise 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 $f_\alpha : \mathcal{X} \to \mathcal{Y}$ such that $Y_\alpha = f_\alpha(X_\alpha)$, then $B(Y_\mathcal{A}) = B(X_\mathcal{A})$.



(a) $H(X_{1..4})$ (b) $I(X_{1..4})$ (c) $B(X_{1..4})$

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 implied in (15) 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 $\alpha, \alpha' \in \mathcal{A}$, $X_\alpha$ is some known function of $X_{\alpha'}$, 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, it is easy to show that $B(X_\mathcal{A}) = H(X_\mathcal{A}) = H(X_\alpha)$ for any $\alpha \in \mathcal{A}$, which, as we will see, is relatively low compared with what is possible as soon as $N$ becomes appreciably large. 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.

In the following sections, we will 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:

$$I(X_\mathcal{A}) \triangleq -H(X_\mathcal{A}) + \sum_{\alpha\in\mathcal{A}} H(X_\alpha). \tag{18}$$

## 4. Bounds on binding and multi-information

In the following we confine our attention to finite sets of discrete random variables taking values in a common alphabet containing $K$ symbols. This subsumes Ising ($K = 2$) and Potts ($K > 2$) spin systems with arbitrary high-order (i.e. not just pairwise) interactions. 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.

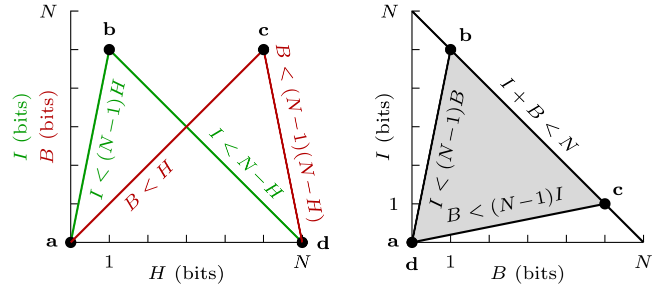


Figure 4: Constraints on multi-information $I(X_{1..N})$ and binding information $B(X_{1..N})$ for a system of $N = 6$ binary random variables. The labelled points represent identifiable distributions over the $2^N$ states that this system can occupy: (a) *known state*, the system is deterministically in one configuration; (b) *giant bit*, one of the $P_B^6$ processes; (c) *parity*, the parity processes $P_{2,0}^6$ or $P_{2,1}^6$; (d) *independent*, the system of independent unbiased random bits.

**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*

$$I(X_\mathcal{A}) \leq N \log K - H(X_\mathcal{A}) \qquad (19)$$

$$\textit{and} \quad I(X_\mathcal{A}) \leq (N-1)H(X_\mathcal{A}). \qquad (20)$$

*Proof.* The multi-information is $I(X_\mathcal{A}) = \sum_{\alpha\in\mathcal{A}} H(X_\alpha) - H(X_\mathcal{A})$. Since each variable $X_\alpha$ 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 (19) follows directly. We also have, for all $\alpha \in \mathcal{A}$, $H(X_\alpha) \leq H(X_\mathcal{A})$, and so

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

□

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

$$B(X_\mathcal{A}) \leq H(X_\mathcal{A}) \qquad (21)$$

$$\textit{and} \quad B(X_\mathcal{A}) \leq (N-1)(N \log K - H(X_\mathcal{A})). \qquad (22)$$

*Proof.* The first inequality comes directly from the definition of the binding information (17) or (16), since $H(X_\alpha | X_{\mathcal{A}\setminus\{\alpha\}}) \geq 0$ for any discrete random variable $X_\alpha$. To obtain the second inequality, we expand the conditional entropies of (16):

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

But, for all $\alpha$, $H(X_{\mathcal{A}\setminus\{\alpha\}}) \leq (N-1)\log K$ 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. 4.

Next, we examine the relationship between $I(X_\mathcal{A})$ and $B(X_\mathcal{A})$.

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

$$I(X_\mathcal{A}) + B(X_\mathcal{A}) \leq N \log K. \qquad (23)$$

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

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

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. □

**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:*

$$I(X_\mathcal{A}) \leq (N-1)B(X_\mathcal{A}) \qquad (24)$$

$$\textit{and} \quad B(X_\mathcal{A}) \leq (N-1)I(X_\mathcal{A}). \qquad (25)$$

We will not prove Proposition 1 here but a sketch of a potential proof can be found in [17]. It is based on a result that can be obtained for $N = 3$; in this case, it is relatively easy to find that

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

Since both quantities are sums of non-negative mutual informations or conditional mutual informations, the inequalities (24) and (25) for $N = 3$ follow directly. The proof sketch in [17] rests on finding a similar decomposition into non-negative terms when $N > 3$.

