A Bivariate Measure of Redundant Information

Malte Harder,[*] Christoph Salge,[†] and Daniel Polani[‡]

Adaptive Systems Research Group
University of Hertfordshire
(Dated: July 23, 2012)

We define a measure of redundant information based on projections in the space of probability distributions. Redundant information between random variables is information that is shared between those variables. But in contrast to mutual information, redundant information denotes information that is shared about the outcome of a third variable. Formalizing this concept, and being able to measure it, is required for the non-negative decomposition of mutual information into redundant and synergistic information. Previous attempts to formalize redundant or synergistic information struggle to capture some desired properties. We introduce a new formalism for redundant information and prove that it satisfies all the properties necessary outlined in earlier work, as well as an additional criterion that we propose to be necessary to capture redundancy. We also demonstrate the behaviour of this new measure for several examples, compare it to previous measures and apply it to the decomposition of transfer entropy.

I. INTRODUCTION

In this paper we present a new formalism for *redundant information*; measuring for three (finite) random variables X, Y and Z how much information the random variable X contains about Z that is also contained in Y. *Information*, in this paper, is based on Shannon entropy [25], formalizes how much information one variable contains about another, where *mutual information* is the established formalism to quantify this (see [9] for a detailed account).

A naive extension of mutual information to information shared among multiple variables faces several problems. Since mutual information only measures the amount of information one variable contains about another it is unclear if two variables X and Y, which both contain information about Z, actually contain the "same" information. Alternatively, we could ask how much additional information (e.g. reduction in entropy) about Z would we get from X, if we already knew Y? This can be formalized as conditional mutual information $I(Z; X|Y) = I(Z; X, Y) - I(Z; Y)$. Thus one might think that $I(Z; X) - I(Z; X|Y)$, also called *interaction information* [7], is a candidate for a measure of redundant information, but the problem here is that it also captures the synergy between X and Y in the same measurement: in some cases, e.g. for binary variables, with Z being the outcome of an XOR combination of X and Y, each variable by itself contains no information about Z, but both taken together do contain information, which would be detected by the conditional mutual information. But we want redundant information only to be present if this information about Z is present in each variable on its own.

Redundant as well as synergistic information is information about the output variable contained in both variables; redundant information on the one hand is directly available in each input variable, whereas synergistic information is only available in the joint variable of the inputs. As we saw, interaction information cannot distinguish between redundant information and synergistic information, and is therefore ill-suited for this purpose.

In general, we want a redundant information formalism that quantifies how much Shannon information about the outcome of a multivariate mechanism a variable provides on its own that is also provided by all other variables as well.

II. RELATED WORK

Studies of synergies and redundancies have received attention in several areas including computational neuroscience [6, 8, 14, 20] and genetic regulatory networks [21, 22]. However, there seems to be no agreement how to best measure redundancy and synergy. A detailed overview of the requirements for a measure of synergy and redundancy, as well as a comprehensive overview of possible candidate measures can be found in [15].

Generalizations of mutual information have been proposed as measures of redundant information in the literature: One of them is *total correlation* also called *multi-information* which measures all dependencies among the individual variables [4]. Another generalization is called *interaction information* (as used in the introductory example in Section I), measuring the information that is shared among the variables of the system, but not shared by any subset of the variables [7]. However, both measures do not explain the structure of multivariate information in terms of atomic information quantities shared between variables. The former only quantifies the dependencies, where the latter has the problem of possibly be-

[*] m.harder@herts.ac.uk
[†] c.salge@herts.ac.uk
[‡] d.polani@herts.ac.uk

ing negative. Therefore, interaction information cannot distinguish between a system of independent variables and a system where redundancies and synergies between variables compensate each other. Thus, it also fails to capture the precise structure of multivariate mutual information [15, 30].

Other measures, like *interaction complexity* [18] give a good insight into the structure of interactions among random variables, however interactions and redundancy, though related, are not the same, as interaction complexity does not fulfill the criteria stated in [29]. Moreover, measures of *information flow* [5, 17] which are able to measure the overall amount of causal information flow, still struggle with over-determination (i.e. the measurement of redundant causal information flow), which is closely related to the problem of identifying redundant information.

A new approach addressing these problems was introduced by Williams and Beer [30]. It introduces a non-negative decomposition of multivariate mutual information terms $I(Z; X_1, ..., X_k)$. The decomposition captures all redundancies and synergies between all possible subsets of the variables $X_1, ..., X_k$ with respect to another random variable Z. Thus, the decomposition is able to reveal the atomic structure of the information that is shared by the variables $X_1, ..., X_k$ and Z.

Williams and Beer's decomposition can be applied to other information theoretic measures like transfer entropy as well. This allows to get further insight into the information transfer between processes by distinguishing state-independent information transfer from state dependent information transfer [31].

The information decomposition relies on a measure of redundancy [30]. Redundancy quantities then become the "building blocks" of the construction. Information in the sense of Shannon's information theory, as used here, always denotes a measure of information that one variable contains about another. The notion of redundancy then translates to information theoretic terms as the information that two variables share about another variable.

We will argue that the redundancy measure proposed by Williams and Beer, while exhibiting a number of essential properties needed to formalize redundancy, is not capturing the concept of redundancy in a fully satisfactory way. These problems have been noted by Griffith [15], who recently proposed [16] a synergy/redundancy measure based on *intrinsic conditional information* [23], which shares similarities with an *information bottleneck* [26].

We propose a different measure for the bivariate case which addresses our concerns and we compare it to the existing measures [16, 30]. The measure is based on a geometric argument and we will show that it fulfils all axioms required by Williams for a redundancy measure [29]. We also demonstrate that the non-negativity of the information decomposition is still guaranteed when using our measure. Furthermore, we will argue in favour of an additional axiom that any measure of redundancy has to fulfil.

A. Minimal Information as a Measure of Redundancy

As mentioned above, the term redundancy has been used in several contexts denoting different quantities. Here, we specificly consider information about another random variable that is shared among several random variables and we mean the same "piece" of information. A candidate measure for this quantity is called *minimal information* and denoted by $I_{\min}$ [30].

Given a set of finite random variables $X_V = \{X_1, ..., X_n\}$, the index set $V = \{1, ..., n\}$ and a finite random variable Z with values from $\mathcal{X}_1 \times ... \times \mathcal{X}_n$ and $\mathcal{Z}$ respectively, we denote the mutual information between Z and X_V as follows:

$$I(Z; X_V) := I(Z; X_1, ..., X_n). \tag{1}$$

Following [30], we now define the (non-negative) *specific information* [12], the increase in likelihood (or reduction in surprise) of the outcome of a specific event, where $A \subseteq V$, by

$$I(Z = z; A) := \sum_{x_A} p(x_A|z) \left[\log \frac{1}{p(z)} - \log \frac{1}{p(z|x_A)} \right] \tag{2}$$

$$= D_{\text{KL}} \left(p(x_A|z) \, \| \, p(x_A) \right), \tag{3}$$

where $D_{\text{KL}} \left(\cdot \, \| \, \cdot \right)$ is the usual Kullback-Leibler divergence. This is then be used by Williams and Beer to define the *minimal information* a set of random variables contains about the outcome as

$$I_{\min}(Z; A_1, ..., A_k) := \sum_z p(z) \min_i I(Z = z; A_i). \tag{4}$$

This measure is obviously non-negative and, in fact, positive if all variables contain some information about a specific outcome (for outcomes having probabilities which do not vanish).

For the bivariate case we will change the notation slightly and use the random variables directly instead of the index set notation, so instead of $I_{\min}(Z; A_1, A_2)$, where A_1 and A_2 are index sets of some collection of random variables, we will directly write $I_{\min}(Z; X, Y)$.

B. Redundancy Axioms

In [29], Williams states three axioms any redundancy measure has to fulfill. For any redundancy measure $I_\cap(Z; A_1, ..., A_k)$ the following must hold:

Symmetry: $I_\cap$ is symmetric with respect to the A_i's.

Self Redundancy: $I_\cap(Z; A) = I(Z; X_A)$.

Monotonicity:

$$I_\cap(Z; A_1, ..., A_{k-1}, A_k) \leq I_\cap(Z; A_1, ..., A_{k-1})$$

with equality if $A_{k-1} \subseteq A_k$.

From these axioms follows the non-negativity of the redundancy measure, and that it is bounded above by the mutual information between Z and each source. To prove this, note that A_i are subsets of V that could be empty, and for consistency $I_\cap(Z; \emptyset) = 0$ by definition. It is easy to check that all three axioms are fulfilled by the measure $I_{\min}$ [29].

C. Why Minimal Information is not Capturing Redundancy

This measure contradicts a basic intuition about redundancy. Let us consider the case with two binary input variables X, Y (i.e. $\mathcal{X} = \mathcal{Y} = \{0, 1\}$) that are independent, uniformly distributed and where $Z = (X, Y)$ is an unaltered copy of both variables, i.e. the joint distribution of X and Y. Now we expect that there should be no redundancy between X and Y with regard to Z because we know that X and Y are independent, so the information contained about Z in X and Y respectively is clearly not the same. However, we have $I_{\min}(Z; X, Y) = 1$ bit.

This happens because for each outcome of X or Y we observe a reduction of entropy regarding an outcome z (i.e. the specific information between X and z as well Y and z is positive). However, we ignore that even though X and Y give the same amount of information about an outcome z, they tell something different about the change of the distribution $p(z)$ after an observation in X or Y has been made. In this particular example X gives information about the first component of Z while, Y gives information about the second component of Z.

