Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Information Decomposition Based on Cooperative Game Theory

by

Nihat Ay, Daniel Polani, and Nathaniel Virgo

Preprint no.: 81 2020

Information Decomposition Based on Cooperative Game Theory

Nihat Ay[1,2,3], Daniel Polani[4], Nathaniel Virgo[1,5]

July 21, 2020

[1]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[2]University of Leipzig, Leipzig, Germany
[3]Santa Fe Institute, Santa Fe, NM, USA
[4]University of Hertfordshire, UK
[5]Earth-Life Science Institute (ELSI), Tokyo, Japan

Abstract

We offer a new approach to the *information decomposition* problem in information theory: given a 'target' random variable co-distributed with multiple 'source' variables, how can we decompose the mutual information into a sum of non-negative terms that quantify the contributions of each random variable, not only individually but also in combination? We define a new way to decompose the mutual information, which we call the *Information Attribution* (IA), and derive a solution using cooperative game theory. It can be seen as assigning a "fair share" of the mutual information to each combination of the source variables. Our decomposition is based on a different lattice from the usual 'partial information decomposition' (PID) approach, and as a consequence the IA has a smaller number of terms than PID: it has analogs of the synergy and unique information terms, but lacks separate terms corresponding to redundancy, instead sharing redundant information between the unique information terms. Because of this, it is able to obey equivalents of the axioms known as 'local positivity' and 'identity', which cannot be simultaneously satisfied by a PID measure.

Keywords: partial information decomposition, information geometry, cooperative game theory

1 Introduction

Consider a random variable Y, called the *target*, and suppose it is co-distributed with a set of random variables $X_1, \ldots, X_n$, which may also be correlated with each other. We call the variables $X_1, \ldots, X_n$ *input variables* or *inputs*, though this should not be taken to imply a causal interpretation. Learning the values of the inputs provides information about Y, and this can be quantified using the standard tools of information theory. In particular one may consider quantities such as $I(X_1, X_2; Y)$ (the information that the inputs X_1 and X_2 together provide, on average, about Y), or $I(X_1; Y|X_2)$ (the information provided by X_1 when X_2 is known).

However, it has been clear for some time that the mutual information and conditional mutual information are relatively blunt tools, and it would be useful in applications to have a more fine-grained picture. This led in particular to the *Partial Information Decomposition* (PID) framework, proposed by Williams and Beer (2010). The original proposal was to divide the mutual information $I(X_1, \ldots, X_n; Y)$ into a sum of non-negative terms in a particular way. However, this has proven difficult to do satisfactorily for more than two inputs, and in particular a no-go theorem, due to Rauh et al. (2014), has shown that no decomposition of the mutual information can satisfy both Williams and Beer's axioms and a certain set of intuitively desirable properties.

Here we propose a different way to decompose the mutual information, which we call the *Information Attribution* (IA) problem. Let $V = \{X_1, \ldots, X_n\}$ be the set of all the input variables we are interested in,

and consider the power set 2^V. Members of 2^V (i.e. sets of input variables) are termed *predictors*, since in general each one provides some information about the target. Given a predictor A we wish to *attribute* to it a certain fraction of the total mutual information. These should be non-negative and sum to $I(X_1, \ldots, X_n; Y)$. In addition we wish to avoid double-counting: the information attributed to a set of inputs should tell us something meaningful about how much more information that set provides, beyond that provided by each of its subsets individually. This has much in common with the notion of *synergy* in the PID framework, but differs in that the IA framework does not have distinct concepts of unique and redundant information.

Because of this difference, we are able to offer a solution to the IA problem that remains non-negative for an arbitrary number of input variables, while exhibiting a number of desirable properties. Our solution makes use of both information geometry and cooperative game theory, allowing us first to assess the information provided by each *set* of predictors (i.e. set of sets of input variables), and then to share the total mutual information among the individual predictors in a way that can be considered uniquely fair.

In more detail, in the case of two inputs, the PID framework decomposes the mutual information $I(X_1, X_2; Y)$ into a sum of four terms of three different kinds. These terms are (*i*) the information that the two sources provide redundantly about the target (known as redundant information, shared information or common information); (*ii*) two terms corresponding to the information provided uniquely by each input, and (*iii*) the synergistic or complementary information, which can only be obtained by knowing both of the sources simultaneously.

However, the axioms proposed by Williams and Beer do not completely determine these quantities. As a result, many PID measures have been proposed in the literature, each satisfying different additional properties beyond the ones given by Williams and Beer. Several approaches have been proposed. Among these are several that are based on information geometry (Harder et al., 2013; Bertschinger et al., 2014; Perrone and Ay, 2016; Olbrich et al., 2015; Griffith and Koch, 2014; James et al., 2019), which we build upon here.

Generalizing towards the case of three or more input variables has turned out to be more problematic under the PID framework. One of the most intuitive additional axioms proposed is known as the identity axiom, proposed by Harder et al. (2013), but it was shown by Rauh et al. (2014) that no measure can exist that obeys both Williams and Beer's axioms (including "local positivity") and the identity axiom. Because of this, there are a number of proposed PID measures that relax either the identity axiom or the local positivity axiom of Williams and Beer, or both. Such approaches include (Ince, 2017; Finn and Lizier, 2018; Kolchinsky, 2019). Another promising class of approaches involve changing to a slightly different perspective, for example, by considering the full joint distribution between multiple random variables, rather than singling out a single variable as the target (Rosas et al., 2016; James and Crutchfield, 2017; Finn and Lizier, 2020). The present work falls into a third class of approaches, which involves decomposing the mutual information in a different way, using a different lattice from the one defined by Williams and Beer. Rauh et al. (2014) argued for such an approach after proving their no-go theorem.

In contrast to the partial information decomposition, for two predictors our Information Attribution (IA) framework decomposes the mutual information $I(X_1, X_2; Y)$ into only three terms, all non-negative, corresponding to information that can be attributed to the predictors $\{X_1\}$, $\{X_2\}$ and $\{X_1, X_2\}$. The third term is similar to synergy in the PID framework, since it corresponds to information that can only be attributed to X_1 and X_2 together, and not to either of them individually. the singleton $\{X_1\}$ and $\{X_2\}$ terms behave like a combination of the redundant and unique information terms, in that information shared between the two sources is split between the $\{X_1\}$ and $\{X_2\}$ terms.

In the general case of n inputs, with $V = \{X_1, \ldots, X_n\}$, we write

$$I(X_1, \ldots, X_n; Y) = \sum_{A \in 2^V} I_A(X_1, \ldots, X_n; Y) \; .$$

For a given predictor A, the term $I_A(X_1, \ldots, X_n; Y)$ indicates the proportion of the total mutual information attributed to A, beyond what is already provided by its subsets. (The information attributed to the empty set is always 0.)

Despite being based on a different lattice, our decomposition obeys analogs of both the local positivity and identity axioms. These properties hold for any number of inputs.

There are a number of other approaches to multivariate information besides PID, some of which are closely related to our appraoch. These include in particular an approach called *reconstructability analysis* (Zwick, 2004), which we draw on extensively, as well as Amari's *hierarchical decomposition* (Amari, 2001), Ay's measure of complexity (Ay, 2015), and several measures that have arisen in the context of Integrated Information Theory (IIT), such as (Oizumi et al., 2016). This family of measures is reviewed in (Amari et al., 2016), which describes their relationships in terms of information geometry.

We now outline the structure of the argument before proceeding to the details. Throughout the paper we restrict ourselves to the case where all of the variables, $X_1, \ldots, X_n$ and Y, have finite state spaces, although we expect our measure to generalise well to cases such as Gaussian models in which the state spaces are continuous. (Several previous PID measures have been extended to the Gaussian case by Barrett (2015), who also proves some important results about the behaviour of PID measures on Gaussian models.)

To construct our Information Attribution measure we proceed in two steps. We begin by defining the mutual information provided by certain *sets* of predictors, i.e. sets of sets of input variables. We do this via a sublattice of the lattice of probability distributions that James et al. (2019) termed the "constraint lattice." The same lattice has appeared in the literature previously, within the topic of reconstructability analysis (Zwick, 2004). Having established the information contribution of each set of predictors, we then attribute a contribution to each individual predictor by a method that involves summing over the maximal chains of the constraint lattice.

We then show that this procedure of averaging over maximal chains can be derived using cooperative game theory. We can conceptualise our measure in terms of a cooperative game, in which each set of predictors is thought of as a coalition of players. Each coalition is assigned a worth corresponding to the information they jointly provide about the target. Our measure can then be derived via a known generalisation of the Shapley value due to Faigle and Kern (1992), which assigns a payoff to each individual player (i.e. predictor) based on its average performance among all the coalitions in which it takes part, while respecting additional precedence constraints.

Since our measure is based on the constraint lattice, we review this concept in depth in section 2. We approach the constraint lattice from the perspective of information geometry and state its relationship to known results in that field. In section 3 we consider a sublattice of the constraint lattice which we term the *input lattice*, which allows us to define a quantity corresponding to the information that a set of predictors provides about the target. From this we derive our measure by summing over the maximal chains of the input lattice. After proving some properties of our Information Attribution measure and giving some examples (sections 5 and 6), we then make the connection to cooperative game theory in section 7, proving that our measure is equivalent to the generalised Shapley value of Faigle and Kern (1992).

2 Background: the constraint lattice

We begin by defining the so-called "constraint lattice" of James et al. (2019), which has also been defined previously in the context of reconstructability analysis (Zwick, 2004). This section serves to summarise previous work and to establish notation for the following sections.

The constraint lattice, along with its sublattice that we term the input lattice, are used in reconstructability analysis, as described by Zwick. It is used both as a tool for constructing probabilistic models from limited data and as a way to quantify the information in higher-order correlations by calculating the information lost when those correlations are excluded from the model. This corresponds to the family of information measures that we derive in sections 2.2 and 3. Our contribution, in section 4, is to turn this family of separate measures into a single decomposition of the mutual information between a set of source variables and a target variable, the Information Attribution measure, by summing over a set of chains in the lattice.

The constraint lattice is also used by James et al. (2019) to define a partial information decomposition (PID) measure. This is done by calculating the same distributions on each node as Zwick (which we also calculate here). Their PID measure is derived from the mutual information between the source variables and

the target according to these probability distributions. This is a different procedure from the one derived here, and results in a PID measure rather than an Information Attribution measure.

2.1 Lattices of random variables

In this section we introduce several concepts from partial order theory and introduce the elements of the constraint lattice as defined in the literature. We state several standard results from partial order theory, which can be found, for example, in (Stanley, 2011), and for the most part we follow the terminology of that reference (and see also Grabisch, 2016).

Suppose we have a set W of co-distributed random variables,

$$W = \{Z_1, Z_2, \ldots, Z_m\}.$$

Subsets of W may also be considered as random variables. For example, $\{Z_1, Z_2\}$, which we also write Z_1Z_2, can be thought of as a random variable whose sample space is the Cartesian product of the sample spaces of Z_1 and Z_2. We therefore consider the power set 2^W, whose members are to be thought of both as sets and as random variables in this sense. We think of 2^W as a partially ordered set, ordered by set inclusion. That is, given $s, t \in 2^W$, considered as sets, we write $s \leq t$ if $s \subseteq t$. With this partial order, 2^W forms a distributive lattice, in fact a Boolean algebra.

The idea behind the constraint lattice is that given a *set of sets* of random variables $\mathcal{S}$, that is, a subset of 2^W, we can form a new joint distribution on the sample space of all the random variables, such that the marginals of all the members of $\mathcal{S}$ match those in the true distribution, but subject to that constraint the new distribution is as decorrelated as possible, in a maximum entropy sense, which we define in the next section. To do this, it is natural to define the following partial order on sets of sets of random variables (that is, on 2^{2^W}):

$$\mathcal{S} \leq \mathcal{T} \quad \text{if and only if} \quad A \in \mathcal{S} \Rightarrow \exists B \in \mathcal{T} : A \subseteq B \,, \tag{1}$$

for $\mathcal{S}, \mathcal{T} \in 2^{2^W}$. This reflects the idea that $\mathcal{T}$ should lie above $\mathcal{S}$ if it specifies at least all of the same marginals as $\mathcal{S}$. However, this represents the situation in a redundant way, because, for example, the sets $\{\{Z_1, Z_2\}\}$ and $\{\{Z_1, Z_2\}, \{Z_1\}\}$ are distinct but specify the same information, since specifying the joint marginal for Z_1Z_2 also specifies the marginal for Z_1. Because of this, we want to put some restrictions on which subsets of 2^W are permitted as elements of the lattice. This may be done in terms of two different concepts, *antichains* or *down-sets*. The resulting lattices have different elements but the same partial order. It is standard in the Partial Information Decomposition literature to define lattices in terms of antichains. However, in making the connection to cooperative game theory it will be more convenient to talk in terms of down-sets instead. For this reason, we define both terms here, but define the constraint lattice in terms of down-sets.

Given any poset P, an *antichain* of P is a subset $\mathcal{A} \subseteq P$, such that no two members of $\mathcal{A}$ are comparable. That is, given $a, b \in \mathcal{A}$, we have neither $a \leq b$ nor $b \leq a$. For a power set lattice like 2^W, this corresponds to a set $\mathcal{A}$ of subsets of W, such that no member of $\mathcal{A}$ is a subset of any other. For example, if W has at least four elements then $\{\{Z_1, Z_2\}, \{Z_2, Z_3\}, \{Z_4\}\}$ is an antichain of 2^W.

On the other hand, given a poset P, a *down-set* of P is a subset $\mathcal{S} \subseteq P$, such that

$$a \in \mathcal{S}, \;\; b \leq a \quad \Rightarrow \quad b \in \mathcal{S} \,. \tag{2}$$

That is, if an element a of P is included in $\mathcal{S}$, then so are all the elements below it in the partial order. For example, $\{\{Z_1, Z_2\}, \{Z_2, Z_3\}, \{Z_1\}, \{Z_2\}, \{Z_3\}, \{Z_4\}, \emptyset\}$ is a down-set of 2^W.