Subsequent to our initial work on binding information [17], we learned that Proposition 1 follows from results presented by Han [21], in which Han analyses the space of information measures that can be expressed as linear combinations of entropies. He discovered a duality relation on this space, and identified the *dual total correlation* as the formal dual of Watanabe's [22] *total correlation*. The multi-information and the binding information are precisely the total correlation and its dual respectively. The duality is hinted at by the symmetry between the bounds illustrated in fig. 4.

## 5. 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 binary variables) is surprisingly simple.

**Theorem 4.** *If $\{X_1,\ldots,X_N\}$ 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' $P^N_{2,0}$ and $P^N_{2,1}$. For $m \in \{0,1\}$, the probability of observing $\mathbf{x} \in \{0,1\}^N$ under each process is*

$$P^N_{2,m}(\mathbf{x}) = \begin{cases} 2^{1-N} & : \textit{if } \left(\sum_{i=1}^N x_i\right) \bmod 2 = m, \\ 0 & : \textit{otherwise}. \end{cases} \qquad (26)$$

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

$P^N_{2,0}$ is the 'even' process, which assigns equal probability to all configurations with an even number of 1s and zero to all others. $P^N_{2,0}$ is the 'odd' process, which assigns uniform probabilities over the complementary set. When a parity process is observed sequentially in *any* order, the finite-sequence form of the PIR (14) yields the maximum possible 1 bit for each observation except the last, which cannot provide any predictive information as there is nothing left to predict.

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 (8), 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 disjoint sets $\mathcal{B}$ and its complement $\overline{\mathcal{B}} = 1..N \setminus \mathcal{B}$, and the probabilities assigned to configurations $\mathbf{x} \in \{0,1\}^N$ as follows:

$$P^N_{\mathcal{B}}(\mathbf{x}) = \begin{cases} \frac{1}{2} & : \text{if } \forall i \in 1..N \,.\, x_i = \mathbb{I}(i \in \mathcal{B}), \\ \frac{1}{2} & : \text{if } \forall i \in 1..N \,.\, x_i = \mathbb{I}(i \in \overline{\mathcal{B}}), \\ 0 & : \text{otherwise}, \end{cases} \qquad (27)$$

where $\mathbb{I}(\cdot)$ is 1 if the proposition it contains is true and 0 otherwise. Hence, there are only two equiprobable global configurations, which can be obtained from each other by 'flipping' all the bits. 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 Figure 4: 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. In [17], we prove Theorem 4 as a corollary of a more general theorem for discrete-valued random variables, where we find that the parity processes generalise to the *modulo-$K$ processes*. We also prove that, for systems of binary random variables, the converse is also true: namely, that the parity processes are *uniquely* the only two processes which maximise the binding information when $K = 2$.

## 6. Binding information in spin systems

In this section we consider a system of $N$ binary random variables $X_i$ for $i \in 1..N$, such as can be used to model a system of interacting spin $\frac{1}{2}$ particles. This system has $2^N$ distinct configurations and so the set of probability distributions over these states is the standard $(2^N-1)$-simplex embedded in $2^N$ dimensions. For small $N$, we can represent such probability distributions explicitly, compute their joint entropy $H(X_{1..N})$, multi-information $I(X_{1..N})$, and binding information $B(X_{1..N})$ numerically, and thereby represent them as points in a 3-dimensional space. We will model these distributions using an Ising model with random interactions, that is, a spin glass, initially with just pairwise interactions, then gradually adding in higher-order interactions. To facilitate this, we let the variables take values in $\{-1,+1\}$ rather than $\{0,1\}$.

### 6.1. Up to pairwise interactions

With pairwise interactions and a non-uniform external field, the Hamiltonian of the Ising model, as a function of the system's configuration $\mathbf{x} \in \{-1,+1\}^N$, is

$$\mathcal{H}_2(\mathbf{x}) = \sum_{i \in 1..N} B_i x_i + \sum_{\{i,j\} \subseteq 1..N} J_{ij} x_i x_j. \qquad (28)$$