More precisely the *a posteriori* distributions of Z, $p(z|x)$ and $p(z|y)$, when either X or Y have been observed, give a different kind of information (have different content) even though they give the same amount of information. The core idea therefore is to separate the contributions of X and Y by adopting a geometric view in the space of probability distributions over Z.

III. A NEW MEASURE OF REDUNDANT INFORMATION

To define a new (bivariate) redundancy measure we will take a geometric view on informational quantities. Information geometry is a powerful tool-set to investigate information theoretic question in the context of Riemannian manifolds [1, 2]. Geometric arguments and algorithms have profound application to information theory, statistics [11] and have been successfully employed to construct information theoretic multivariate interaction measures [18]. Information geometry deals with statistical manifolds of probability distributions equipped with the Fisher metric [1]. The Kullback-Leibler divergence is now a divergence function on the statistical manifold and thus certain helpful properties and theorems, such as the Pythagorean Theorem, can be used. Here, we will introduce concepts of information geometry only as needed as most arguments can be done on an ad-hoc basis.

A. Additional Axiom

Before we start with the construction of the measure, we want to address the shortcoming identified above. For this purpose, we propose to add an additional axiom to the axioms from Section II B. We call it the *identity property*, as it states how redundancy should behave with respect to a joint random variable of identical copies of the two source variables. It requires that for any redundancy measure $I_\cap$

$$I_\cap\left((X_{A_1}, X_{A_2}); A_1, A_2\right) = I(X_{A_1}; X_{A_2}) \qquad (5)$$

The idea behind this additional axiom is, that if the (bivariate) mechanism we are considering is just copying the input, the redundancy must be exactly the mutual information between the variables. Given a multivariate redundancy measure the monotonicity automatically states that the multivariate redundancy is then bounded above by the minimum of pairwise mutual information terms.

B. Construction of a Redundant Information Measure

The redundancy measure we will construct is based on the notion of *projected information* which we will introduce shortly. We will begin with the definition of a bivariate redundancy measure I_{red}, i.e. we will measure the redundancy between two sources X and Y with respect to Z denoted by $I_{\text{red}}(Z; X, Y)$.

1. Preliminaries

In what follows, let $\Delta(Z)$ denote the space of all probability distributions over Z. An information projection is now defined as the minimization of the Kullback-Leibler divergence between a probability distribution in $p \in \Delta(Z)$ and a subset $B \subset \Delta(Z)$:

$$\pi_B(p) := \arg\min_{r \in B} D_{\text{KL}}\left(p \,\|\, r\right). \qquad (6)$$

The Kullback-Leibler divergence is not symmetric, therefore it is possible to define a dual projection $\pi_B{}^*(p)$ where the parameters of $D_{\text{KL}}\left(\cdot \,\|\, \cdot\right)$ are reversed (in [10], $\pi_B(p)$

is called reverse information projection and $\pi_B{}^*(p)$ information projection). Here we will exclusively use the projection $\pi_B(p)$.

For $B \subset \Delta(Z)$, we denote the convex closure of B in $\Delta(Z)$ by

$$C_{\text{cl}}(B) = \{\lambda p + (1-\lambda)q \,|\, p, q \in B, \lambda \in [0,1]\}. \quad (7)$$

As $\Delta(Z)$ is convex we have $C_{\text{cl}}(B) \subseteq \Delta(Z)$. Observing an event x in X or y in Y leads to a distribution over Z, $p(\cdot|x) \in \Delta(Z)$ and $p(\cdot|y) \in \Delta(Z)$ respectively. Let

$$\langle X \rangle_Z := \{p(\cdot|x) : x \in \mathcal{X}\} \quad (8)$$

denote the set of all conditional distributions of Z for the different events of X. Because the marginal distributions over Z are a convex combination of the conditional distributions, namely

$$p(z) = \sum_x p(z|x)p(x), \quad (9)$$

we have that the space of distributions over X, i.e. $\Delta(X)$, is embedded in $\Delta(Z)$ by the convex set

$$C_{\text{cl}}(\langle X \rangle_Z) = C_{\text{cl}}\left(\{p(\cdot|x) : x \in \mathcal{X}\}\right). \quad (10)$$

The convex closure of $\langle X \rangle_Z$ in $\Delta(Z)$ now contains all possible marginals $p(z)$ if we do not know the actual distribution of X, but where the mechanism (the conditional distribution) is known. For example, the problem of finding the channel capacity between two random variables X and Z can now be translated to find the point in the convex closure that maximizes its Kullback-Leibler divergence from all extremal points $p(\cdot|x)$ of the convex closure (weighted by the respective probabilities $p(x)$), as this is equivalent to maximizing the mutual information between X and Z.

2. Projective Information

Using information projections we can now project the conditionals of one variable onto the convex closure of the other. We denote this projection by

$$p_{(x \searrow Y)}(\cdot) := \pi_{C_{\text{cl}}(\langle Y \rangle_Z)}(p(\cdot|x)). \quad (11)$$

The projection is not guaranteed to be unique (for uniqueness, the set we are projecting onto would need to be log-convex and not convex [10]), however this does not matter for our purposes as we will see in the next lemma. Now, we define the *projected information* of X onto Y with respect to Z as

$$I_Z^\pi (X \searrow Y) := \sum_{z,x} p(z,x) \log \frac{p_{(x \searrow Y)}(z)}{p(z)}. \quad (12)$$

The rationale behind this construction is that the projected information quantifies the amount of information

that two variables share with each other, here X and Z, that can be expressed in terms of the information Y shared with Z (we are projecting onto Y). This is illustrated for binary input variable in FIG. 1.

Lemma 1. *Projected information $I_Z^\pi (X \searrow Y)$ is well-defined, finite and non-negative.*

Proof. First, note that projected information can be written as the difference of two Kullback-Leibler divergences

$$I_Z^\pi (X \searrow Y) = \sum_x p(x)[D_{\text{KL}}\left(p(z|x) \,\|\, p(z)\right)$$
$$- D_{\text{KL}}\left(p(z|x) \,\|\, p_{(x \searrow Y)}(z)\right)].$$

Therefore, if the projection is not unique, projected information only takes the KL-divergence into account which is the same for all possible solutions of the minimization problem in (6). Now we have $D_{\text{KL}}\left(p(z|x) \,\|\, p_{(x \searrow Y)}(z)\right) \leq D_{\text{KL}}\left(p(z|x) \,\|\, p(z)\right)$ for all $x \in \mathcal{X}$ because of $p(z) \in C_{\text{cl}}(\langle Y \rangle_Z)$ and the definition of $p_{(x \searrow Y)}(z)$ as the distance minimizing distribution to $p(\cdot|x)$ in $C_{\text{cl}}(\langle Y \rangle_Z)$. Hence $I_Z^\pi (X \searrow Y) \geq 0$. Furthermore $I(X; Z) = \sum_x p(x)D_{\text{KL}}\left(p(z|x) \,\|\, p(z)\right) < \infty$. $\square$

3. Definition of Bivariate Redundancy

The (bivariate) redundancy measure is now simply defined as the minimum of both projected information terms

$$I_{\text{red}}(Z; X, Y) := \min\{I_Z^\pi (X \searrow Y), I_Z^\pi (Y \searrow X)\}. \quad (13)$$

At this point we can take the minimum over both values because we already corrected for the change of the distributions in different directions by projecting the conditionals. This is different to the approach taken by Williams and Beer [30], where the minimization does
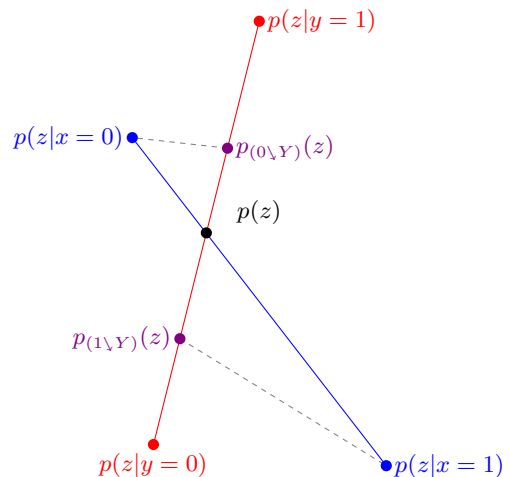

FIG. 1. Construction of projective information for binary input variables.

not consider that events in different source variables may change the distribution of the outcome in different directions in the geometrical space of distributions. Moreover, we define self-redundancy explicitly as

$$I_{\mathrm{red}}(Z;X) := I_{\mathrm{red}}(Z;X,X) \tag{14}$$
$$= I_Z^\pi (X \searrow X). \tag{15}$$

to prove:

$$I_{\mathrm{red}}(Z;X) = I_Z^\pi (X \searrow X) \tag{16}$$
$$= \sum_{z,x} p(z,x) \log \frac{p_{(x \searrow X)}(z)}{p(z)} \tag{17}$$
$$= \sum_{z,x} p(z,x) \log \frac{p(z|x)}{p(z)} \tag{18}$$
$$= I(Z;X). \tag{19}$$

For the monotonicity axiom we first need to show $I_{\mathrm{red}}(Z;X,Y) \leq I(Z;X)$. Using the expression of projected information as a difference of Kullback-Leibler divergences we get

$$I_{\mathrm{red}}(Z;X,Y) \leq I_Z^\pi (X \searrow Y) \tag{20}$$
$$= \sum_x p(x) \left[D_{\mathrm{KL}} \left(p(z|x) \,\|\, p(z) \right) \right.$$
$$\left. - D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \right] \tag{21}$$
$$= I(Z;X) - \sum_x p(x) D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right).$$