For finite posets there is a one-to-one correspondence between antichains and down-sets: given an antichain $\mathcal{A} \subseteq P$ of a poset, one can form a down-set $\mathcal{S}$ from the members of P that are below some member of $\mathcal{A}$ in the partial order, that is, $\mathcal{S} = \{a \in P \mid a \leq b \text{ for some } b \in \mathcal{A}\}$. On other other hand, given a down-set $\mathcal{S}$ one can recover the corresponding antichain by taking the maximal members of $\mathcal{S}$, that is, $\mathcal{A} = \{a \in \mathcal{S} \mid \nexists\, b \in \mathcal{S} : a < b\}$. This correspondence allows us to use the two concepts somewhat interchangeably.

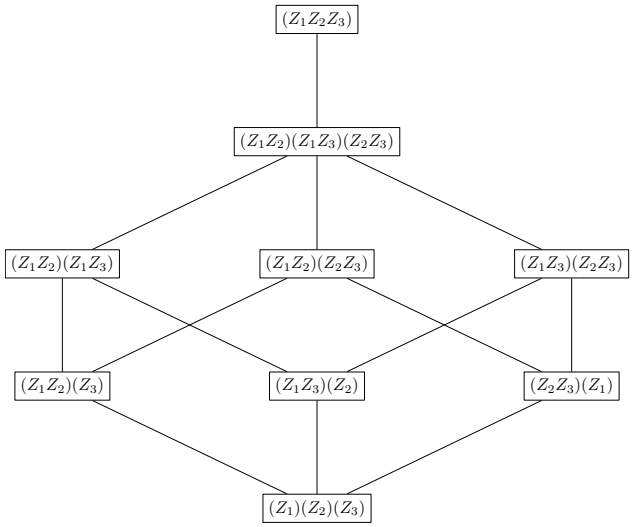


Figure 1: The Hasse diagram for the constraint lattice, as defined by (Zwick, 2004; James et al., 2019), for three random variables, $W = \{Z_1, Z_2, Z_3\}$.

The correspondence also allows us to define the following shortcut notation for down-sets of power set lattices like 2^W: we take the corresponding antichain, write its elements as lists surrounded by parentheses, and concatenate them. For example, the notation $(Z_1Z_2)(Z_3)$ refers to the down-set $\{\{Z_1, Z_2\}, \{Z_1\}, \{Z_2\}, \{Z_3\}, \emptyset\}$. This convention is used in the figures and elsewhere.

For any poset P, one can form the lattice of down-sets of P, denoted $J(P)$, whose elements are all of the down-sets of P. For down-sets, the partial order (1) reduces to set inclusion. For any poset P, the poset $J(P)$, ordered by set inclusion, is a distributive lattice. For the rest of the paper we will largely be concerned with sublattices of $J(2^W)$, that is, lattices whose elements are down-sets of 2^W, ordered by set inclusion.

We define the constraint lattice in terms of *down-set covers* of W, meaning those down-sets $\mathcal{S}$ of 2^W for which each element of W appears at least once in one of the members of $\mathcal{S}$. That is, $\mathcal{S} \in 2^W$ is a down-set cover of W if $\bigcup_{a \in \mathcal{S}} a = W$.

With this terminology in place, we can define the elements of the constraint lattice as those elements of $J(2^W)$ that are down-set covers of W, ordered by set inclusion. This gives an equivalent lattice to the one defined by James et al. and Zwick, but with elements that are down-sets instead of the corresponding antichains. For $m = 3$, the resulting lattice is illustrated in fig. 1. The definition of the constraint lattice will be completed by assigning a probability distribution to each node in a particular way, which we do in the next subsection.

We will make use of a few more definitions from partial order theory below. Given a lattice P and members $a, b \in P$, we write $a < b$ if $a \leq b$ and $a \neq b$. We say that a lattice element b *covers* an element a, written $a \prec b$, if $a < b$ and there exists no $c \in P$ such that $a < c < b$. A sequence of elements $a_1, \ldots, a_k$ is called a *chain* in P if $a_1 < a_2 < \cdots < a_k$. A chain is called a *maximal chain* if it is not a proper subset of any other chain. This implies that a_1 is the bottom element of the lattice, a_k is the top element, and $a_1 \prec a_2 \prec \cdots \prec a_k$.

In fig. 1, the relationship $a \prec b$ is indicated by drawing b above a and connecting the elements with an edge. The resulting graph is called the Hasse diagram of the lattice. The maximal chains are the directed paths from the bottom node in fig. 1 to the top node.

2.2 Constraints and split distributions

Let $p = p(Z_1, \ldots, Z_m)$ be the joint probability distribution of the members of W. We call this the *true distribution*. Following (James et al., 2019) and (Zwick, 2004), we now wish to associate with element of

the constraint lattice a joint distribution $p_{\mathcal{S}} = p_{\mathcal{S}}(Z_1, \ldots, Z_m)$. In the spirit of (Ay, 2015; Oizumi et al., 2016; Amari et al., 2016) we term these *split distributions*. Each split distribution captures only some of the correlations present in the true distribution, and we can think of the remaining correlations as being split apart, or forced to be as small as possible.

Specifically, given an element $\mathcal{S}$ of the constraint lattice, which is a down-set cover of W (and hence a set of sets of random variables), the split distribution $p_{\mathcal{S}}$ is constructed so that it captures the correlations associated with the members of $\mathcal{S}$, in the sense that $p_{\mathcal{S}}(A) = p(A)$, for every $A \in \mathcal{S}$. This defines a family of distributions, and from this family we choose the one with the maximum entropy. Intuitively, the maximum entropy distribution is the least correlated one in the family, so it excludes any additional correlations aside from those specified by $\mathcal{S}$.

In the remainder of this section, we define the split distributions more rigorously, alongside some related objects, and we point out an important property, which follows from the so-called Pythagorean theorem of information geometry. This section is largely a review of previous work, and makes a connection between the constraint lattice of (James et al., 2019; Zwick, 2004) and the language of information geometry (Ay et al., 2017, chapter 2).

Let Δ be the set of all joint probability distributions of the random variables in W. For a down-set cover $\mathcal{S}$ of W, let

$$M_{\mathcal{S}} = \{q \in \Delta : \forall A \in \mathcal{S}, q(A) = p(A)\} \ .$$

That is, $M_{\mathcal{S}}$ is the set of all probability distributions for which the members of $\mathcal{S}$ have the same marginal distributions as in the true distribution p. Note that if the constraint $q(A) = p(A)$ holds for some $A \subseteq W$, then it will also automatically hold for $B \subseteq A$. This is the reason for considering down-sets, rather than arbitrary subsets of 2^W. $M_{\mathcal{S}}$ is a mixture family, and we have that $\mathcal{S} \leq \mathcal{T} \implies M_{\mathcal{S}} \supseteq M_{\mathcal{T}}$.

We can now define the split distribution $p_{\mathcal{S}}$ as

$$p_{\mathcal{S}} = \operatorname*{argmax}_{q \in M_{\mathcal{S}}} H(q). \tag{3}$$

Equivalently, we can instead define the split distributions in terms of the Kullback-Leibler divergence, as we will see below. This has the advantage that it is likely to generalise to cases such as Gaussian models in which the state space is not discrete.

There is another interpretation of the split distributions, which is interesting to note. In addition to the mixture family $M_{\mathcal{S}}$, we can also define an exponential family corresponding to a given node in the constraint lattice. This can be seen as a family of *split models*, i.e. probability distributions in which some kinds of correlation are forced to be absent. The split distribution $p_{\mathcal{S}}$ can be seen as the closest member of this exponential family to the true distribution.

To see this, we define the exponential family

$$E_{\mathcal{S}} = \left\{ q \in \Delta : q(z_1, \ldots, z_m) = \prod_{A \in \mathcal{S}} \mu_A(z_1, \ldots z_m), \text{for some set of functions } \mu_A \right\} \ , \tag{4}$$

where the functions μ_A have the additional requirements that $\mu_A(z_1, \ldots, z_m)$ depends only on z_i for $Z_i \in A$, and $\mu_A(z_1, \ldots, z_m) > 0$. $E_{\mathcal{S}}$ is an exponential family, and we have $\mathcal{S} < \mathcal{T} \implies E_{\mathcal{S}} \subseteq E_{\mathcal{T}}$.

Finally, we let $\overline{E}_{\mathcal{S}}$ be the topological closure of the set $E_{\mathcal{S}}$, meaning that $\overline{E}_{\mathcal{S}}$ contains every member of $E_{\mathcal{S}}$, and in addition also contains all the limit points of sequences in $E_{\mathcal{S}}$. The difference is that $E_{\mathcal{S}}$ does not contain distributions with zero-probability outcomes, whereas the closure $\overline{E}_{\mathcal{S}}$ does.

Note that we cannot obtain $\overline{E}_{\mathcal{S}}$ by simply relaxing the condition that $\mu_A(z_1, \ldots, z_m) > 0$. This is because although every member of $E_{\mathcal{S}}$ must factorise according to eq. (4), the limit points on the boundary of the simplex can fail to factorise in the same way. An example of this is given by (Lauritzen, 1996, Example 3.10). These limit points must be included in order to make sure the split distribution is always defined.

It is a known result in information geometry (Ay et al., 2017, Theorem 2.8) that for any $\mathcal{S}$, the sets $M_{\mathcal{S}}$ and $\overline{E}_{\mathcal{S}}$ intersect at a single point. In fact this point is the split distribution $p_{\mathcal{S}}$. With the Kullback-Leibler

divergence

$$D_{\mathrm{KL}}(q\|p) = \sum_{z_1,\ldots,z_m} q(z_1,\ldots,z_m) \log \frac{q(z_1,\ldots,z_m)}{p(z_1,\ldots,z_m)} \ ,$$

we can equivalently characterise $p_{\mathcal{S}}$ by

$$p_{\mathcal{S}} = \operatorname*{argmin}_{q \in M_{\mathcal{S}}} D_{\mathrm{KL}}(q\|u) \ , \tag{5}$$

where u denotes the uniform distribution. This directly follows from (3). A further equivalent characterisation of $p_{\mathcal{S}}$ is given by

$$p_{\mathcal{S}} = \operatorname*{argmin}_{q \in E_{\mathcal{S}}} D_{\mathrm{KL}}(p\|q) \ . \tag{6}$$

In the terminology of information geometry, eq. (5) is an I-projection (information projection) and eq. (6) is an rI-projection (reverse I-projection). The classical theory of these information projections has been greatly extended by Csiszár and Matúš (2003, 2004).

We also have the so-called *Pythagorean theorem* of information geometry (Amari and Nagaoka, 2007), which in our notation says that for elements $\mathcal{S} < \mathcal{T} < \mathcal{U}$ of the constraint lattice,

$$D_{\mathrm{KL}}(p_{\mathcal{U}}\|p_{\mathcal{S}}) = D_{\mathrm{KL}}(p_{\mathcal{U}}\|p_{\mathcal{T}}) + D_{\mathrm{KL}}(p_{\mathcal{T}}\|p_{\mathcal{S}}) \ . \tag{7}$$

Equation (7) can be extended to any chain in the constraint lattice $\mathcal{S}_1 < \mathcal{S}_2 < \cdots < \mathcal{S}_k$, to give

$$D_{\mathrm{KL}}(p_{\mathcal{S}_k}\|p_{\mathcal{S}_1}) = \sum_{i=2}^{k} D_{\mathrm{KL}}(p_{\mathcal{S}_i}\|p_{\mathcal{S}_{i-1}}) \ . \tag{8}$$

This will be crucial in defining our information contribution measure below.

Consider the top node in the constraint lattice, given by $(Z_1,\ldots,Z_m)$, which we denote $\top$. We have $p_\top = p$. That is, the split distribution corresponding to $\top$ is equal to the true distribution.

Since we are considering only down-set covers of W, the bottom node of the lattice is given by $(Z_1)\ldots(Z_m)$, which we denote $\bot$. We have $p_\bot(z_1,\ldots,z_m) = p(z_1)\ldots p(z_m)$. That is, its split distribution is given by the product of the marginal distributions for all the members of W.

Together with eq. (5), this allows us to interpret $p_{\mathcal{S}}$ as the distribution that is *as decorrelated as possible* (i.e. closest to the product distribution, in the Kullback-Leibler sense), subject to the constraint that the marginals of the members of $\mathcal{S}$ match those of the true distribution. Alternatively, via eq. (6), we can see it as the distribution that is as close to the true distribution as possible, subject to the constraint that it lies in the closure of the exponential family $\overline{E}_{\mathcal{S}}$.

For a general down-set cover $\mathcal{S}$ of W, the split distribution $p_{\mathcal{S}}$ may not have an analytical solution, and instead must be found numerically. One family of techniques for this is iterative scaling (Csiszár and Shields, 2004, chapter 5), which was used to calculate the examples below. Alternatively, one may solve eq. (5) as a numerical optimisation problem, starting from an element such as $\bot$ with a known split distribution. This yields a convex optimisation problem with linear constraints, but it is not always well conditioned.

Finally, given an element $\mathcal{S}$ of the constraint lattice, we define $I_{\mathcal{S}} \coloneqq D_{\mathrm{KL}}(p_\top\|p_{\mathcal{S}})$. This can be thought of as the amount of information that is present in the true distribution $p_\top$ but is not present in $p_{\mathcal{S}}$. Note that due to the Pythagorean relation (eq. (7)) we have $D_{\mathrm{KL}}(p_{\mathcal{T}}\|p_{\mathcal{S}}) = I_{\mathcal{S}} - I_{\mathcal{T}}$, for any antichain covers $\mathcal{S} \leq \mathcal{T}$. The quantity $I_{\mathcal{S}}$ turns out to be a useful generalisation of the mutual information, as shown in the following examples.

Example 2.1. Independence. Suppose $W = \{Z_1, Z_2\}$, and let $\mathcal{S} = (Z_1)(Z_2)$. Then $\overline{E}_{\mathcal{S}}$ is the set of distributions q that can be expressed as a product $q(z_1,z_2) = \mu_1(z_1)\mu_2(z_2)$, which we may also write $q(z_1,z_2) = q(z_1)q(z_2)$. So $\overline{E}_{\mathcal{S}}$ is the set of distributions for which Z_1 and Z_2 are independent. We have that $p_{\mathcal{S}} = p(z_1)p(z_2)$, and consequently, it is straightforward to show that in this example, $D_{\mathrm{KL}}(p_\top\|p_{\mathcal{S}}) = I(X_1;X_2)$.