where the $J_{ij}$ (with $i < j$) are the interaction strengths between each pair of spins and the $B_i$ are the external field strengths at each site. The probability $P(\mathbf{x})$ of any configuration $\mathbf{x}$ is proportional to the exponential of the Hamiltonian: $P(\mathbf{x}) \propto e^{-\mathcal{H}_2(\mathbf{x})/T}$, where $T$ is the temperature (which we will set to 1 in the remainder). To complete the model, we need a scheme for sampling the values of the $B_i$ and $J_{ij}$; here, we will sample them independently from two student's t distributions with scale factors $\sigma_1$ and $\sigma_2$ respectively:

$$B_i \sim \sigma_1 \mathcal{T}(\nu), \quad J_{ij} \sim \sigma_2 \mathcal{T}(\nu), \qquad (29)$$

where $\mathcal{T}(\nu)$ denotes the student's t distribution with $\nu$ degrees of freedom, allowing for a non-Gaussian distribution of interactions. Unlike the Gaussian interactions of the Sherrington-Kirkpatrick spin glass [23], the student's t model when $\nu$ is small induces *sparsity*; for example, at $\nu = 0.25$, the pairwise component of the Hamiltonian will be dominated by relatively few strong 'bonds'. In practise, we find that with fixed $\sigma_1$ and $\sigma_2$, this sampling scheme is unable to reach simultaneously the limits of high and low entropy that can be obtained with the model, and so we sample $\sigma_1^2$ and $\sigma_2^2$ themselves from two independent inverse-gamma distributions: for $k \in \{1, 2\}$,

$$\sigma_k^2 \sim \mathcal{IG}(\alpha, \beta_k), \tag{30}$$

where the $\alpha$ is the (global) shape parameter and the $\beta_k$ are the scale parameters. These allow the various orders of interaction to be individually enabled or disabled by setting the corresponding $\beta_k$ to nonzero or zero values.

The scatter plot in fig. 5(a) was obtained by sampling 3000 systems with $N = 8$, $\nu = 0.25$, $\alpha = 0.5$, $\beta_1 = 0$ and $\beta_2 = 10^{-8}$, that is, with only pairwise interactions and no external field. Each system was plotted as a single point according to its 3 information coordinates. The resulting cloud of points stretches between the 'giant bit' and 'independent bits' processes illustrated in fig. 4 as points **b** and **d** respectively. However, it does not reach all the way to the 'parity' process at point **c**, which cannot be modelled using only pairwise interactions. Instead, the binding information appears to reach a maximum of 4 bits at $H = B = I = 4$. Examination of the processes that reach this point shows that they all consist of 4 independent copies of 2-bit parity processes, with 4 strong bonds forming 4 disconnected pairs.

What may not be obvious from the scatter plot is that all the points lie in the plane $I = N - H$, which implies (referring to the proof of Theorem 1) that the marginal entropy $H(X_\alpha)$ of each variable is 1 bit. This also means that the net expected magnetisation $\sum_{\alpha\in 1..8} \langle X_\alpha \rangle = 0$.

By varying $\beta_1$ and $\beta_2$, the relative strengths of the pairwise interactions and the external field can be adjusted. With $\beta_2 = 0$, we obtain the relatively uninteresting case where all the generated processes have independent (but biased) bits, which lie on the line $I = B = 0$, between points **a** and **d** in fig. 4. Setting $\beta_1 = \beta_2 = 2 \times 10^{-8}$, we obtain fig. 5(b). The point cloud here fills a volume between the planar region in fig. 5(a) and point **a**, respecting all the bounds established in § 4.