Hence it follows that $I_{\mathrm{red}}(Z;X,Y) \leq I(Z;X)$ as the KL-divergence is non-negative. To show equality holds if $X \subseteq Y$ we will first need the following two lemmas

4. The Proposed Measure is a Bivariate Redundancy Measure

To show that this is actually a redundancy measure, we have to show that it fulfils the four axioms (symmetry, self-redundancy, monotonicity and identity). Symmetry is obviously fulfilled, self-redundancy is also very quick

Lemma 2. *For all $x \in \mathcal{X}$ and random variables Y and W,*

$$\sum p(z|x) \left(\log p_{(x \searrow (Y,W))}(z) - \log p_{(x \searrow Y)}(z) \right) > 0. \tag{22}$$

Proof. Let $x \in \mathcal{X}$, as $C_{\mathrm{cl}}(\langle Y \rangle_Z) \subseteq C_{\mathrm{cl}}(\langle (Y,W) \rangle_Z)$ (note that $p(y|z) = \sum_w p(y,w|z)$) we have due to the definition of the projection that

$$\sum p(z|x) \log \frac{p(z|x)}{p_{(x \searrow (Y,W))}(z)} \leq \sum p(z|x) \log \frac{p(z|x)}{p_{(x \searrow Y)}(z)} \tag{23}$$
$$\iff \sum p(z|x) \log p_{(x \searrow (Y,W))}(z) \geq \sum p(z|x) \log p_{(x \searrow Y)}(z) \tag{24}$$

□

Lemma 3. *For all $(y,w) \in \mathcal{Y} \times \mathcal{W}$*

$$\sum p(z|y,w) \left(\log p_{((y,w) \searrow X)}(z) - \log p_{(y \searrow X)}(z) \right) > 0. \tag{25}$$

Proof. By definition, we have that $r = p_{((y,w) \searrow X)}$ is minimizing $D_{KL}(p(z|y,w)\|r)$ therefore

$$\sum p(z|y,w) \log \frac{p(z|y,w)}{p_{((y,w) \searrow X)}(z)} \leq \sum p(z|y,w) \log \frac{p(z|y,w)}{p_{(y \searrow X)}(z)} \tag{26}$$
$$\iff \sum p(z|y,w) \log p_{((y,w) \searrow X)}(z) \geq \sum p(z|y,w) \log p_{(y \searrow X)}(z) \tag{27}$$

□

Now the following proposition proves the missing piece for the monotonicity.

Proposition 4. $I_{\mathrm{red}}(Z;X,Y) \leq I_{\mathrm{red}}(Z;X,(Y,W))$

Proof. From Lemma 2 it follows directly that $I_Z^\pi (X \searrow Y) \leq I_Z^\pi (X \searrow (Y,W))$, furthermore from Lemma 3, $I_Z^\pi (Y \searrow X) \leq I_Z^\pi ((Y,W) \searrow X)$ respectively. Hence, we conclude $I_{\mathrm{red}}(Z;X,Y) \leq I_{\mathrm{red}}(Z;X,(Y,W))$.

□

Now it is only left to show that the measure also fulfils our new identity property, namely

$$I_{\text{red}}((X,Y);X,Y) = I(X;Y). \tag{28}$$

First we need the following lemma

Lemma 5. *If $Z = (X,Y)$ and (x',y') denote an event of Z then $p_{(y' \setminus X)}(x',y') = p_{(x' \setminus Y)}(x',y') = p(x'|y')p(y'|x')$*

Proof. Let $r \in C_{\text{cl}}(\langle X \rangle_Z)$, it is of the form

$$r(x',y') = \sum_x \alpha_x p(x',y'|x) = \alpha_{x'} p(y'|x'), \tag{29}$$

where $\alpha_x \geq 0$ and $\sum \alpha_x = 1$. We also have

$$D_{\text{KL}}\left(p(\cdot|y) \,\|\, r\right) = \sum_{x',y'} p(x',y'|y) \log \frac{p(x',y'|y)}{\alpha_{x'} p(y'|x')} \tag{30}$$

$$= \sum_{x'} p(x'|y) \log \frac{p(x'|y)}{\alpha_{x'} p(y|x')} \tag{31}$$

A simple calculation shows that the point $\alpha_{x'} = p(x'|y)$ fulfills the Karush-Kuhn-Tucker (KKT) conditions [19] for the minimization of Eq. (31) with respect to the vector $\alpha_{x'}$ and the simplex constraints. The KL-divergence is convex in the second parameter and thus it follows from the KKT conditions that $\alpha_{x'} = p(x'|y)$ is a global solution for the constrained minimization of the KL-divergence $D_{\text{KL}}\left(p(\cdot|y) \,\|\, r\right)$ parametrized by α_x as in Eq. (31) and in turn $r(x',y') = p(x'|y)p(y|x')$. If we now set $y' = y$ then we get $p_{(y' \setminus X)}(x',y') = p(x'|y')p(y'|x')$ and $p_{(x' \setminus Y)}(x',y') = p(x'|y')p(y'|x')$ respectively. $\square$

And hence we can conclude our proof with the following proposition:

Proposition 6. $I^\pi_{X,Y}(X \setminus Y) = I^\pi_{X,Y}(Y \setminus X) = I(X;Y)$

Proof. Without loss of generality,

$$I^\pi_{X,Y}(X \setminus Y) \tag{32}$$

$$= \sum_{x',y',x} p(x',y',x) \log \frac{p_{(x \setminus Y)}(x',y')}{p(x',y')} \tag{33}$$

$$= H(X,Y) + \sum_{x',y'} p(x',y') \log p_{(x' \setminus Y)}(x',y')$$

$$= H(X,Y) + \sum_{x,y} p(x,y) \log[p(x|y)p(y|x)]$$

$$= H(X,Y) - H(X|Y) - H(Y|X) \tag{34}$$

$$= I(X;Y). \tag{35}$$

$\square$

Thus I_{red} is a good candidate for measuring redundancy (in terms of redundancy with respect to some target variable).

IV. COMPARISONS

Now that we have constructed a bivariate redundancy measure, we will present a few examples of redundancy calculations.

A. Relation to Minimal Information

There are some cases where I_{red} and I_{min} coincide and we will have a look at some of these cases later in Section IV C. In general there is a tendency of I_{min} to overestimate redundancy and in our examples it seems that I_{min} is an upper bound for I_{red} in most cases. There are a few exceptions, but it is not yet clear for which cases these exceptions appear or whether they are due to numerical instabilities. The overestimation of redundancy by I_{min} becomes predominant if the dimension of Z is increased (see FIG. 3). The explanation for this is that, the higher the dimension of the space gets, the larger the error becomes which results from not taking directionality into account.

B. Decomposition of Mutual Information

In [30] Williams and Beer introduce partial information atoms (PI-atoms) as a way to decompose multivariate mutual information into non-negative terms. These terms can be defined for any multivariate redundancy measure and denote redundant and synergistic contributions between several variables of a set of random vari-
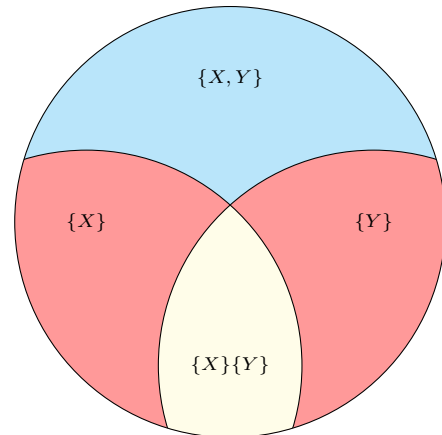