Example 2.2. Conditional independence. Suppose $W = \{Z_1, Z_2, Z_3\}$, and let $\mathcal{S} = (Z_1Z_3)(Z_2Z_3)$. Then $\overline{E}_{\mathcal{S}}$ is the set of distributions q that can be expressed as a product

$$q(z_1, z_2, z_3) = \mu_1(z_1, z_3)\mu_2(z_2, z_3).$$

These are the distributions for which $q(z_1, z_2, z_3) = q(z_3)q(z_1|z_3)q(z_2|z_3)$, i.e. for which $Z_1 \perp\!\!\!\perp_q Z_2 \mid Z_3$. So in this case $E_{\mathcal{S}}$ can be seen as a conditional independence constraint. It is straightforward to show that that $p_{\mathcal{S}}(z_1, z_2, z_3) = p(z_3)p(z_1|z_3)p(z_2|z_3)$, and consequently $D_{\mathrm{KL}}(p_\top \| p_{\mathcal{S}}) = I(X_1; X_2|X_3)$.

Example 2.3. Amari's triplewise information. Suppose $W = \{Z_1, Z_2, Z_3\}$, and let $\mathcal{S} = (Z_1Z_2)(Z_1Z_3)(Z_2Z_3)$. Then $E_{\mathcal{S}}$ is the set of distributions q that can be expressed as a product

$$q(z_1, z_2, z_3) = \mu_1(z_1, z_2)\mu_2(z_1, z_3)\mu_3(z_2, z_3).$$

Unlike the previous two examples, there is no analytic expression for μ_1, μ_2 and μ_3 in terms of the probabilities $q(z_1, z_2, z_3)$. However, Amari (2001) argued that $\overline{E}_{\mathcal{S}}$ can be interpreted as the set of distributions in which there are no three-way, or "triplewise" interactions between the variables Z_1, Z_2 and Z_3, beyond those that are implied by their pairwise interactions. The split distribution $p_{\mathcal{S}}$ can be calculated numerically as described above, in order to obtain the quantity $D_{\mathrm{KL}}(p_\top \| p_{\mathcal{S}})$, which quantifies the amount of information present in the triplewise interactions. Amari (2001) gives a straightforward generalisation, allowing n-way interactions to be quantified, among n or more random variables. As an example of triplewise information, consider the case where Z_1 and Z_2 are uniformly distributed binary variables, and $Z_3 = Z_1 \operatorname{XOR} Z_2$. In this case, in the split distribution $p_{(Z_1Z_2)(Z_1Z_3)(Z_2Z_3)}$ all three variables are independent. The split distribution has 8 equally likely outcomes while the true distribution has 4 equally likely outcomes, leading to a triplewise information of 1 bit.

3 The input lattice

The elements of the constraint lattice are formed from an arbitrary set of random variables $W = \{Z_1, \ldots, Z_m\}$. We are interested specifically in the case where W is composed of a set of input variables $X_1, \ldots, X_n$ and a target variable Y. We write V for the set of input variables, so $W = V \cup \{Y\}$.

We wish to decompose the mutual information $I(X_1, \ldots, X_n; Y)$ into a sum of terms $I_A(X_1, \ldots, X_n; Y)$, one for each subset A of the input variables. To do this, we start by noting that

$$I(X_1, \ldots, X_n; Y) = D_{\mathrm{KL}}(p_{(X_1,\ldots,X_n,Y)} \| p_{(X_1,\ldots,X_n)(Y)}) \,.$$

Because of this, we can use the constraint lattice to derive decompositions of the mutual information.

Consider the set of elements $\mathcal{S}$ of the constraint lattice such that $(X_1, \ldots, X_n)(Y) \leq \mathcal{S}$. This set forms a sublattice of the constraint lattice, i.e. a lattice under the same partial order. We call this sublattice the input lattice. The input lattice is highlighted in red in fig. 2, left.

The input lattice can also be thought of in a different way. Each element of the input lattice contains the element $\{Y\}$ and the element $\{X_1, X_2, \ldots, X_n\}$ along with its subsets. In addition, an element of the input lattice may also contain elements such as $\{X_1, X_2, Y\}$, which contain Y along with some members of V. In fact the input lattice consists of exactly those down-sets of 2^W that have this form. As we will now show, this means that the elements of the input lattice are in one-to-one correspondence with nonempty down-sets of 2^V, which means that members of the input lattice can be thought of as sets of sets of the input variables, rather than sets of sets of all variables.

As a matter of technical bookkeeping, some care needs to be taken over the empty set. Both $\emptyset$ and $\{\emptyset\}$ are down-sets of 2^V. However, in our current context it is necessary to temporarily exclude the empty set from consideration, because, as we show shortly, the input lattice is in a 1-1 correspondence with $J(2^V) \setminus \{\emptyset\}$ (that is, $J(2^V)$ with the empty down-set removed). We will need to consider the empty set again in section 7 once we make the connection to cooperative game theory.

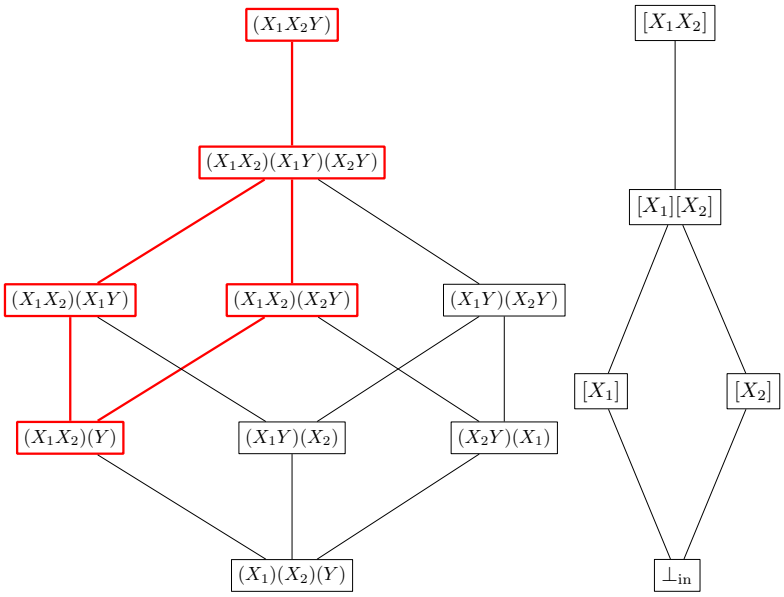


Figure 2: (Left) the Hasse diagram for the constraint lattice for $W = \{X_1, X_2, Y\}$. Highlighted in red bold is the sublattice that we call the input lattice, which provides decompositions of $I(X_1, X_2, \ldots, X_n; Y)$. (Right) the input lattice alone, with the nodes labelled using down-sets of 2^V (that is, sets of sets of input variables only), rather than 2^W. The square brackets indicate down-sets of 2^V. The two lattices are related by the mapping σ, defined in the text.

To see the connection between the input lattice and $J(2^V) \setminus \{\emptyset\}$, note that if an element $\mathcal{S}$ of the input lattice contains an element of the form $A \cup \{Y\}$, where A is a subset of the input variables, then it must also contain every element of the form $A' \cup \{Y\}$, where $A' \subseteq A$. Therefore these sets of input variables must by themselves form a down-set of 2^V, in order for $\mathcal{S}$ to be a down-set of 2^W.

To go in the other direction, let $\mathfrak{S}$ be a non-empty down-set of 2^V. The corresponding member of the constraint lattice is given by

$$\sigma(\mathfrak{S}) = (X_1 \ldots X_n) \cup \left\{ A \cup \{Y\} : A \in \mathfrak{S} \right\} .$$

We restrict ourselves to non-empty down-sets of 2^V because both $\emptyset$ and $\{\emptyset\}$ are down-sets of 2^V and there is no way to map $\emptyset$ to a member of the constraint lattice distinct from $\sigma(\{\emptyset\}) = (X_1 \ldots X_n)(Y)$. With this restriction, σ is a bijection. We have that $\sigma(\mathfrak{S})$ is a member of the constraint lattice for any $\mathfrak{S}$, and $\sigma(\mathfrak{S}) \geq (X_1 \ldots X_n)(Y)$. This allows us to think of the elements of the input lattice as corresponding to down-sets of 2^V. The mapping is illustrated in fig. 2.

In the following sections, unless otherwise noted, we will refer to the elements of the input lattice as down-sets of 2^V rather than as down-set covers of W. When writing members of the input lattice as down-sets of 2^V we use square brackets. So for example, the notation $[X_1][X_2]$ corresponds to the element $(X_1X_2)(X_1Y)(X_2Y)$ of the constraint lattice. We write the bottom node of the input lattice as $\perp_{\text{in}}$, which is equal to $\{\emptyset\}$ when considered as a down-set of 2^V, or $(X_1 \ldots X_n)(Y)$ when considered as a member of the constraint lattice.

We now consider chains from $\perp_{\text{in}}$ to $\top$ in the input lattice. Since these are also chains in the constraint lattice, each one provides a decomposition of $I(X_1, \ldots, X_n; Y)$ into a sum of non-negative terms. An example of such a decomposition is the chain rule for mutual information,

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) ,$$

which can be derived by applying the Pythagorean theorem to the (not maximal) chain

$$\perp_{\text{in}} < [X_1] < [X_1 X_2] \,.$$

This corresponds to

$$(X_1X_2)(Y) < (X_1X_2)(X_1Y) < (X_1X_2Y)$$

when considered as elements of the constraint lattice. Applying Equation 7, we have

$$D_{\text{KL}}(p_{(X_1X_2Y)} \| p_{(X_1X_2)(Y)}) = D_{\text{KL}}(p_{(X_1X_2)(X_1Y)} \| p_{(X_1X_2)(Y)}) + D_{\text{KL}}(p_{(X_1X_2Y)} \| p_{(X_1X_2)(X_1Y)}), \quad (9)$$

which corresponds term-by-term to the chain rule for mutual information. This chain is not maximal, but considering a maximal chain yields a more fine-grained decomposition:

$$\perp_{\text{in}} < [X_1] < [X_1][X_2] < [X_1X_2]$$

corresponds to a decomposition of the mutual information with three non-negative terms,

$$I(X_1, X_2; Y) = I(X_1; Y) + \big(I(X_2; Y|X_1) - I_3(X_1, X_2, Y)\big) + I_3(X_1, X_2, Y) \,,$$

where $I_3(X_1, X_2, Y)$ is Amari's triplewise information. (See example 2.3 above.)

In this way we can write $I(X_1, \ldots, X_n; Y)$ as a sum of non-negative terms in many different ways. However, these decompositions in general treat the input variables asymmetrically. The decompositions are "path-dependent," in the sense that they depend on which particular chain is chosen. In the next section we turn these path-dependent decompositions into a single path-independent one by suitably averaging over the maximal chains in the input lattice.

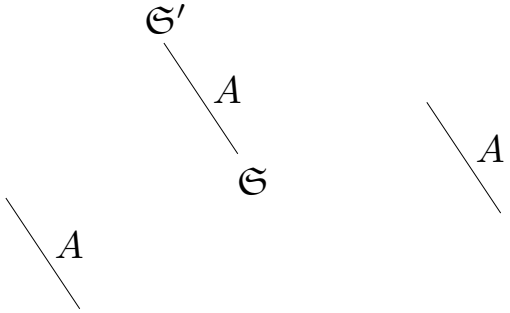
4 Defining the Information Attribution as a sum over chains

We have defined a decomposition of the mutual information for each chain from $\perp_{\text{in}}$ to $\top$ in the input lattice. We now define from this a path-independent Information Attribution measure. This decomposition will define a separate information contribution for each of the non-empty subsets A of V, that is, to the elements of 2^V rather than its lattice of down-sets.

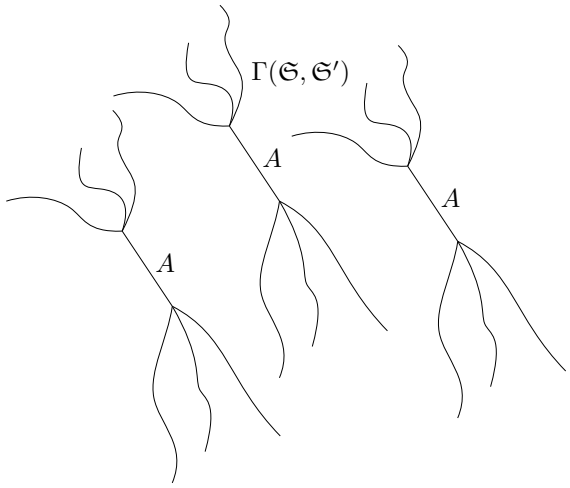
In order to do so, consider the set Γ of all maximal chains in the input lattice, that is, all directed paths from $\perp_{\text{in}}$ to $[X_1, \ldots, X_n]$. Consider a maximal chain $\gamma \in \Gamma$. For any index l in the chain, the collection $\gamma(l)$ of subsets forms a non-empty down-set of V and for each transition from $\gamma(l)$ to $\gamma(l+1)$ a subset A of V is added to the set $\gamma(l)$ until the topmost element, which ends up containing all subsets of V. In particular, the chain has the property that all non-empty subsets A of V are added at some point along a chain γ.

In particular, this ensures that there is exactly one $l_\gamma(A)$ that satisfies the following condition: all the elements $\gamma(l)$, $0 \leq l < l_\gamma(A)$ of the input lattice do not contain A, and all elements $\gamma(l)$, $l_\gamma(A) \leq l \leq 2^n - 1$, do contain A; i.e. $l_\gamma(A)$ denotes the step in the chain γ at which A is added. (The empty set $\emptyset$ is necessarily contained in the first complex of each chain, i.e. $l_\gamma(\emptyset) = 0$).