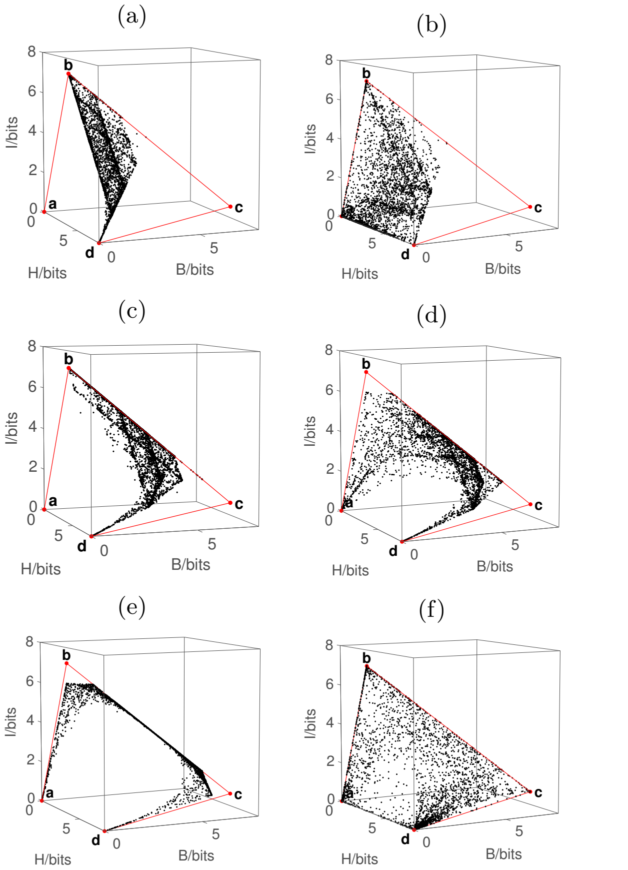


Figure 5: Joint entropy (H), binding information (B) and multi-information (I) plotted for systems of $N = 8$ binary random variables, generated using the Ising model with high-order interactions (31). For each scatter plot, 4000 systems were generated, using the following parameters (all unspecified $\beta_k$ default to zero): (a) pairwise interactions only, $\nu = 0.25$, $\alpha = 0.5$, $\beta_2 = 10^{-8}$; (b) $\nu = 0.25, \alpha = 0.2, \beta_1 = \beta_2 = 2 \times 10^{-8}$; (c) $\nu = 0.25, \alpha = 0.5, \beta_4 = 10^{-10}$; (d) $\nu = 0.25, \alpha = 0.25, \beta_5 = 8 \times 10^{-11}$; (e) $\nu = 0.25, \alpha = 0.25, \beta_7 = 5 \times 10^{-4}$; (f) $\nu = 1, \alpha = 0.15, \beta_k = 10^{-9}\ \forall k \in 1..8$. The points labelled **a**, **b**, **c** and **d** correspond with those in fig. 4.

### 6.2. *Higher-order interactions*

It is possible to extend the model by adding higher order interactions modulated by scale factors $\sigma_3$, $\sigma_4$ etc., up to $\sigma_N$. The new Hamiltonian is

$$\mathcal{H}_N(\mathbf{x}) = \sum_{k=1}^{N} \left[ \sum_{\alpha\in[N]^{(k)}} J_\alpha^{(k)} \prod_{i\in\alpha} x_i \right], \tag{31}$$

where the $J_\alpha^{(k)}$, with $k \in 1..N$, represent all the interaction strengths for all orders of interaction, and $[N]^{(k)}$ denotes the set of all subsets of $1..N$ of cardinality $k$, that is, $[N]^{(k)} = \{\alpha \subseteq 1..N : |\alpha| = k\}$. The model of (28) can be recovered by setting $J_{\{i\}}^{(1)} = B_i$, $J_{\{i,j\}}^{(2)} = J_{ij}$ and all other interactions to zero. The interaction strengths are again independent with student's t distributions and linked scale factors $\sigma_k$, again sampled according to (30), for each order of interaction:

$$J_\alpha^{(k)} \sim \sigma_k \mathcal{T}(\nu), \quad \forall \alpha \in [N]^{(k)}. \tag{32}$$

Fig. 5 illustrates the effect of bringing in progressively higher-order interactions on the binding information. For example, when only fourth-order interactions are included, the maximum reachable binding information appears to be 6 bits, which is acheived by processes consisting of 2 independent 4 bit parity processes. The absolute maximum binding information of $N-1$ bits is obtained when order-$N$ interactions are the *only* ones present. This is not surprising when we consider that the order-$N$ term of the Hamiltonian, $\prod_{i\in 1..N} x_i$ on the domain $\{-1, 1\}^N$ is isomorphic to the parity function on $\{0, 1\}^N$. Note also that in all cases where only even-order interactions are present, the points are confined to the $I = N - H$ plane.

While these examples give a mostly qualitative picture of how binding information behaves for systems of binary random variables, they do show that an information theoretic characterisation of such systems is enriched by including the binding information along with the entropy and the multi-information, as well as suggesting quantitative properties for further investigation. There is a strong relationship between binding information and the presence of strong high-order interactions, which is not apparent when the entropy and multi-information are examined alone.

## 7. Discussion

Our examination of the extended Ising model with high-order interactions shows that, unlike the entropy or the multi-information, the binding information is sensitive to the presence of high-order interactions. Such levels of stochastic dependence are discussed by Studenỳ and Vejnarovà [14, §4], who formulate a level-specific measure of dependence which captures the dependency visible when fixed size subsets of variables are examined in isolation. Studenỳ and Vejnarovà 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. [24], 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.