FIG. 2. PI-diagram for the decomposition of the mutual information between Z and X,Y into PI-atoms. $\{X,Y\}$ denotes the synergistic, $\{X\},\{Y\}$ the unique and $\{X\}\{Y\}$ the redundant part of the mutual information.
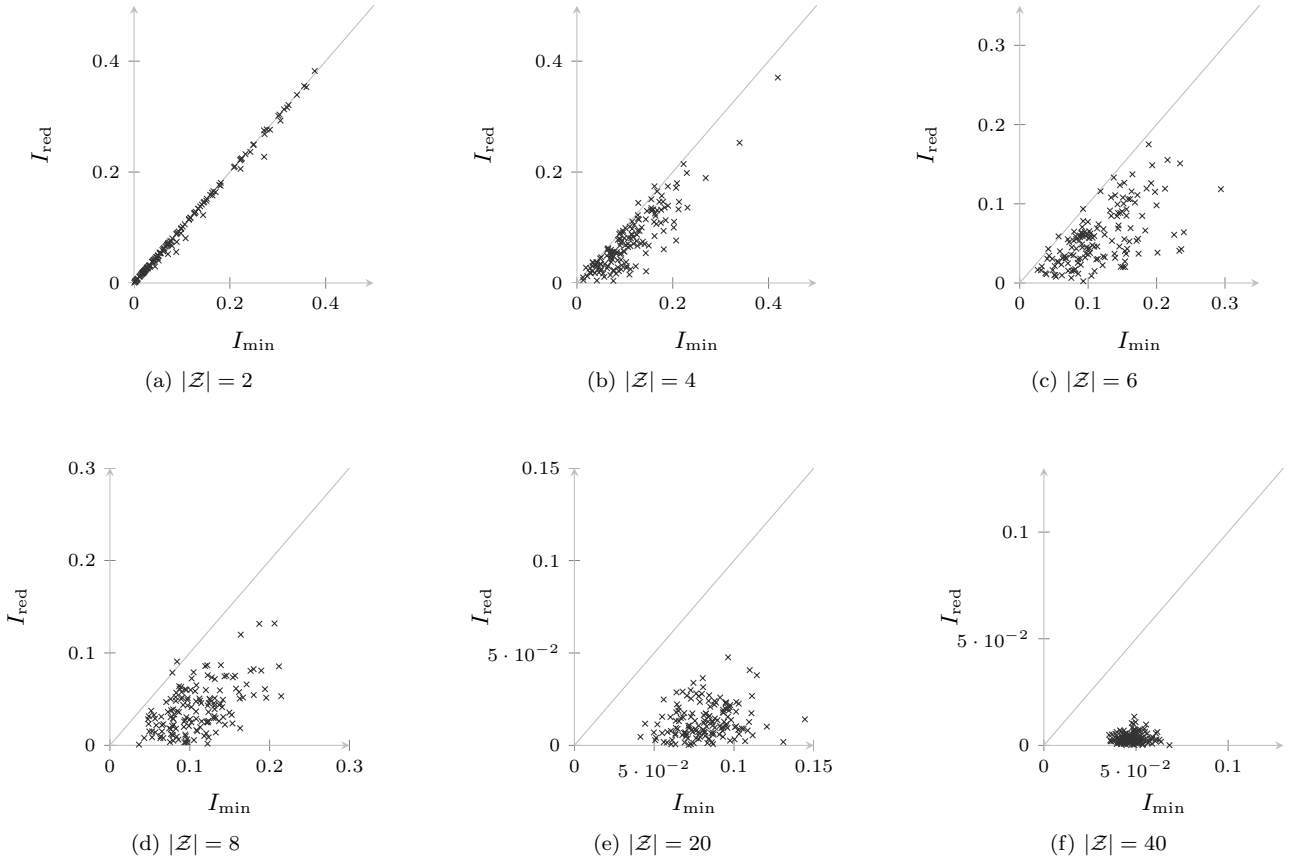
FIG. 3. Comparison of $I_{\min}$ and I_{red} for random distributions $p(x,y,z)$ and $|\mathcal{X}| = |\mathcal{Y}| = 3$ and different sizes of $\mathcal{Z}$. Note that as the dimension goes up, $I_{\min}$ gets larger in comparison to I_{red}. The distributions are initialized with random values, additionally the probability of each event being set to 0 with probability 0.5.

ables $\mathbf{R}$ towards another random variable Z. They are denoted by $\Pi_{\mathbf{R}}(Z; \alpha)$ where α is a set of subsets of the base set of random variables $\mathbf{R}$. As this construction is possibly with any redundancy measure, we will use $\Pi_{\mathbf{R}}(Z; \alpha)$ denoting the PI-atoms based on $I_{\min}$ as a redundancy measure and thereby staying consistent in the notation with [30]. The primed version $\Pi'_{\mathbf{R}}(Z; \alpha)$ on the other hand will denote the decomposition using the redundancy measure I_{red} introduced here.

In the bivariate case, this leads to the decomposition of mutual information $I(Z; X, Y)$ into four partial information atoms. Here we have $\mathbf{R} = \{X, Y\}$. Now, following [30] there are four atomic terms,

- $\Pi'_{\mathbf{R}}(Z; \{X\}\{Y\})$ which is the redundant information contained in X and Y about Z,

- $\Pi'_{\mathbf{R}}(Z; \{X\})$ and $\Pi'_{\mathbf{R}}(Z; \{Y\})$ are the unique information about Z, which is only contained in X or Y respectively,

- and $\Pi'_{\mathbf{R}}(Z; \{X, Y\})$, synergistic information, the information about Z that is only available if X and Y are both known.

The sum of these terms is exactly the mutual information

between Z and all sources, i.e.

$$I(Z; X, Y) = \Pi'_{\mathbf{R}}(Z; \{X\}\{Y\}) + \Pi'_{\mathbf{R}}(Z; \{X\})$$
$$+ \Pi'_{\mathbf{R}}(Z; \{Y\}) + \Pi'_{\mathbf{R}}(Z; \{X, Y\}). \quad (36)$$

as well as

$$I(Z; X) = \Pi'_{\mathbf{R}}(Z; \{X\}\{Y\}) + \Pi'_{\mathbf{R}}(Z; \{X\}) \quad (37)$$

and for Y respectively. Still following [30], but having replaced $I_{\min}$ by I_{red} we get $\Pi'_{\mathbf{R}}(Z; \{X\}\{Y\}) = I_{\mathrm{red}}(Z; X, Y)$ and $\Pi'_{\mathbf{R}}(Z; \{X\}) = I(Z; X) - I_{\mathrm{red}}(Z; X, Y)$. Finally, for the synergistic term

$$\Pi'_{\mathbf{R}}(Z; \{X, Y\}) = I(Z; X, Y) - \Pi'_{\mathbf{R}}(Z; \{X\})$$
$$- \Pi'_{\mathbf{R}}(Z; \{Y\})$$
$$- \Pi'_{\mathbf{R}}(Z; \{X\}\{Y\}) \quad (38)$$
$$= I(Z; X, Y) - I(Z; X)$$
$$- I(Z; Y) + I_{\mathrm{red}}(Z; X, Y). \quad (39)$$

Now this decomposition is not non-negative by default and this needs to be shown for the specific redundancy measure used. It is shown by Williams in [29] for the decomposition using $I_{\min}$. Here, we will show it for the bivariate case with I_{red} as redundancy measure: Firstly,

$I_{\mathrm{red}}(Z; X, Y)$ is non-negative, as shown earlier, furthermore it follows from the axioms of the redundancy measure that $I_{\mathrm{red}}(Z; X, Y) \leq I(X; Z)$ and with the same argument $I_{\mathrm{red}}(Z; X, Y) \leq I(Y; Z)$ which immediately implies that the unique information terms are non-negative. The following lemma now gives the non-negativity of the synergistic term:

Lemma 7. $I(Z; X, Y) - I(Z; X) - I(Z; Y) + I_Z^\pi (X \searrow Y) \geq 0$

Proof. We can reformulate the left hand side

$$I(Z; X, Y) - I(Z; X) - I(Z; Y) + I_Z^\pi (X \searrow Y) \tag{40}$$

$$= I(Z; X, Y) - I(Z; Y) - \sum_x p(x) D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \tag{41}$$

$$= \sum_{x,y} p(x, y) D_{\mathrm{KL}} \left(p(z|x, y) \,\|\, p(z|y) \right) - \sum_x p(x) D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \tag{42}$$

$$= \sum_x p(x) \left(\left(\sum_y p(y|x) D_{\mathrm{KL}} \left(p(z|x, y) \,\|\, p(z|y) \right) \right) - D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \right) \tag{43}$$

and now by the convexity of the Kullback-Leibler divergence:

$$\geq \sum_x p(x) \left(D_{\mathrm{KL}} \left(\sum_y p(y|x) p(z|x, y) \,\Big\|\, \sum_y p(y|x) p(z|y) \right) - D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \right) \tag{44}$$

$$= \sum_x p(x) \left(D_{\mathrm{KL}} \left(p(z|x) \,\|\, r(z|x) \right) - D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \right) \tag{45}$$

where $r(z|x) := \sum_y p(y|x) p(z|y) \in C_{\mathrm{cl}}(\langle Y \rangle_Z)$ and thus

$$D_{\mathrm{KL}} \left(p(z|x) \,\|\, r(z|x) \right) - D_{\mathrm{KL}} \left(p(z|x) \,\|\, p_{(x \searrow Y)}(z) \right) \geq 0 \text{ for all } x \in \mathcal{X}. \tag{46}$$

□

Given the non-negativity of the decomposition, we can visualize it using a PI-diagram as seen in FIG. 2. The whole circle represents the mutual information $I(Z; X, Y)$ and the colored/shaded regions represent redundant (yellow/light shaded), unique (red/dark shaded) and synergistic (blue/medium shaded) information.

C. Examples

We will now go through some examples for the bivariate measure, in particular those discussed in [15], which are a good selection of test cases for the desired properties of a redundancy/synergy measure.

1. Copying - From Redundancy to Uniqueness

Our first example is a very simple mechanism which simply copies the binary input variables X and Y into Z, i.e. $Z = (X, Y)$. However, we also add a control parameter $\lambda \in [0, 1]$ which determines how correlated X and Y are, as follows: Let W be a uniformly distributed binary random variable, $p(x|w) = \lambda \frac{1}{2} + (1 - \lambda) \delta_{xw}$ and $p(y|w) = \lambda \frac{1}{2} + (1 - \lambda) \delta_{yw}$. For $\lambda = 1$ we have that X and

Y are independent and we recover the example "UNQ (Unique Information)" from [15]. On the other extreme $\lambda = 0$ we have that X and Y are identical copies of W and therefore Z is equivalent to W from an information theoretic point of view. This is also reflected in the decomposition as in this case $I(Z; X, Y) = I(W; X, Y)$ and $I_{\mathrm{red}}(Z; X, Y) = I_{\mathrm{red}}(W; X, Y)$, so we can see that this is the example "RDN (Redundant Information)" from [15].