Based on this, we now derive a decomposition of the mutual information between inputs and output "aligned" with respect to a particular subset A of inputs. For this purpose, consider the set $\mathcal{E}_A$ of all edges $(\mathfrak{S}, \mathfrak{S}')$ where $\mathfrak{S}'$ is obtained from $\mathfrak{S}$ by adding A, i.e. where $\mathfrak{S}' = \mathfrak{S} \uplus \{A\}$:



We furthermore now subdivide the set Γ into classes of maximal chains in the input lattice, grouped by specific edges $(\mathfrak{S}, \mathfrak{S}')$. Denote by $\Gamma(\mathfrak{S}, \mathfrak{S}')$ the set of all maximal chains $\gamma \in \Gamma$ that contain this particular edge $(\mathfrak{S}, \mathfrak{S}')$:



Then, for any non-empty subset A, one has the following partition:

$$\Gamma = \biguplus_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} \Gamma(\mathfrak{S}, \mathfrak{S}') \,.$$

Every maximal chain is accounted for in this disjoint union, because for every maximal chain there is exactly one step (edge) at which the set A is added. This is illustrated in fig. 3 for the case of three input variables.

We now consider a probability weighting over the maximal chains, that is, a set of weights $\mu(\gamma)$ such that $\sum_{\gamma\in\Gamma} \mu(\gamma) = 1$. We obtain

$$\begin{aligned}
I(X_1, \ldots, X_n; Y) = & \quad D_{\mathrm{KL}}(p_{[X_1,\ldots,X_n]} \,\|\, p_{\perp_{\mathrm{in}}}) && (10)\\
= & \quad \sum_{\gamma\in\Gamma} \mu(\gamma) \sum_{l=1}^{2^n-1} D_{\mathrm{KL}}(p_{\gamma(l)} \,\|\, p_{\gamma(l-1)}) && (11)\\
= & \quad \sum_{\gamma\in\Gamma} \mu(\gamma) \sum_{\emptyset\neq A\subseteq V} D_{\mathrm{KL}}(p_{\gamma(l_\gamma(A))} \,\|\, p_{\gamma(l_\gamma(A)-1)}) && (12)\\
= & \quad \sum_{\emptyset\neq A\subseteq V} \sum_{\gamma\in\Gamma} \mu(\gamma) D_{\mathrm{KL}}(p_{\gamma(l_\gamma(A))} \,\|\, p_{\gamma(l_\gamma(A)-1)}) && (13)\\
= & \quad \sum_{\emptyset\neq A\subseteq V} \sum_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} \underbrace{\left\{ \sum_{\gamma\in\Gamma(\mathfrak{S},\mathfrak{S}')} \mu(\gamma) \right\}}_{=1} D_{\mathrm{KL}}(p_{\mathfrak{S}'} \,\|\, p_{\mathfrak{S}}) && (14)\\
= & \quad \sum_{\emptyset\neq A\subseteq V} \sum_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} \mu(\mathfrak{S}, \mathfrak{S}')\, D_{\mathrm{KL}}(p_{\mathfrak{S}'} \,\|\, p_{\mathfrak{S}}) && (15)
\end{aligned}$$

Here, we used the short-hand notation $\mu(\mathfrak{S}, \mathfrak{S}')$ for $\mu(\Gamma(\mathfrak{S}, \mathfrak{S}')) = \sum_{\gamma\in\Gamma(\mathfrak{S},\mathfrak{S}')} \mu(\gamma)$. The equality (11) follows because of the Pythagorean theorem (eq. (8)) and the normalization of the weights μ. Note, via (14), that the non-negative weights in the decomposition (15) satisfy the following condition:

$$\sum_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} \mu(\mathfrak{S}, \mathfrak{S}') \;=\; 1 \,. \qquad (16)$$

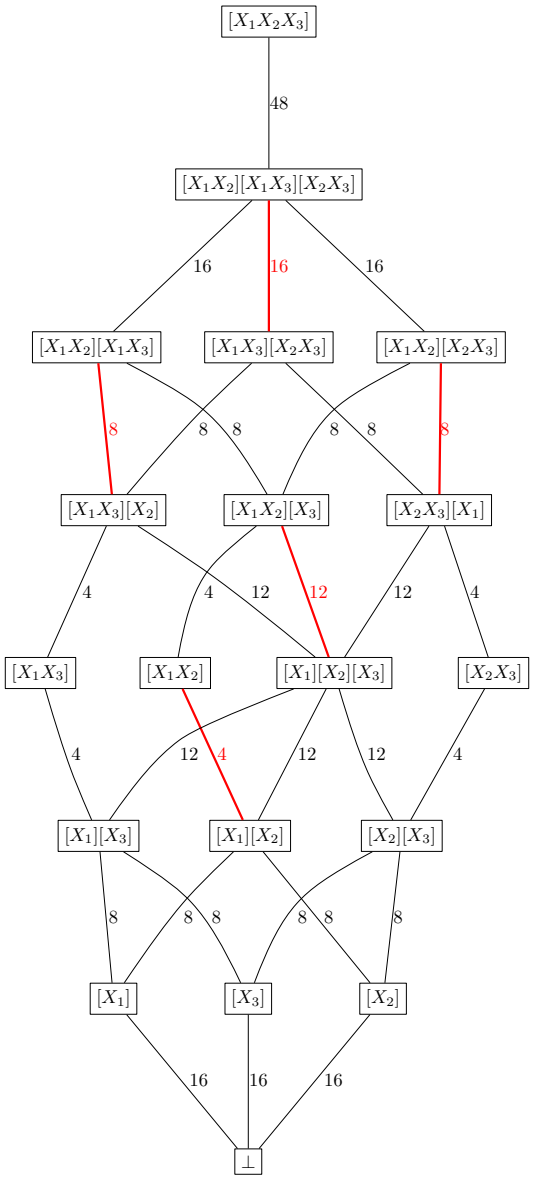


Figure 3: The input lattice for three inputs. Each edge is labelled with the total number of maximal chains that pass through that that edge. The edges where the subset $\{X_1, X_2\}$ appears for the first time are highlighted in red. Each maximal chain passes through exactly one of these edges. The contribution of $\{X_1, X_2\}$ to the total information is calculated by averaging over these edges, weighted by their path counts (the numbers in red.) In this lattice there are 48 maximal chains in total.

This allows us to interpret

$$I_A^{(\mu)}(X_1, \ldots, X_n\,;\, Y) \;:=\; \sum_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} \mu(\mathfrak{S},\mathfrak{S}')\, D_{\mathrm{KL}}(p_{\mathfrak{S}'} \,\|\, p_{\mathfrak{S}}) \tag{17}$$

as the mean information in A that is not contained in a proper subset of A.

This gives us a non-negative decomposition of $I(X_1, \ldots, X_n; Y)$ into terms corresponding to each subset of the input variables, but note that this decomposition is dependent on the choice of weights μ. This non-negative decomposition is related and refines the works (Olbrich et al., 2015; Perrone and Ay, 2016), where only particular down-sets in the input lattice were considered. More precisely, for each k, $1 \leq k \leq n$, the set of subsets of $\{1, \ldots, n\}$ with cardinality at most k forms a down-set which represents a distinguished vertex in the input lattice. This defines a particular (non-maximal) chain, indexed by k, from which we can define an information decomposition similar to the one proposed by Amari (2001) (see example 2.3 above). This chain corresponds to a single hierarchy of exponential families, which does not require weights μ of chains γ as we are considering here. In this sense, the present approach generalises the previous ones, which can be considered as a special case of ours by choosing μ to be concentrated on the specific chain outlined above.

A natural choice for the weights μ would be simply to choose the uniform distribution, i.e. $\mu(\gamma) = 1/|\Gamma|$ for all γ. It is not completely straightforward to justify the uniform distribution over maximal chains, because there is no obvious symmetry that transforms one maximal chain into another. Note, for example, that the connectivity of $[X_1][X_2][X_3]$ in the Hasse diagram in figure fig. 3 is different from that of other elements of the same rank.

Nevertheless, we will now proceed with the uniform distribution as a reasonable intuitive choice. It will be shown in section 5 that choosing μ this way gives rise to a decomposition of $I(X_1, \ldots, X_n; Y)$ that has some intuitively desirable properties. For the special choice of μ as the uniform distribution we will write $I_A^{(\mu)}(X_1, \ldots, X_n\,;\, Y)$ simply as $I_A(X_1, \ldots, X_n\,;\, Y)$. We refer to this as the share of the mutual information *attributed* to A. In section 7 we will then proceed to show that above originally merely intuitive choice of μ as uniform distribution finds a deeper justification in the theory of cooperative game theory.

For the practical calculation of I_A we first calculate the number $n_{(\mathfrak{S},\mathfrak{S}')}$ of maximal chains that pass through each edge $(\mathfrak{S}, \mathfrak{S}')$ in the Hasse diagram of the input lattice. These numbers are shown in fig. 3, as well as the total number of maximal chains, $|\Gamma|$. For each node $\mathfrak{S}$ in the lattice we calculate the distribution $p_{\mathfrak{S}}$ by iterative scaling (Csiszár and Shields, 2004, chapter 5), from which we obtain $D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\perp_{\mathrm{in}}})$. We then find the set $\mathcal{E}_A$ of edges in which a given predictor A is added for the first time in a maximal chain (for the example of $A = \{X_1, X_2\}$ this is shown in red in fig. 3). We then calculate our measure I_A from the changes in Kullback-Leibler divergence along these paths by adding the set A of interest, weighted by the chain counts $n_{(\mathfrak{S},\mathfrak{S}')}$ of the respective edges:

$$I_A(X_1, \ldots, X_n\,;\, Y) = \frac{1}{|\Gamma|} \sum_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} n_{(\mathfrak{S},\mathfrak{S}')} \big(D_{\mathrm{KL}}(p_{\mathfrak{S}'} \| p_{\perp_{\mathrm{in}}}) - D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\perp_{\mathrm{in}}}) \big)\,. \tag{18}$$

5 Properties of the Information Attribution

We now prove the following properties of our Information Attribution measure $I_A(X_1, \ldots, X_n\,;\, Y)$, as a decomposition of the mutual information.

Theorem 1. For $A \subseteq \{X_1, \ldots, X_n\}$ we have

I. $I_A(X_1, \ldots, X_n\,;\, Y) \geq 0$ (*nonnegativity*)

II. $\sum_{A\in 2^V} I_A(X_1, \ldots, X_n\,;\, Y) = I(X_1, \ldots, X_n\,;\, Y)$ (*completeness*)

III. $I_A(X_1, \ldots, X_n\,;\, Y)$ is invariant under permutations of $X_1, \ldots, X_n$. (*symmetry*)

IV. $I_A(X_1, \ldots, X_n\,;\, (X_1, \ldots, X_n)) = 0$ if $|A| > 1$. (*singleton*)

V. if $X_i = (X_i', X_i'')$ for all i, $Y = (Y', Y'')$, and

$$p(x_1, \ldots, x_n, y) = p(x_1', \ldots, x_n', y')\, p(x_1'', \ldots, x_n'', y'') \;,$$

then

$$I_A(X_1, \ldots, X_n\,;\, Y) = I_{A'}(X_1', \ldots, X_n'\,;\, Y') + I_{A''}(X_1'', \ldots, X_n''\,;\, Y'') \;,$$

where $A' = \{X_i' : (X_i', X_i'') \in A\}$ and $A'' = \{X_i'' : (X_i', X_i'') \in A\}$. (*additivity*)

As we discuss below, the singleton property is somewhat analogous to the identity axiom proposed by (Harder et al., 2013) for partial information decomposition measures, which effectively says that there should be no synergy terms if the output is simply an indentical copy of the input.

Proof. (I) follows from the nonnegativity of the Kullback-Leibler divergence. (II) is proved in Section 4 above. (III) is true by construction, since the values of the Kullback-Leibler divergences do not depend on the order in which the input variables are considered, and the uniform distribution over maximal chains in the input lattice is invariant to reordering the input variables.

To prove (IV), write $Y = (\overline{X}_1, \ldots, \overline{X}_n)$, where $\overline{X}_i$ is considered to be a copy of X_i, in the sense that X_i and $\overline{X}_i$ are separate random variables but we have

$$p(x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_n) = \begin{cases} p(x_1, \ldots, x_n) & \text{if } x_1 = \overline{x}_1,\ \ldots,\ x_n = \overline{x}_n \;, \\ 0 & \text{otherwise} \;, \end{cases} \tag{19}$$

which implies that $p(x_i, \overline{x}_i) = \delta_{x_i, \overline{x}_i} p(x_i)$, for every i. We then have

$$p_{\perp_{\text{in}}}(X_1, \ldots, X_n, \overline{X}_1, \ldots, \overline{X}_n) = p(X_1) \ldots p(X_n) p(\overline{X}_1, \ldots, \overline{X}_n) \;.$$

Consider now any edge $(\mathfrak{S}\backslash\{A\}, \mathfrak{S})$ in the Hasse diagram of the input lattice. There are two cases to consider:

(i) $A = \{X_i\}$ for some i. In this case $\sigma(\mathfrak{S})$ contains the element $\{X_i, Y\}$. Therefore, from its definition, the marginal $p_{\mathfrak{S}}(X_i, Y)$ must match the true marginal $p(X_i, Y)$, which implies that $p_{\mathfrak{S}}(x_i, \overline{x}_i) = \delta_{x_i, \overline{x}_i} p(x_i)$. However, $\sigma(\mathfrak{S} \setminus \{A\})$ does not contain the element $\{X_i, Y\}$, and so the marginal $p_{\mathfrak{S}\backslash\{A\}}(x_i, \overline{x}_i)$ may in general differ from $\delta_{x_i, \overline{x}_i} p(x_i)$, and $D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\mathfrak{S}\backslash A})$ can be nonzero.