As noted in § 1, Bialek et al. [8] argue that the predictive information $\mathcal{I}_{\text{pred}}(N)$, being the sub-extensive component of the entropy, is the unique measure of complexity that satisfies certain reasonable desiderata, including transformation invariance for continuous-valued variables. While lack of space precludes a full discussion, we note that transformation invariance does *not*, as Bialek *et al* state [8, p. 2450], demand sub-extensivity: binding information *is* transformation invariant, since it is a sum of conditional mutual informations, and yet it *can* have an extensive component, since its intensive counterpart, the PIR, can have a well-defined value, e.g., in stationary Markov chains [9].

As Verdú and Weissman [18] note, the erasure entropy is relevant to single-site Gibbs sampling methods for simulating a random process: the erasure entropy is the entropy rate, per sweep through the variables, of the Gibbs sampling process itself. Since the binding information is the joint entropy minus the erasure entropy, processes with high binding information will be harder to simulate via Gibbs sampling, with long mixing times due to the low rate of entropy generation as compared with the entropy of the desired distribution.

Van Enter [25] gives analytical methods for computing the erasure entropy density in the thermodynamic limit for the 2-D Ising model with nearest-neighbour interactions on rectangular and honeycomb grids. In combination with a method for computing the entropy density, these methods can be used to compute the binding information density in such systems. Where analytical methods are not available, Yu and Verdu's [27] erasure entropy estimation method based on bidirectional context tree weighting (CTW) could be used instead. Alternatively, for smaller systems, a brute-force histogram-based entropy estimator (e.g. [26]) could be used.

## 8. Acknowledgements

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

## References

[1] T. M. Cover, J. A. Thomas, Elements of Information Theory, John Wiley and Sons, New York, 1991.

[2] P. Grassberger, International Journal of Theoretical Physics 25 (1986) 907–938.

[3] J. P. Crutchfield, K. Young, Physical Review Letters 63 (1989) 105–108.

[4] K. Lindgren, M. Nordahl, Complex Systems 2 (1988) 409–440.

[5] R. Lopez-Ruiz, H. Mancini, X. Calbet, Physics Letters A 209 (1995) 321–326.

[6] D. P. Feldman, J. P. Crutchfield, Physics Letters A 238 (1998) 244–252.

[7] J. S. Shiner, M. Davison, P. T. Landsberg, Physical Review E 59 (1999) 1459–1464.

[8] W. Bialek, I. Nemenman, N. Tishby, Neural Computation 13 (2001) 2409–2463.

[9] S. A. Abdallah, M. D. Plumbley, Connection Science 21 (2009) 89–117.

[10] W. McGill, Information Theory, IRE Professional Group on 4 (1954) 93–111.

[11] J. Crutchfield, N. Packard, Physica D: Nonlinear Phenomena 7 (1983) 201–223.

[12] J. P. Crutchfield, Physica D: Nonlinear Phenomena 75 (1994) 11–54.

[13] W. Li, Complex systems 5 (1991) 381–399.

[14] M. Studenỳ, J. Vejnarovà, in: M. I. Jordan (Ed.), Learning in Graphical Models, MIT Press, 1998, pp. 261–297.

[15] I. Erb, N. Ay, Journal of Statistical Physics 115 (2004) 949–976.

[16] J. P. Crutchfield, D. P. Feldman, Physical Review E 55 (1997) 1239R–1243R.

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

[18] S. Verdú, T. Weissman, in: IEEE International Symposium on Information Theory (ISIT 2006), pp. 98–102.

[19] C. H. Bennett, in: W. H. Zurek (Ed.), Complexity, Entropy, and the Physics of Information, Addison-Wesley, 1990, pp. 137–148.

[20] J. P. Crutchfield, D. P. Feldman, C. R. Shalizi, Physical Review E 62 (2000) 2996–2997.

[21] T. Han, Information and Control 36 (1978) 133–156.

[22] S. Watanabe, IBM Journal of research and development 4 (1960) 66–82.

[23] D. Sherrington, S. Kirkpatrick, Phys. Rev. Lett. 35 (1975) 1792–1796.

[24] Y. Xiang, S. Wong, N. Cercone, in: Proc. 12th Conf. on Uncertainty in Artificial Intelligence, pp. 564–571.

[25] A. Van Enter, E. Verbitskiy, Arxiv preprint arXiv:1001.3122 (2010).

[26] P. Grassberger, Physics Letters A 128 (1988) 369–373.

[27] J. Yu, S. Verdú, Information Theory, IEEE Transactions on 55 (2009) 350–357.