(a) Bayesian model

(b) PI-diagram for $\lambda = 0$, complete redundancy (RDN)

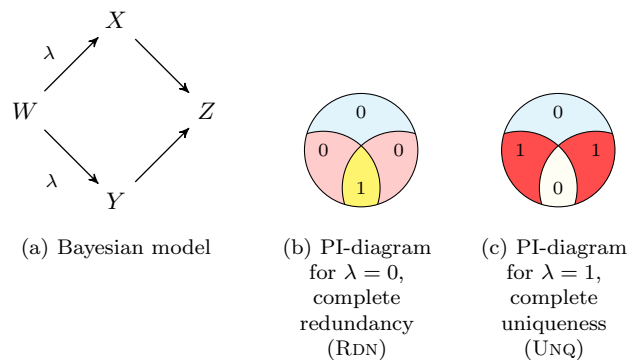(c) PI-diagram for $\lambda = 1$, complete uniqueness (UNQ)

FIG. 4. Copy Example. Complete redundancy and complete uniqueness using I_{red}.
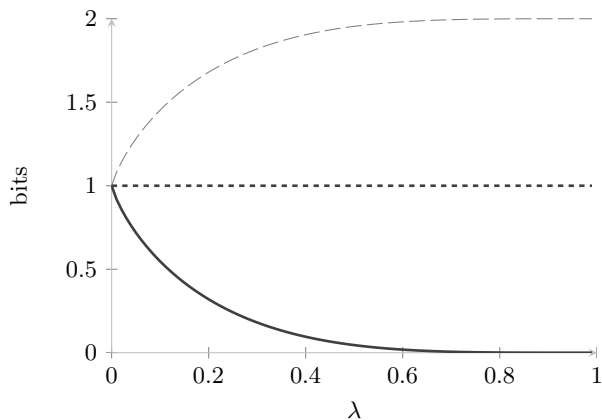
FIG. 5. Comparison of total mutual information $I(Z;X,Y)$ $(--)$, our redundancy measure I_{red} (———) and $I_{\min}$ $(\cdots\cdots)$ for varying values of λ, where λ controls the correlation between X and Y. It can be seen $I_{\min}$ measures a constant amount of redundancy and therefore does not distinguish between redundancy and uniqueness with varying λ as desired, whereas I_{red} does.

By varying λ we can vary the entropy of the outcome Z and at the same time exchange unique information for redundancy. FIG. 4 illustrates the decomposition at both extremal values of λ and it can be seen that the resulting values of I_{red} coincide with the proposed values in [15]. The effect of changing λ is shown in FIG. 5.

2. XOR

The XOR gate ($\oplus$), is a classical example for the appearance of synergy, in the sense of the whole being more than the sum of the individuals. We expect to only observe synergistic information, as the result is only known if both inputs are available, and the uncertainty given one input is the same as giving no input at all. Again the inputs are uniformly distributed binary random variables and $Z = X \oplus Y$. In fact, in this case we have $I_{\mathrm{red}}(Z;X,Y) = I_{\min}(Z;X,Y) = 0$ and get the purely synergistic decomposition as illustrated in FIG. 6. Note that I_{red} defines the redundancy, other terms are all derived by the decomposition.

3. AND - Mechanisms at Work

We now come to the AND gate, $Z = X \wedge Y$. This turns out to be an interesting case, because it demonstrates the subtle difference between redundant information that is due to the "ignorance" of the mechanism with respect to the source, and redundancy that is already apparent in the sources. In [15, 16] it is argued that vanishing mutual information between the sources X and
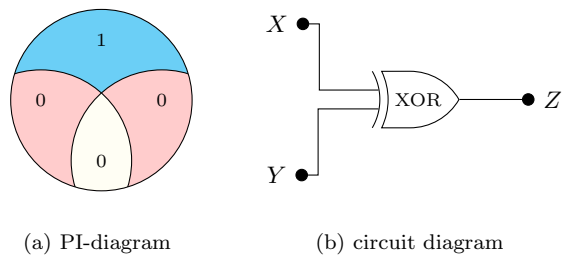

(a) PI-diagram   (b) circuit diagram

FIG. 6. XOR Example. A purely synergistic mechanism.

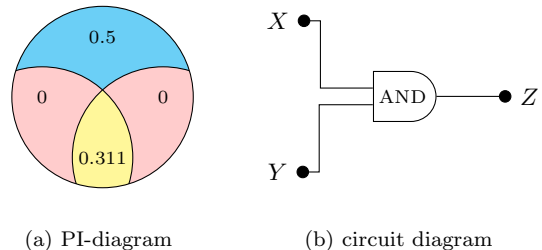(a) PI-diagram   (b) circuit diagram

FIG. 7. AND Example. The total mutual information is $I(Z;X,Y) = 0.811278$.

Y themselves implies vanishing redundant information[1]. This feature is also shared by the synergy measure introduced in [16]. However, here we would like to embrace a different view on redundant information: even if the sources are independent, there can be a correlation in the change of the distribution over Z given observations in X and Y respectively. Observing one input does not give any information about the other input, but part of the information gain about the distribution of the output can be the same as one gets from the other input alone. In particular in the case of the AND gate, observing a 0 in either input leads to $p(z = 0) = 1$. As a result of calculating the redundancy for this example we get $I_{\mathrm{red}}(Z;X,Y) = I_{\min}(Z;X,Y) = 0.311278$, so this is another example where minimal and redundant information coincide. FIG. 7 illustrates the decomposition of the total mutual information for this example.

We denote redundant information that is only due to the mechanism, as it is the case here, *mechanistic redundancy*. Contrary to this we call redundant information that already appears in the inputs *source redundancy*. Redundancy in the source must already manifest itself in the mutual information between the inputs. We do not give a rigorous definition for these terms, as it can be seen in the next example, there are cases where it is not clear how to separate both. However, if there is positive redundant information $I_{\mathrm{red}} > 0$ but vanishing mutual

---

[1] "However, because X_1 and X_2 are independent, [...], thus necessitating there is zero redundant information [...].",[15]

information between the sources, we will attribute all redundant information to mechanistic redundancy.

4. Summing Dice

Let us now consider an example where we throw two dice (cubic dice, with numbered sides from 0 to 5), represented by the random variables D_1, D_2 and sum their results. There are several ways to sum the results, we could simply add the two results — this would lead to results ranging from 0 to 10 where 5 is the most probable result and 0 or 10 the least probable results — or we multiply the result of the first die by 6 to get a uniform distribution of all numbers ranging from 0 to 35. Indeed, we will also look at all intermediate summations defined by $R = \alpha D_1 + D_2$ where $\alpha \in \{1, 2, 3, 4, 5, 6\}$. Our hypothesis was that for the direct summation ($\alpha = 1$) there is a positive amount of redundancy between D_1 and D_2 with respect to R, because knowing the roll of one die gives "overlapping" information (in the same direction in the space of distributions) with the roll of the other die about the final result. The redundancy should then decrease if α is increased, up to the point where $\alpha = 6$ and the sum of both dice rolls is isomorphic to the joint variable of the two dice rolls, i.e. $6D_1 + D_2 \simeq (D_1, D_2)$. Indeed, this is reflected in the redundancy $I_{\text{red}}(R; D_1, D_2)$. In FIG. 8 we added an additional parameter λ that controls how correlated the two dice are, in the same way as λ was introduced in the copy example in Section IV C 1 to control the correlation between the input variables. For $\lambda = 1$ they are independent and it can be seen that the redundancy increases with decreasing α, on the other extreme $\lambda = 0$ the dice are completely correlated. In this case we can see that the redundancy is already existent in the source ($I(D_1, D_2) \approx 2.58$) shadows all redundancy otherwise induced through the mechanism and hence there is no difference in the redundancy value for all values of α.

5. Composition of Mechanisms

The last three examples from [15] are compositions of the already shown examples. The first one RDNXOR combines the redundant copy example ($\lambda = 0$) with an XOR gate: (X, W) and (Y, W) are the inputs and $Z = (W, X \oplus Y)$ is the output. With our redundancy measure, this results in the required composite of one bit of redundant and one bit of synergistic information, the same as measured with I_{min}.

The second example RDNUNQXOR, combines an XOR gate with the two extremal copy cases. The inputs are (X_1, X_2, W) and (Y_1, Y_2, W), all independent and uniformly distributed. The output is $Z = (X_1 \oplus Y_1, (X_2, Y_2), W)$. Here we get the intended 1 bit of information in every partial information term, i.e. 1 bit of redundant, 1 bit synergistic information and 1 bit unique
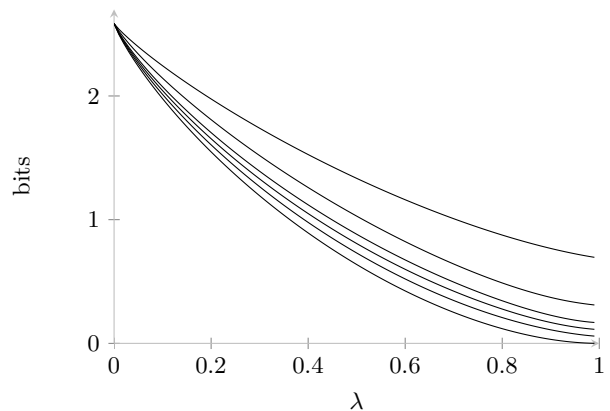

FIG. 8. Plot of the redundant information $I_{\text{red}}(R; D_1, D_2)$ depending on the correlation λ between the two dice D_1 and D_2. From top to bottom the summation coefficient is $\alpha = 1, ..., 6$. It can be seen that for independent dice $\lambda = 1$ the amount of redundancy depends on the mechanism that is used to sum the results, whereas on the other extreme, all redundancy comes from the correlation of the sources.

information per input, and a total 4 bits of mutual information.