(ii) $|A| > 1$. Consider first the case that $A = \{X_i, X_j\}$. Because $\mathfrak{S}$ is a down-set of V, we have that $\{X_i\} \in \mathfrak{S}\backslash\{A\}$ and $\{X_j\} \in \mathfrak{S}\backslash\{A\}$. Therefore $p_{\mathfrak{S}\backslash\{A\}}$ has to match the constraints $p_{\mathfrak{S}\backslash\{A\}}(x_i, \overline{x}_i) = \delta_{x_i, \overline{x}_i} p(x_i)$ and $p_{\mathfrak{S}\backslash\{A\}}(x_j, \overline{x}_j) = \delta_{x_j, \overline{x}_j} p(x_j)$. We also have, from eq. (19), that $p_{\mathfrak{S}\backslash\{A\}}(\overline{x}_i, \overline{x}_j) = p(\overline{x}_i, \overline{x}_j)$. From these constraints we have

$$p_{\mathfrak{S}\backslash A}(x_i, x_j, \overline{x}_i, \overline{x}_j) = p(\overline{x}_i, \overline{x}_j) \delta_{x_i, \overline{x}_i} \delta_{x_j, \overline{x}_j} = p(x_i, x_j, \overline{x}_i, \overline{x}_j) \;.$$

Therefore $p_{\mathfrak{S}\backslash A}$ already meets the constraint that the marginals for X_i, X_j, Y match those of the true distribution and minimising $D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\mathfrak{S}\backslash A})$ subject to this constraint must result in zero. The proof of this is similar if $|A| > 2$.

Therefore every term in eq. (15) will be zero if $|A| > 1$, but in general they can be nonzero if $|A| = 1$.

To prove (V) we first note the following general additivity property of the Kullback-Leibler divergence. Let Z' and Z'' be two co-distributed random variables, let $p_0(z', z'') = p_0(z') p_0(z'')$ for each z', z'' in the sample spaces of Z', Z'', that is, render the two random variables independent according to the distribution p_0. Then let M be a mixture family defined by constraints that depend only on either Z' or Z''. That is,

$$\begin{aligned} M = \Big\{ q : \sum_{z'} q(z') f^{(i)}(z') = F^{(i)} \quad (i = 1, \ldots, r) \;, \\ \sum_{z''} q(z'') g^{(j)}(z'') = G^{(j)} \quad (j = 1, \ldots, s) \Big\} \;. \end{aligned} \tag{20}$$

Calculating $\operatorname{argmin}_{p \in M} D_{\mathrm{KL}}(p \| p_0)$, introducing Lagrange multipliers in the usual way, gives us

$$p(z', z'') = p_0(z') p_0(z'') e^{\sum_i \lambda_i f^{(i)}(z') + \sum_j \eta_j g^{(j)}(z'') - \psi} = p(z') p(z'') \ ,$$

where $p(z') = p_0(z') e^{\sum_i \lambda_i f^{(i)}(z') - \psi'}$ and $p(z'') = p_0(z'') e^{\sum_j \eta g^{(j)}(z'') - \psi''}$. Note that these are the same distributions that would be obtained if the projection were performed on each of the marginals rather than the joint distribution. We have both that Z' and Z'' remain independent after projecting onto M, and also that $D_{\mathrm{KL}}(p(Z', Z'') \,\|\, p_0(Z', Z'')) = D_{\mathrm{KL}}(p(Z') \,\|\, p_0(Z')) + D_{\mathrm{KL}}(p(Z'') \,\|\, p_0(Z''))$.

Now consider constructing a system of random variables $X_i = (X_i', X_i'')$, $Y = (Y', Y'')$, according to the condition of property V. Each of the split distributions is defined as a projection from the product distribution onto a mixture family. By construction, all of these mixture families satisfy eq. (20). Because of this, every term in eq. (17) can be written as a sum of the corresponding terms for the systems $\{X_1', \ldots, X_n', Y'\}$ and $\{X_1'', \ldots, X_n'', Y''\}$. The additivity property follows from this.

□

We note that these properties do not uniquely determine the Information Attribution measure. In particular, one could choose a different measure μ over the maximal chains of the input lattice besides the uniform measure; there are in general many such measures that would yield an information measure satisfying theorem 1. To see this, note that the proofs of properties I, II, IV and V do not depend on the choice of measure μ, and hence don't constrain it. Property III does restrict the choice of measure, but for more than two inputs the number of paths in the lattice is greater than the number of inputs, and consequently the symmetry axiom does not provide enough constraint to uniquely specify μ. However, as argued above, the uniform measure is a natural choice, and we will show below that its use can be more systematically justified from the perspective of cooperative game theory.

5.1 Comparison to partial information decomposition

As noted above, the Information Attribution I_A is not a partial information decomposition (PID) measure, because it decomposes the mutual information into a different number of terms than the latter. In the case of two input variables X_1 and X_2, the PID has four terms (synergy, redundancy, and two unique terms), whereas the information contribution has only three, $I_{\{X_1\}}$, $I_{\{X_2\}}$ and $I_{\{X_1, X_2\}}$. The joint term, $I_{\{X_1, X_2\}}$, behaves somewhat similarly to a synergy term, and the two singleton contributions $I_{\{X_1\}}$ and $I_{\{X_2\}}$ have some similarity to the two unique information terms, but there is no separate term corresponding to shared/redundant information. Instead, the singleton terms behave as if they capture a combination of unique and redundant information. For more than two inputs, the terms of a partial information measure can be expressed in terms of a lattice known as the redundancy lattice (Williams and Beer, 2010), which is different from the constraint lattice or the input lattice discussed above.

Within the PID framework, (Harder et al., 2013) introduced the *identity axiom*, which states that a measure of redundant information $I_\cap$, should satisfy

$$I_\cap(X_1, X_2; (X_1, X_2)) = I(X_1; X_2) \ .$$

This is equivalent to saying that the corresponding measure of synergy, $I_\cup$, should be zero in the case where the output is a copy of its two input variables:

$$I_\cup(X_1, X_2; (X_1, X_2)) = 0 \ . \tag{21}$$

It was proven by Rauh et al. (2014) that there can be no non-negative PID measure that satisfies all of Williams and Beer's axioms together with the identity axiom. This can be achieved if we restrict ourselves to two input variables, but for three or more inputs there are distributions for which it cannot be achieved. (See Example RBOJ below.)

While the Information Attribution measure is not a PID measure, if we take the joint term $I_{\{X_1, X_2\}}$ to be analogous to a synergy term, then the singleton decomposition property (property IV), for two inputs,

is roughly analogous to eq. (21). Therefore our measure obeys an analog of the identity axiom for PID measures, alongside analogs of the non-negativity and symmetry axioms for PID measures. This is possible only because the Information Attribution is not a PID measure, and hence does not have to obey the precise set of Williams-Beer lattice axioms.

It is also worth comparing our Information Attribution measure with the framework proposed by James and Crutchfield (2017), which seeks a different kind of information decomposition from PID. In this framework, instead of decomposing the mutual information between a set of sources and a target, one instead wishes to decompose the joint entropy $H(Z_1, \ldots, Z_n)$ of several jointly distributed random variables, into a sum of terms corresponding to each subset of the variables. Our framework sits somewhere between this approach and PID, since we have the distinction between the inputs and the target, but we decompose $I(X_1, \ldots, X_n; Y)$ into a sum of terms corresponding to subsets of the inputs, in a similar manner to James and Crutchfield's proposal.

It is an open question whether our Information Attribution measure obeys analogs of other axioms that have been proposed for PID measures. Such axioms include the so-called target chain rule, which was proposed by Bertschinger et al. (2013) under the name "left chain rule" and has been discussed for example by Finn and Lizier (2018).

6 Examples

We now explore a few examples of our Information Attribution measure. Here, we apply it to joint distributions between a target and two or three inputs. Note that our framework does not require any restrictions on these joint distributions. In particular, it is expressly not assumed that the inputs are independent of one another. Importantly, the measures will in general be affected by dependent inputs, which is a desirable property of such a measure, because it has been observed before that appropriate attributions of joint interactions should depend on input correlations (see the discussion on *source* vs. *mechanistic* redundancy in Harder et al., 2013).

We take most of our examples from the literature on partial information decomposition, in particular (Williams and Beer, 2010; Griffith and Koch, 2014; Harder et al., 2013; Bertschinger et al., 2014). These examples are relatively standardised, and give some intuition for how our measure compares to PID measures.

We first explore some basic examples with two predictors, which are presented in table 1. For each of these examples, the method attributes an amount of information to the predictors $\{X_1\}$, $\{X_2\}$ and $\{X_1, X_2\}$. The numbers assigned to these sets are nonnegative and, together, they sum up to the mutual information $I(X_1, X_2; Y)$.

In the example RDN in table 1, the two inputs share a single bit of information about the target. In the PID framework, this typically corresponds to one bit of shared or redundant information. However, our Information Attribution measure does not try to identify redundancy as a separate term, and instead assigns half a bit to each of the predictors. The joint predictor $\{X_1, X_2\}$ is assigned a zero contribution. This reflects the fact that once the correlations between Y and the two individual predictors are known the three-way correlations are already determined, and so learning them does not reveal any extra information.

In the second example, XOR, we have $Y = X_1 \oplus X_2$, where $\oplus$ is the exclusive-or function. In this example, no contribution is assigned to the individual predictors X_1 and X_2, but one bit is assigned to the joint predictor $\{X_1, X_2\}$. This can be seen as a kind of synergy measure — it says that all of the information that the predictors give about the target is found in the three-way correlations between X_1, X_2 and Y, and none in the pairwise correlations between either predictor and the target. Interpreting this causally, it means that the causal influences of X_1 and X_2 on Y are strongly tied together. While the Information Attribution does not have a separate term corresponding to redundancy, we see that it characterises synergy in a rather intuitive way.

Our third and fourth examples are discussed in (Harder et al., 2013). The "two bit copy" operation plays an important role in the PID literature in the context of the identity axiom. The Information Attribution assigns one bit each to both of the predictors and none to the joint predictor, reflecting the fact that the two inputs each provide a different piece of information about the target. This can be compared to the

RDN

X_1	X_2	Y	p
0	0	0	1/2
1	1	1	1/2

predictor	contribution (bits)
$\{X_2\}$	1/2
$\{X_1\}$	1/2
$\{X_1, X_2\}$	0

XOR

X_1	X_2	Y	p
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

predictor	contribution (bits)
$\{X_2\}$	0
$\{X_1\}$	0
$\{X_1, X_2\}$	1

2 BIT COPY

X_1	X_2	Y	p
0	0	0	1/4
0	1	1	1/4
1	0	2	1/4
1	1	3	1/4

predictor	contribution (bits)
$\{X_2\}$	1
$\{X_1\}$	1
$\{X_1, X_2\}$	0

AND

X_1	X_2	Y	p
0	0	0	1/4
0	1	0	1/4
1	0	0	1/4
1	1	1	1/4

predictor	contribution (bits)
$\{X_2\}$	0.40563765
$\{X_1\}$	0.40563762
$\{X_1, X_2\}$	0

SYNRDN

X_1	X_2	Y	p
0	0	0	1/8
0	1	1	1/8
1	0	1	1/8
1	1	0	1/8
2	2	2	1/8
2	3	3	1/8
3	2	3	1/8
3	3	2	1/8

predictor	contribution (bits)
$\{X_2\}$	1/2
$\{X_1\}$	1/2
$\{X_1, X_2\}$	1

Table 1: Examples of the Information Attribution for several simple two-predictor cases. For each example the joint distribution is shown on the left, and on the right we tabulate $I_{\{X_1\}}$, $I_{\{X_2\}}$ and $I_{\{X_1,X_2\}}$, the contributions made by the two singleton predictors $\{X_1\}$ and $\{X_2\}$ and the joint predictor $\{X_1, X_2\}$. These three values always sum to the mutual information $I(X_1, X_2; Y)$. All logarithms are taken to base 2, so that the numbers are in bits. The interpretation of these examples is given in the text.

PID framework, since it is usually seen as desirable for a PID measure to assign zero bits of synergy in this case. Note, however, that because our Information Attribution measure does not separate redundancy from unique information, it does not distinguish between this case and the case of RDN, where the information is also shared equally between the two predictors. The results for the AND distribution are similar, telling us that there is also no synergy in this case. This is because for AND the joint distribution can be inferred completely by knowing the marginals (X_1, Y), (X_2, Y) and (X_1, X_2), and consequently there is no triplewise information.

Our final two-predictor example is SYNRDN, which can be formed by combining the XOR example with an independent copy of the RDN example. The values assigned to the two predictors and the joint predictor are simply the sum of their values in the original two examples, which is a result of the additivity property (theorem 1, part V).

Table 2 shows the results for three input variables. In this case the method assigns an amount of information to every non-empty subset of $\{X_1, X_2, X_3\}$, representing the share of the mutual information provided by that set of inputs. The first example, PARITY, is a three-input analog of the XOR example, since $Y = X_1 \oplus X_2 \oplus X_3$. In this example it is not possible to infer anything about the value of Y until the values of all three inputs are known. Correspondingly, the method assigns all of the total mutual information (1 bit) to the predictor $\{X_1, X_2, X_3\}$ and none to the others.

Our second example, XORMULTICOAL (which we take from (Griffith and Koch, 2014)) has the property that knowing any single input gives no information about the target, but any pair of predictors completely determines it. This is reflected in the contributions assigned by the Information Attribution: the singleton predictors $\{X_1\}$, $\{X_2\}$ and $\{X_3\}$ each make no contribution to the total. Instead, the total one bit of mutual information is shared equally between the three two-input predictors, $\{X_1, X_2\}$, $\{X_1, X_3\}$ and $\{X_2, X_3\}$. The three-input predictor $\{X_1, X_2, X_3\}$ makes no contribution, because the target is already fully determined by knowing any of the pairwise predictors.

The third example, RBOJ, played an important role in the literature on PID, because it was used in (Rauh et al., 2014) to prove that no partial information decomposition is possible that obeys the so-called identity axiom, along with the axioms of Williams and Beer (2010) and local-positivity. In particular, no such decomposition is possible for this distribution. In this joint distribution, the inputs X_1, X_2 and X_3 are related by the exclusive-or function, and the target Y is in a one-to-one relationship with its inputs. As a result, each input provides one bit (in the usual sense) of information about the target, and each pair of inputs provides two bits, which completely determine the target. Consequently, learning the third input adds no new information about the target, if the other two are already known. Because Information Attribution is different from PID, it is able to assign non-negative values to each of the predictors. It shares out the total two bits of mutual information equally between the three singleton predictors, $\{X_1\}$, $\{X_2\}$ and $\{X_3\}$. This can be seen as a compromise between the fact that the contributions of each member of a pair of input variables are independent (similarly to the 2-bit copy) and that they, at the same time, need to be fairly allocated to three variables.