The third example XORAND, combines an XOR gate with an AND gate, i.e. $Z = (X \wedge Y, X \oplus Y)$. This obviously leads to a different result than in [15], as the same effect of mechanistic redundancy appears in the AND gate, as mentioned in Section IV C 3.

6. Summary

In summary, these examples show that I_{red} captures proposed the concept of redundancy very well. Furthermore the resulting decomposition is in agreement with the desired examples in [15] except for the case where what we call mechanistic redundancy appears, which was

Example	Expected	I_{red}	I_{min}
Copy ($\lambda = 0$) / RDN	1	1	1
Copy ($\lambda = 1$) / UNQ	0	0	1
XOR	0	0	0
AND	0.311	0.311	0.311
RDNXOR	1	1	1
RDNUNQXOR	1	1	2
XORAND	0.5	0.5	0.5
Copy ($\lambda < 1$)	I(X;Y)	I(X;Y)	1

TABLE I. Summary of the bivariate redundancy examples. Results for the calculations of the examples using I_{red} and I_{min}, as well as the expected value that results from considerations of the desired properties of a redundancy measure, cf. [15].

not accounted for in the comparison of current measures of synergy. TABLE I summarises the comparison of $I_{\min}$ and I_{red}.

D. Information Transfer

In [31] the partial information decomposition is used to introduce new measures of information transfer. The measures are based on a decomposition of transfer entropy. Transfer entropy, introduced by Schreiber [24], is defined for two random processes X_t and Y_t as

$$T_{Y \to X} = I(X_{t+1}; Y_t | X_t). \tag{47}$$

It measures the influence of the process Y at time t on the state of the process X in the next time step. One can also take a longer history instead of Y_t and X_t into account. Conditional mutual information is defined as

$$I(X_{t+1}; Y_t | X_t) = I(X_{t+1}; Y_t, X_t) - I(X_{t+1}; X_t). \tag{48}$$

As the conditional entropy is the difference of two mutual information terms, the PI-decomposition can be used to decompose transfer entropy into two non-negative components. The decomposition is illustrated in FIG. 9. Let $\mathbf{R} = \{X_t, Y_t\}$ then it follows from (36) and (37) that

$$T_{Y \to X} = \Pi'_{\mathbf{R}}(X_{t+1}; \{Y_t\}) + \Pi'_{\mathbf{R}}(X_{t+1}; \{X_t, Y_t\}). \tag{49}$$

The first term denotes all information that uniquely comes from Y_t, called *State Independent Transfer Entropy* (SITE) by Williams and Beer [31]. The second term on the other hand denotes information that comes from Y_t but depends on the state of X_t and thus is called *State Dependent Transfer Entropy* (SDTE) in [31]. We now apply both measures $I_{\min}$ (with corresponding PI-atoms $\Pi_{\mathbf{R}}$) and I_{red} (with corresponding PI-atoms $\Pi'_{\mathbf{R}}$) as the underlying redundancy measure for the decomposition and compare the results.

We will consider two examples to show the difference of the decomposition when using I_{red} instead of $I_{\min}$. The first one revisits an example from [31] where X and Y are two binary, coupled Markov random processes. The process Y is uniformly i.i.d. and $x_{t+1} = y_t$ if $x_t = 0$, moreover

$$p(x_{t+1} = y_t | x_t = 1) = 1 - d, \tag{50}$$
$$p(x_{t+1} = 1 - y_t | x_t = 1) = d. \tag{51}$$

So $d \in [0, 1]$ controls whether there is any dependence on the previous state of X. If d vanishes X is simply a copy of Y. For this example and $d = 0$ shows only state-independent transfer while $d = 1$ shows only state dependent transfer and most importantly the decompositions of transfer entropy using either measure ($I_{\text{red}}, I_{\min}$) coincide (compare with FIG. 10).

The second example, though constructed for this specific purpose, is more intricate. First of all it shows the difference between the two measures, but it is also a good
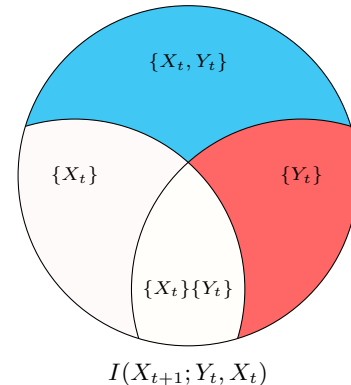

FIG. 9. PI-diagram for the decomposition of transfer entropy into PI-atoms. The coloured areas denote the transfer entropy.
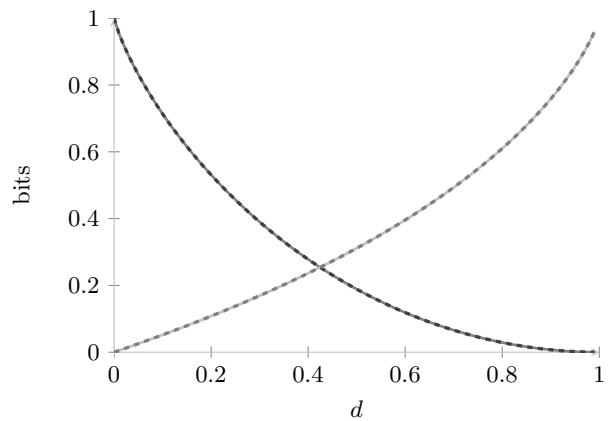

FIG. 10. Decomposition of transfer entropy $T_{Y \to X}$ for the first example process. The plot shows SITE (····· using $I_{\min}$, —— using I_{red}) and SDTE (····· using $I_{\min}$, —— using I_{red}) given d. It can be seen that both decompositions coincide for this process.

example of the subtlety of redundancy in mechanisms. Let us consider the following two processes (X_t, Y_t) and Z_t where Z_t are uniformly i.i.d. random variables, X_{t+1} is a copy of X_t and

$$p(y_{t+1} | y_t, z_t) = (1 - d)\delta_{y_t y_{t+1}} + d\delta_{z_t y_{t+1}}. \tag{52}$$

The process Y_t, copies with probability d the value of Z_{t-1} and with probability $(1 - d)$ the value of Y_{t-1}. We now measure the transfer entropy $T_{Z \to (X, Y)}$, see FIG. 11 for a Bayesian network of the process.

It can be seen in FIG. 12 that the two decompositions coincide for $d \le 0.5$. For $d = 0$ the two processes are completely independent which is reflected in the vanishing overall transfer entropy in this case. On the other extreme using $d = 1$, the decomposition using I_{red} gives complete state-independent transfer entropy while the decomposition using $I_{\min}$ sees total state-dependent
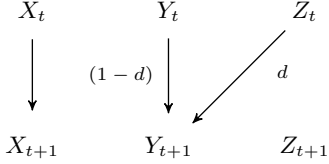
FIG. 11. Bayesian network of the second example process. X_t is a parallel and independent process, the only information transfer between the processes is from Z_t to Y_{t+1}.

FIG. 12. Decomposition of transfer entropy $T_{Z\to(X,Y)}$ for the second example process. The plot shows SITE (------ using $I_{\min}$, —— using I_{red}) and SDTE (▪▪▪▪▪ using $I_{\min}$, ▬▬ using I_{red}).

transfer entropy. In this case the decompositions disagree completely and we argue that our measure reflects the process much better. With $d=1$ the process always copies Z_t to Y_{t+1}, which is completely independent of (X_t, Y_t). Specifically, $I_{\min}$ mistakenly sees redundancy between X_t and Z_t in the evolution of one timestep. Following (39) and (37) this is then reflected in the vanishing state-independent transfer entropy for all d (larger redundancy means more synergy and less unique information, given that the mutual information stays constant).

The fact that $I_{\min}$ measures more redundancy has the same reason why $I_{\min}$ measures redundancy between independent X and Y with respect to $Z=(X,Y)$, namely it compares changes in different direction in the space of distributions. The parallel and independent process X_t lets $I_{\min}$ see a dependency between the two processes X_t and Z_t that does not exist. If we consider the transfer entropy $T_{Z\to Y}$ from Z_t to Y_t only, ignoring the process X_t completely, we can see in FIG. 13 that the decomposition (------,▪▪▪▪▪) now coincides with the decomposition of $T_{Z\to(X,Y)}$ using I_{red} (——,▬▬ in FIG. 12).

Nonetheless, we have not yet explained the quite unusual non-differentiable shape of the state-independent transfer entropy, which only is positive for $d>0.5$. This is surprising because up to $d=0.5$ all transfer entropy is considered to be state-dependent, even though with probability d the state of Y_{t+1} takes on the state of Z_t. As the process X_t was only used to demonstrate that using $I_{\min}$ for the decomposition measures state dependencies in the transfer-entropy that are not there, we will now leave X_t aside and only consider the process (Y_t, Z_t) as described above.

To understand the shape of the graph of state-dependent transfer entropy of this process, we need to have a look at the mutual information $I((Y_{t+1}); Z_t)$ (----- in FIG. 14) and the redundancy $I_{\mathrm{red}}(Y_{t+1}; Y_t, Z_t)$ (---- in FIG. 14). From (37) it follows that the state-independent transfer entropy (—— in FIG. 12 and ------ in FIG. 13) is now the difference of these two terms (compare with FIG. 9).

The increase of mutual information $I(Y_{t+1}; Z_t)$ is obvious from the definition of the process. For $d=0$ we have independence between both processes and for $d=1$ we have $Y_{t+1}=Z_t$. It is also clear that the redundant information with respect to Y_{t+1} needs to be zero at the extremal points $d\in\{0,1\}$, because at these points the
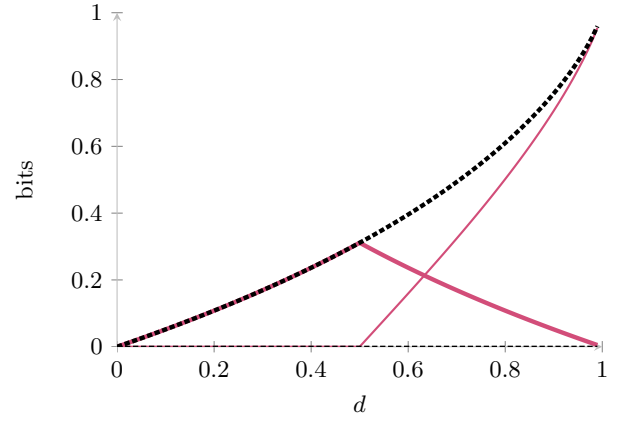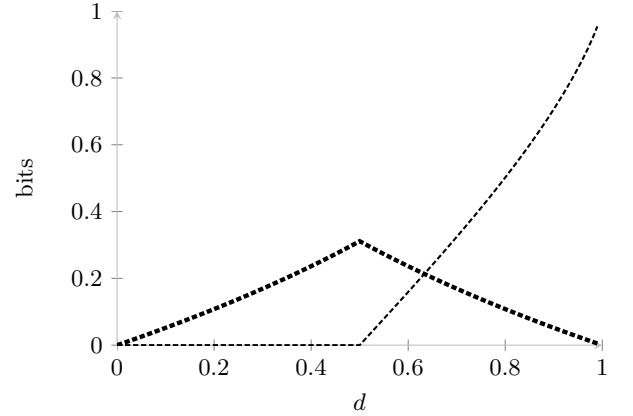

FIG. 13. Decomposition of transfer entropy $T_{Z\to Y}$ for the second example process. The plot shows SITE (------ using $I_{\min}$), SDTE (▪▪▪▪▪ using $I_{\min}$).
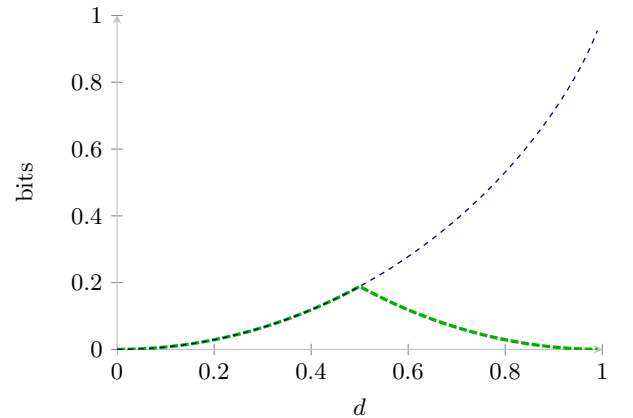

FIG. 14. The plot shows $I(Y_{t+1}; Z_t)$ (-----) and $I_{\mathrm{red}}(Y_{t+1}; Y_t, Z_t)$ (----) for the second example process.

value of Y_{t+1} depends either on Y_t ($d = 0$) or Z_t ($d = 1$) and therefore either $I(Y_{t+1}; Z_t) = 0$ or $I(Y_{t+1}; Y_t) = 0$ which both are upper bounds for the redundancy.

On the other hand for $d = 0.5$ the state of either process at time t tells us something about the distribution of Y_{t+1} and because the space of distributions of Y_{t+1} is one-dimensional, this must be information about a change in the same direction, so there is positive redundancy. Observing one of the outcomes necessarily contributes to some extent to the prediction of the outcome of Y_{t+1}. We can now show this more rigourously, we have

$$p(y_{t+1}|y_t) = \frac{d}{2}\delta_{y_{t+1}(1-y_t)} + \left(1 - \frac{d}{2}\right)\delta_{y_{t+1}y_t}, \quad (53)$$

$$p(y_{t+1}|z_t) = \frac{1-d}{2}\delta_{y_{t+1}(1-z_t)} + \frac{1+d}{2}\delta_{y_{t+1}z_t}. \quad (54)$$

as the conditional distributions given the current state of either Y_t or Z_t. To calculate $I_{\mathrm{red}}(Y_{t+1}; Y_t, Z_t)$ we need to calculate the projected information $I^\pi_{Y_{t+1}}(Z_t \searmark Y_t)$ and $I^\pi_{Y_{t+1}}(Y_t \searmark Z_t)$ as the redundancy is the minimum of both terms. Because the space of distributions $\Delta(Y_{t+1})$ is one dimensional (it is simply the unit interval) we can make a simple illustrative argument to compute $p_{(z_t=0\searmark Y_t)}, p_{(z_t=1\searmark Y_t)}, p_{(y_t=0\searmark Z_t)}$ and $p_{(y_t=1\searmark Z_t)}$, which are the terms that are needed to calculate projected information. From the illustration in FIG. 15 it can be seen that for $d \le 0.5$, $p_{(z_t=0\searmark Y_t)}(y_{t+1}) = p_{(y_t=0\searmark Z_t)}(y_{t+1}) = p(y_{t+1}|z_t = 0)$ and $p_{(z_t=1\searmark Y_t)}(y_{t+1}) = p_{(y_t=1\searmark Z_t)}(y_{t+1}) = p(y_{t+1}|z_t = 1)$. If we insert this into (12) we get that $I^\pi_{Y_{t+1}}(Z_t \searmark Y_t) = I^\pi_{Y_{t+1}}(Y_t \searmark Z_t) = I(Y_{t+1}; Z_t)$ for $d \le 0.5$.

Conversely for $d \ge 0.5$ we get $I^\pi_{Y_{t+1}}(Z_t \searmark Y_t) = I^\pi_{Y_{t+1}}(Y_t \searmark Z_t) = I(Y_{t+1}; Y_t)$ for $d \le 0.5$. As $I(Y_{t+1}; Z_t)$ and $I(Y_{t+1}; Y_t)$ are perfectly symmetric, this then explains the form of the redundant information as in (----- in FIG. 14). Thus, even though Z_t and Y_t are completely independent, the mechanism, which is a random read-out (with distribution $d, (1 - d)$), creates redundancy with respect to Y_{t+1}. Furthermore, this explains why we have no state-independent transfer entropy for $d \le 0.5$.

1. Open Loop Controllability

Ashby [3] proposed and Touchette and Lloyd [27] confirmed that there is a natural link between control theory and information theory. As shown by Touchette and Lloyd [28], for a process, with initial state X and final state X', and a controller C which are linked by the probability distribution $p(x'|x, c)$, the conditional mutual information $I(X'; C|X)$ (which is the transfer entropy from the controller to the system) is a measure of controllability. Williams and Beer show in [31] that the decomposition of transfer entropy using $I_{\min}$ as a redundancy measure has a close relation to the notion of open-loop controllability. We will now show, that this is still the case if I_{red} is used to decompose transfer entropy.

Perfect controllability, as defined in [28], means that for all initial states $x \in \mathcal{X}$ and final states $x' \in \mathcal{X}$ there exists a control state $c \in \mathcal{C}$ such that $p(x'|x, c) = 1$. The following equivalence is then shown in [31]

Lemma 8. *A system is perfectly controllable iff for any x' there exists a distribution $p(c|x)$ such that $p(x') = 1$ for any distribution $p(x)$.*

It follows also that if a system is perfectly controllable, there exists an x' such that $p(x'|x) = 1$ for each $x \in \mathcal{X}$, see [31] for a proof. Now, a system has perfect open-loop controllability iff it has perfect controllability and $I(X; C) = 0$. Moreover, in [31] it is shown that the following theorem holds:

Theorem 9 (Williams and Beer). *A system is perfectly open-loop controllable iff it is perfectly controllable with vanishing state-dependent transfer entropy (using $I_{\min}$) from C to X'.*

We will now also show that this theorem still holds in the case where the decomposition using our measure of redundant information I_{red} is used. To prove the theorem we will use the following lemma. It is shown in [31] that the condition of the lemma is fulfilled for any perfect open-loop controller and thus proves the direct part of the theorem (perfect open-loop controllability implies perfect controllability with zero SDTE using I_{red} as a redundancy measure):

Lemma 10. *If*

$$p(x'|x, c) = p(x'|c) \quad x' \in \mathcal{X}, \forall x \in \mathcal{X}, c \in \mathcal{C}$$

then the STDE from C to X' is zero.