We finish with an example, THREE WAY AND, in which the decomposition is less intuitive. In this case, the target is 1 if and only if all three inputs are 1. Similarly to the AND example, our measure divides the information contributions between the three singleton predictors, assigning none to the two- or three-input predictors. The reason for this is similar to the AND example. Because of this, from the perspective of our measure, this example looks similar to the RBOJ example.

7 Cooperative game theory and weighted path summation

In section 4 we defined our Information Attribution measure $I_A(X_1, \ldots, X_n ; Y)$ based on a uniform weighting of the maximal chains in the input lattice. In this section we return to the question of how this uniform distribution would be justified.

To do so, we use the notion of the *Shapley value* (Shapley, 1953) from cooperative game theory. Informally, the idea of the Shapley value is that one has a set of players $N = \{A_i, i = 1 \ldots |N|\}$. Subsets of the players are called *coalitions*, and each coalition is assigned a *total worth*, which is to be interpreted as how well that

PARITY

X_1	X_2	X_3	Y	p
0	0	0	0	1/8
0	0	1	1	1/8
0	1	0	1	1/8
0	1	1	0	1/8
1	0	0	1	1/8
1	0	1	0	1/8
1	1	0	0	1/8
1	1	1	1	1/8

predictor	contribution (bits)
$\{X_1\}$	0
$\{X_2\}$	0
$\{X_3\}$	0
$\{X_1, X_2\}$	0
$\{X_1, X_3\}$	0
$\{X_2, X_3\}$	0
$\{X_1, X_2, X_3\}$	1
total	1

XORMULTICOAL

X_1	X_2	X_3	Y	p
4	0	4	0	1/8
0	2	2	0	1/8
1	1	0	0	1/8
5	3	6	0	1/8
5	1	4	1	1/8
1	3	2	1	1/8
0	0	0	1	1/8
4	2	6	1	1/8

predictor	contribution (bits)
$\{X_1\}$	0
$\{X_2\}$	0
$\{X_3\}$	0
$\{X_1, X_2\}$	1/3
$\{X_1, X_3\}$	1/3
$\{X_2, X_3\}$	1/3
$\{X_1, X_2, X_3\}$	0
total	1

RBOJ

X_1	X_2	X_3	Y	p
0	0	0	0	1/4
0	1	1	1	1/4
1	0	1	2	1/4
1	1	0	3	1/4

predictor	contribution (bits)
$\{X_1\}$	2/3
$\{X_2\}$	2/3
$\{X_3\}$	2/3
$\{X_1, X_2\}$	0
$\{X_1, X_3\}$	0
$\{X_2, X_3\}$	0
$\{X_1, X_2, X_3\}$	0
total	2

THREE WAY AND

X_1	X_2	X_3	Y	p
0	0	0	0	1/8
0	0	1	0	1/8
0	1	0	0	1/8
0	1	1	0	1/8
1	0	0	0	1/8
1	0	1	0	1/8
1	1	0	0	1/8
1	1	1	1	1/8

predictor	contribution (bits)
$\{X_1\}$	0.18118725
$\{X_2\}$	0.18118724
$\{X_3\}$	0.18118724
$\{X_1, X_2\}$	0
$\{X_1, X_3\}$	0
$\{X_2, X_3\}$	0
$\{X_1, X_2, X_3\}$	0
total	0.54356444

Table 2: Some examples of our measure, applied to joint distributions between a target and three inputs. The interpretation of these examples is given in the text.

Shapley Theory	**Information Attribution**
player A, B, C	set $A, B, C \subseteq V$ of input variables
coalition	down-set of 2^V, a set of sets of input variables
set of feasible coalitions	the set of down-sets of 2^V, written $\mathcal{D} = J(2^V)$
empty coalition $\emptyset$	empty down-set $\mathfrak{S}^{(0)} = \emptyset \in J(2^V)$
coalition of all players $\mathfrak{N}$	set of all subsets of V, i.e. $2^V \in J(2^V)$
worth v of a coalition $\mathfrak{S}$	$D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\perp_{\mathrm{in}}})$, or 0 if $\mathfrak{S} = \emptyset$.
Shapley value $\phi_A(v)$	information contribution $I_A(X_1, \ldots, X_n\,;\, Y)$

Table 3: Correspondence between our quantities and coalitional game theory. Note that the empty coalition/down-set is included at the bottom of the lattice.

set of players could do at some task (measured in terms of some payoff), without the participation of the remaining non-coalition players. Given this data, the problem is to assign a payoff to each individual player, such that the payoffs of each individual player sum up to the total worth of the coalition. The players' individual payoffs should reflect their "fair" contribution in achieving the total worth.

For this assignment of payoffs to be uniquely characterized, Shapley postulates that these payoffs assigned to players should be a linear function of the coalitions' worths, a notion of relevance (explained below) and a notion of symmetry amongst the players (where players whose contribution to the coalitional worth cannot be distinguished via a symmetrical exchange of players should attain the same Shapley value). The basic Shapley value assumes that all subsets of N are possible as coalitions. Since Shapley's original work, many generalizations of the Shapley value have been developed (Bilbao, 1998; Bilbao and Edelman, 2000; Lange and Grabisch, 2009; Faigle and Grabisch, 2012, 2013).

The purpose of our measure $I_A(X_1, \ldots, X_n\,;\, Y)$ is to attribute to each predictor A (that is, each member of 2^V) a uniquely characterized share of the mutual information $I(X_1, \ldots, X_n\,;\, Y)$. Equation (18) calculates this as a linear function of the quantities $D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\perp_{\mathrm{in}}})$, which can be thought of as that part of the total information that can be seen to have been contributed to a given set $\mathfrak{S}$ of predictors. This is closely reminiscent to the central problem in cooperative game theory of identifying the contribution of a particular player to a total worth when the worth of all valid coalitions of players are known and which is solved by the concept of the Shapley value.

In fact, we can apply cooperative game theory directly to our problem, in the following way: we consider a cooperative game in which each player corresponds to a predictor. That is, each player is a set of input variables, or member of 2^V, the empty set included. In this game, a worth v is assigned not to individual players but to *coalitions*, which are sets of players, that is, sets of sets of input variables. To apply it for the purpose of information decomposition, we consider a coalitional game whose payoff function (i.e. coalitional worth) is given by

$$v(\mathfrak{S}) = D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\perp_{\mathrm{in}}}), \tag{22}$$

and, in particular, the worth achieved by the whole set of input variables becomes $I(X_1, \ldots, X_n\,;\, Y)$.

In formulating this particular cooperative game, we now encounter the additional complication that not every set of players forms a viable coalition, since we want the coalitions to correspond to the elements of the input lattice. This constrains each coalition $\mathfrak{S}$ to be a down-set of 2^V. Explicitly, if a given set of input variables is a member of a coalition, then all of its subsets must be members of the coalition as well. We thus need a modified formulation which permits us to restrict the possible coalitions in this way while still allowing a payoff to be uniquely assigned to each player. This restriction necessitates a modification of the symmetry axiom of the original Shapley value to guarantee that the generalized Shapley allocation becomes uniquely determined.

Concretely, here we argue that the specific quantity in Eq. (17) can be interpreted precisely as the generalized Shapley value under *precedence constraints* in the sense of Faigle and Kern (1992).

For consistency and to highlight the parallels, we will use for cooperative games a notation similar to the notation we have used so far for the information quantities. We use the symbols $A, B, C, \ldots$ for

players (members of 2^V), and similarly $\mathfrak{S}, \mathfrak{S}', \ldots$ for coalitions (down-sets of 2^V), and $\mathcal{D}$ for the set of all feasible coalitions, to keep notation coherent. Finally, let $\mathfrak{N}$ denote the set of all players. Table 3 gives the relationship between game-theoretic quantities and the quantities defined in previous sections. To simplify the exposition and render it coherent with respect to existing literature on cooperative game theory, we allow both the empty coalition $\emptyset$ as well as the coalition consisting only of the empty player, $\{\emptyset\}$, to be valid coalitions coalitions. Thus the set of feasible coalitions becomes $\mathcal{D} = J(2^V)$ rather than $J(2^V) \setminus \{\emptyset\}$ as in previous sections. We assign a worth of 0 to the empty coalition $\emptyset$. This does not affect any of the following results.

In what follows, we introduce Faigle and Kern's extension of the Shapley value, and then show that that applying it to this 'information game' is indeed equivalent to eq. (17) with μ taken as the uniform distribution over maximal chains, resulting in eq. (18). This demonstrates that our measure obeys the axioms of Linearity, Carrier and Hierarchical strength, described below, which are used to derive Faigle and Kern's result.

7.1 Shapley Value under Precedence Constraints

We now proceed to define the (generalized) Shapley value under precedence constraints as defined in (Faigle and Kern, 1992). For brevity, when we henceforth say "Shapley value", we will refer to this variant unless stated otherwise.

Let $\mathfrak{N}$ be an arbitrary finite partially ordered set of players, where for $A, B \in \mathfrak{N}$ the relation $B \leq A$ enforces that, if $A \in \mathfrak{S}$ for any coalition $\mathfrak{S} \subseteq \mathfrak{N}$, one also has $B \in \mathfrak{S}$. In other words, the permitted coalitions, known as *feasible* coalitions, are down-sets of the poset $\mathfrak{N}$. The set $\mathcal{D}$ of feasible coalitions is closed under intersection and union operations, but not necessarily under the complement operation.

A *cooperative game* on $\mathfrak{N}$ is now a function

$$v : \mathcal{D} \to \mathbb{R} \tag{23}$$

such that $v(\emptyset) = 0$. $v(\mathfrak{S})$ is called the *worth* of the coalition $\mathfrak{S}$, and is to be interpreted as the total payoff assigned to it in the game. It measures the total "achievement" of the players in this coalition.

The above definition of the game assigns a worth to all possible coalitions. While there are also single-player coalitions, cooperative game theory asks the question how to quantify the performance of a given player across all possible coalitions. One natural way to do is the *Shapley value*. The idea of the Shapley value is, given the overall worth function v, to assign a payoff to individual players in such a way that it fairly reflects each player's overall contribution across all the coalitions. To do this formally, consider the vector space Υ of all cooperative games $v : \mathcal{D} \to \mathbb{R}$ on $\mathfrak{N}$, where the vector space structure on Υ is given by elementwise addition and scalar multiplication of each coalition's worth. Then the Shapley value is a function

$$\Phi : \Upsilon \to \mathbb{R}^{\mathfrak{N}} \tag{24}$$

which defines, for each player A from $\mathfrak{N}$, their share $\Phi_A(v)$ of the total worth $v(\mathfrak{N})$ (the worth for the coalition consisting of the complete set of players) for the game v in a way that fulfils a particular set of axioms. The axioms for the Shapley value are as follows:

Axiom 1 (Linearity)**.** For all $c \in \mathbb{R}, v, w \in \Upsilon$, demand

$$\Phi(c\, v) = c\, \Phi(v)$$
$$\Phi(v + w) = \Phi(v) + \Phi(w)$$

Axiom 2 (Carrier)**.** Call a coalition $\mathfrak{U} \in \mathcal{D}$ a *carrier* of $v \in \Upsilon$ if $v(\mathfrak{S}) = v(\mathfrak{S} \cap \mathfrak{U})$ for all $\mathfrak{S} \in \mathcal{D}$. Then, if $\mathfrak{U}$ is a carrier of v, we have

$$\sum_{A \in \mathfrak{U}} \Phi_A(v) = v(\mathfrak{U})\ . \tag{25}$$

The carrier axiom needs a brief explanation. It unifies two intuitive axioms that are sometimes used instead, the *dummy axiom* (a player that does not affect the worth — or payoff — of any coalition is

irrelevant for these and thus attains a neutral Shapley value of 0) and the *efficiency axiom* (the sum of the Shapley values of all players sums up to the total payoff of the whole set of players). Note that, as a special case, this axiom guarantees that $v(\mathfrak{N})$ is the sum of the Shapley values of all individual players, i.e. it distributes the total available worth across the players.

The third axiom of the traditional Shapley value postulates that players whose contribution to coalition payoffs are equivalent with respect to a symmetric permutation will also receive the same Shapley allocation. This axiom cannot be directly used in our case, because, while for the traditional Shapley value v all possible subsets of $\mathfrak{N}$ are permitted as coalitions, we here have the additional restriction that the coalitions of $\mathfrak{N}$ must be *feasible*. To still obtain the a unique characterization, a generalized Shapley value is used which imposes a stronger requirement. There are several axiom sets which are equivalent on the ordered coalitional games discussed here (see introduction of section 7 above). We follow Faigle and Kern (1992) in choosing the formulation via hierarchical strength.

We need a number of definitions. Call an injective map (which in the special setups considered here is also bijective)

$$\pi : \mathfrak{N} \to \{1, 2, \ldots, |\mathfrak{N}|\}$$

a *(feasible) ranking* of the players in $\mathfrak{N}$ if for all $A, B \in \mathfrak{N}$ we have that $A < B$ (i.e. $A \leq B$ and $A \neq B$) implies $\pi(A) < \pi(B)$.

The ranking π of $\mathfrak{N}$ induces a ranking $\pi_{\mathfrak{S}} : \mathfrak{S} \to \{1, 2, \ldots, |\mathfrak{S}|\}$ on all (feasible) coalitions $\mathfrak{S} \in \mathcal{D}$ via $\pi_{\mathfrak{S}}(A) < \pi_{\mathfrak{S}}(B)$ if and only if $\pi(A) < \pi(B)$ for all $A, B \in \mathfrak{S}$. Note that $\pi_{\mathfrak{S}}$ only inherits the order from π, but in general an induced rank $\pi_{\mathfrak{S}}(A)$ will differ from $\pi(A)$, the original one.

We generalize the concept of feasible rankings to any subset $\mathfrak{P} \subseteq \mathfrak{N}$ of the players with partial order analogously, namely as a bijection $\pi : \mathfrak{P} \to \{1, 2, \ldots, |\mathfrak{P}|\}$ that preserves that partial order. For any such set $\mathfrak{P}$ of players, we denote the set of all feasible rankings of $\mathfrak{P}$ by $\mathcal{R}(\mathfrak{P})$.