Proof. From (13) and (36) it follows that

$$\Pi'(X'; \{C, X\}) \le I(X'; X, C) - I(X'; X) \\ - I(X'; C) + I^\pi_{X'}(X \searmark C). \quad (55)$$

The synergy is non-negative and now the right hand side can be reformulated as in (43). But with $p(x'|x, c) = p(x'|c) \, \forall x, x' \in \mathcal{X}, c \in \mathcal{C}$ the positive Kullback-Leibler divergences in (43) all vanish. Therefore $\Pi'(X'; \{C, X\}) = 0$. $\square$

For the converse direction, perfect controllability and vanishing STDE (from C to X') imply perfect open-loop controllability, we first need to prove the following lemma:

Lemma 11. *If a system is perfectly controllable with a distribution $p(c|x)$ then $I_{\mathrm{red}}(X'; X, C) = 0$.*

Proof. From Lemma 8 it follows that $p(x') = 1$ for some $x' \in \mathcal{X}$ as well as $p(x'|x) = 1$ for all $x \in \mathcal{X}$ and therefore $C_{\mathrm{cl}}(\langle X \rangle_Z)$ in $\Delta(X')$ is just $\{p(x')\}$ which implies $I^\pi_{X'}(C \searmark X) = 0$. Thus it follows that $I_{\mathrm{red}}(X'; X, C) = 0$. $\square$

(a) $\Delta(Y_{t+1})$ for $d \leq 0.5$
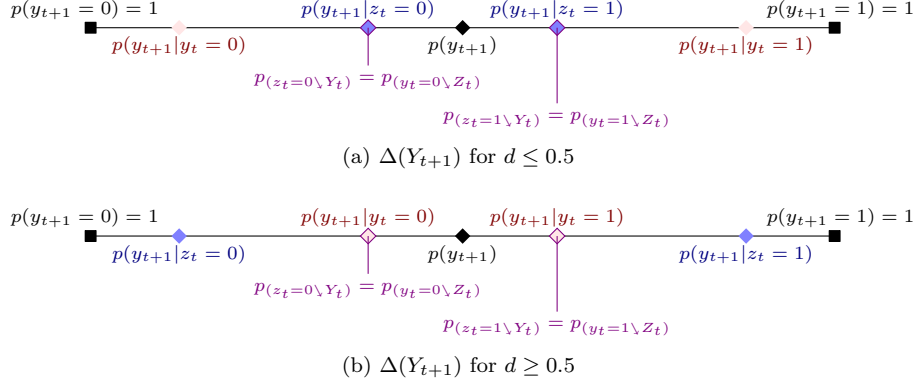

(b) $\Delta(Y_{t+1})$ for $d \geq 0.5$

FIG. 15. Illustration of the conditional distributions of Y_{t+1} for the second example process in the two cases $d \leq 0.5$ and $d \geq 0.5$. The line represents the one dimensional simplex, i.e. the space of probability distributions over Y_{t+1} denoted by $\Delta(Y_{t+1})$ where Y_{t+1} is a binary valued random variable. The black diamond represents the marginal distribution of $p(y_{t+1})$ and the shaded diamonds the conditionals given specific values of Y_t and Z_t. It can now be seen that the projections are always equal to the conditional distributions closer to the marginal of Y_{t+1}. In particular, the projections are the same, no matter in which direction the projection is done (from Y_t to Z_t or vice versa).

Thus, for the converse direction, starting with perfect controllability and vanishing STDE, we have the following equality

$$0 = \Pi'(X'; \{C, X\}) \tag{56}$$
$$= I(X'; X, C) - I(X'; X)$$
$$- I(X'; C) + I_{\text{red}}(X'; X, C) \tag{57}$$
$$= I(X'; X, C) - I(X'; X) - I(X'; C) \tag{58}$$
$$= \sum_{x,c,x'} p(x', x, c) \log \frac{p(x'|x,c)p(x')}{p(x'|c)p(x'|x)}, \tag{59}$$

as we also have $p(x'|x) = p(x')$ because of perfect controllability,

$$= \sum_{x,c,x'} p(x', x, c) \log \frac{p(x'|x,c)}{p(x'|c)}. \tag{60}$$

We also know that for every $x \in \mathcal{X}$ there exists $x' \in \mathcal{X}$ and $c \in \mathcal{C}$ such that $p(x'|x,c) = 1$. Thus for any $x' \in \mathcal{X}$ there exists a $c \in \mathcal{C}$ such that $p(x'|c) = 1$. It is shown in [31] that this is equivalent to open-loop controllability.

Hence, we have shown that Theorem 9 also holds if we apply I_{red} as the underlying redundancy measure and the relation between open-loop controllability and decomposition of transfer entropy is transferable to our new measure.

V. DISCUSSION

The motivation for this paper was to overcome the shortcomings of current measures of redundancy and synergy. We introduced a new measure for bivariate redundant information. Redundant information between two random variables is information that is shared between two variables. In contrast to mutual information, redundant information denotes information with respect to the outcome of a third variable. Our measure is conceptually motivated by measuring similarities in the direction of change in the outcome distribution, depending on which input is observed. We proved that the construction adheres to properties of redundancy as stated in the literature, and can be used for a non-negative decomposition of mutual information. The measure is closely related to the concept of *minimal information* as introduced in [30].

We demonstrated in several examples that I_{red} follows several intuitions about redundancy. Furthermore, it is possible to decompose *transfer entropy* as considered in [31]; in particular we showed that using *minimal information* instead of *redundant information* to decompose *transfer entropy* can lead to the detection of fake state-dependent transfer entropy. We were able to prove that the results about open-loop controllability from [31] are also applicable to the decomposition using I_{red}. Thus our measure is able to serve as a replacement for the bivariate version of minimal information.

A particular insight of our definition is the emphasis of mechanisms in the concept of redundant information, which has been rather neglected in the literature so far. Firstly, we linked bivariate redundant information in the case of a copying mechanism to the mutual information between the input variables. We identify redundant information that already appears in the inputs with *source redundancy*, contrary to redundant information that is only due to the mechanism, as demonstrated in the AND-gate or the 50:50-readout. We identify this kind of redundancy with *mechanistic redundancy*. This is in contrast to the redundancy measure proposed in [16] which does not capture *mechanistic redundancy*. The separation of both kinds of redundancy is not explicit at this point, and currently we do not yet propose a clear and obvi-

ous separation of mechanistic and source contributions of redundant information.

Future work will show whether it is possible to separate the two concepts of mechanistic and source redundancy when they appear simultaneously. Another limitation we currently have is the restriction to a bivariate measure. In general, however, there are applications where it is interesting to be able to compute redundant information between more than two variables [13, 30]. However, the geometric structure for this problem gets significantly more complex, and it is, for example, not entirely clear by what the *identity property* should be replaced in the multivariate case. There are several ways to generalize mutual information to a multivariate measure, none of which seems to be fitting in this case. The construction of a multivariate measure of redundant information, as well as a generalization to continuous random variables is part of ongoing research.

ACKNOWLEDGMENTS

[1] Amari, S. and Nagaoka, H. (2007). *Methods of information geometry*, volume 191. Amer Mathematical Society.

[2] Amari, S.-I. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711.

[3] Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall Ltd.

[4] Ay, N., Olbrich, E., Bertschinger, N., and Jost, J. (2006). A unifying framework for complexity measures of finite systems. *Proceedings of ECCS06. Oxford, UK: European Complex Systems Society*.

[5] Ay, N. and Polani, D. (2008). Information Flows in Causal Networks. *Advances in Complex Systems*, 11(1):17–41.

[6] Balduzzi, D. and Tononi, G. (2008). Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Comput Biol*, 4(6):e1000091.

[7] Bell, A. (2003). The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003*.

[8] Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2000). Adaptive rescaling optimizes information transmission. *Neuron*, 26:695–702.

[9] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience.

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

[11] Csiszar, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial. In Verdú, E.-i.-c. S., Notredame, D. C., Stanford, T. C., Maryland, A. E., and Stanford, A. G., editors, *Foundations and Trends in Communications and Information Theory*, volume 1.

[12] DeWeese, M. and Meister, M. (1999). How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 10(4):325–340.

[13] Flecker, B., Alford, W., Beggs, J., Williams, P., and Beer, R. (2011). Partial information decomposition as a spatiotemporal filter. *Chaos*, 21(3).

[14] Gat, I. and Tishby, N. (1999). Synergy and redundancy among brain cells of behaving monkeys. *Advances in neural information processing systems*, pages 111–117.

[15] Griffith, V. (2011). Quantifying synergistic information remains an unsolved problem. *Arxiv preprint arXiv:1122.1680v3*.

[16] Griffith, V. and Koch, C. (2012). Quantifying synergistic mutual information. *Arxiv preprint arXiv:1205.4265*.

[17] Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schoelkopf, B. (2012). Quantifying causal influences. *ArXiv preprint*.

[18] Kahle, T., Olbrich, E., Jost, J., and Ay, N. (2009). Complexity measures from interaction structures. *Physical Review E*, 79(2):11.

[19] Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium*, pages 481–492.

[20] Latham, P. and Nirenberg, S. (2005). Synergy, redundancy, and independence in population codes, revisited. *Journal of Neuroscience*.

[21] Liang, K. and Wang, X. (2008). Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008.

[22] Margolin, A. a., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7.

[23] Maurer, U. and Wolf, S. (1999). Unconditionally secure key agreement and the intrinsic conditional information. *Information Theory, IEEE Transactions on*, 45(2):499 – 514.

[24] Schreiber, T. (2000). Measuring information transfer. *Arxiv preprint nlin/0001042*.

[25] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(October):379–423.

[26] Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.

[27] Touchette, H. and Lloyd, S. (2000). Information-theoretic limits of control. *Physical review letters*, 84(6):1156–1159.

[28] Touchette, H. and Lloyd, S. (2004). Information-theoretic approach to the study of control systems. *Physica A*, 331:140.

[29] Williams, P. L. (2011). *Information Dynamics: Its Theory and Application to Embodied Cognitive Systems.* PhD thesis, Indiana University.

[30] Williams, P. L. and Beer, R. D. (2010). Nonnegative Decomposition of Multivariate Information. *Arxiv preprint arXiv:1004.2515.*

[31] Williams, P. L. and Beer, R. D. (2011). Generalized Measures of Information Transfer. *Arxiv preprint arXiv:1102.1507.*