We say that player $C \in \mathfrak{S}$ is $\mathfrak{S}$-maximal in the ranking π if $\pi_{\mathfrak{S}}(C) = |\mathfrak{S}|$ which is the same as saying that $\pi(C) = \max_{A \in \mathfrak{S}} \pi(A)$ or that there is no player A in coalition $\mathfrak{S}$ with $C < A$.

We are now ready to express the concept of hierarchical strength: the *hierarchical strength* $h_{\mathfrak{S}}(C)$ of the player C in $\mathfrak{S}$ is defined as the proportion of (total) rankings π in which C is $\mathfrak{S}$-maximal. Formally,

$$h_{\mathfrak{S}}(C) := \frac{1}{|\mathcal{R}(\mathfrak{N})|} |\{\pi \in \mathcal{R}(\mathfrak{N}) \mid C \text{ is } \mathfrak{S}\text{-maximal for } \pi\}| \tag{26}$$

where $\mathcal{R}(\mathfrak{N})$ is the set of all (feasible) rankings for the set $\mathfrak{N}$ of players.

Define now a particular fundamental game type, the *unanimity game* centered on a coalition $\mathfrak{S} \neq \emptyset$, which we call $\zeta_{\mathfrak{S}}$ via:

$$\zeta_{\mathfrak{S}}(\mathfrak{T}) := \begin{cases} 1 & \text{if } \mathfrak{S} \subseteq \mathfrak{T} \\ 0 & \text{otherwise .} \end{cases} \tag{27}$$

In other words, the payoff of the game is 1 if the tested coalition $\mathfrak{T}$ encompasses a given reference coalition $\mathfrak{S}$ and vanishes otherwise. We mention without proof that these games form a basis of Υ and thus it is sufficient to define the Shapley value over all unanimity games on $\mathfrak{N}$.

Finally, we can now define

Axiom 3 (Hierarchical Strength (Equivalence))**.** For any $\mathfrak{S} \in \mathcal{D}$, $A, B \in \mathfrak{S}$, we demand:

$$h_{\mathfrak{S}}(A)\Phi_B(\zeta_{\mathfrak{S}}) = h_{\mathfrak{S}}(B)\Phi_A(\zeta_{\mathfrak{S}}) \tag{28}$$

Informally, this means that the Shapley value of a player B in a coalition $\mathfrak{S}$ for the unanimity game is weighted against that of another player A in the same coalition via their hierarchical strength. Everything else being equal, the Shapley values of the two players relate to each other as their hierarchical strengths — a larger value of the hierarchical strength corresponds to a larger Shapley value, i.e. larger allocation of payoff.

Faigle and Kern (1992) note that the hierarchical strength emphasizes the given player being *on top* of its respective coalition in the given ranking rather than, say, considering its average rank. This is insofar an

intuitive choice for the generalized Shapley value, since it is only the top-ranked player in a coalition which determines whether that particular coalition is formed at all. In other words, it is a measuring in how many rankings (relative to the total number of rankings) that particular player has the power to decide whether the given coalition will be formed or not.

It turns out this has a straightforward reinterpretation and generalization in the context of Markovian coalition processes (Faigle and Grabisch, 2012) In addition, there are many other formulations equivalent with it (see references mentioned in the introduction to the present chapter section 7). We opted for the formulation via hierarchical strength since it is the most widely established generalization of the symmetry axiom for the classical Shapley value in the literature.

We state here without proof that the unique payoff allocation of this generalized Shapley value is given by

$$\Phi_C(v) = \frac{1}{|\mathcal{R}(\mathfrak{N})|} \sum_{\substack{\mathfrak{T}\in\mathcal{D}:\\ C\in\mathfrak{T}^+}} |\mathcal{R}(\mathfrak{T}\setminus\{C\})| \cdot |\mathcal{R}(\mathfrak{N}\setminus\mathfrak{T})| \, \big(v(\mathfrak{T}) - v(\mathfrak{T}\setminus\{C\})\big) \tag{29}$$

where we trivially assume $|\mathcal{R}(\emptyset)| = 1$ and where $\mathfrak{T}^+$ denotes the set of maximal players of $\mathfrak{T}$. The sum therefore sums over all coalitions $\mathfrak{T}$ for which the player C is maximal. The (generalized) Shapley value of C is thus given by the marginal contribution of C to all coalitions $\mathfrak{T}$ for which it is maximal, weighted by the proportion of rankings for which this is the case.

We now show that our definition of the Information Attribution $I_A(X_1, \ldots, X_n; Y)$ of a set of input variables A, eq. (18), is equivalent to the generalized Shapley value under precedence constraints where the worth of each coalition $\mathfrak{S}$ is given by $D_{\mathrm{KL}}(p_{\mathfrak{S}} \| p_{\perp_{\mathrm{in}}})$ (and 0 for the special coalition $\emptyset$). Thus, the Information Attribution has a natural interpretation in the context of game-theoretic Shapley value allocation.

7.2 Equivalence of Generalized Shapley Value and the Sum over Maximal Chains

We now return to definition of our Information Attribution measure, which is defined by eq. (17), together with the choice of a uniform measure μ over the set Γ of maximal chains. The choice of the uniform measure is justified by the following theorem, which uses eq. (29) to show that when the uniform measure is chosen, the Information Attribution measure becomes equal to the Shapley value under precedence constraints.

Theorem 2. Under the identifications of table 3, the information attributed to a predictor (i.e. set) A is identical with its Shapley value under precedence constraints, with A now interpreted as a player. More precisely:

$$\sum_{(\mathfrak{S},\mathfrak{S}')\in\mathcal{E}_A} \mu(\mathfrak{S},\mathfrak{S}')\, D_{\mathrm{KL}}(p_{\mathfrak{S}'} \,\|\, p_{\mathfrak{S}}) = \sum_{\substack{\mathfrak{S}\in\mathcal{D}:\\ A\in\mathfrak{S}^+}} \frac{|\mathcal{R}(\mathfrak{S}\setminus\{A\})| \cdot |\mathcal{R}(2^V\setminus\mathfrak{S})|}{|\mathcal{R}(2^V)|} \left[D_{\mathrm{KL}}(p_{\mathfrak{S}}||p_{\{\emptyset\}}) - D_{\mathrm{KL}}(p_{\mathfrak{S}\setminus\{A\}}||p_{\{\emptyset\}})\right] , \tag{30}$$

where the weighting μ of the lattice chains γ is chosen as the uniform distribution over the set Γ of maximal chains, that is, $\mu(\gamma) = 1/|\Gamma|$ and the rankings $\mathcal{R}(\,.\,)$ of sets of players are with respect to the order relation "$\leq$" induced by inclusion $B \leq A :\Longleftrightarrow B \subseteq A$.

To permit consistency between lattice and Shapley model, we furthermore define the bracketed term on the right side to be 0 if $\mathfrak{S} = \emptyset$ or $\mathfrak{S} = \{\emptyset\}$.

Proof. Consider $\mathfrak{N} = 2^V$. Identify the elements $A \in \mathfrak{N}$, i.e. the subsets of V, with the players in a Shapley coalitional game with partial ordering defined via the subset relation, i.e. via

$$B \leq A :\Longleftrightarrow B \subseteq A\,.$$

Per definition, the partial order-compatible coalitions $\mathfrak{S}$ then precisely constitute the down-sets of $\mathfrak{N}$.

We now show that, under these identifications, the (feasible) rankings of players define precisely the maximal chains over the down-sets of $\mathfrak{N}$. In other words, there is a one-to-one correspondence between the rankings of the ordered coalitional game and the maximal chains over its corresponding down-sets.

First, we establish that the orders in which the predictors are added in a maximal chain correspond precisely to the feasible rankings of the predictors interpreted as Shapley players.

A (feasible) ranking π of $\mathfrak{N}$ can be interpreted, in the terminology of Stanley (2011), as a bijective *order-preserving map* $\pi : \mathfrak{N} \to [2^{|V|}]$ where $[2^{|V|}]$ is the ordered set of natural numbers $\{1, \ldots, 2^{|V|}\}$. We note that $[2^{|V|}] = [|2^V|])$. In other words, π is a topological sorting of $\mathfrak{N}$. According to the proof of Proposition 3.5.2 (see also proof of Proposition 3.5.1) in Stanley (2011), there is a one-to-one relation between the maximal chains

$$\begin{aligned} \mathfrak{S}_0 := \emptyset, \mathfrak{S}_1, \mathfrak{S}_2, \ldots, \mathfrak{S}_{|\mathfrak{N}|} \\ = (\mathfrak{S}_k)_{k=0,\ldots,|\mathfrak{N}|} = (\mathfrak{S}_k)_k \end{aligned} \tag{31}$$

of $\mathcal{D}$ and the (feasible) rankings $\pi \in \mathcal{R}(\mathfrak{N})$.

For the sake of self-containedness, we additionally give a proof in Appendix A that relies only on terms introduced in the present paper. The intuition of the proof is that in the Hasse diagram for the input lattice (section 4), each maximal chain is formed by successively adding each of the predictors, one at a time, in such a way that each step of the chain remains a down-set.

Let $(\mathfrak{S}, \mathfrak{S}') \in \mathcal{E}_A$ be given, i.e. an edge where $\mathfrak{S}' = \mathfrak{S} \cup \{A\}$ is obtained by adding A to $\mathfrak{S}$. Consider now the set $\Gamma(\mathfrak{S}, \mathfrak{S}')$ of (maximal) chains $(\mathfrak{S}_k)_k$ that pass through this edge, i.e. for which $\mathfrak{S}_j = \mathfrak{S}$ and $\mathfrak{S}_{j+1} = \mathfrak{S}'$ for some j.

In analogy of the one-to-one relation between maximal chains and their feasible rankings established above, we can find a one-to-one map between $\Gamma(\mathfrak{S}, \mathfrak{S}')$ and

$$\mathcal{R}(\mathfrak{S}' \setminus \{A\}) \times \mathcal{R}(\mathfrak{N} \setminus \mathfrak{S}'), \tag{32}$$

the set of pairs of rankings over $\mathfrak{S}' \setminus \{A\}$ and $\mathfrak{N} \setminus \mathfrak{S}'$.

This is seen by replacing the full ranking with two subrankings, one over the lower sublattice with $\mathfrak{S}$ as top element instead of $\mathfrak{N}$ and one over the upper one which has $\mathfrak{S}'$ as bottom element replacing $\mathfrak{S}_0$ and applying the same argument from Stanley (2011) as above on the two sublattices.

It follows that we have

$$|\Gamma(\mathfrak{S}, \mathfrak{S}')| = |\mathcal{R}(\mathfrak{S}' \setminus \{A\})| \cdot |\mathcal{R}(\mathfrak{N} \setminus \mathfrak{S}')|. \tag{33}$$

Consider a particular edge $(\mathfrak{S}, \mathfrak{S}') \in \mathcal{E}_A$. We note that this edge corresponds precisely to the down-set $\mathfrak{S}' \in \mathcal{D}$ where A is maximal in $\mathfrak{S}'$, i.e. $A \in \mathfrak{S}'^+$. We had earlier the short-hand notation $\mu(\mathfrak{S}, \mathfrak{S}') = \sum_{\gamma \in \Gamma(\mathfrak{S}, \mathfrak{S}')} \mu(\gamma)$ (where $\Gamma(\mathfrak{S}, \mathfrak{S}')$ again ranges over all maximal chains containing a particular edge). If all chains/paths γ over the full lattice are equally weighted, their weight is given by

$$\frac{1}{|\Gamma|} = \frac{1}{|\mathcal{R}(\mathfrak{N})|} = \frac{1}{|\mathcal{R}(2^V)|} \tag{34}$$

where 2^V is considered as a poset ordered by set inclusion. Finally, note that

$$D_{\mathrm{KL}}(p_{\mathfrak{S}'} \,\|\, p_{\mathfrak{S}}) = D_{\mathrm{KL}}(p_{\mathfrak{S}'} || p_{\perp_{\mathrm{in}}}) - D_{\mathrm{KL}}(p_{\mathfrak{S}' \setminus \{A\}} || p_{\perp_{\mathrm{in}}}) \tag{35}$$

because of the Pythagorean relation (8) (and noting that in our case $\perp_{\mathrm{in}} = \{\emptyset\}$). This completes the proof of (30). □

Note that, when constructing the correspondence between the input lattice to the Shapley value, for the former we had the maximal chains start at $\{\emptyset\}$ rather than at $\emptyset$ as bottom of the lattice. However, the property from theorem 2 continues to hold in this case, since the bottom step from $\emptyset$ to $\{\emptyset\}$ is unique and does not affect the path counts.

One may ask whether the function v defined in eq. (22) would fulfil some additional desirable structural properties, such as submodularity or supermodularity, which would permit the establishment of additional

relations. However, it turns out that, when instantiated with the information-based worth functions, neither of these are guaranteed, as the following remark shows.

Remark. Given a distributive lattice L, a function $f\colon L \to \mathbb{R}$ is called *submodular* if

$$f(a) + f(b) \geq f(a \vee b) + f(a \wedge b)\,, \quad \forall a, b \in L\,,$$

and f is called *supermodular* if $-f$ is submodular (Simovici, 2014).

We consider the examples RDN and AND, defined in section 6 below. In each case, let $a = [X_1]$ and $b = [X_2]$, so that $a \wedge b = \bot_{\text{in}}$ and $a \vee b = [X_1][X_2]$. Then, by calculating eq. (22) directly, we obtain, for the RDN example,

$$v(a) = 1\,, \quad v(b) = 1\,, \quad v(a \wedge b) = 0\,, \quad v(a \vee b) = 1\,,$$

so that $v(a) + v(b) > v(a \vee b) + v(a \wedge b)$. For the AND example it happens that the split distribution $p_{[X_1][X_2]}$ is given by $p_{[X_1 X_2]}$. (That is, knowing the marginals X_1X_2, X_1Y and X_2Y is enough to deduce the full joint distribution.) This allows us to calculate

$$v(a) = v(b) = \frac{3}{2} - \frac{3}{4}\log_2(3) \approx 0.311\,, \quad v(a \wedge b) = 0\,, \quad v(a \vee b) = 2 - \frac{3}{4}\log_2(3) \approx 0.811\,, \tag{36}$$

so that, in contrast to the RDN example, $v(a) + v(b) < v(a \vee b) + v(a \wedge b)$. This shows that, in our context, the function v, derived from the Kullback-Leibler divergences between split distributions, is neither submodular nor supermodular in general.

8 Discussion

In the search for a partial information measure that attributes informational contributions to various input variable sets (i.e. predictors) we relinquished the demand to quantify redundancy separately from unique information and instead applied the Pythagorean decomposition to characterize the additional contribution of an input variable set as it is added onto the relevant maximal chains. This "longitudinal" contribution is chain-dependent, though. To be able to talk about a contribution of an individual predictor, though, we need to express this contribution independently of the particular chain.

Intuitively, this can be done by assigning a probability distribution over the chains and averaging a predictor's contribution over all these chains; most naturally, the equidistribution could be chosen for this purpose. A more justified reasoning for this choice can be derived by observing that the setup of information contribution precisely matches the situation of a coalitional game where the worth of a coalition is the contribution of that player/predictor to the overall worth of the coalition, i.e. the information of the coalition about the target variable; and that contribution can be fairly assigned via the Shapley value concept. Of course, having a natural precedence order of predictors, not all coalitions of predictors (i.e. players when viewed through the eyes of game theory) are viable. We needed to resort to the variant of the Shapley value under *precedence constraints* which, as it turns out, corresponds precisely to the averaging over all maximal chains of the input lattice, strengthening both the confidence in the appropriateness of the proposed measure and the intuition behind it.

While the view of a predictor contribution stemming from averaging over chains (paths) through the lattice seems abstract and artificial, the Shapley value-based interpretation justifies its use. In fact, this perspective finds, again, additional justification from more recent coalitional game theory in which coalitions are not considered as immutable, but can change as per a stochastic process via local incentives (Faigle and Grabisch, 2012). In our context, this would correspond to a dynamically chosen path in an input lattice. At this stage, however, we are interested in the static contributions of the predictors; whether there will be an incentive to invoke a complex trajectory in the input lattice over which the contributions will be averaged, remains a question for the future.

Acknowledgement

DP would like to acknowledge support by H2020-641321 socSMCs FET Proactive project. NA and NV acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

A Correspondence between Feasible Rankings of $\mathfrak{N}$ and Maximal Chains in $\mathcal{D}$

With the machinery from Stanley (2011), the proof of existence of a bijection between feasible rankings on the set of players $\mathfrak{N}$ and maximal chains on the set of down-sets, has been straightforward. However, for the sake of self-containedness, we provide an explicit proof using only developments from the present paper.

Proof. We first show well-definedness, i.e. that each ranking defines a maximal chain. Let π, a (feasible) ranking over the set of players $\mathfrak{N}$, be given (we remind that each player is a subset of V). Define the sequence

$$\mathfrak{S}_0 := \emptyset, \mathfrak{S}_1, \mathfrak{S}_2, \dots, \mathfrak{S}_{|\mathfrak{N}|} \tag{37}$$

where for $k = 1, \dots, |\mathfrak{N}|$

$$\mathfrak{S}_k := \{A \in \mathfrak{N} \mid \pi(A) \leq k\} \tag{38}$$

$$= \pi^{-1}(\{1, \dots, k\})\,. \tag{39}$$

We now need to show now that this sequence $(\mathfrak{S}_k)_{k=0,\dots,|\mathfrak{N}|}$ is, first, a chain of down-sets (equivalently, feasible coalitions) and, second, maximal.

If $k = 0$, then $\mathfrak{S}_k = \emptyset$ is trivially a down-set. Else, let $1 \leq k \leq |\mathfrak{N}|$. Consider now $A \in \mathfrak{S}_k$, and any $B \in \mathfrak{N}$ with $B \subseteq A$. We have $\pi(B) \leq \pi(A) \in \{1, \dots, k\}$ per ranking property, and thus $\pi(B) \in \{1, \dots, k\}$, and thus $B \in \mathfrak{S}_k$ and $\mathfrak{S}_k$ is a down-set.

From (39) it follows that, for $k \leq l$, $\mathfrak{S}_k \subseteq \mathfrak{S}_l$. Therefore, if $A \in \mathfrak{S}_k$, also $A \in \mathfrak{S}_l$ and thus $\mathfrak{S}_k \leq \mathfrak{S}_l$ and the $(\mathfrak{S}_k)_k$ form a chain.

This chain is maximal. To show this, consider successive down-sets $\mathfrak{S}_k, \mathfrak{S}_{k+1}$, $k = 0, \dots, |\mathfrak{N}| - 1$ in the sequence. Consider $\widetilde{\mathfrak{S}}$ such that $\mathfrak{S}_k \leq \widetilde{\mathfrak{S}} \leq \mathfrak{S}_{k+1}$ according to the natural partial order $\leq$ on down-sets. If $\mathfrak{S}_k \neq \widetilde{\mathfrak{S}}$, then there exists a $B \in \widetilde{\mathfrak{S}} \setminus \mathfrak{S}_k$ and, since $\widetilde{\mathfrak{S}} \leq \mathfrak{S}_{k+1}$, one has $B \subseteq C$ for some $C \in \mathfrak{S}_{k+1}$. This means that $\pi(B) \leq \pi(C)$. Since $B \notin \mathfrak{S}_k$, also $\pi(B) \notin \{1, \dots, k\}$, so, per construction of $\mathfrak{S}_{k+1}$, necessarily $\pi(B) = k + 1$ and $B = \pi^{-1}(k + 1) = C \in \mathfrak{S}_{k+1}$. It follows that $\widetilde{\mathfrak{S}}$ must be either $\mathfrak{S}_k$ or $\mathfrak{S}_{k+1}$, thus, $\mathfrak{S}_k \prec \mathfrak{S}_{k+1}$ and the chain is maximal. This shows that the mapping from rankings to maximal chains is well-defined.

We show now that mapping rankings to maximal chains (37) via (39) is injective. For this, consider two rankings $\pi \neq \rho$. We have to show that they induce different maximal chains.

Consider B with $\pi(B) \neq \rho(B)$. Assume, without loss of generality, $\pi(B) < \rho(B)$. If we consider the chain $(\mathfrak{S}^\pi_k)_k$ induced by π (and analogously $(\mathfrak{S}^\rho_k)_k$ for ρ), then observe that the chain can be written in the form of inclusion chain as

$$\emptyset \subseteq \mathfrak{S}^\pi_0 \subseteq \mathfrak{S}^\pi_1 \subseteq \dots \subseteq \underset{\substack{\uparrow \\ \text{first time where } B \text{ appears in } (\mathfrak{S}^\pi_k)_k}}{\mathfrak{S}^\pi_{\pi(B)}} \subseteq \dots \subseteq \mathfrak{S}^\pi_{|\mathfrak{N}|} = \mathfrak{N}\,. \tag{40}$$

In this chain, the first down-set to contain B is the one with index $\pi(B)$. Under the same consideration for the chain induced by ρ, the first member of the chain to contain B is the one with index $\rho(B)$. However, $\pi(B) < \rho(B)$ and therefore the chains must differ and assigning chains to rankings via (37) is injective.

Show now surjectivity: for each maximal chain, there is a ranking that produces it. Let

$$\emptyset = \mathfrak{S}_0 \subseteq \mathfrak{S}_1 \subseteq \cdots \subseteq \mathfrak{S}_{|\mathfrak{N}|} = \mathfrak{N} \tag{41}$$

be a maximal chain. We show now that each step adds exactly one $C \in \mathfrak{N}$. Assume none of the steps in the sequence is trivial, i.e. we always have $\mathfrak{S}_j \subsetneq \mathfrak{S}_{j+1}$. All $\mathfrak{S}_k$ are at the same time down-sets as well as — equivalently — order-compatible coalitions. Choose $C \in \mathfrak{S}_{j+1} \setminus \mathfrak{S}_j$ minimal (i.e. such that for any $B \in \mathfrak{S}_{j+1} \setminus \mathfrak{S}_j$ with $B \subseteq C$, we have $B = C$).

Since $C \in \mathfrak{S}_{j+1}$, for any $B \subseteq C$, we have $B \in \mathfrak{S}_{j+1}$. It follows that either $B \in \mathfrak{S}_j$ or $B \in \mathfrak{S}_{j+1} \setminus \mathfrak{S}_j$; in the latter case, however, because of minimality of C in $\mathfrak{S}_{j+1} \setminus \mathfrak{S}_j$, it follows $B = C$. Thus $\mathfrak{S}_j \cup \{C\}$ is a down-set, and because of maximality of the chain, it must be identical to $\mathfrak{S}_{j+1}$. In summary, in each step of the maximal chain precisely one down-set is added.

Finally, given a maximal chain

$$\emptyset \prec \mathfrak{S}_1 \prec \mathfrak{S}_2 \prec \cdots \prec \mathfrak{S}_{|\mathfrak{N}|} = \mathfrak{N}\,, \tag{42}$$

define for every $j = 1, \ldots, |\mathfrak{N}|$ the inverse ranking $\pi^{-1}(j)$ to map onto the unique set (player) in $\mathfrak{S}_j \setminus \mathfrak{S}_{j-1}$. The maximal chain (42) is induced by the ranking π; we have thus shown the mapping (39) of rankings to maximal chains to be surjective (for every maximal chain there is a ranking that is mapped to it). With the injectivity shown earlier, this mapping is thus bijective. In short, we have shown that to each maximal chain corresponds one and only one feasible ranking. □

References

Amari, S.-i. (2001). Information geometry on hierarchical decomposition of stochastic interactions. *IEEE Transaction on Information Theory*, 47:1701–1711.

Amari, S.-i. and Nagaoka, H. (2007). *Methods of information geometry*, volume 191. American Mathematical Soc.

Amari, S.-i., Tsuchiya, N., and Oizumi, M. (2016). Geometry of information integration. In *Information Geometry and its Applications IV*, pages 3–17. Springer.

Ay, N. (2001/2015). Information geometry on complexity and stochastic interaction. *Entropy*, 17(4):2432–2458. Originally published in 2001 as MiS-Preprint 95/2001. Journal version published 2015. Preprint URL https://www.mis.mpg.de/publications/preprints/2001/prepr2001-95.html.

Ay, N., Jost, J., Vân Lê, H., and Schwachhöfer, L. (2017). *Information Geometry*. Springer.

Barrett, A. B. (2015). Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Phys. Rev. E*, 91:052802.

Bertschinger, N., Rauh, J., Olbrich, E., and Jost, J. (2013). Shared information—new insights and problems in decomposing information in complex systems. In *Proceedings of the European conference on complex systems 2012*, pages 251–269. Springer.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. (2014). Quantifying unique information. *Entropy*, 16(4):2161–2183.

Bilbao, J. M. (1998). Axioms for the Shapley value on convex geometries. *European Journal of Operational Research*, 110(2):368–376.

Bilbao, J. M. and Edelman, P. H. (2000). The Shapley value on convex geometries. *Discrete Applied Mathematics*, 103(1-3):33–40.

Csiszár, I. and Matúš, F. (2003). Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490.

Csiszár, I. and Matúš, F. (2004). On information closures of exponential families: counterexample. *IEEE Transactions on Information Theory*, 50(5):922–924.

Csiszár, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528.

Faigle, U. and Grabisch, M. (2012). Values for Markovian coalition processes. *Economic Theory*, 51(3):505–538.

Faigle, U. and Grabisch, M. (2013). A Concise Axiomatization of a Shapley-type Value for Stochastic Coalition Processes. *Economic Theory Bulletin*, pages 189–199.

Faigle, U. and Kern, W. (1992). The Shapley value for cooperative games under precedence constraints. *International Journal of Game Theory*, 21(3):249–266.

Finn, C. and Lizier, J. (2018). Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, 20(4):297.

Finn, C. and Lizier, J. T. (2020). Generalised measures of multivariate information content. *Entropy*, 22(2):216.

Grabisch, M. (2016). *Set Functions, Capacities and Games*. Springer.

Griffith, V. and Koch, C. (2014). Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pages 159–190. Springer.

Harder, M., Salge, C., and Polani, D. (2013). Bivariate measure of redundant information. *Phys. Rev. E*, 87:012130. http://arxiv.org/abs/1207.2080.

Ince, R. A. (2017). The partial entropy decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702.01591*.

James, R. and Crutchfield, J. (2017). Multivariate dependence beyond shannon information. *Entropy*, 19(10):531.

James, R. G., Emenheiser, J., and Crutchfield, J. P. (2019). Unique information via dependency constraints. *Journal of Physics A: Mathematical and Theoretical*, 52(1):014002.

Kolchinsky, A. (2019). A novel approach to multivariate redundancy and synergy. *arXiv preprint arXiv:1908.08642*.

Lange, F. and Grabisch, M. (2009). Values on regular games under Kirchhoff's laws. *Mathematical Social Sciences*, 58:322–340.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Science Publications.

Oizumi, M., Tsuchiya, N., and Amari, S.-i. (2016). Unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, 113(51):14817–14822.

Olbrich, E., Bertschinger, N., and Rauh, J. (2015). Information decomposition and synergy. *Entropy*, 17(5):3501–3517.

Perrone, P. and Ay, N. (2016). Hierarchical quantification of synergy in channels. *Frontiers in Robotics and AI*, 2:35.

Rauh, J., Bertschinger, N., Olbrich, E., and Jost, J. (2014). Reconsidering unique information: Towards a multivariate information decomposition. In *2014 IEEE International Symposium on Information Theory*, pages 2232–2236. IEEE.

Rosas, F., Ntranos, V., Ellison, C., Pollin, S., and Verhelst, M. (2016). Understanding interdependency through complex information sharing. *Entropy*, 18(2):38.

Shapley, L. S. (1953). A Value for n-person Games. In Kuhn, H. and Tucker, A., editors, *Annals of Mathematics Studies*, volume 28, pages 307–317. Princeton University Press.

Simovici, D. A. (2014). On submodular and supermodular functions on lattices and related structures. In *Proceedings of The International Symposium on Multiple-Valued Logic*, pages 202–207.

Stanley, R. P. (2011). *Enumerative Combinatorics*. Cambridge University Press.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Zwick, M. (2004). An overview of reconstructability analysis. *Kybernetes*, 33(5/6):877–905.